Draft. Please do not cite without permission.

Illusions of Influence in Newcomb's Problem[*]

Dilip Ninan ○ MIT

September 14, 2006

Abstract I argue that the one-boxing intuition in Newcomb's Problem arises from the fact that it wouldn't be epistemically rational for an agent in a Newcomb Problem to be certain that her decision would not affect the contents of the opaque box. I then show that a very small amount of credence in the hypothesis that one's choice will affect the contents of the opaque box is enough to make one-boxing rational, according to causal decision theory. The best argument for this account is that it offers a fairly precise explanation of why changing certain parameters in the case alters our intuitions in systematic ways.

1 Newcomb's Problem and Decision Theory

1.1 The Flagship Newcomb Problem

Here is a description of the original Newcomb Problem:

> The Predictor is a being who is able to predict your choices with great accuracy. The Predictor has accurately predicted your choices in the past, and has accurately predicted the choices of others in the past. The Predictor has also had great success at making predictions about the choices people make in the following situation: There are two boxes, a transparent box b_1 and an opaque box b_2. You can see that b_1 contains \$1,000; b_2 contains either \$1,000,000 or nothing. You have two choices: either you take what is in both boxes, or you take what is in the opaque box b_2 alone. If the Predictor predicted that you will take what is in both boxes, he does not put \$1,000,000 in b_2; but if the Predictor predicted that you will take what is in

[*] Versions of this paper were presented at the M.A.T.T.I. Group at MIT, the 2006 Princeton-Rutgers and Columbia-NYU Graduate Conferences, and the 2006 meeting of the Society for Exact Philosophy in San Diego. For helpful comments, thanks to these audiences and to Selim Berker, Tom Dougherty, Ned Hall, Caspar Hare, Adam Hosein, Gabriel Love, Vann McGee, Robert Stalnaker, Eric Swanson, and Seth Yalcin.

> b_2 alone, he puts \$ 1,000,000 in b_2. You value more money to less money. What should you do?[1,2]

It seems like taking b_2 alone (*one-boxing*) is a good way of ensuring that you get \$1,000,000, whereas taking both boxes (*two-boxing*) is a good way of ensuring that you only get \$1,000. Since you value more money to less money, you should take b_2 alone. Pre-theoretically, one-boxing is an option that enjoys some intuitive support.

Of course, there is a well-known – and, in my view, very compelling – argument for taking both the boxes. Since the Predictor has already made his move, either the \$1,000,000 is in b_2 right now or it's not. Whether you take one box or two will have no effect on the contents of b_2. Either the Predictor has predicted that you'll take one box or he predicted you'll take two boxes. Suppose he has predicted that you'll take one box. Then there's \$1,000,000 in b_2 and \$1,000 in b_1. If you take both, you'll get \$1,001,000; if you take b_2 alone, you'll get \$1,000,000. So you should take both boxes. Suppose now that the Predictor has predicted that you'll take both boxes. That means there is nothing in b_2. If you take both, you'll get \$1,000; if you take b_2 alone, you'll get nothing. So you should take both boxes. So no matter what the Predictor has predicted, you'll be better off if you take two boxes rather than one.

A reasonable requirement on any "solution" to Newcomb's Problem is that it not only tell us what the right choice is in Newcomb's Problem, but also that it explain why the other option looks attractive. So the two-boxing argument aired above does not by itself constitute a solution to the Problem. It needs to be supplemented by an explanation of why one-boxing possesses any intuitive appeal at all.

That one-boxing possesses substantial pre- and post-theoretical appeal is hard to deny. Certainly, some philosophers have defended taking one box in the Newcomb case (e.g. Bar-Hillel and Margalit (1972), Horgan (1981) and Leslie (1991)). And when Newcomb's Problem was presented in *Scientific American* in 1974, of the first 148 letters, 89 people said they would take one box, while only 37 people thought two-boxing was the correct option, a ratio of about 2.5 to 1 (the remaining group rejected the problem for one reason or another).[3] (This is after readers were presented with the argument in favor of two-boxing.) So Newcomb's Problem gives rise to disagreement between persons.

Even if one finds the argument for two-boxing compelling (as I do), it is still difficult not to feel at least *tempted* by the one-boxing option in Newcomb's case. There is deliberative tension in the Newcomb case, a feeling of being pulled in two directions. I feel especially pulled towards one-boxing when I do my best

[1] Let us be more specific about the Predictor's great success. The Predictor has made two thousand predictions prior to your choice. Of those, one thousand people took one box and the Predictor predicted this 99% of the time; the other thousand took two boxes and the Predictor predicted this 99% of the time. This additional information is important to circumvent an objection to the case made by Isaac Levi (1982).

[2] The case is due to the physicist William Newcomb, but was introduced into the literature by Nozick (1969).

[3] See Nozick (1974).

to vividly imagine actually being in the Newcomb situation, facing down the two boxes. I'm tempted simply to ignore the above argument for taking both boxes and grab only the opaque box, confident that if I do so, I'll walk away a millionaire. So Newcomb's Problem gives rise to intra-personal tension, when two-boxers discover the one-boxer within.

In this paper, I take it for granted that the two-boxing argument is correct; my interest is in the question of why one-boxing possesses any intuitive appeal at all.

1.2 The Great Decision Theory Debate

As is well-known, Newcomb's Problem gave rise to a debate over what the right decision theory is, and it is natural to look to this debate for an answer to our question. Newcomb's Problem is seen by two-boxers to be a counterexample to Richard Jeffrey's formulation of *evidential decision theory* (Jeffrey 1983), and is used to motivate an alternative theory of rational decision—*causal decision theory*.[4] Evidential decision theory is a way of calculating the expected utility of an action. According to evidential decision theory, the expected utility of an action is the sum of the weighted utilities of the possible outcomes; the utility of each outcome is weighted by the agent's subjective conditional probability that the outcome will obtain, given that the action is performed.

We assume a relevant space of possible worlds, and construe actions as propositions (sets of possible worlds) that the agent can make true, and think of states simply as propositions. Both the set of actions and the set of states form partitions on the space of possibilities. For any action A and state partition $\mathbf{S}$, evidential decision theory calculates the expected utility of A as follows:

$$\mathrm{V}(A) = \sum_{S} \mathrm{P}(S|A)\mathrm{u}(A \wedge S) \tag{EDT}$$

(where $\mathrm{V}(A)$ is the evidential expected utility of A, and outcomes are understood simply as act-state conjunctions). Bayesians take $\mathrm{P}(S|A)$ as an indication of what the agent's degree of confidence in S would be were he to learn A. Thus, $\mathrm{V}(A)$ is a measure of the extent to which learning that A is true, i.e. that he is about to make A true, would provide the agent with evidence that desirable outcomes will ensue. According to evidential theory, we should perform the action we'd be happiest to learn that we were about to perform, as if the fact that the action was about to be performed came as news to us. For this reason, $\mathrm{V}(A)$ is sometimes called the *news value* of A.

Why do two-boxers take Newcomb's Problem to be a counterexample to Jeffrey's decision theory? The reason for this is that on the most straightforward application of this theory to Newcomb's Problem, the theory tells us to take b_2

[4] Jeffrey's own position is a bit more complicated, and it is not clear that Newcomb's Problem should be regarded as a counterexample to his interpretation of EDT. Instead, we should really say that the Problem is a (potential) counterexample to a certain well-known version of EDT, a version which may not be Jeffrey's own.

alone.[5] Why is this? Given the setup of Newcomb's Problem, that an agent will take b_2 alone is good *evidence* that there is \$1,000,000 in b_2; whereas, that an agent will take both boxes is good evidence that b_2 is empty. (Imagine yourself watching another decision maker facing Newcomb's choice. Suppose you learn that he is about to take both boxes. Would you bet on, or against, there being \$1,000,000 in b_2?) Suppose, for example, that P(b_2 *contains \$1,000,000*| *The agent takes one box*) = 0.99, and suppose utility is proportional to money. Let A_1 be the proposition that the agent takes b_2 alone; let A_2 be the proposition that the agent takes both boxes; and let M be the proposition that there is \$1,000,000 in b_2. Then, the evidential expected utility of taking b_2 alone is:

$$\begin{aligned} \mathrm{V}(A_1) =& \mathrm{P}(M|A_1) \times \mathrm{u}(M \wedge A_1) + \mathrm{P}(\neg M|A_1) \times \mathrm{u}(\neg M \wedge A_1) \\ &= (0.99 \times \mathrm{u}(\$1,000,000)) + (0.01 \times \mathrm{u}(\$0)) \\ &= 990,000 \end{aligned}$$

Two-boxing, on the other hand, has the following evidential expected utility:

$$\begin{aligned} \mathrm{V}(A_2) =& \mathrm{P}(M|A_2) \times \mathrm{u}(M \wedge A_1) + \mathrm{P}(\neg M|A_2) \times \mathrm{u}(\neg M \wedge A_2) \\ &= (0.01 \times \mathrm{u}(\$1,001,000)) + (0.99 \times \mathrm{u}(\$1,000)) \\ &= 10,010 + 990 \\ &= 11,000 \end{aligned}$$

Thus, the evidential expected utility of taking b_2 alone exceeds that of taking both boxes. Since evidential decision theory counsels decision makers to maximize evidential expected utility, it counsels agents in the Newcomb case to take b_2 alone.

Two-boxers seek to build a decision theory that respects the causal intuition that undergirds the two-boxing argument. The causal intuition is that since my choice has no effect on the contents of b_2, the fact that one-boxing is highly correlated with getting \$1,000,00 and that two-boxing is highly correlated with getting \$1,000 should be irrelevant to my decision. Jeffrey's decision theory ignores this intuition, since the relevant notions of dependence and independence his theory employs are not causal but *probabilistic* (= evidential). What one does in the Newcomb case is causally, but not probabilistically, independent of the relevant states of the world. But this, according to two-boxers, is precisely the problem with evidential decision theory.

The basic idea behind causal decision theory is that rational agents ought to perform the action that has the best causal consequences. We formulate causal decision theory in terms of the counterfactual/subjunctive conditional,

[5] Some defenders of evidential decision theory have argued that 'the most straightforward' application of their theory to Newcomb-type problems is too naive. Once a more nuanced understanding of these cases is in place, evidential decision theory does not yield recommendations that contradict causal decision theory. See Eells (1982, 1984) and Jeffrey (1983). For some responses to this move, see Lewis (1981), Sobel (1994), and Joyce (1999).

following Stalnaker (1981) and Gibbard and Harper (1978).[6] Read '...>...' as 'If it were the case that..., it would be the case that...'. The counterfactual must be understood in the 'causal', rather than the 'back-tracking' sense.[7] Here is our formulation:

$$\mathrm{U}(A) = \sum_S \mathrm{P}(A > S)\mathrm{u}(A \wedge S) \tag{CDT}$$

In this formulation, the A's are the relevant propositions which the agent is able to make true; the S's form a *rich partition* in the sense of Lewis (1981, 317) (i.e., each $A \wedge S$ completely determines an outcome that the agent cares about). The idea behind using counterfactuals in the formulation of causal decision theory is that counterfactuals appropriately track causal influence.

The probability that there is \$1,000,000 in b_2 on the subjunctive supposition that I take both boxes should just be my unconditional subjective probability in the proposition that there is \$1,000,000 in b_2, because whether or not the money is in b_2 is independent of my actions. Indeed, one would think that $\mathrm{P}(A_1 > M) = \mathrm{P}(A_2 > M) = \mathrm{P}(M)$ (where A_1, A_2, and M stand, as before, for *The agent takes* b_2 *alone*, *The agent takes both boxes*, and *There is \$1,000,000 in* b_2 respectively). The reader can verify that under these conditions, the causal expected utility of two-boxing exceeds that of one-boxing.[8]

1.3 Evidential Decision Theory and One-Boxing

Given the cogency of the two-boxing argument, what explains the appeal of one-boxing? In light of the foregoing discussion, the answer might be thought to be that one-boxing is the action with the greatest news value. Perhaps what explains the one-boxing intuition is simply that we do have intuitions that support evidential decision theory.

I have two things to say about this suggestion: first, it is not clear that causal decision theorists can accept it; and second, it is objectionable on independent grounds.

First point: Although some causal decision theorists themselves (e.g. Gibbard and Harper (1978, 183) and Joyce (1999, 154)) have accepted the idea that EDT explains the appeal of one-boxing, it is not clear that they should. I say this because of the role intuitions about cases have played in the dialectic between evidential and causal decision theory. In particular, intuitions about Newcomb-type problems have been important in motivating causal decision theory. But if the one-boxing intuition just *is* an evidential decision theory intuition, then it seems difficult to see why causal decision theorists are justified in ignoring

[6] Other formulations of causal decision theory, which are more or less equivalent, are possible. See Lewis (1981) and Joyce (1999, Ch.5) for discussion.

[7] See Lewis (1975, 1981).

[8] See Gibbard and Harper (1978, 181) for details. The proof relies on picking a fixed but arbitrary value for $\mathrm{P}(M)$, something Levi objects to. See Skyrms (1990) and Collins (2001) for discussion.

the maxim *maximize evidential expected utility*, since that maxim enjoys some intuitive support.

Nozick (1993, 41-50) takes this line of thought seriously, and argues that the task of the decision theorist is not to decide between causal and evidential decision theory, but rather to incorporate both causal and evidential intuitions into a single decision theory. Reflection on certain variations of Newcomb's Problem leads Nozick to endorse a 'hybrid' decision theory of this sort. To make this move is to give up the sort of 'pure' causal decision theory that has attracted many philosophers since the discovery of Newcomb's Problem.

I think that a causal decision theorist who wants to explain the appeal of one-boxing by appealing to its higher news value need not go Nozick's route. But the initial move by causal decision theorists of explaining the attraction of one-boxing by appeal to news value is in any case *prima facie* puzzling, given the dialectical context. One-boxers and two-boxers argue over what the right decision in Newcomb's Problem is. Each side gives its respective arguments for the choice it recommends (the arguments I gave in §1.1). Now after these initial arguments are produced, either side might try to bolster its position by showing how its theory has the resources to explain the appeal of the opposing view. But such an explanation is dialectically ineffective if it appeals to the *very considerations* the opponent relies on in her initial case. So causal decision theorists have a reason to reject the idea that EDT is the source of the one-boxing intuition.

Second point: We have independent reason to reject the idea that evidential decision theory is the source of the one-boxing intuition. The main reason for rejecting this explanation of the appeal of one-boxing is that there are a variety of cases that are structurally identical to Newcomb's Problem, but in which the EDT-supported option has no intuitive appeal. But if evidential decision theory were the source of the one-boxing intuition in Newcomb's case, then the EDT-supported option ought to have appeal in these other cases as well. Since it doesn't, EDT isn't the source of the one-boxing intuition.

What sort of cases do I have in mind? I will discuss two variations on Newcomb's Problem in §3 which are structurally identical to the original case, but in which one-boxing is not at all appealing. But there is another sort of case which makes the point more clearly. Here is a typical example of a 'medical' Newcomb case:

> It is discovered that the reason for the correlation between smoking and lung cancer is not that smoking tends to cause lung cancer. Rather, the cause of lung cancer is a certain genetic factor, and a person gets lung cancer if and only if he has that factor. Everyone enjoys smoking; but the the reason for the correlation of lung cancer with smoking is that the same genetic factor gives one an *effective desire* to smoke, i.e. a desire that is effective in getting one to smoke. You are a smoker who knows these facts and are trying to decide whether to give up smoking. You like to smoke, but you want much more to avoid cancer than to continue to smoke. What should you

do?[9]

Now almost no one thinks that it would be rational to refrain from smoking in this case. And yet, if evidential decision theory is correct, then one ought to refrain from smoking. Refraining from smoking is to one-boxing as smoking is to two-boxing: EDT recommends refraining from smoking and one-boxing, whereas CDT recommends smoking and two-boxing. But since refraining from smoking has no intuitive pull here, the fact that EDT recommends one-boxing can't explain its appeal.

The point carries over to the one-boxing argument I offered at the beginning of the paper. That argument can be set out as follows:

1. If I take b_2 alone, I'm very likely to get $1,000; but if I take both boxes, I'm very likely to get $1,000,000.

2. If (1), then I should take b_2 alone.

3. So, I should take b_2 alone.

Does this argument explain the intuitive appeal of one-boxing? I don't think so. For consider the parallel argument in the smoking case:

$1'$. If I refrain from smoking, it is very unlikely that I will get lung cancer; but if I smoke, I'm very likely to get lung cancer.

$2'$. If ($1'$), then I should refrain from smoking.

$3'$. So, I should refrain from smoking.

But this latter argument seems like a terrible reason to refrain from smoking: ($2'$) just seems false. So neither EDT nor the simple one-boxing argument is what explains the one-boxing intuition, at least for those who agree that refraining from smoking in the smoking case is irrational.

Causal decision theorists have theory-internal reasons not to accept the claim that EDT is what explains the one-boxing intuition; all of us have an independent reason to reject that claim as well. But in spite of the argument I just gave against the claim that EDT is the source of the one-boxing intuition, I think the lack of a clear, causal decision-theoretic explanation of the appeal of one-boxing is one of the reasons evidential decision theory still lingers around as a live option.[10] The goal of this paper is to fill this gap and provide a compelling causal decision theoretic explanation of the one-boxing intuition in the original Newcomb Problem.

[9] Adapted from Gibbard (1979) as cited in Horgan (1981, 178).

[10] Another major reason – which is to some extent in tension with the reason just given – is the existence of the 'Tickle Defense'. See footnote 5 for references.

2 The Epistemic Explanation

So how should causal decision theorists explain the appeal of one-boxing in the original Newcomb Problem? My view is that the one-boxing intuition has its source in the fact that it wouldn't be epistemically rational for an agent in a Newcomb Problem to be *certain* that she was in a genuine Newcomb case. That is, an agent in a Newcomb Problem ought to give some amount of credence to the hypothesis that her choice will cause have a causal effect on the contents of b_2, i.e. the hypothesis that one-boxing will cause b_2 to contain \$1,000,000 and that two-boxing will cause b_2 to be empty. I argue for this claim in §2.1. This is connected to the one-boxing intuition, since, as I show in §2.2, causal decision theory recommends one-boxing if the agent has a small but non-negligible degree of belief in this *causal hypothesis*. The upshot of these two claims is that CDT recommends one-boxing for an epistemically rational agent in Newcomb's Problem. My claim then is that we have the one-boxing intuition because one-boxing is (according to CDT) the *all things considered* rational action in Newcomb's Problem. In §3, I offer what I take to be the best evidence for this approach: the explanation correctly predicts that our intuitions about the case will be unstable across certain variations on the original Problem, variations which alter some of the parameters of the case without altering the Problem's decision-theoretic structure.

My claim that the one-boxing intuition is tracking what it would be rational *tout court* for an agent in a Newcomb Problem to do. But this claims raises a few questions: Why do our intuitions track this fact? Why do we also have the two-boxing intuition? And: isn't it a stipulation of the case that the agent is *certain* that she doesn't have causal control over the contents of b_2? I take these questions up in §4.

2.1 Credence and the Causal Hypothesis

Here is how we tend to imagine what it's like to be in a Newcomb case: You enter a room and are greeted by a credible-looking person who tells you the following:

> You have a choice between taking two boxes, b_1 and b_2, and taking b_2 alone. b_1 is transparent and contains \$1,000. b_2 is opaque and contains either \$1,000,000 or \$0. The contents of b_2 are determined as follows. A powerful Predictor has observed each person who will face this choice for the past week. Based on his observations, he has made a prediction as to what each person will do. If he predicted that a particular person will take both boxes, he left the opaque box b_2 empty. If he predicted that that person will take b_2 alone, he put \$1,000,000 in it. Almost everyone who took b_2 alone got \$1,000,000; almost everyone who took both b_1 and b_2 got \$1,000 (i.e. b_2 was empty).

You then watch two thousand people go before you, choosing between the two options after the credible-looking person tells each one what she told you. You observe that of the one thousand people who take both boxes, all but (say) ten of them leave the room with exactly \$1,000 more than they had when they entered, and that of the one thousand people who take the opaque box b_2 alone, all but ten of them leave the room with exactly \$1,000,000 more than they had when they entered.

What should you believe about the situation you are in? It seems that there are two salient hypotheses about the causal structure of the situation for you to consider: first, there is the *official story*, which is the hypothesis contained in the testimony of the credible-looking person; and second there is the *causal hypothesis* which says that taking b_2 alone causes it to contain \$1,000,000, whereas taking both boxes causes b_2 to be empty.

My main contention in this section is that you should give a small but non-negligible amount of credence to the causal hypothesis. Why do I say this? In the Newcomb case, you know that certain correlations hold: one-boxing is highly correlated with getting \$1,000,000 and two-boxing is highly correlated with getting \$1,000. How you distribute your credence over the official story and the causal hypothesis will reflect how you think these correlations should be explained. What I wish to argue is that, for almost any way of specifying your evidence for the official story that we can actually imagine, it would still be irrational for you not to give some credence to the hypothesis that the correlations are explained by the causal hypothesis.

It is important to keep in mind that the conclusion we wish to establish is *not* that, on balance, you should find the official story less plausible than the causal hypothesis. It will be enough for my purposes if it is rationally permissible for you to give some small but non-negligible amount of credence to the causal hypothesis. And this weak claim seems hard to deny.

The description of Newcomb's Problem I gave at the beginning of this section is fairly representative of how the Problem is usually presented in the literature. It should be clear, I think, that relative to normal ways of filling in the details of this case you should have less than full credence in the testimony of the credible-looking person. Why? The reason is simple: the existence of a Predictor who is able to predict, with great accuracy, human choices in this sort of situation is *prima facie* implausible. If the testimony, combined with the evidence you gain from watching what happens to the choosers who go before you, is all you have to go on, then it seems obvious that conditionalizing on this evidence won't lead a rational person to a degree one belief that the official story is true. For there is, after all, another salient alternative hypothesis immediately suggests itself: that taking b_2 alone causes that box to contain \$1,000,000 and taking both boxes causes it to be empty. Relative to this body of evidence, it seems obvious that one ought to give the causal hypothesis some small amount of credence.

There are several ways the causal hypothesis could be true, and you might give each of these ways a small amount of credence. For example, one way the causal hypothesis could be true is if there is merely some clever 'cheating' going on, a bit of sleight-of-hand. This possibility, one would think, should get *some* credence, even if only a very small amount. One reason to give cheating some credence is that it seems relatively easy to accept that someone can cleverly fool you in the relevant way. We've all seen this sort of thing before, since we've all seen magicians and con-men. Any moderately street-wise person will not be certain that the official story is true.

A natural response to this suggestion is that we have simply under-described the case. After all, one could imagine having much more evidence that there is no cheating going on. For example, maybe when the other participants choose, you get to stand behind the boxes and b_2 has a transparent back, so that you can see whether or not it is empty. And then when your turn comes, some trusted friends stand behind the box and promise to alert you if there's any funny business going on. This evidence might be good enough to make you certain that there was no cheating going on. But even in this evidentially rich scenario, you might still have non-negligible credence in the causal hypothesis, though not because you think there's any cheating going on. I will suggest two other reasons for giving some credence to the causal hypothesis.

First, you may find something plausible about the following line of thought:

> If the official story is true, then the Predictor can reliably predict what I'm going to choose—indeed it seems that he *knows* what I'm going to do. But how can he? I haven't even *decided* what I'm going to do, and prior to my deciding what to do, there is no fact of the matter about what I'm going to do, since I, by my decision, create the relevant fact. So, right now, there is no relevant fact to be known. So the official story *can't* be true, since there simply can't be such a Predictor.

I think all of us find this sort of thought at least *tempting* when we consider our decision from our deliberative perspective. Given that it's up to me what I'm going to do, and since I haven't decided what to do yet, how can the Predictor *already know* what I'm going to do? If you find this thought (or something close to it) at all plausible, then you should give some amount of credence to it. But to do that is to give some amount of credence to the impossibility of the official story. If the story is impossible, then it isn't actual, and if it isn't actual, then you're not in a genuine Newcomb case. In that case, you ought to give some credence to the causal hypothesis, since it is the main salient alternative to the official story. Note that one can give the causal hypothesis some credence even if one has no idea what the relevant causal mechanism is; one can have reason to suspect that there is a causal relation between two factors even if one has no idea what *explains* the relation.

The second reason you might give some credence to the causal hypothesis – a reason which is consistent with the reason just mentioned – is that you might have non-negligible credence in the hypothesis that there is some sort of backwards causation going on. Perhaps your choice determines what the Predictor predicts even though the prediction occurs prior to the choice. Backwards causation may seem like a far-out possibility; perhaps it is, but think about the situation you're in. If you're prepared to grant the possibility of the Predictor story, then the possibility of backwards causation shouldn't sound crazy to you. Given the apparently exotic situation you're in, I think backwards causation would become a salient possibility to which you might give *some* small amount credence. Again, you might be fairly confident that you are in a genuine Newcomb case, since you might find the idea of backwards causation implausible. We can concede that: all I want to point out is that it would probably be reasonable for you to give *some* small amount of credence to the possibility of backwards causation.

I want to stress one last time that the point of considering these possibilities is not that, on the balance of all your evidence, you ought to *believe* the causal hypothesis. I'm not arguing for anything nearly that strong. All I want to point out is that these various lines of support for the causal hypothesis make it reasonable for you to be less than completely certain that the official story is true. More could be said about all this, but I think it's clear that you ought to give the causal hypothesis *some* credence, even if it isn't clear how much credence you ought to give it. Although it is not possible to say what precise credences one ought to attach to each hypothesis, we can offer the following constraint on a rational agent's credences: a rational agent in a Newcomb case will have beliefs that can be represented by a probability function P, such that $0.5 <$ P(official story) < 1 (and so, $0 <$ P(causal hypothesis) < 0.5). One ought to think that the official story is more likely to be true than the causal hypothesis, but one shouldn't be certain that it's true. It is important to note that this constraint is intended only to rule out unacceptable credence functions, and that not any credence function that meets it will be acceptable. Important for our purposes is the claim that one ought to give a *non-negligible* (and not merely non-zero) amount of credence to the causal hypothesis; in the next subsection, we'll attempt to make this claim a bit more precise.[11],[12]

2.2 How Plausible Does the Causal Hypothesis Have to Be?

The argument for two-boxing that I presented at the beginning of the paper relied on the premise *My choice will not affect the contents of* b_2. But reasoning from that premise would be illegitimate for an agent who thinks that there is some chance that her choice will affect the contents of b_2. And as I've been suggesting, when we imagine being the Newcomb agent, this is the type of agent

[11] I slide between saying that you are rationally permitted to give the causal hypothesis some small amount of credence and saying that you are rationally required to do so. If epistemic permissibility comes apart from epistemic obligation, then these two claims are different. But I think I can give my underlying explanation of the one-boxing intuition in either case.

[12] The motivating idea behind the epistemic explanation of the one-boxing intuition is not without precedent. See Bar-Hillel and Margalit (1972, 300-302), Mackie (1977), McKay (2004), and Schmidtz and Wright (2004). Most of these authors focus on which of the official story and the causal hypothesis is on balance more plausible, and most are more concerned with the justification of one-boxing, rather than the explanation of the one-boxing intuition. None of these authors consider variation cases discussed in §3.

we imagine being. How we answer the question of what such an agent ought to do in Newcomb's Problem depends on how the agent ought to distribute her credence over the various possibilities about the causal structure of her situation.

What it is rational to do in the Newcomb situation depends on how you should distribute your credence between the official story and the causal hypothesis. I've argued that one shouldn't believe the official story to degree 1; one should place some amount of credence in the causal hypothesis as well. We now want to know what someone with the correct credence distribution should do in a Newcomb case. We approach this by first asking the question: what degree of belief in the causal hypothesis is enough to make one-boxing rational? Once we know this, we will be in a position to determine whether an agent with a responsible credence distribution would take one or two boxes in the Newcomb case.

Recall our earlier formulation of causal decision theory:

$$\mathrm{U}(A) = \sum_{S} \mathrm{P}(A > S)\mathrm{u}(A \wedge S) \tag{CDT}$$

We will need a slightly different (but equivalent) formulation for our purposes. Note that one can be uncertain whether or not one can causally influence some outcome that one cares about. In that case, one might have a few different hypotheses about what the causal structure of the world is like. Following Brian Skyrms (1980, 136-138), let the H's be the agent's various hypotheses about what outcomes might be in her control. Then (CDT) is equivalent to:

$$\mathrm{U}(A) = \sum_{H} \mathrm{P}(H) \sum_{S} \mathrm{P}(A > S|H)\mathrm{u}(A \wedge S \wedge H) \tag{CDT}$$

(We will assume the H's will be chosen so that $\mathrm{u}(A \wedge S \wedge H) = \mathrm{u}(A \wedge S)$.)[13]

[13] Readers familiar with Lewis's (1981) version of causal decision theory might wonder what the relationship is between Skyrms's H-hypotheses and Lewis's *dependency hypotheses*. Lewis's dependency hypotheses are equivalent to maximally specific consistent conjunctions of counterfactuals, so that, where D is a Lewisian dependency hypothesis, $\mathrm{P}(A > S|D)$ is always one or zero (since D either entails, or is incompatible with, $A > S$). Think for a moment of the official story and the causal hypothesis, both of which are H-hypotheses. Let the official story be H_1. H_1 is true just in case the agent's choice doesn't influence the contents of the opaque box b_2. In terms of counterfactuals ('$\implies$' is entailment):

$$H_1 \implies (A_1 > M \wedge A_2 > M) \vee (A_1 > \neg M \wedge A_2 > \neg M)$$

Each disjunct is a dependency hypothesis; H_1 entails that one of those dependency hypotheses is true. But the entailment doesn't go the other way; H_1 isn't simply equivalent to the disjunction of these two dependency hypotheses. Why not? Because even if the causal hypothesis (H_2) is true, either of these dependency hypotheses might be true (though each is very unlikely to obtain if the causal hypothesis is true).

If the causal hypothesis (H_2) obtains, the most probable dependency hypothesis is: $(A_1 > M \wedge A_2 > \neg M)$. But this isn't the only possibility. There are three others, all of which might result if the causal mechanism messes up:

- $A_1 > M \wedge A_2 > M$
- $A_1 > \neg M \wedge A_2 > \neg M$
- $A_1 > \neg M \wedge A_2 > M$

As before, we let A_1 stand for *You take b_2 alone*, A_2 for *You take both boxes*, and M for *There is \$1,000,000 in b_2*. And we let H_1 stand for *The official story is true*, and H_2 for *The causal hypothesis is true.* Let $\mathrm{P}(M|H_1) = \mu$.[14]

Given the payoffs, what is the smallest number n $(0 < n < 1)$ such that if $\mathrm{P}(H_2) > n$, then $\mathrm{U}(A_1) > \mathrm{U}(A_2)$? That is, what degree of belief do you need to have in the causal hypothesis in order to make one-boxing the rational choice? We assume that money is proportional to utility and that 99% of one-boxers get \$1,000,000, 99% of two-boxers get exactly \$1,000.

To answer this question, we first set $\mathrm{U}(A_1)$ equal to $\mathrm{U}(A_2)$. Since the H's form a partition, we have $\mathrm{P}(H_1) + \mathrm{P}(H_2) = 1$. This gives us two equations with two unknowns, namely $\mathrm{P}(H_1)$ and $\mathrm{P}(H_2)$. Solving for $\mathrm{P}(H_2)$ will give us the n such that if $\mathrm{P}(H_2) = n$, then the causal expected utility of one-boxing is equal to that of two-boxing. Thus, when $\mathrm{P}(H_2) > n$, the causal expected utility of one-boxing exceeds that of two-boxing.

$\mathrm{U}(A_1)$ is the sum of the following two summands:

1. $\mathrm{P}(H_1)\big(\mathrm{P}(A_1 > M|H_1)\mathrm{u}(M \wedge A_1) + \mathrm{P}(A_1 > \neg M|H_1)\mathrm{u}(\neg M \wedge A_1)\big)$

2. $\mathrm{P}(H_2)\big(\mathrm{P}(A_1 > M|H_2)\mathrm{u}(M \wedge A_1) + \mathrm{P}(A_1 > \neg M|H_2)\mathrm{u}(\neg M \wedge A_1)\big)$

Assume: $\mathrm{P}(A_1 > M|H_1) = \mathrm{P}(M|H_1) = \mu$; $\mathrm{u}(M \wedge A_1) = 1,000,000$; and $\mathrm{u}(\neg M \wedge A_1) = 0$. Then, (1) is equivalent to:

1′. $\mathrm{P}(H_1)\mu(1,000,000)$

Assume: $\mathrm{P}(A_1 > M|H_2) = 0.99$; $\mathrm{P}(A_1 > \neg M|H_2) = 0.01$; $\mathrm{u}(M \wedge A_1) = 1,000,0000$; and $\mathrm{u}(\neg M \wedge A_1) = 0$. Then (2) is equivalent to:

2′. $\mathrm{P}(H_2)(990,000)$

So $\mathrm{U}(A_1) = \mathrm{P}(H_1)\mu(1,000,000) + \mathrm{P}(H_2)(990,000)$.

$\mathrm{U}(A_2)$ is the sum of the following two summands:

Each of these is possible on the causal hypothesis, just extremely unlikely. They're all possible because not everyone who took one box got \$1,000,000 and not everyone who took two got only \$1,000. And there seems to be no reason to exclude any particular one of them from being possible. Each of the three is possible, because the causal mechanism might fail in your case.

So Skyrms's H's are not Lewisian dependency hypotheses. One way to think of the H's is that they tell you how to distribute your credence over dependency hypotheses. They don't tell you *exactly* how to do this, but they offer constraints. For example, the official story tells you to distribute all of your credence over two possibilities: the possibility that there is \$1,000,000 in b_2 no matter what you do and the possibility b_2 will be empty no matter what you do. But the official story itself doesn't tell you how much credence to give to each of these dependency hypotheses—it simply tells you to give each some (non-zero) credence. Skyrms's H-hypotheses are more useful than Lewis's dependency hypotheses for our purposes, since we want to know how probable the causal hypothesis needs to be in order to make one-boxing rational, and neither the causal hypothesis nor the official story are dependency hypotheses.

[14] I follow Gibbard and Harper in picking a fixed, arbitrary value μ for $\mathrm{P}(M|H_1)$ (though they use μ as the unconditional probability of M, since they aren't considering multiple H-hypotheses). See footnote 8 on this assumption.

3. $\mathrm{P}(H_1)\big((\mathrm{P}(A_2 > M|H_1)\mathrm{u}(M \wedge A_2)+\mathrm{P}(A_2 > \neg M|H_1)\mathrm{u}(\neg M \wedge A_2)\big)$

4. $\mathrm{P}(H_2)\big((\mathrm{P}(A_2 > M|H_2)\mathrm{u}(M \wedge A_2))+\mathrm{P}(A_2 > \neg M|H_2)\mathrm{u}(\neg M \wedge A_2)\big)$

Assume: $\mathrm{P}(A_2 > M|H_1) = \mathrm{P}(M|H_1) = \mu$; $\mathrm{u}(M \wedge A_2) = 1,001,000$; and $\mathrm{u}(\neg M \wedge A_2) = 1,000$. Then (3) is equivalent to:

3′. $\mathrm{P}(H_1)\big(\mu(1,001,000) + (1 - \mu)(1,000)\big)$

And assume: $\mathrm{P}(A_2 > M|H_2) = 0.01$; $\mathrm{P}(A_2 > \neg M|H_2) = 0.99$; $\mathrm{u}(M \wedge A_2) = 1,001,000$; and $\mathrm{u}(\neg M \wedge A_2) = 1,000$. Then (4) is equivalent to:

4′. $\mathrm{P}(H_2)(11,000)$

So $\mathrm{U}(A_2) = \mathrm{P}(H_1)\big((\mu(1,000,000) + 1,000)+ \mathrm{P}(H_2)(11,000)\big)$

If the causal expected utility of taking both boxes was equal to that of taking one box, we'd have:

$$\mathrm{P}(H_1)(\mu(1,000,000))+\mathrm{P}(H_2)(990,000) = \mathrm{P}(H_1)(\mu(1,000,000) + 1,000) + \mathrm{P}(H_2)(11,000))$$

which simplifies to:

$$\mathrm{P}(H_2)(979,000) = \mathrm{P}(H_1)(1,000)$$

This gives us our first equation. Since $\{H_1, H_2\}$ is a partition, $\mathrm{P}(H_1) = 1-$ $\mathrm{P}(H_2)$ is our second equation. Thus,

$$\mathrm{P}(H_2)(979) = 1- \mathrm{P}(H_2)$$

$$\mathrm{P}(H_2) = \frac{1}{980} \approx 0.00102$$

So if your subjective probability for the causal hypothesis is any amount greater than 0.00102, then causal decision theory says that it is rational for you to take only b_2, the opaque box. A very slim chance that your action might causally influence the contents of the opaque box is enough to make it rational for you to take only that box. Upon reflection, this isn't surprising, of course. If there is a small chance that performing an action will bring about a large payoff, it is often worth taking the chance. It all depends on how big the small chance is, how large the payoff is, and what your other options are. More specifically: if the official story is true, then causal decision theory makes U(two-boxing) greater than U(one-boxing) by $1,000; but if the causal hypothesis is true, causal decision theory makes U(one-boxing) greater than U(two-boxing) by $979,000. So P(official story) has to be very large in order for two-boxing to come out on top.

In §2.1 we argued that an epistemically rational agent in a Newcomb Problem would not be certain of the official story: she would give a non-negligible amount of credence to the causal hypothesis. If we assume that 'non-negligible' means 'greater than 0.00102', then we have the result that an epistemically rational agent in a Newcomb Problem ought to one-box (assuming the truth of CDT). The precision of the number is misleading of course; the thing to focus on is the underlying explanation: a small amount of doubt about the official story is sufficient to make one-boxing the CDT-recommended choice in Newcomb's Problem. Thus, we conjecture that the intuitive appeal of one-boxing can be traced to the fact that, according to CDT, the all-things-considered rational act in Newcomb's Problem is to take b_2 alone.

3 Newcomb Variations

In this section, we present further evidence for the epistemic explanation, by way of two variations on Newcomb's Problem. Our intuitions about these cases seem puzzling on the assumption that evidential decision theory is what gives rise to the one-boxing intuition, since the cases are structurally identical to Newcomb's Problem, and yet most people do not find one-boxing appealing in these cases. As we will see, the epistemic explanation explains why our intuitions are unstable across these variations. What the variation cases strongly suggest is that our intuitions are tracking the choices of an epistemically rational agent who is guided by causal decision theory. The phenomena we are about to discuss are rather puzzling, and the few explanations of them in the literature are unsatisfying.[15] Thus, the ability of the epistemic explanation to handle these cases is perhaps the best argument for its truth.

3.1 Variation #1: Messing with the Money

Nozick (1993, 44-45) notes the following interesting Newcomb-related phenomenon: that changing the amount of money in the boxes – changing the ratio between the amounts – affects our intuitions about what to do in a surprising way. The basic phenomenon is this: Let Y be the amount of money in b_1 and let X be the amount of money the Predictor puts in b_2 if he predicts that you'll take b_2 alone. Then as $X - Y$ gets smaller, the appeal of one-boxing decreases. And as $X - Y$ gets larger, the appeal of two-boxing decreases.

Consider the case in which $X - Y$ gets smaller. Suppose X is 1,000,000 as before, but that b_1 now contains $900,000. If utility is proportional to money, then EDT recommends one-boxing as long as b_1 contains less than $980,000

[15] The only discussions of these phenomena that I know of are Nozick (1993, 44-48), which discusses Variation #1, and Hurley (1994) which discusses both. Nozick advocates a 'hybrid' decision theory which incorporates both evidential and causal elements; this move strikes me as inelegant and *ad hoc*. Hurley's solution requires imposing an interpretation on Necomb's Problem that goes far beyond what is stipulated in the case—in particular, it requires assuming that the Predictor has a particular preference ranking over the outcomes. Hurley's proposal, like Nozick's, also requires us to essentially give up 'pure' causal decision theory; from my point of view, this is a heavy cost of accepting either of them.

(Nozick 1993, 44). But if b_1 contains \$900,000, hardly anyone would pass it up and take b_2 alone. I'm tempted by the one-boxing option in the original Newcomb Problem, but I'm not tempted by it here.

The epistemic explanation actually predicts that we will react to this case in precisely this way. To see this, note the following ways in which the causal expected utility of your options depends on certain parameters of the case:

i. If the amount in b_1 (i.e. Y) goes up, U(two-boxing) goes up (and U(one-boxing) stays the same).

ii. If P(causal hypothesis) goes up, U(one-boxing) goes up (and U(two-boxing) goes down).

The reason (i) holds is that two-boxing has two possible outcomes: either you will get $\$Y$ (the amount in b_1), or you will $\$(X + Y)$. So if Y increases, then both possible outcomes of two-boxing go up, and so U(two-boxing) has to go up. But since you simply don't get what b_1 contains if you take b_2 alone, the amount in b_1 (Y) has no effect on the two possible outcomes of one-boxing. So U(one-boxing) stays the same as Y increases. The reason (ii) holds is simple: the more confident you are that one-boxing causes b_2 to contain \$1,000,000, the higher the causal expected utility of one-boxing.

Since Y in the variation case is greater than Y in the original case, U(two-boxing) in the variation is greater than U(two-boxing) in the original case. Since all other factors are held fixed, U(one-boxing) in the variation case is equal to U(one-boxing) in the original case, *unless P(causal hypothesis) goes up*, i.e. unless P(causal hypothesis) in the variation case is greater than P(causal hypothesis) in the original case..

In §2.2 we learned that, in the original case, when P(causal hypothesis) $>$ 0.00102, then U(one-boxing) $>$ U(two-boxing). But since in this variation case, U(two-boxing) has gone up, the only way for U(one-boxing) to exceed U(two-boxing) in this case is for P(causal hypothesis) to also go up (since we are holding all other factors fixed). It turns out that P(causal hypothesis) will have to go up a lot if U(one-boxing) is to exceed U(two-boxing) in the variation case. The calculation to determine this is the same as the one we did in §2.2 except that $\mathrm{u}(M \wedge A_2) = 1,900,000$ and $\mathrm{u}(\neg M \wedge A_2) = 900,000$. The reader can verify that in order for one-boxing to be rational in this case, one's degree of belief in the causal hypothesis (H_2) must exceed $\frac{90}{98} \approx 0.918$.

If Nozick is right that few people would feel comfortable taking one box in this case, I suggest that this is because our intuitions tell against assigning such a high probability to the causal hypothesis. In §2.1, we argued that a rational agent in the original case would have had a credence distribution P with P(causal hypothesis) $<$ 0.5. But one's relevant evidence (i.e. evidence bearing on the truth of the causal hypothesis and the official story) is exactly the same in this variation as it was in the original case. So a rational agent in the variation case would still have less than 0.5 degree of belief in the causal hypothesis. Since $0.5 < 0.918$, P(causal hypothesis) will obviously be less than 0.918 in this variation. This is why the one-boxing intuition loses its force in this case: it is no longer the CDT-recommended option for an epistemically rational agent.

3.2 Variation #2: The Good-but-not-Great Predictor

Nozick (1969, 70) writes that "...it is not the expected-utility principle [i.e. evidential decision theory] which leads some people to to choose only what is in the second box."[16] His argument for this is that, if we consider a case in which the Predictor is only *pretty good* at predicting, then most people would choose two boxes, even though, if the Predictor's accuracy is sufficiently good (despite not being great), evidential decision theory still advises us to one-box. Suppose, for instance, that the probability of the Good-but-not-Great Predictor's predicting correctly is only 0.6. Then two-boxing has the following evidential expected utility: (let A_1 and A_2 be as before; C stands for "He predicts correctly"):

$$\begin{aligned} \mathrm{V}(A_2) =& \mathrm{P}(C|A_2) \times \mathrm{u}(C \wedge A_2) + \mathrm{P}(\neg C|A_2) \times \mathrm{u}(\neg C \wedge A_2) \\ &= (0.6 \times \mathrm{u}(\$1,000)) + (0.4 \times \mathrm{u}(\$1,001,000)) \\ &= 401,000 \end{aligned}$$

One-boxing, on the other hand, has the following evidential expected utility:

$$\begin{aligned} \mathrm{V}(A_1) =& \mathrm{P}(C|A_1) \times \mathrm{u}(C \wedge A_1) + \mathrm{P}(\neg C|A_1) \times \mathrm{u}(\neg C \wedge A_1) \\ &= (0.6 \times \mathrm{u}(\$1,000,000)) + (0.4 \times \mathrm{u}(\$0)) \\ &= 600,000 \end{aligned}$$

As Lewis (1979, 302) points out, if we assume (as we have been) that utility is proportional to money, then evidential decision theory advocates taking one box in any case in which the Predictor's reliability is greater than 0.5005.[17] Thus, in all such cases, the expected utility of one-boxing exceeds that of two-boxing, when expected utility is calculated according to evidential decision theory. But, Nozick correctly notes, almost no one would one-box in these sort of cases.[18]

Why does the one-boxing intuition wear off as the Predictor's reliability drops? What makes the original Newcomb case exotic is the Predictor's uncanny ability to predict our choices, and this is also what makes the official story hard to believe. We argued earlier that an agent in the original Newcomb case shouldn't be virtually certain that the official story is true, given the evidence she possesses. In the original case, she should give some credence to the causal hypothesis.

[16] Nozick offers no explanation of this variation case.

[17] More generally, if r is the ratio of the utility of \$1,000 to the utility of \$1,000,000, then evidential decision theory advises one-boxing any time the Predictor's reliability exceeds $\frac{1+r}{2}$.

[18] But see Leslie (1991, 80) for a dissenting voice.

The question we must ask in the Good-but-not-Great situation is: how should we distribute our credence over the causal hypothesis and the official story in this situation? I think that the official story is much more believable in this situation than in the original Newcomb case. Why? Because the existence of a Good-but-not-Great Predictor is not particularly surprising; after all, he's only right 60% of the time, which isn't very impressive. That someone could predict human choices with *that* sort of accuracy doesn't come as a surprise. It certainly doesn't challenge my view of myself as a free chooser. More generally, weaker correlations are simply worse evidence for causal connections. All this suggests that the causal hypothesis isn't all that likely in the Good-but-not-Great case, and that the official story is much more probable in this case than it is in the original Newcomb case. But if one is fairly confident of the official story, then one ought to two-box. For if one accepts the official story, then one is in in a position to reason from the premise *What I choose has no effect on the contents of* b_2.

The epistemic explanation ties the appeal of one-boxing to the implausibility of there being a *preternatural* Predictor. The natural prediction of the epistemic explanation, then, is that if one holds the *structure* of the case fixed but makes the scenario more realistic by making the Predictor less accurate, then the one-boxing intuition should wear off. This prediction is confirmed by the fact that the one-boxing loses much of its appeal in the Good-but-not-Great case.

It is worth mentioning here that we can give a parallel explanation of why we don't have the 'refrain from smoking' intuition in the smoking case discussed in §1.3, even though EDT supports refraining from smoking. 'Medical' Newcomb cases are like the Good-but-not-Great variation in that they are much more realistic than the original Problem. It is quite easy to imagine being in a genuine smoking case, i.e. accepting that smoking and lung cancer are two effects of a common cause. After all, surprising medical discoveries are, by now, familiar occurrences. Thus, an epistemically rational agent in a smoking case will accept that she is in a genuine smoking case; for such an agent, the CDT-recommended option is to smoke.

I want to emphasize the ability of the epistemic explanation to account for our reactions to these cases. Each of the foregoing cases is structurally identical to Newcomb's Problem. They each change one of the relevant quantities – payoffs or probabilities – of the case but neither changes them enough for our decision-theoretic principles to notice, i.e. both theories offer the same advice in the variations that they respectively offer in the original Newcomb Problem. The variations thus strongly suggest that the one-boxing intuition is not an EDT intuition. Our intuitions about Newcomb's Problem are not simply sensitive to decision-theoretic principles; they are tracking something other than the decision-theoretic structure of the Problem. Something is interfering, and the epistemic explanation offers an explanation of the interference that is flexible in just the right way.

4 What About Two-Boxing?

The epistemic explanation claims that we find one-boxing appealing because one-boxing is the CDT-recommended option for a epistemically rational agent in Newcomb's Problem. But this might seem to prove too much, since our investigation began with the conviction that the two-boxing argument is a good one. The two-boxing argument still is a good argument, of course; what I've been trying to press is that an agent in a Newcomb case may not be in a position to avail herself of one of its premises. Specifically, she may not be in a position to reason from the premise *What I choose has no effect on the contents of* b_2, since she will have to give the negation of that premise some non-zero credence. But this still leaves an important question unanswered: where does the two-boxing intuition fit into the story we have been telling? The question that motivated this investigation was why we feel pulled in two directions when considering Newcomb's Problem; but now it looks as if we should only be pulled in the one-boxing direction, and that seems odd.

There are, I think, a number ways of explaining where the two-boxing intuition comes from that are consistent with the epistemic explanation of the one-boxing intuition. I will elaborate on one that I find illuminating, but I want to emphasize that I think the epistemic explanation does not stand or fall with the general diagnosis of Newcomb's Problem that I'm about to offer. One can accept the epistemic explanation without buying the rest of my story.

4.1 Agents and Theorists, Inside and Out

When we consider decision problems like Newcomb's Problem there are always two perspectives that we can and do take on the hypothetical case: that of the theorist or author who stipulates the facts of the case and that of the agent in the case who faces the choice. The distinction is between considering the case as an impersonal possibility, on the one hand, and imagining the case from the perspective of an agent inside the case, on the other. It is the difference between imagining 'from the outside' *that* an agent (who one might identify as oneself) faces a certain choice and imagining 'from the inside' *being* in the situation and *facing* the choice.[19]

It is obvious that we have the first perspective, since it is obvious that it is we who invent the case and so get to stipulate what we want about it. That we also consider the case from the perspective of the agent is perhaps less obvious. But that we do this in Newcomb's case is shown by the following fact: when one describes a Newcomb Problem one *never* stipulates whether or not there is \$1,000,000 in b_2 or not.

Any Newcomb Problem is really one of two hypothetical situations: it is either (a) a situation in which b_2 contains \$1,000,000, or it is (b) a situation in which b_2 is empty. We settle that it is one of (a) or (b) by stipulation; but

[19] This distinction is discussed in various places in the philosophical literature on imagination; see, e.g, Williams (1966), Peacocke (1985), Walton (1990, 28-35), Martin (2002, 402-413), Higginbotham (2003) and Ninan (2006).

we do not stipulate which one it is. From the perspective of the theorist, it should be irrelevant whether we stipulate (a) or stipulate (b) or fail to make either stipulation. It should be irrelevant because the important feature of a decision problem is not the way the world is, but how the agent takes the world to be. Once we stipulate that the agent doesn't know whether she's in an (a)-type situation or in a (b)-type situation, that is enough to set up the decision problem. If we did not feel it was important to consider the case from the perspective of the agent, then either stipulating that the agent is in an (a)-type situation or stipulating that the agent is in a (b)-type stipulation wouldn't matter. We could make either stipulation and then ask, what should the agent do?

The difficulty with this, however, is that once one makes either of these stipulations, the one-boxing intuition vanishes. If we stipulate (a) or stipulate (b), it becomes obvious that the agent should take both boxes. For example, suppose we stipulate (a), that b_2 contains \$1,000,000. Now it seems obvious that the agent should take both boxes. The same would be true if we were to stipulate (b). The only way for the one-boxing intuition to get a grip is for us to refrain from stipulating (a) and refrain from stipulating (b). This suggests that in order for us to generate a case that is intuitively puzzling we have to consider the Problem from the perspective of the agent, i.e. we have to imagine ourselves with *the agent's* evidence. And this is typically what we do: we imagine actually being in the agent's shoes, wondering what to do. We test our intuitions by asking ourselves, *What would/should I do if I were actually in a Newcomb Problem?*

Why is the one-boxing intuition linked in this way to the agent's perspective? Here is what I suggest: When we consider Newcomb's Problem from the perspective of a theorist who stands outside the hypothetical case, we know that the case is a Newcomb case (since we stipulated that it was), and we know that two-boxing is the action which will in fact maximize the agent's utility. But when we imagine ourselves as the agent facing Newcomb's choice, we imagine ourselves without enough evidence to be certain that the case is a Newcomb case. From the agent's perspective, we feel the pull of one-boxing, since one-boxing is the rational option for an agent with a non-negligible amount of doubt about the official story. As theorists outside the case, we are pulled towards two-boxing; as agents inside the case, we are pulled towards one-boxing. Hence the feeling of being pulled in two directions. Each of our conflicting intuitions about the case corresponds to a different way of imagining the hypothetical scenario.

4.2 Missing the Point?

This diagnosis raises a question of its own: why, when we imagine being the agent in Newcomb's Problem, do we imagine being less than certain of the official story? The question motivates an objection to the epistemic explanation:

> Newcomb's Problem contains an implicit stipulation to the effect that the agent in the case is certain that she is in a genuine Newcomb

> Problem. It is part of the definition of the case that the agent in the case is certain that the official story is true. The interesting question about the case, then, is what such an agent ought to do. So the epistemic explanation, in considering what it would be the rational choice for an agent who is less than certain of the official story, just misses the point of the Problem.

This objection misunderstands the nature of the epistemic explanation, though it is a natural misunderstanding. I agree that an implicit stipulation of the case is that the agent is certain that the official story is true. When we imagine the case from the outside, we accept that stipulation, and reasoning in accord with that stipulation leads to the two-boxing conclusion. The two-boxing argument is a cogent argument, and so two-boxing is the correct response to Newcomb's Problem for an agent who is certain of the official story.

But all that is only part of the story. One's feeling of puzzlement about the case is not eliminated by convincing oneself that two-boxing is the rational option, given the stipulations of the case. One still wants to know why one feels pulled in the other direction. The epistemic explanation is aimed at explaining this feeling of puzzlement. Let me put it this way: the epistemic explanation does not offer a *justification* of one-boxing, given the stipulations of the case, so much as an *explanation* of the one-boxing intuition.[20]

According to the epistemic explanation, when we imagine actually being in a Newcomb Problem, we *don't* imagine the scenario from the perspective of an agent who is certain that the official story is true. So when we imagine being in a Newcomb Problem, we don't accept the stipulation that the agent in the case is certain of the official story. So far, I've argued that the one-boxing intuition tracks our sense of what an epistemically rational agent ought to do in Newcomb's Problem by showing how this hypothesis explains the instability of our intuitions across structurally identical cases. But this account raises a question of its own: given that it is stipulated that the agent in the case is certain that the official story is true, why is it that when we imagine the case from the inside, we don't imagine the case from the perspective of an agent who is certain that the official story is true?

Here's a tentative answer to that question. It seems to me that when we imagine being in a situation, what we can imagine believing in that situation is constrained by two factors: (i) what evidence we imagine having in that situation, and (ii) our actual opinions about what sort of credence distribution that evidence rationally permits. I think that if you believe that evidence e does not rationally permit credence distribution C, then you can't imagine having evidence e and being in a belief state represented by C. Here is a simple example of what I'm talking about: imagine a case in which you have an experience as of a bright red beach ball in normal lighting conditions. Suppose further that you have evidence that your visual system is working normally and that the lighting conditions are normal too. Suppose, that is, that in the imaginary scenario,

[20] But the explanation does, of course, provide a justification for one-boxing in this sense: it is what an epistemically rational agent in a Newcomb Problem ought to do.

you don't any evidence against the claim that the ball is red, and substantial evidence that it is. I submit that you will find it exceedingly difficult or even impossible to imagine being in this situation and having a degree one belief that the ball is green. As you imagine the case, you imagine having evidence that doesn't permit you to be certain that the ball is green. How you imagine your evidence constrains what you are able to imagine believing. (You can, of course, imagine from the outside *that* you believe the ball is green—what you have difficulty doing is imagining this 'from the inside'.)[21]

Earlier we argued that an agent in a Newcomb case wouldn't be justified in completely ruling out the causal hypothesis. The above example then suggests the following explanation of why we fail to imagine being certain of the official story: When we imagine the case, we imagine ourselves in a situation in which it wouldn't be rational to be certain of the official story, for reasons discussed in §2.1. But then if our above conjecture is correct, it will be exceedingly difficult or even impossible to imagine being certain that the official story is true. What you can imagine believing in Newcomb's case is constrained by what you think it would be rational for you to believe in the case; since you do not think it would be rational to be *certain* of the official story, you do not – and perhaps cannot – imagine being in Newcomb's Problem and being certain of the official story.

5 Summary

We began with the following question: Given the compelling argument for taking both boxes in Newcomb's Problem, why are we nevertheless tempted to one-box? I argued for the following answer: We are tempted to one-box because that is the action CDT recommends to an agent in Newcomb's Problem who responds rationally to her evidence. The best argument for this approach is how it handles the two variations on Newcomb's Problem discussed in §3; these variations are difficult to explain if we assume that our intuitions about these cases are responding only to their decision-theoretic structure. I went on to provide a diagnosis of Newcomb's Problem that suggests that the two-boxing intuition is a result of considering the hypothetical scenario from the external viewpoint of the theorist, and that the one-boxing intuition arises only when we consider actually being the Newcomb agent, facing the boxes. Finally, I tentatively suggested that the one-boxing intuition is tied to the agent's perspective because, when we imagine actually being in Newcomb's Problem, we are unable to imagine being certain that the official story is true.

[21] Why are our imaginings constrained in this way? I think the answer to this question is explained by three considerations: first, our first-order beliefs are tightly constrained by our beliefs about what rationality permits us to believe; second, we cannot imagine that our normative epistemic judgments are false, i.e. we experience *imaginative resistance* with respect to our normative epistemic judgments (Weatherson 2004, 3); and third, that when we cannot imagine a proposition p for reasons of imaginative resistance, we also cannot imagine (from the inside) believing p.

References

Bar-Hillel, M. and Margalit, A. (1972). ‘Newcomb’s Paradox Revisited’. *The British Journal for the Philosophy of Science* **23**:295–304. 2, 11

Campbell, R. and Sowden, L. (eds.) (1985). *Paradoxes of Rationality and Cooperation*. University of British Columbia Press, Vancouver, BC. 23

Collins, J. (2001). ‘Newcomb’s Problem’. In N. Smelser and P. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier. `http://collins.philo.columbia.edu/newcomb.pdf`. 5

Eells, E. (1982). *Rational Decision and Causality*. Cambridge University Press, Cambridge, UK. 4

——— (1984). ‘Newcomb’s Many Solutions’. *Theory and Decision* **16**. 4

Gibbard, A. (1979). ‘Decision Matrices and Instrumental Expected Utility’ Paper presented to a conference at the University of Pittsburgh. 7

Gibbard, A. and Harper, W. L. (1978). ‘Counterfactuals and Two Kinds of Expected Utility’. In C. Hooker, J. Leach and E. McClennan (eds.), *Foundations and Applications of Decision Theory, Vol. 1: Theoretical Foundations*. Reidel, Dordrecht and Boston. Reprinted in Campbell and Sowden (1985). 5

Higginbotham, J. (2003). ‘Remembering, Imagining, and the First Person’. In A. Barber (ed.), *Epistemology of Language*, 496 – 533. Oxford University Press, Oxford and New York. 19

Horgan, T. (1981). ‘Counterfactuals and Newcomb’s Problem’. *Journal of Philosophy* **78** (6):331–56. Reprinted in Campbell and Sowden (1985). 2, 7

Hurley, S. (1994). ‘A New Take from Nozick on Newcomb’s Problem and Prisoner’s Dilemma’. *Analysis* **54** (2):65–72. 15

Jeffrey, R. (1983). *The Logic of Decision*. University of Chicago Press, Chicago, second edition. First published in 1964. 3, 4

Joyce, J. (1999). *Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge, UK and New York. 4, 5

Leslie, J. (1991). ‘Ensuring Two Bird Deaths With One Throw’. *Mind* **100**:73–86. 2, 17

Levi, I. (1982). ‘A Note on Newcombmania’. *Journal of Philosophy* **79**:337–42. 2

Lewis, D. K. (1975). ‘Counterfactual Dependence and Time’s Arrow’. *Noûs* **13**:455–476. Reprinted in Lewis (1986). 5

——— (1979). ‘Prisoners’ Dilemma is a Newcomb Problem’. *Philosophy and Public Affairs* **9**:235–240. Reprinted in Lewis (1986). 17

——— (1981). 'Causal Decision Theory'. *Australasian Journal of Philosophy* **59**. Reprinted in Lewis (1986). 4, 5, 12

——— (1986). *Philosophical Papers, Volume II.* Oxford University Press, New York and Oxford. 23, 24

Mackie, J. (1977). 'Newcomb's Paradox and the Direction of Causation'. *Canadian Journal of Philosophy* **7** (2). Reprinted in Mackie (1985). 11

Martin, M. (2002). 'The Transparency of Experience'. *Mind and Language* **17** (4):376 – 425. 19

McKay, P. (2004). 'Newcomb's Problem: the Causalists Get Rich'. *Analysis* **64** (2):187–89. 11

Ninan, D. (2006). 'Imagination, Inside and Out'. Unpublished manuscript, MIT. 19

Nozick, R. (1969). 'Newcomb's Problem and Two Principles of Choice'. In D. D. Rescher, N. and C. Hempel (eds.), *Essays in Honor of Carl G. Hempel.* D. Reidel Publishing Company, Dordrecht, Holland. Reprinted in Nozick (1997). 2, 17

——— (1974). 'Reflections on Newcomb's Problem'. *Scientific American* **230**:102 – 106. Reprinted in Nozick (1997). 2

——— (1993). *The Nature of Rationality.* Princeton University Press, Princeton, NJ. 6, 15, 16

——— (1997). *Socratic Puzzles.* Harvard University Press, Cambridge, MA. 24

Peacocke, C. (1985). 'Imagination, Experience, and Possibility: A Berkeleian View Defended'. In J. Foster and H. Robinson (eds.), *Essays on Berkeley*, 19 – 35. Clarendon Press, Oxford. 19

Schmidtz, D. and Wright, S. (2004). 'What Nozick Did for Decision Theory'. *Midwest Studies in Philosophy* **28**:282 – 294. 11

Skyrms, B. (1980). *Causal Necessity.* Yale University Press, New Haven. 12

——— (1990). *The Dynamics of Rational Deliberation.* Harvard University Press, Cambridge, MA. 5

Sobel, J. H. (1994). *Taking Chances: Essays on Rational Choice.* Cambridge University Press, Cambridge, UK. 4

Stalnaker, R. (1981). 'A Letter to David Lewis'. In W. Harper, R. Stalnaker and G. Pearce (eds.), *Ifs.* Reidel, Dordrecht. 5

Walton, K. (1990). *Mimesis as Make-Believe.* Harvard University Press, Cambridge, MA and London. 19

Weatherson, B. (2004). 'Morality, Fiction, and Possibility'. *Philosophers' Imprint* **4** (3). http://www.philosophersimprint.org/004003/. 22

Williams, B. (1966). 'Imagination and the Self'. *Proceedings of the British Academy* Reprinted in Williams (1973). 19

——— (1973). *Problems of the Self : Philosophical Papers 1956-1972*. Cambridge University Press, Cambridge. 25