The Implicit Decision Theory of Non-Philosophers

The authors declare that there is no conflict of interest that would affect the content of the material in this paper.

Abstract

This paper empirically investigates non-philosophers’ intuitions about the Newcomb problem. We first aim to determine whether non-philosophers are one-boxers or two-boxers, and we then explore two potential explanations of *why* people have different intuitions about the case. While our first study found, like previous studies, that a majority of people are one-boxers, our second study found that a majority are two-boxers. We suggest that this difference lies in the fact that our second study better controls for beliefs about backwards causation than previous studies. Generally, we suggest that the more backwards causation is successfully ruled out as an explanation for the predictor’s accuracy, the more likely non-philosophers are to be two-boxers. We then aim to determine whether non-philosopher two-boxers, without prompting and without prior education, make a distinction that is crucial to causal decision theorists: the distinction between preferring something *as a news item* and preferring it *as an object of choice*. Next, we note interesting parallels between a popular argument for two-boxing and a popular argument for the rationality of *future bias*, and we investigate whether differences in people’s response to the Newcomb problem correlate with differences in their level of future bias. We find that two-boxers *do not* make the distinction between news items and objects of choice, and that there is little to no correlation between people’s future bias and their view of the Newcomb problem. We end by discussing the philosophical upshots of this work.

1. Introduction

Newcomb’s problem is one of the most widely discussed cases in decision theory. Here is a standard formulation:

You must choose between taking (and keeping the contents of) (i) an opaque box now facing you or (ii) that same opaque box *and* a transparent box next to it containing $1000. Yesterday, a being with an excellent track record of predicting human behaviour in this situation made a prediction about your choice. If it predicted that you would take only the opaque box (‘one-boxing’), it placed $1M in the opaque box. If it predicted that you would take both (‘two-boxing’), it put nothing in the opaque box. (Ahmed, 2018, 1)

In Newcomb’s problem, the available acts are one- or two-boxing, and the states are that the $1 Million is present or it is not. There is an evidential dependence between $1 Million being present and one-boxing, but there is no causal dependence. While some (e.g., evidential decision theorists) see the evidential dependence between the state and the act as decision-relevant and thus think that one-boxing is rational,[[1]](#footnote-2) others (e.g., causal decision theorists) do not see this dependency as decision-relevant. They focus their attention on the causal dependencies alone, and from that focus, they see no good reason to one-box: the causal effect of one-boxing is merely to leave one a thousand dollars poorer than if one had two-boxed.

Our aim in this paper is to probe non-philosophers’ judgements about Newcomb-style cases. Bourget and Chalmers (2020) recently probed *philosophers’* judgements about the Newcomb case. They found that ~31% of philosophers accept or lean towards one-boxing, while ~40% accept or lean towards two-boxing (with the remaining ~30% being unsure, unfamiliar with the issue, or having some other view).

There is also some much older research that looks at non-philosophers’ judgements regarding Newcomb cases. In 1974, after discussion of the Newcomb problem in the Scientific American by Martin Gardner, readers were invited to send in their opinion, and 652 responded. Of these, 483 (i.e. ~74%) said that they would one-box, while the remaining said they would two-box. There are several notable features of this data. First, the article included discussion of an argument for one-boxing and an argument for two-boxing. Thus, in reporting their choice, respondents were in part *directly* reporting which of the two arguments they found most persuasive. Second, this was not a controlled study, and respondents were surely not a random sample. In 1979, MacCrimmon and Larson ran the first controlled study. They found that of 19 participants, all but two would one-box (i.e., ~89.5%). This study, however, has several notable drawbacks. First, it used a very small sample size, and second, it was unclear whether participants understood that there was no causal relationship between their choice and the prediction. Specifically, it seemed that participants might have supposed that there was some kind of *backwards causation* between choice and prior prediction. Considering this, in 1992 Shafir and Tversky ran a follow-up study that aimed to present participants with a version of the Newcomb Problem that more clearly did not involve backwards causation. To do this, in their vignette the role of the predictor was played by a fictitious computer program that predicted participant’s choices based on a previously established database. They found that of 40 participants, 14 (35%) chose both boxes, while the remaining 65% chose one box.

Shafir and Tversky claimed that in the absence of backwards causation, there are no grounds to one-box, and they hypothesised that what explains people’s choice is *quasi-magical thinking.* They take quasi-magical thinking to be thinking in which people act as though they have the magical beliefs in question, even though they do not explicitly endorse these beliefs. Thus, Shafir and Tversky concluded, on the basis of their empirical data, that 65% of people had these quasi-magical beliefs.

In all, we see that across the three sets of data, ~89.5%, ~74%, and ~65% of participants chose to one-box. While in all these cases this is a majority, there are substantial differences in these results. This may in part be the result of (a) different samples and sample sizes, (b) different vignettes which better control, or not, for beliefs about backwards causation between choice and prediction, and (c) differences in whether the arguments for each position are mentioned or not. We take it to be of interest to run a new version of the Newcomb problem, which, like that of Shafir and Tversky, attempts to better rule out the possibility of backwards causation between choice and prediction, but which has larger sample sizes than the previous studies and does not make any mention of the arguments in favour of each choice. We also take it to be desirable to run a study that includes comprehension questions to ensure that participants understand what they are being asked.

Given the results of these previous studies, we predict that non-philosophers are more likely to be one-boxers. Thus, our first hypothesis is:

H1: A majority of people will one-box.

It is natural to wonder, in addition, whether two-boxers (even if it turns out that they are in the minority) draw a conceptual distinction between *news-item preferences* and *choice preferences*. Consider that we can ask someone what they *decide* in a Newcomb-style case. This is to ask them for their choice preference. But we can also ask them about their news-item preference. Suppose you wake up after a Newcomb-type decision but cannot remember what choice you made. What do you prefer to learn that you chose? Even if you are a two-boxer, you may well prefer to learn that you one-boxed. For preferring to learn that you one-boxed is learning that you are a certain kind of person, namely, a person who is now almost certainly rich! (More on this below.) Thus, reflective two-boxers will treat choice preferences and news-item preferences differently. Could this help explain why two-boxers choose as they do?

The distinction between news-item and choice preferences is critical for causal decision theorists because it helps them explain why they think two-boxing is rational even though it would be *good news* to learn that one is a one-boxer. This is because when making decisions, causal decision theorists only aim to causally promote good outcomes, and, unlike evidential decision theorists, they do not aim to create good news. More specifically, the crucial difference between causal decision theoryandevidential decision theory is found in their views on which types of dependencies between states of the world and acts are decision-relevant. Evidential decision theory views *all* *evidential dependencies* between states and acts as relevant to evaluating the rationality of the agent’s decision, whereas causal decision theory takes a more restricted view. Causal decision theory views *only causal dependencies* between states and acts as decision-relevant. Since causal dependencies are also evidential dependencies (if A causally promotes B, then A also provides evidence for B), causal decision theory claims that only a subset of the evidential dependencies is decision-relevant. This subset is captured by the agent’s *causal dependency hypothesis*. A causal dependency hypothesis picks out the dependencies that are causal.

Evidential decision theorists, like Jeffrey (1965), do not distinguish between the “news value” of an act and its value as an object of choice. As Jeffrey (1965, 73–4) writes, “If the agent is deliberating about performing act *A* or act *B*, and if *AB* is impossible, there is no effective difference between asking whether he prefers *A* to *B* as a news item or as an act, for he makes the news.” In contrast, causal decision theorists, like Joyce (1999), believe that it is crucial to distinguish between the news value of acts and their value as objects of choice. Joyce calls the news value of an act its “auspiciousness” and its value as an object of choice its “efficacy.” According to causal decision theorists like Joyce, only efficacy matters when it comes to *choosing* an act, though auspiciousness plays a role in evaluating how desirable an act is.

Joyce (1999, 151–4) devotes several pages to explaining this distinction through an imagined dialogue between a causal decision theorist and a one-boxer. Joyce’s protagonist explains: It would have been “better for me” to be the one-boxing type, but that gives me no reason to one-box when actually faced with the decision. Joyce (1999, 154) concludes:

When [the causal decision theorist] wishes she would refuse the extra $1,000 she is wishing for good news about herself, and when she decides not to take it she is deciding on the basis of efficacy. There is nothing wrong with simultaneously evaluating acts in both these ways. It should not be any part of causal decision theory to deny that acts have auspiciousness values or that these play a central role in our thinking about the desirability of prospects. Quite to the contrary, it turns out to be quite useful for causal decision theorists to have news values around since they help to explain why people so often feel “torn” when thinking about what to do in Newcomb-type problems. The deliberative tension is the result of auspiciousness pulling one way and efficacy pulling the other. The only thing we causal decision theorists are committed to is that efficacy should always win out in this tug-of-war when the issue is one of deciding how to act.[[2]](#footnote-3)

An implication of Joyce’s discussion—and, indeed, a suggestion made by causal decision theorists more generally—is that two-boxers choose the way they do partly because they comprehend the distinction between news-item preferences and choice preferences. The distinction allows two-boxers like Joyce to acknowledge that one-boxing is desirable insofar as they’d prefer to learn that they are one-boxers. It is thus natural to wonder whether non-philosopher two-boxers make this distinction. In short, does seeing the difference between news-item preferences and choice preferences help explain why some people are two-boxers? Hence, we introduce the following exploratory hypothesis:

H2: People who choose to two-box will prefer to learn that they had one-boxed (i.e., there will be a difference between two-boxers’ choice preferences and their news-item preferences in response to our Newcomb vignette).

1. Decision and Future bias

People are sometimes sensitive not only to the relative values and probabilities of events, but also to whether the events are in the future or the past. Most important for the purposes of this paper is the fact that, all else being equal, many people prefer positively valenced events to be located in the future rather than the past, and negatively valenced events to be located in the past rather than the future (Greene, Latham, Miller, and Norton 2021a). This preference is known as future bias.[[3]](#footnote-4)

Over the last few years, we have come to learn a good deal about future-biased preferences. We know not only that people have such preferences, but also that they are often *strong*. People’s preference for positive events to be in the future persists even when the future positive event is of less value than the past event (Greene, Latham, Miller, and Norton forthcoming). The same is true, *mutatis mutandis,* of negative events. One study found that people prefer ten units of pain in the past to a single unit of pain in the future (Greene, Latham, Miller, and Norton 2021b) while another found that people prefer one unit of pleasure in the future, to two in the past (Greene, Latham, Miller, and Norton forthcoming). Having said that, results here have been mixed. Lee, Hoerl, Burns, Fernandes, O’Connor and McCormack (2020 asked both children and adults whether they would prefer to be someone who experienced a pleasurable or painful state of affairs in the past or someone who will have that same experience in the future. They found that when the experience is equally painful or pleasurable, people prefer to be someone with pain in the past and pleasure in the future. However, when the amount of pleasure or pain in the past would be greater, this preference was abandoned by a majority of children, and roughly half of adults.[[4]](#footnote-5)

At the very least, we can say that many people prefer that, overall, from a time-neutral perspective, they are worse off. There are also various arguments that future bias can influence choice preferences, such as when future-biased agents are also risk averse (Dougherty 2011), regret averse (Greene and Sullivan 2015), or evidential decision theorists (Tarsney 2017).

Of particular interest to us is the way that future bias may interact with choice preferences in the standard Newcomb case. According to evidential decision theory, there is sometimes a decision-relevant dependency between an act and a past state of the world. For example, consider the standard Newcomb problem introduced above. Evidential decision theorists point out that one-boxing is evidence that a prediction has been made in the past, and they claim that this evidence is relevant in deciding what to do. In contrast, causal decision theorists claim that the decision-relevant dependencies are always between acts and future states (the possibility of backwards-causation aside). Causal decision theorists do not care about dependencies between acts and past states, because the past cannot be causally influenced. Thus, when deciding what to do, evidentialists will sometimes consider how choices bear on past events, whereas causalists will only consider how choices bear on future events.

Since two-boxing is associated with causal decision theory and one-boxing boxing is associated with evidential decision theory, we might hypothesize that two-boxers will display greater levels of future bias than one-boxers. Two-boxers often point out that whether one has been allocated a million dollars has already occurred and cannot be causally influenced, and is thus an irrelevant consideration for decision making. In this way, their argument bears a similarity to that of philosophers who draw a connection between future bias and the causal inaccessibility of the past. Some philosophers have argued in support of future bias by appealing to the idea that the quality of past states is irrelevant because they are ‘over and done with’ (Prior 1959; Schlesinger 1976; Craig 1999; Pearson 2018), while others have argued that this characteristic of past states of affairs explains why we are future biased, even though it does not rationalise our having that preference (Horwich (1987, pp. 194-196) and developed by Maclaurin & Dyke (2002) and Suhler & Callender (2012).

Those who think that the fact that past states of affairs are causally inaccessible explains why we are future biased hold that we attach less evaluative weight to past events because there is nothing that we can do to affect the past, which means that past events cannot count for, or against, present choices in the way that potential future events can. This has become known as the *practical irrelevance explanation* (Latham, Miller, Norton and Tarsney 2020).[[5]](#footnote-6)

By contrast, philosophers who defend the rationality of future bias often point to a somewhat different respect in which states of affairs are ‘over and done with’: namely the sense in which we are moving through time from the past toward the future, and therefore past experiences lie ‘behind us’ while future ones lie ‘ahead of us’. This explanation has become known as the *temporal metaphysics explanation* since it connects the presence of temporal metaphysical facts—irreducibly tensed facts about which states of affairs are objectively past, present, and future—to the presence of future bias (Latham, Miller, Norton and Tarsney 2020; Latham, Miller, Tarsney and Tierney 2021, 2022).

This explanation, in turn, is often connected to the presence of tensed emotions, which are emotions that are differentially elicited depending on where in time a state of affairs is represented as being located. For instance, we *anticipate* future states of affairs, not past ones; we *regret* past states of affairs, not future ones, and we feel a certain sort of distinctive *relief* that certain negative states of affairs are ‘over and done with’ only when they are past, and not when they are future. This sort of relief is what Hoerl (2015) calls *temporal relief.* The idea, then, is that it is because past states of affairs are, metaphysically speaking, ‘over and done with’, that we experience temporal relief and associated tensed emotions, and that it is because we experience such emotions that we are future biased (Prior 1959; Pearson 2018; Schlesinger 1976; and Craig 1999).[[6]](#footnote-7)

If the practical irrelevance explanation is true, then we might find that two-boxers are more likely to be future biased than are one-boxers. By contrast, if the temporal metaphysics explanation is correct, then we would have little reason to make this prediction. Since there is some evidence in favour of this explanation (Latham, Miller, Norton, and Tarsney (2020) and little evidence in favour of the temporal metaphysics explanation (Latham, Miller, Tarsney, and Tierney 2021, 2022) our third hypothesis was as follows:

H3: More two-boxers will be future biased than one-boxers.

If this hypothesis is vindicated, it could provide additional insight into how people make decisions. Just as H2 probes whether two-boxing is correlated with a recognition of the distinction between news-item preferences and choice preferences, H3 probes whether two-boxing is correlated with future bias. In short, does future bias help explain why some people are two-boxers?

In what follows we report the methodology and results of the two experiments that we ran. The experiments are very similar and differ only in two respects. First, our first experiment simply talks about a predictor making a prediction about what will occur, but does not specify how the predictor does so, leaving open that perhaps the predictor has information about the future via backwards causation. Second, our first experiment used only two comprehension questions rather than four.

In Section 3, we outline the methodology of these two experiments and present their results. In Section 4, we consider the upshot of those results for theorising about both decision theory and future bias.

3 Methodology and Results

3.1 Experiment 1 Methodology

*3.2.1 Participants*

343 people participated in the study. Participants were U.S. residents, recruited and tested online using Amazon Mechanical Turk, and compensated $1 for their time. Participants had a HIT (task) approval rate of at least 95% and had at least 1000 HITs (tasks) approved. This means that all our participants had already successfully completed at least 1000 other tasks and received at least a 95% approval rating on these tasks. 244 participants had to be excluded for failing to follow task instructions and attention checks, or for failing to correctly answer *all* comprehension questions for the Newcomb vignette and the future-bias vignette. The remaining sample was composed of 99 participants (48 female; aged 21-69 mean age 41.03 (SD = 11.49)). Ethics approval for these studies was obtained from the [blanked] Human Research Ethics Committee. Informed consent was obtained from all participants prior to testing. The survey was conducted online using Qualtrics.

*3.1.2* *Materials and Procedure*

Participants saw one of two vignettes in random order. A Newcomb vignette (below) modelled after Joyce’s (1999, 146-7) presentation,[[7]](#footnote-8) and a future-bias vignette, which will be presented shortly, and which is either positive or negatively valenced.

****

After reading the vignette, participants responded to two comprehension questions in random order. “If you refuse the $1000 dollars and the predictor predicted that you would refuse the $1000, then you will receive”. To which they could respond:

1. $ 1 million
2. $1000
3. Nothing
4. $1.01 million

And “If you take the $1000 dollars and the predictor predicted that you would take the $1000, then you will receive”. To which they could respond:

1. $ 1 million
2. $1000
3. Nothing
4. $1.01 million

Participants were then asked: “What do you do?”. To which they could respond:

1. Take the thousand dollars.
2. Refuse the thousand dollars.

Participants then read the following text: “You wake up the morning after a hard night on the town, and for a moment you cannot remember what decision you made.”

They were then asked: “What would you prefer to learn that you had done?” To which they could respond:

1. Take the thousand dollars.
2. Refuse the thousand dollars.

The other vignette was the future-bias vignette, which was either positively or negatively valenced. Since the positive and negative vignettes differ only minimally, we present them together below.

****

After reading the vignette, participants responded to four comprehension questions in random order. “In order to remain disease free, you must take the pill”. To which they could respond:

1. At some point during the 12-month period after the treatment
2. The week after treatment
3. Twice in six months
4. Not more than 6 months after treatment

“If you took the pill in the previous 6 months, then”. To which they could respond:

1. The pill caused three days of pain
2. The pill caused three days of pleasure
3. The pill caused one day of pain
4. The pill caused one day of pleasure

“If you will take the pill in the next 6 months then”. To which they could respond:

1. The pill will cause three days of pain
2. The pill will cause three days of pleasure
3. The pill will cause one day of pain
4. The pill will cause one day of pleasure.

And “When you awake in the morning”. To which they could respond:

1. You remember that you already took the pill
2. You remember that you need to take the pill
3. You cannot remember whether you took the pill already or not
4. You remember that you are about to take the pill

Finally, participants were asked to “Please indicate your preference using one of the following statements:

1. I would prefer to learn that I will take the pill in the next 6 months and did not take it in the last 6 months.
2. I would prefer to learn that I took the pill in the last 6 months, and will not take it in the next 6 months.
3. I have no preference between these options.

3.2 Results

Before presenting our analysis, we will start by summarising our main findings regarding each hypothesis. We first hypothesised that (H1) most people will be one-boxers. This hypothesis was supported, with the vast majority of people choosing to one-box rather than two-box. Next, we hypothesised (H2) that people who chose to two-box will prefer to learn that they had one-boxed. This hypothesis was not supported. People who had chosen to two-box preferred to learn that they had two-boxed (and equally, we found that people who had chosen to one-box also preferred to learn that they had one-boxed). There was no difference between choice and news-preferences. Finally, we hypothesised (H3) that two-boxers will be more future biased than one-boxers. This hypothesis was not supported. There was no association between being a two-boxer (or not) and future-biased preferences.

To assess whether most people were one-boxers (H1) we conducted a one-way chi-square test.[[8]](#footnote-9) As predicted, the results of this test showed that the majority of people were one-boxers (84; 84.8%) rather than two-boxers (15; 15.2%), χ2(1, N = 99) = 48.091, *p* < .001).

Next, to assess whether people that choose to two-box would prefer to learn that they had one-boxed, we conducted a chi-square test of independence. If there is an association between what people choose and what they prefer to learn, then this test will produce a significant result. The result of this test was significant, χ2(1, *N* = 99)= 22.799, *p* < .001, however, the association was not the one that we predicted. Instead, choosing to two-box was associated with preferring to learn that you had two-boxed. Similarly, choosing to one-box was associated with preferring to learn that you had one-boxed (see Table 1 below).

*Table 1. People’s Choice and News Preferences*

|  |  |  |
| --- | --- | --- |
|  | **Learn (Two-Box)** | **Learn (One-Box)** |
| **Choose (Two-Box)** | 9 (60.0%) | 6 (40.0%) |
| **Choose (One-Box)** | 8 (9.5%) | 76 (90.5%) |

Finally, to assess whether people who choose to two-box are more future biased than people who choose to one-box, we conducted another chi-square test of independence. For the purposes of this analysis, we combined past-biased preferences and time-neutral preference into a single new category: non-future-biased. The result of this test was not significant, χ2 (1, *N* = 99)= .101, *p* = .751, which suggests there is no evidence of an association between people’s decision to one-box or two-box and future-biased preferences (see Table 2 below).[[9]](#footnote-10)

*Table 2: People’s Decision to One-Box or Two-Box and People’s Future-Bias Preferences.*

|  |  |  |
| --- | --- | --- |
|  | Future biased | Non-Future-biased |
| Two-Box  | 6 (40.0%) | 9 (60.0%) |
| One-Box  | 30 (35.7%) | 54 (64.3%) |

3.2 Experiment 2 Methodology

*3.2.1 Participants*

320 people participated in the study. Participants were U.S. residents, recruited and tested online using Prolific, and compensated $1.50 for their time. 228 participants had to be excluded for failing to follow task instructions and attention checks, or for failing to correctly answer *all* four comprehension questions for the Newcomb vignette and *all* four comprehension questions for the future-bias vignette. The remaining sample was composed of 92 participants (47 female, 5 trans or non-binary; aged 18-69 mean age 36.57 (SD = 12.95)). Ethics approval for these studies was obtained from the [blanked] Human Research Ethics Committee. Informed consent was obtained from all participants prior to testing. The survey was conducted online using Qualtrics.

*3.1.2* *Materials and Procedure*

Participants saw one of two vignettes in random order. A Newcomb vignette (below) modelled after Joyce’s (1999, 146-7) presentation, and a future-bias vignette, which will be presented shortly, and which is either positively or negatively valenced. We have highlighted the portion of the Newcomb Vignette that is different from the version we ran in Experiment 1.

**Newcomb Vignette**

You are on a game show. The host offers you one thousand dollars and you must choose whether to take it or refuse it. If you choose to take it, then you will receive that thousand dollars. However, the show also has a special ‘predicting’ machine. The machine predicts whether you will take or refuse the thousand dollars. **The machine collects information about you prior to arriving on the gameshow, and then uses that information to predict what you will do when you go on the show.** The predictor has made 2 million predictions, and it has yet to make an incorrect prediction. It is an extremely reliable predictor. The show works like this. Before the show, the predictor made a prediction about which choice you would make. If the predictor predicted that you would refuse the thousand dollars, then $1 million was transferred to your bank account. If the predictor predicted that you would take the thousand dollars, then no money was transferred. You are not allowed to check your bank account before the show starts, so you do not know what the predicting machine predicted and thus you do not know whether you have received $1 million or not in your bank account.

After reading the vignette, participants responded to *four* comprehension questions in random order:

“If you refuse the $1000 dollars and the predictor predicted that you would take the $1000, then how much money in total will you have made by appearing on the game show”

“If you refuse the $1000 dollars and the predictor predicted that you would refuse the $1000, then how much money in total will you have made by appearing on the game show”.

“If you take the $1000 dollars and the predictor predicted that you would refuse the $1000, then how much money in total will you have made by appearing on the game show”.

“If you take the $1000 dollars and the predictor predicted that you would refuse the $1000, then how much money in total will you have made by appearing on the game show”.

In response to each of these questions participants could respond:

1. $ 1 million
2. $1000
3. Nothing
4. $1million plus $1000

Participants were then asked: “What do you do?”. To which they could respond:

1. Take the thousand dollars.
2. Refuse the thousand dollars.

Participants then read the following text: “You wake up the morning after a hard night on the town, and for a moment you cannot remember what decision you made.”

They were then asked: “What would you prefer to learn that you had done?” To which they could respond:

1. Take the thousand dollars.
2. Refuse the thousand dollars.

The other vignette was the future-bias vignette, which was either positively or negatively valenced. Since the positive and negative vignettes differ only minimally, we present them together below.

The future bias vignette and questions were the same as in experiment 1.

3.2 Results

Before presenting our analysis, we will start by summarising our main findings regarding each hypothesis. We first hypothesised that (H1) most people will be one-boxers. This hypothesis was not supported. In fact, we found that most people chose to two-box rather than one-box. Next, we hypothesised (H2) that people who chose to two-box will prefer to learn that they had one-boxed. This hypothesis was not supported. People who chose to two-box preferred to learn that they had two-boxed (and equally, people who chose to one-box also preferred to learn that they had one-boxed). There was no difference between choice and news-preferences. Finally, we hypothesised (H3) that two-boxers will be more future biased than one-boxers. This hypothesis was not supported. There was no association between being a two-boxer (or not) and future-biased preferences.

To assess whether most people were one-boxers (H1) we conducted a one-way chi-square test. Contra our prediction, the result of this test showed that most people were two-boxers (61; 66.3%) rather than one-boxers (31; 33.7%), χ2(1, N = 92) = 9.783, *p* = .002).

Next, to assess whether people that choose to two-box would prefer to learn that they had one-boxed, we conducted a chi-square test of independence. If there is an association between what people choose and what they prefer to learn, then this test will produce a significant result. The result of this test was significant, χ2(1, *N* = 92)= 68.420, *p* < .001, however, the association was not the one that we predicted. Instead, choosing to two-box was associated with preferring to learn that you had two-boxed. Similarly, choosing to one-box was associated with preferring to learn that you had one-boxed (see Table 1 below).

*Table 1. People’s Choice and News Preferences*

|  |  |  |
| --- | --- | --- |
|  | **Learn (Two-Box)** | **Learn (One-Box)** |
| **Choose (Two-Box)** | 56 (91.8%) | 5 (8.2%) |
| **Choose (One-Box)** | 1 (3.2%) | 30 (96.8%) |

Finally, to assess whether people who choose to two-box are more future biased than people who choose to one-box, we conducted another chi-square test of independence. For the purposes of this analysis, we combined past-biased preferences and time-neutral preference into a single new category: non-future-biased. The result of this test was not significant, χ2 (1, *N* = 92)= .154, *p* = .694, which suggests there is no evidence of an association between people’s decision to one-box or two-box and future-biased preferences (see Table 2 below).[[10]](#footnote-11)

*Table 2: People’s Decision to One-Box or Two-Box and People’s Future-Bias Preferences.*

|  |  |  |
| --- | --- | --- |
|  | Future biased | Non-Future-biased |
| Two-Box  | 23 (37.7%) | 38 (62.3%) |
| One-Box  | 13 (41.9%) | 18 (58.1%) |

4. Discussion

There are several notable aspects of our results.

First, in our first study we found that a majority (84%) of people were one-boxers, while in our second study we found that a majority (66%) were two-boxers.

There were three differences between these two studies. First, the Newcomb Vignettes were identical except that in the second study we included the sentence “The machine collects information about you prior to arriving on the gameshow, and then uses that information to predict what you will do when you go on the show.” Thus, while our first study did not suggest that the predictor used information about the future to inform the prediction, our second study went further in attempting to rule this out. Having said this, one might still think that even our second vignette is *compatible* with some kind of backward causation.[[11]](#footnote-12) For instance, perhaps the information that the machine collects includes “facts from the future.” On this reading, although there is no direct backwards causation between your choice and the machine’s prediction, there might be backwards causation involved in producing the information on which the machine’s prediction is based. In other words, participants might read “The machine collects information about you prior to arriving on the gameshow” as “Before you arrive on the gameshow, the machine collects pre-existing information, caused by future events, about what you will do in the future.” We do not think that this is a particularly natural reading and doubt that a significant number of participants understood our vignette this way. So, we think that our second experiment goes significantly further in controlling for the effect of backwards causation than previous ones, even if it does not entirely rule it out.

Second, in our first study we presented participants with two rather than four comprehension questions. They were presented only with the questions, “If you refuse the $1000 dollars and the predictor predicted that you would refuse the $1000, then you will receive…” and “If you take the $1000 dollars and the predictor predicted that you would take the $1000, then you will receive…”. In our second study we included the further two comprehension questions regarding what would happen if one made a choice which was not the one the predictor predicted. This is because it may be that by presenting only these two comprehension questions, we primed participants to choose a one-boxing option over a two-boxing option. Third, in this study we used Prolific as a platform, whereas in our first study we used MTurk.

These three small differences made a huge difference to the results. We think this suggests two things. First, it suggests that more people are two-boxers than was previously supposed once we better control for relevant factors. Second, it suggests that how people respond to the Newcomb problem is very sensitive to its presentation, and to which factors are salient in making the decision.

Our results show that under conditions in which backwards causation is better controlled for, and in which all four outcomes are made salient, a majority of people choose to two-box. This suggests that the extent to which it has been assumed that people are ‘naturally’ one-boxers has been unwarranted. We return to this idea shortly.

Having said this, we also think that the totality of results of empirical work in this area show that people’s decision in Newcomb cases is much more sensitive to the presentation of the case than one might have expected. The difference in results between our current study and prior studies is extreme. Indeed, the difference between our results across our two studies is very large. The extra sentence added to the vignette in the second study to better control for backwards causation between choice and prediction might play some role in altering the results. However, given that Shafir and Tversky also included material in the vignette that was very similar to the sentences we added, but still found that a majority of people chose to one-box, we are inclined to the view that the presentation of all four outcomes, via the comprehension questions, is the larger factor in explaining why in our second study we found that a majority of people chose to two-box.

Consider our first study, which presented only two comprehension questions. By doing so we draw attention to the outcomes in which the predictor is correct. So, presenting only these questions may have primed people to choose to one-box by making salient only those outcomes. We think it plausible that the results of the other earlier studies might also be partly explained by the differential salience of the outcomes. Those studies did not include comprehension questions. However, it is plausible that participants tend to focus almost exclusively on the two outcomes in which the predictor predicts correctly, given that they are told that the predictor almost always predicts correctly. By including four comprehension questions in our second study, we made salient to participants the two outcomes in which the predictor is incorrect in their prediction, something that we take was not salient in any of the previous experiments. Drawing attention to the fact that the predictor can be incorrect, and what would happen if that were so, makes salient the fact that the money is now either in the opaque box (or in your account) or not, regardless of what you now choose. While Shafir and Tversky’s study (and our second study) do better at ruling out backwards causation, only our second study draws attention to why it is that the absence of backwards causation matters: namely the fact that whether the money is there or not does not causally depend on what you choose.

In all, we think that these results jointly tend to suggest that non-philosophers’ one-boxing intuitions about the Newcomb case are primarily driven by the, likely tacit, idea that the location of the money causally depends on the choice made. Shafir and Tversky, recall, think that people choose to one-box because they are engaged in quasi-magical thinking. We see no reason to think that people explicitly endorse quasi-magical thinking. Rather, we suspect that people probably tacitly employ a heuristic according to which if an extremely strong correlation obtains between x and y, then x and y are taken to be causally connected. In such cases people may not explicitly token a belief that there is a causal relation present, but act and choose as though this is the case. If that is right, then even if a vignette managed to completely rule out backwards causation in the description of the Newcomb case, because it is a case of extremely strong correlation people may still act and choose as though there is a causal connection present, thus explaining why they one-box. By drawing attention to the outcomes in which the predictor is wrong, however, our second study undermined the use of that heuristic by making clear that the location of the money is not causally dependent on choice. In those conditions, with this heuristic removed, people tend to two-box.

If this is right, then the totality of results would suggest that it is much more difficult than previously assumed to design studies that test non-philosophers’ intuitions about the Newcomb case under the assumption that there is no causal dependence between the location of the money and participant’s choice. However, the more that this can be done, the more likely non-philosophers are to be two-boxers.

At this point we want to return to the question of how reflecting on these results is relevant to theorising about decision theory.

There are three ways our results could potentially be put to use in this regard. First, they could be used to respond to an argument from the descriptive facts regarding people’s tendency to one--box, to a normative claim regarding which decision theory is the correct one. Here is a stab at how that argument might go.

*The Theory Choice Argument*

1. Choosing to one-box reveals an evidential decision theory intuition.
2. If most people have a particular decision theory intuition, then that is a reason to adopt a normative decision theory that vindicates that intuition.
3. Most people choose to one-box.
4. Therefore, most people have an evidential decision theory intuition (from 1, 3)
5. Therefore, we have a reason to adopt evidential decision theory (from 2, 4).

Consider (1). You might think that choosing to one-box just manifests an evidential decision theory intuition. Indeed, not only have most evidentialists accepted this, but so have some causal decision theorists (Gibbard and Harper (1978, 183) and Joyce (1999, 154). Of course, one might reject (1). First, there are other normative theories that vindicate one-boxing.[[12]](#footnote-13) Second, Ninan (2014) argues that the fact that people are drawn to one-boxing does not show that they have evidential decision theory intuitions. First, he argues that this cannot be so, since in so-called medical Newcomb cases we no longer have the same intuitions even though the cases are structurally similar with respect to evidential and causal factors. This is taken to show that our intuitions in the original Newcomb case are not really evidential intuitions at all. Second, Ninan argues that we can explain people’s tendency to one-box without supposing that they have evidential intuitions. His idea is that faced with the Newcomb problem, anyone ought to have some, at least small, credence in the hypothesis that their choice *will* have a causal effect on the contents of the boxes. And if people do give some such small credence, then causal decision theory tells them that they should one-box.

Next, consider (2). Various philosophers working in decision theory have accepted something like (2). They hold that intuitions about Newcomb’s problem—to wit, the sorts of considerations we adumbrated earlier with respect to the money either being or failing to be, in the opaque box—have been taken to motivate causal decision theory. Nozick (1993, 41-50) goes further, and argues that we should incorporate both causal and evidential intuitions into a single decision theory, in part based on the idea that Newcomb’s problem elicits both sets of intuitions. If that’s right, then we have reason to accept (2).

Of course, one might resist (2) on the grounds that intuition is not a good guide to normative theorising either (a) in general or (b) at least when it comes to theorising about decision theory. For one might be inclined to say that intuitions about decision theory itself are likely to be of little value since there is good evidence that, across various domains of decision-making, people are systematically irrational (see Kahneman 2011 for a good summary). It may be, then, that our intuitions about decision making are a poor guide to any such normative theory.

Our current results, however, show that even if one accepts (1) and (2) there is reason to reject (3). Thus, our results give us a reason to reject this argument for the conclusion that we should endorse evidential decision theory.

That brings us to another argument on which our results shed light.

*The Evaluation Argument*

1. Causal decision theory is correct.
2. If causal decision theory is correct, then we should two-box in a Newcomb case.
3. Most people choose to one-box in a Newcomb case.
4. Therefore, most people make the wrong choice in a Newcomb case (from 13-15).

We take (9) to be philosophically interesting. Again, one might resist this argument. (6) is controversial. And against (7), we note that while most causal decision theorists accept that if causal decision theory is correct then we should two-box in a Newcomb case[[13]](#footnote-14) some do not.[[14]](#footnote-15) Notably though, if one accepts (6) and (7), then previous empirical results suggest that this argument is sound. Our current results, however, suggest otherwise. Instead, the picture they paint of people’s decisions in Newcomb cases is much more nuanced than (8). Granting (6) and (7), the strongest conclusion our results support is that most people seem to make the “wrong” choice in a Newcomb case only because they are assuming a causal connection between choice and prediction. But, of course, if there really were such a causal connection, then one-boxing would be the correct choice even according to causal decision theory.

Moving on to other aspects of our results. We did not find that people who choose to two-box will prefer to learn that they had one-boxed. Rather, most people who chose to two-box preferred to learn that this is what they had done. Moreover, there was no significant difference between these two conditions (i.e., there was no significant difference between two-boxers’ responses in the choice and news-item conditions).

This result suggests that non-philosopher two-boxers generally do not make the distinction between choice preferences and news-item preferences. There are two ways this finding could be used to advance debate in this area. First, non-causalists might argue that it undercuts a potential argument in favour of causal decision theory: that non-philosopher two-boxers decide as they do *because* they understand the difference between choice and news-item preferences. If this were the case, then it could be argued that two-boxing amongst non-philosophers is associated with a higher level of conceptual sophistication when evaluating choices. Given the prominent role this distinction plays in philosophical motivations for causal decision theory (e.g., in Joyce, 1999), it would be impressive if non-philosopher two-boxers decide as they do for seemingly similar reasons. Instead, it seems that something other than the choice-versus-news-item distinction motivates two-boxing for non-philosophers.

Next, let’s turn to what our results tell us about the connection between decision and future bias. First, it is worth noting that we found lower levels of future bias in this study than have been found in some studies. Depending on the exact features of the vignettes used, prior studies have found somewhat different levels of future bias in the population. Some of that difference reflects a difference in whether the state of affairs over which people are forming the preference are of equal value (see Greene, Latham, Miller and Norton 2022, forthcoming). But even focussing only on studies in which the past state of affairs is of the same magnitude as the future state of affairs, we find significant differences in future bias reported. Greene, Latham, Miller, and Norton (2021) found that ~80% of people were future biased about hedonic events (76% for positive events and 86% for negative ones), while Latham, Miller, and Norton (2023) and Latham, Miller, Tarsney and Tierney (2022) both found that approximately 75% of people were future biased. These studies all used similar vignettes in which participants were to imagine receiving either a pleasant or unpleasant food. Other studies have found much more variation using different vignettes. Latham, Miller, Shpall, Yu, and Oh (2023) compared future bias with respect to various kinds of sensations and with respect to mood, and they found that about 50% of people were future biased about positive/negative sensations and 40% about positive/negative mood. Our results are at the lower end of these findings. We think that is almost certainly due to some differences in the vignettes we used. These vignettes attempt to control for certain factors, including (a) the probability of the event occurring and (b) its value/magnitude. We suspect that one reason levels of future bias in this study are somewhat lower than in some other studies is that the affect generated by imagining either the positive or the negative effect of the pill is partially swamped by the positive affect that is generated by imagining that the fatal genetic disease has *already* been successfully treated. Future bias is thought to be in part the product of differential affect produced by imagining past versus future states of affairs (Ramos, Caruso, and Van Boven 2022; Molouki, Hardisty, and Caruso 2019)–namely, the product of imagining past states of affairs generating less affect than imagining future states of affairs. Given this, if the affect generated by imagining the positive/negative effects of the pill regardless of where those effects are located is in general diminished by being swamped by imagining the curing of the fatal genetic disease, we would expect to see lower levels of future bias, which is indeed what we find. In all, our results are consistent with previous results which suggest that the degree to which people display future bias varies depending on the kind of state of affairs that is being imagined, and, very likely is in part a function of the affect generated by imagining the states of affairs.

Returning to our hypotheses then, we found no difference between future bias amongst one-boxers and two-boxers. If the practical irrelevance explanation for future bias is correct, we might have expected to find that two-boxers were more future biased than one-boxers because according to this explanation of future bias, the fact that past states of affairs are causally inaccessible means that we attach less evaluative weight to them (Horwich 1987 (pp. 194-196); Maclaurin & Dyke 2002; and Suhler & Callender 2012). By contrast, if the temporal metaphysics explanation of future bias is correct, we would not have expected to find any association between future bias and one- or two-boxing.

While this finding is interesting, we do not take it to provide strong evidence against the practical irrelevance explanation and in favour of the temporal metaphysics explanation. Views on decision-relevant dependencies and future-biased preferences do not necessarily go hand in hand. It is possible for one to view the evidence that choices provide about past states as irrelevant to decision-making, while at the same time caring very much about past states.

Indeed, our results suggest that this is how things are. Our results suggest that if the practical irrelevance explanation is correct, then the fact that people attach less evaluative weight to past states of affairs because they are causally inaccessible[[15]](#footnote-16) does not have any impact on whether past states of affairs are taken to be decision-relevant. This is also an interesting upshot of this research, since it pulls apart two ways in which we might care about past states of affairs: the evaluative way and the decision-relevant way. While we might have thought it likely that these would go together, our data suggests that they do not.

Finally, while our findings suggest that people lack the capacity to distinguish choice from news-item preferences (as per H2), our findings with regard to H3 suggest that people often do distinguish between decision-relevant dependencies and future-biased preferences. This provides some support for the view that people’s decision making (at least when considering future-biased preferences) accords with normative decision theory. According to normative decision theory, decision rules are defined with respect to *previously* specified preferences over outcomes (typically represented by a utility function). Thus, if people are rational in the way specified by normative decision theory, we would expect their future-biased preferences to be prior to and independent of their choice preferences in concrete decision problems like the Newcomb case: people start with preferences, including future-biased preferences, and then use those preferences to make decisions. This is precisely what we find.

5. Conclusion

In our second study we found that a majority of non-philosophers are two-boxers in the Newcomb case. In comparison, previous research, including our own, has found that non-philosophers tend to endorse one-boxing. We have suggested that this is the result of people employing a, quite likely tacit, heuristic that takes there to be a causal connection between very strongly correlated items, and hence on which there is a backwards causal connection between choice and prediction in the Newcomb case. Our study suggests than when we undermine the use of that heuristic by making salient outcomes that rule out the presence of a causal connection, people no longer choose to one-box. In turn, we have argued that the descriptive claim that most people intuitively favour one-boxing cannot be marshalled in favour of evidential decision theory. Further, one cannot claim that if causal decision theory is correct, then most people make the wrong choice in the Newcomb case. In fact, according to our results, people’s intuitive judgments are in line with causal decision theory, since causal decision theorists recommend one-boxing in cases where there is a causal connection between choice and prediction, and two-boxing cases where there is no such connection.

Finally, we did not find any evidence that two-boxing is associated with future bias. We found this surprising given the similarities in the “future focus” of common rationales for both two-boxing and future bias. Instead, non-philosophers’ future bias seems to be independent of their choice preferences in Newcomb cases. In this way, their responses were in line with normative decision theory, according to which decision rules are defined with respect to previously specified preferences over outcomes.

References

Ahmed, A. (2018). ‘Introduction.’ In *Newcomb’s Problem*, ed. A. Ahmed, 1–18. Cambridge University Press.

Caruso, E., D. Gilbert and T. Wilson. (2008). ‘A wrinkle in time: asymmetric valuation of past and future events.’ *Psychological Science* 19: 796–801.

Bourget, D and Chalmers, D (2020). “Philosophers on Philosophy: The 2020 PhilPapers Survey” https://philpapers.org/rec/BOUPOP-3

Craig, W.L. (1999). ‘Tensed time and our differential experience of the past and future.’ *Southern Journal of Philosophy* 37: 515–37.

Dougherty, T. (2011). ‘On Whether to Prefer Pain to Pass.’ *Ethics* 121: 521–37.

Dougherty, T. (2015). ‘Future bias and Practical Reason’. *Philosophers’ Imprint* 15.

Egan, A. (2007). “Some counterexamples to Causal Decision Theory.” Philosophical Review 116 (1):93-114.

Gardner, M. (1974, March). Reflections on Newcomb’s problem: a prediction and free-will dilemma. *Scientific American*, Vol. 230, No. 3, pp. 102-109.

Gauthier, D., (1986). Morals By Agreement. OUP.

Gibbard, A. and Harper, W. (1978). “Counterfactuals and Two Kinds of Expected Utility.” In A. Hooker, J. J. Leach & E. F. McClennen (eds.), [*Foundations and Applications of Decision Theory*](https://philpapers.org/rec/HOOFAA-2). D. Reidel. pp. 125-162.

Greene, P. (2018). ‘Success-First Decision Theories.’ In *Newcomb’s Problem*, ed. A. Ahmed, 115–37. Cambridge University Press.

Greene, P. (Forthcoming). ‘“Pure” Time Preferences Are Irrelevant to the Debate over Time Bias: A Plea for Zero Time Discounting as the Normative Standard.’ *Australasian Journal of Philosophy*.

Greene, P. and M. Sullivan. (2015). ‘Against Time Bias.’ *Ethics* 125: 947–70.

Greene, P., Latham, A. J., Miller, K and Norton, J. (2021a). ‘Hedonic and Non-Hedonic Bias Towards the Future.’ *Australasian Journal of Philosophy* <https://doi.org/10.1080/00048402.2019.1703017>

Greene, P., Latham, A. J., Miller, K., & Norton, J. (2021b). “On Preferring that Overall, Things are Worse.” *Philosophy and Phenomenological Research.* <https://doi.org/10.1111/phpr.12819>

Greene, P., Latham, A. J., Miller, K, & Norton, J. (forthcoming). “How much do we discount past pleasures?” *American Philosophical Quarterly.*

Greene, P., and Levinstein, B. (2020) “Act Consequentialism Without Free Rides.” Philosophical Perspectives 34 (1):88-116.

Jeffrey, R. C, (1965). *The Logic of Decision.*  University of Chicago Press.

Joyce, J, M. (1999). *The Foundations of Causal Decision Theory.* Cambridge University Press.

Kahneman, D. (2011). *Thinking Fast and Slow*. Penguin Books.

Lee, R., Hoerl, C., Burns, P., Fernandes, A. S., O'Connor, P. A., & McCormack, T. (2020). Pain in the past and pleasure in the future: The development of past–future preferences for hedonic goods. Cognitive Science, 44(9), Article e12887. [https://doi.org/10.1111/cogs.12887](https://awspntest.apa.org/doi/10.1111/cogs.12887)

Levinstein, B. and Soares, N. (2020). “Cheating Death in Damascus” *Journal of Philosophy* 117(5): 237-266.

MacCrimmon, K.R. and Larsson, S. (1979) Utility Theory: Axioms versus ‘Paradoxes’. In: Allais, M. and Hagen, O. Eds., Expected Utility Hypotheses and the Allais Paradox, D. Reidel Publishing Company, Dordrecht, 333-409.
http://dx.doi.org/10.1007/978-94-015-7629-1\_15

McClennen, E. F. (1990) *Rationality and Dynamic Choice: Foundational Explorations*. CUP.

Meacham, C. (2010). “Binding and its Consequences.” Philosophical Studies 149 (1):49-71.

Miller, K. (2021). “What Time-Travel Teaches Us about Future bias.” *Philosophies* 2021, 6, 38. <https://doi.org/10.3390/philosophies6020038>

Miller, K. (forthcoming). “Tensed Facts and the Fittingness of our Attitudes.’ *Philosophical Perspectives.*

Molouki, S., Hardisty, D. J., & Caruso, E. M. (2019). The sign effect in past and future discounting. [Psychological Science, 30, 1674–1695.](https://www.google.com/url?q=https%3A%2F%2Fdoi.org%2F10.1177%2F0956797619876982&sa=D&sntz=1&usg=AOvVaw0TXZltpa-T9vdVoQ-rOzFb)

Ninan, D. (2014). “Illusions of Influence in Newcomb’s Problem”

http://www.dilipninan.org/papers/newcomb.pdf

Nozick, R. (1969). ‘Newcomb’s Problem and Two Principles of Choice.’ In N. Rescher, ed., *Essays in Honor of Carl G. Hempel*. D. Reidel: 114­–46.

Pearson, O. (2018). ‘Appropriate Emotions and the Metaphysics of Time.’ *Philosophical Studies* 175(8): 1945–1961.

Prior, A.N. (1959). ‘Thank Goodness That’s Over.’ *Philosophy* 34: 12–7.

Ramos, J., Caruso, E. M., and Van Boven, L. (2022). “Prospection, Retrospection, and Well-Being”. *Temporal Asymmetries in Philosophy and Psychology.* Edited by C Hoerl, T McCormack and A Fernandes. OUP.

Schlesinger, G. (1976). ‘The Stillness of Time and Philosophical Equanimity.’ *Philosophical Studies* 30: 145–59.

Shafir, E and Tversky, A., (1992). “Thinking through uncertainty: Nonconsequential reasoning and choice,” *Cognitive Psychology,* Volume 24, Issue 4: 449-474.

Sobel, J. H. (1985). ‘Circumstances and Dominance in a Causal Decision Theory.’ *Synthese* 63(2): 167–202.

Spohn, W. (2012). “Reversing 30 years of discussion: why causal decision theorists should one-box. *Synthese* 187 (1):95-122.

Sullivan, M. (2018). *Time Biases*. Oxford: Oxford University Press.

Sohn, W. (2012). ‘Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box.’ *Synthese* 187: 95–122.

Tarsney, C. (2017). ‘Thank Goodness That’s Newcomb: The Practical Relevance of the Temporal Value Asymmetry.’ *Analysis* 77(4): 750–9.

Tversky, A., & Shalir, E. (1992). The disjunction effect in choice under uncertainty. Psychological Science, 3(5), 305-309. https://doi.org/10.1111/j.1467-9280.1992.tb00678.x

Yudhowsky, E., and Soares, N. )(2018) “Functional Decision Theory: A New Theory of Instrumental Rationality.” [arXiv:1710.05060](https://arxiv.org/abs/1710.05060)

1. Even though evidential decision theory is the most prominent justification for one-boxing in the philosophical literature, there exist many alternative proposals for the rationale behind one-boxing. For example, *functional decision theory* proposes that one-boxing is rational based on the “algorithmic,” and not evidential, dependencies between state and act (Yudkowsky and Soares, 2018; Levinstein and Soares, 2020; Greene and Levinstein, 2020). Other examples of one-boxing alternatives to evidential decision theory include Gauthier (1986, Chp. 6), McClennen (1990), and Meacham (2010), Spohn (2012), and Greene (2018). [↑](#footnote-ref-2)
2. An anonymous referee has questioned whether Joyce’s distinction between auspiciousness and efficacy plays a justificatory role in his defense of causal decision theory as opposed to a merely dialectical one. We think the passage above illustrates that the distinction plays a justificatory role, since Joyce uses it to explain why people feel torn in Newcomb problems, and a theory’s explanatory power contributes to its justification. [↑](#footnote-ref-3)
3. Future bias refers to a preference for where events are located. Connected to future bias is what is often known as the temporal value asymmetry. This refers to an asymmetry in assigning value to past and future states of affairs, in particular to the fact that people tend to assign more value to a state of affairs when it is located in the future, as opposed to being equidistant in the past (Caruso, 2010; Caruso, Gilbert and Wilson, 2008; Roh and Schuldt, 2014). It has often been thought that future bias is a manifestation of this temporal value asymmetry in that we prefer negative states of affairs to be in the past and not the future because we accord them less value (and hence less negative value) if they are in the past. By the same token, we prefer positive states of affairs to be located in the future, rather than the past, because we accord them more positive value if they are in the future. Not everyone accepts that future bias is a manifestation of the value asymmetry. Hoerl (2022) points to the fact that people only show a value asymmetry in *between-subject designed experiments* and not in *within-subject designed experiments.* In a between-subject condition of the Caruso et al (2008) experiment, each saw a single vignette: that vignette either described future boring work, or past boring work. In the within-subject condition, participants got to see both vignettes and decide on compensation. This effectively allows them to compare their compensation rates. The fact that they found no temporal value asymmetry in the within-subjects design was taken by Caruso et al (2008) to be evidence that people do not take it to be justified to differently value past over future states of affairs, and so they tend to bring their judgements in line with one another. As Hoerl points out, since people probably do think that it is justified to prefer pains to be in the past, and pleasures in the future, this suggests that future-bias may not simply be a manifestation of the value asymmetry. Lee and McCormack (2022) also raise concerns about the idea that future bias is a manifestation of the value asymmetry. They note that a preference for pain to be located in the past rather than the future has been shown to emerge considerably earlier, developmentally, than does the value asymmetry. If that is right, then it’s difficult to see how future bias could simply be a manifestation of the value asymmetry. [↑](#footnote-ref-4)
4. There are two possible explanations for these differences. First, it could be that asking people *who they would prefer to be* (as in the Lee et al study) as opposed to *what experiences they would prefer to have* (as in the Greene et al study) tends to result in people taking a less temporally embedded perspective and hence in reducing future bias. A second possible explanation appeals to autobiographical memory. Lee et al. (2020) leaves open the possibility that the individuals described in the vignette remember the relevant past states of affairs, while in Greene et al. (2021b; forthcoming) the individual is said to temporarily fail to remember whether the past state of affairs occurred. Since the recollection of a memory can itself be a positive or negative hedonic event, the presence of such memories could play a role in mitigating future biased preferences. [↑](#footnote-ref-5)
5. This view has been suggested by Hume (1739, sec. 2.3.7.6), Parfit (1984, p. 186), and Horwich (1987, pp. 194-196), and developed by Maclaurin & Dyke (2002) and Suhler & Callender (2012). [↑](#footnote-ref-6)
6. Some such authors hold that the presence of the relevant metaphysical facts explains the temporal value asymmetry, and that tensed emotions and future bias are both manifestations of that asymmetry. [↑](#footnote-ref-7)
7. “Suppose there is a brilliant (and very rich) psychologist who knows you so well that he can predict your choices with a high degree of accuracy. One Monday as you are on the way to the bank he stops you, holds out a thousand-dollar bill, and says: ‘You may take this if you like, but I must warn you that there is a catch. This past Friday I made a prediction about what your decision would be. I deposited $1,000,000 into your bank account on that day if I thought you would refuse my offer, but I deposited nothing if I thought you would accept.’ Joyce’s version is adapted from that by Sobel (1985, 198–9 n. 6). Sobel points out that the classic formulation of Newcomb’s problem can lead to unnecessary “mechanical complications.” Thus, in Sobel’s version there is no need for money to be described as inside boxes of varying opaqueness, and the agent is asked whether they would like to *take the thousand* or *leave the thousand*, instead of being asked the equivalent, but unnecessarily complex, question of whether they would like to *take the thousand* *and take the million* or *leave the thousand and take the million.* [↑](#footnote-ref-8)
8. We choose the chi-square test statistic because our data is a series of categorical variables. For instance, “choose” (take; refuse), “learn” (take; refuse), and so on. The one-way chi-square tests whether the observed data differs significantly from people being evenly divided between the choice options. The chi-square test of independence tests whether there is a significant association between the categorical variables in people. [↑](#footnote-ref-9)
9. Our reported results do not change if we exclude people that report having a time-neutral preference, χ2 (1, *N* = 74)= .180, *p* = .672. We found no evidence that valence has any effect on the association between people’s decision to one-box or two-box and future-biased preferences (Breslow-Day: *p* > .650). [↑](#footnote-ref-10)
10. Our reported results do not change if we exclude people that report having a time-neutral preference, χ2 (1, *N* = 69)= .261, *p* = .609. We found no evidence that valence has any effect on the association between people’s decision to one-box or two-box and future-biased preferences (Breslow-Day: *p* > .513). [↑](#footnote-ref-11)
11. Thanks to an anonymous referee for raising this concern. [↑](#footnote-ref-12)
12. See Footnote 1. [↑](#footnote-ref-13)
13. Of course, as Ninan (2014) notes, if one gives some small but sufficiently high credence to not being in a genuine Newcomb case, then causal decision theory says that you should one-box. [↑](#footnote-ref-14)
14. Spohn (2012). [↑](#footnote-ref-15)
15. Assuming one accepts that future bias is a manifestation of the temporal value asymmetry, or at least, that the temporal value asymmetry is part of what explains why we are future biased. [↑](#footnote-ref-16)