Capacity for Simulation and Mitigation Drives Hedonic and Non-Hedonic Time-Biases

Philosophical Psychology, 2021

https://doi.org/10.1080/09515089.2021.19

Abstract

Until recently, philosophers debating the rationality of time-biases have supposed that people exhibit a first-person hedonic bias toward the future, but that their non-hedonic and third-person preferences are time-neutral. Recent empirical work, however, suggests that our preferences are more nuanced. First, there is evidence that our third-person preferences exhibit time-neutrality only when the individual with respect to whom we have preferences—the preference target—is a random stranger about whom we know nothing; given access to some information about the preference target, third-person preferences mirror first-person preferences. As a result, the *simulation hypothesis* has been proposed, according to which third-person preferences will mirror first-person preferences when we can simulate the mental states of the preference target. Second, there is evidence that we prefer negative hedonic events to be in our past (we are first-person negatively hedonically future-biased) only when we view future events as fixed and in no way under our control. By contrast, when we perceive it to be within our power to mitigate the badness of future events, we are first-person negatively hedonically past-biased. This is the *mitigation hypothesis*. We distinguish two versions of the mitigation hypothesis, the *squirrelling version* and the *heuristic version*. We ran a study which tested the simulation hypothesis, and which aimed to determine whether the squirrelling or the heuristic version of the mitigation hypothesis enjoys more empirical support. We found support for the heuristic version of the hypothesis, but no support for the squirrelling version.

1. Introduction

Let's say that an agent is *biased toward the future* if they tend to prefer that positively valenced events be in their future and that negatively valenced events be in their past. When the events in question are sensations, such as pleasure or pain, we will say that an agent is *hedonically* future-biased.[1] Further, we will say that an agent is *first-person positively hedonically future-biased* just in case they tend to prefer to have pleasures in their future, and *first-person negatively hedonically future-biased* just in case they tend to prefer to have pains in their past. By contrast, people are time-neutral about a certain kind of event when they have no preference regarding whether it is in their future or past. A common assumption is that humans display future-biased preferences regarding first-person positive and negative hedonic events.[2]

By contrast, philosophers assume that people are time-neutral when it comes to preferences concerning *non-hedonic* events and preferences that are *third-personal.*[3] Non-hedonic events are events that are not pleasures or pains. These are typically events that are not experienced directly by the agent, or where the agent's experience is not tied to the temporal location of the event. For instance, suppose that at some point in your life, someone whose opinion you value loses their respect for you, for a period of one week, after which you regain your prior standing. However, this person keeps this entirely private such that it doesn't change their behaviour at all and you remain ignorant of your temporary loss of standing. This is, we think, a negative event for you, even though it does not cause you any pleasure or pain. Despite this lack of hedonic consequences, this seems

[1] For a more formal characterisation of hedonic time-biases see Greene and Sullivan (2015, 948–9).
[2] See, inter alia, Prior (1959), Hare (2007, 2008), Heathwood (2008), Brink (2011), Greene and Sullivan (2015), Dougherty (2015), Hedden (2015) and Parfit (1984).
[3] See Hare (2013), Brink (2011, 378), Greene and Sullivan (2015, 968), Dougherty (2015, 3, fn. 4), and Parfit (1984, 181).

like the kind of event you might have temporal preferences about: you might prefer it to have already occurred in the past, you might prefer it to occur in the future, or you might be time-neutral and have neither preference. Third-person preferences are the preferences that a person has about events relevant to another person: the *preference target*.

Supporters of the rational permissibility of future-bias, for example, have either explicitly claimed that people are not future-biased about non-hedonic events (Hare, 2013; Parfit, 1984, 160), or focused exclusively on hedonic examples in motivating future-bias (Heathwood, 2008, 57; Prior, 1959).[4] The same is true of third-person preferences (Hare, 2008; Parfit, 1984, Section 69). And these predictions have led supporters of time-neutralism to give arguments for the rational impermissibility of future-bias on the grounds that if people only fail to be time-neutral when reacting to *their own pleasures and pains*, then this makes these future-biased preferences ad hoc, or otherwise suspect.[5]

To date, some empirical work has supported these assumptions and some has called them into question. Consider the supporting evidence first. An early and influential study on future-bias by Caruso, Gilbert, and Wilson (2008) asked participants to determine fair compensation for boring work that either occurred in the past or will occur in the future. Participants assigned themselves 60% more compensation for future work in first-person conditions, suggesting a first-person hedonic future-bias. In contrast, Caruso et al. found that in third-person conditions participants recommended the same compensation regardless of whether the work occurred in the past or will occur in the future, suggesting that participants' third-person hedonic preferences are time-neutral. This supported the predicted first- versus third-person asymmetry.

[4] Parfit (1984) refuses to label future-bias irrational, but ultimately remains neutral about its rational permissibility.

[5] See Brink (2011, 378–9), Greene and Sullivan (2015, 968), and Dougherty (2015, 3, fn. 4).

By contrast, a subsequent study by Greene, Latham, Miller and Norton (2021) found evidence of both first-person *and third-person* hedonic future-bias. They presented participants with vignettes in which a positive or negative hedonic or non-hedonic event either occurred in their immediate past or will occur in their immediate future. They found that a significant majority of participants reported first-person positive and negative hedonic future-biased preferences. They also found that a significant majority of people were future-biased in third-person positive and negative hedonic conditions, alongside symmetries in people's first- and third-person preferences regarding non-hedonic events.

These results contradict the supposed first-person/third-person asymmetry. Greene et al. (2021) thus conclude that, contrary to philosophers' predictions, there is no across-the-board first-person/third-person asymmetry; they instead present a more nuanced picture according to which people will have time-neutral third-person preferences in some circumstances and future-biased third-person preferences in others.

In particular, Greene et al. propose that what explains the difference between their results and those of Caruso et al. is that in the Caruso et al. experiment, participants were asked to determine fair compensation for complete strangers: people about whom they knew nothing. By contrast, in the Greene et al. experiment, participants were asked to report their preference about the timing of events for a particular individual who had a name, a profession, and some likes and dislikes (though notably participants were told nothing of the preference target's time preferences). They hypothesised that this prompted participants to simulate *being* the preference target, and thus participants' third-person preferences mirrored their own future-biased first-person preferences: the *simulation* *hypothesis.*[6]

A follow-up study by Greene, Latham, Miller and Norton (forthcoming) called into question the claim that people exhibit first-person negative hedonic future-bias. This study used the very same first-person positive hedonic and negative hedonic vignettes from the original Greene et al. (2021) study, but elicited participants' preferences in a slightly different way. The original study presented participants with a statement of the form 'I would prefer to learn that [the event] occurred yesterday', or 'I would prefer to learn that [the event] will occur tomorrow' and then asked them to respond to that statement by moving a slider along a Likert scale ranging from 1 'strongly disagree' to 7 'strongly agree'. We call this the A/D (agree/disagree) slider. By contrast, in the follow-up study participants were able to directly indicate their preference by moving a slider on a Likert scale between 'I would strongly prefer to learn that [the event] occurred yesterday' and 'I would strongly prefer to learn that [the event] will occur tomorrow' via a mid-point of 'I have no preference between these two options'. We call this the P/F (past/future) slider.

Greene et al. (forthcoming) predicted that this minor difference in methodology would make no difference to the results of the study. That is not, however, what they found. While this second study replicated Greene et al.'s earlier finding that people are positively hedonically future-biased, it did not replicate the finding that people are negatively hedonically future-biased. Instead, the second study found that people preferred to have negatively valenced hedonic events in their *future* rather than their past: they exhibited a *negative hedonic past-bias.* Past-bias is the inverse of future-bias: people who are past-biased about negatively valenced events prefer that such events be located in their future, while

[6] For discussion of our capacity to project ourselves in such a way as to take on someone else's perspective, see Buckner and Carroll (2007) and Goldman (2006). On episodic simulation more generally, and its connection to memory and to imagining potential novel future events see Szpunar, Spreng and Schacter (2016).

people who are past-biased about positively valenced events prefer that such events be located in their past. Greene et al.'s surprising finding, then, was that people preferred to have *both positive and negative hedonic events* in their future rather than their past.

Greene et al. (forthcoming) hypothesise that the use of the P/F slider in their second study encouraged participants to take an active role in considering ways they might intervene on future events to make things go better overall. They called this the *mitigation hypothesis*. The idea is that the P/F slider is labelled in such a way as to make it resemble a temporal axis which, Greene et al. hypothesise, led participants to imagine that they were *choosing* where along the temporal axis to put the event, rather than merely expressing a preference for where it occurs. In turn, they hypothesise that when participants take themselves to be choosing where to place an event they will take a time-neutral perspective.[7] Why, then, did the study find past-bias rather than time-neutrality? Greene et al. suggest that this was because participants imagined that events located in the future are ones over which they have, or at least might have, some control, and hence events whose relative badness they might be able to mitigate. As such, participants preferred to locate the negative event in the future to leave some chance for mitigation.[8]

[7] It seems plausible to hypothesise that people are more likely to be time-neutral when the outcomes are the result of their temporally extended planning, and that people are more likely to be future-biased when the events are random occurrences outside of their control. Greene et al. (forthcoming) support this hypothesis by presenting a variant of Parfit's influential (1984, 165) *My Past or Future Operations* case in which whether the agent has had a more painful past surgery or will have a less painful future surgery is the result of their own actions. Greene et al. argue that their variant encourages a time-neutral perspective, whereas Parfit's original version—in which the selected surgery is the result of seemingly random events—encourages a future-biased perspective.

[8] An anonymous reviewer suggested that, in a very long journey full of bland meals, the relief from this monotony afforded by a *different* meal—even one's most disliked meal—might be welcome, and even viewed as a positively valenced event in virtue of its novelty. This is an intriguing suggestion, but there are a couple of reasons to doubt that this is what explains Greene et al.'s (forthcoming) findings. Firstly, Greene et al. (2021) found that most people preferred their least favourite meal to be located in the past when asked using the A/D slider. Secondly, since the meal in question is specified to be the participant's most disliked meal (which they are stipulated to "really dislike"; it's not merely their least favourite of the meals they like), we are doubtful that novelty would be sufficient for participants to view this as a positively valenced event.

In effect, then, participants took themselves to be *making a decision* between a past event of *N* units of disutility and a future event of *less-than-N* units of disutility in expectation (i.e., an event of less expected disutility given the chance of mitigation). Consequently, from the time-neutral perspective they inhabited as a result of taking themselves to have a choice (rather than merely expressing a preference), participants quite naturally chose the smaller future expected disutility over the larger past one.

In fact, there seem to be two distinct versions of the mitigation hypothesis that are not clearly differentiated by Greene et al. (forthcoming). In the experiment, participants imagine being an astronaut whose meals are dispensed by the ship. The ship normally dispenses bland meals, but once during the trip it will dispense their most disliked meal. Participants indicate whether they would prefer to learn that their most disliked meal was dispensed yesterday or will be dispensed tomorrow. Sometimes, Greene et al. present participants as engaging in a kind of active problem solving, whereby they consider the specific details of the scenario, and the ways in which they might mitigate the suffering caused by the ship dispensing their least favourite meal tomorrow.

For instance, they imagine participants cleverly thinking about how they would "choose to save and set aside some of today's bland meal to eat tomorrow, in order to mitigate the badness of receiving the disliked meal." Here, the idea seems to be that participants are being led to a more agentive perspective as a result of the labelling of the P/F slider, which in turn leads them to take on a more time-neutral perspective. Then, participants are actively looking for ways in which they can *use* the fact that *if* their most disliked meal is coming tomorrow, *and* they were lucky enough to learn about it today, they then have the opportunity to mitigate. On the other hand, if their most disliked meal arrived yesterday, they have no reason to think that they would have learned of its impending arrival the day before yesterday.

Let's call this the *squirrelling* version of the mitigation hypothesis (inspired by the idea of squirrelling away some of today's bland meal to mitigate tomorrow's suffering). The squirrelling version of the mitigation hypothesis says that when participants take an agentive perspective, they are more inclined toward time-neutrality, and they more closely examine the details of the scenario to look for strategies, which, if they were to learn that the negative hedonic event will happen tomorrow, would enable them to mitigate the event. They then respond as though they are being offered a choice between unequal payoffs (more disutility in the past, versus less expected disutility in the future), and prefer the option that entails the best expected utility overall (i.e., the best from a time-neutral perspective)..

At other times, Greene et al. (forthcoming) seem to have in mind what we call the *heuristic* version of the mitigation hypothesis. This version of the hypothesis shares with the squirrelling version the idea that participants are being led to a more agentive perspective as a result of the labelling of the P/F slider, which leads them to take a more time-neutral perspective. But rather than supposing that participants actively consider problem-solving strategies peculiar to the vignette under consideration, according to this version of the hypothesis taking this agentive perspective leads people to quite generally treat past events as fixed and certain, and future events as unfixed and uncertain: as events that can, in principle, be mitigated, or otherwise intervened upon. If participants suppose that there is some chance that a negative event will not happen, or if they accept that it will happen, but suppose that its badness might in some way, by someone, be mitigated, then from a time-neutral perspective they are likely to prefer that event to be in the future. That is because they take themselves to be choosing between a fixed past event of disutility *N*, and an open future event whose disutility may, somehow, end up being *less-than-N*.

According to both versions of the mitigation hypothesis, participants no longer take themselves to be considering an event with some particular expected utility, and expressing preferences regarding whether they prefer that event to be in the past versus the future. Instead, if that hypothesis is correct then participants are effectively being asked whether they prefer an event that has one expected utility and is located in the past, or an event that has a different expected utility and is located in the future. Participants assign those two events different subjective probabilities or different utilities.

While Greene et al. (forthcoming) bundle these two versions of the mitigation hypothesis together, each individually could explain their results, or they could do so in tandem. Thus, more work is required to determine the best explanation of their surprising finding of past-bias. The present study seeks to make headway on this question.

We take as our starting point the observation that the two versions of the mitigation hypothesis, together with the simulation hypothesis, generate different sets of predictions regarding the pattern of responses we will observe if we present participants with the same set of *non-hedonic* and *third-person* vignettes as those presented in Greene et al. (2021) but use a P/F slider as per Greene et al. (forthcoming).

Overall, we should expect to find that where the valence of the event is positive, our results will be the same as those found in Greene et al. (2021). That is because mitigation is irrelevant in the positive hedonic condition: there's nothing to mitigate. However, where the valence of the event is negative, there is something to mitigate. In negative conditions, different versions of the mitigation hypothesis generate different predictions because the different versions of the hypothesis are ones on which the mitigation takes different forms.

If the *squirrelling* version of the mitigation hypothesis is true, and the simulation hypothesis is true, we should expect to see a shift from future-bias towards past-bias in third-person negative hedonic conditions. That's because the simulation hypothesis predicts that participants will place themselves in the third-person's position and treat that condition as though it were a first-person condition. By contrast, we should *not* expect to see a shift from future-bias towards past-bias in negative non-hedonic conditions, because the non-hedonic vignettes (which we present in §2.1.2) explicitly leave no room for squirrelling. There is, in fact, no opportunity for participants to intervene in order to mitigate the negative non-hedonic events. So even if participants are prompted to search for mitigation strategies, they will quickly realise that none are forthcoming.

The *heuristic* version of the mitigation hypothesis, on the other hand, predicts a shift from future-bias towards past-bias in all the negative conditions, including the negative non-hedonic conditions. That is because if the heuristic version of the mitigation hypothesis is true, then some of those participants who reported future-bias or time-neutrality in Greene et al.'s (2020) first-person negative non-hedonic condition will instead report past-bias, since they will be led to suppose that the badness of negative future events may be mitigated. Despite there being no apparent strategies for participants themselves to mitigate in these particular cases, there is nonetheless still time for some fortuitous occurrence to mitigate the badness of the negative future event. Moreover, if the simulation hypothesis is true, then we would expect to find that third-person preferences mirror first-person preferences. So, wherever we predict a shift towards past-bias in the first-person conditions, we should also predict a shift towards past-bias in the third-person conditions.

In what follows we set out more precisely the predictions of each version of the mitigation hypothesis.

Greene et al. (2021) found that when responding with the A/D slider, third-person positive and negative hedonic conditions mirrored first-person positive and negative hedonic conditions: the mean response was significantly above 4 (participants were, on average, future-biased) and in addition a significant majority of people were future-biased in all four conditions. Greene et al. (forthcoming) found that when responding with the P/F slider, the mean response (and the significant majority) in the first-person negative hedonic condition was past-biased, while the mean response (and the significant majority) in the first-person positive hedonic condition was future-biased. Since mitigation is irrelevant in the positive hedonic condition—parity of reasoning suggests that instead participants would want the opportunity to *enhance* a positive future event, resulting in future-biased preferences—we would expect that use of the P/F slider would yield results that are the same as those afforded using the A/D slider. By contrast, in the third-person negative hedonic condition both the squirrelling and heuristic versions of the mitigation hypothesis predict that people will shift towards past-bias because in that condition squirrelling is possible.

Thus we predict the following:

In the third-person positive hedonic condition:

(H1a) the mean response will be significantly above 4, and

(H1b) a significant majority of participants will be future-biased.

In the third-person negative hedonic condition:

(H2a) the mean response will be significantly below 4, and

(H2b) a significant majority of participants will be past-biased.

Now consider the non-hedonic conditions. In the positive non-hedonic conditions, neither version of the mitigation hypothesis is in play. So we expect to replicate Greene et al.'s (2021) findings that in these conditions the mean response will not differ significantly from 4 (time-neutrality), but that a significant majority of participants will be non-future-biased (i.e. they will be either time-neutral or past-biased). Thus, we predict the following:

In the first-person positive non-hedonic condition:

(H3a) the mean response will not differ significantly from 4, and

(H3b) a significant majority of participants will be non-future-biased.

In the third-person positive non-hedonic condition:

(H4a) the mean response will not differ significantly from 4, and

(H4b) a significant majority of participants will be non-future-biased.

Finally, consider the negative non-hedonic conditions. Since these are negative conditions, mitigation is relevant. In these conditions, however, no squirrelling is possible. So, the squirrelling and heuristic versions of the mitigation hypothesis make different predictions. It is important to note that the hypotheses regarding the different versions of the mitigation hypothesis are exploratory in nature and that future confirmatory studies will be required to replicate the results that we report here.

The squirrelling version of the mitigation hypothesis predicts that we will find no shift towards past-bias since squirrelling is impossible. By contrast, the heuristic version of the mitigation hypothesis predicts that we will still find a shift towards past-bias because even though participants have no way to mitigate the negative non-hedonic event themselves, they will react to the fact that future events are unsettled and hence could by fortuitous chance, be made less bad.

So, the heuristic version of the mitigation hypothesis predicts that where Greene et al. (2021) found that mean response was significantly above 4 (participants were, on average, future-biased) in both negative non-hedonic conditions, we should find that on average participants are past-biased. Moreover, where Greene et al. found that the split between future-biased and non-future-biased participants does not significantly differ from a 50/50 split, we ought find that the split between *past-biased* and non-past-biased participants does not significantly differ from a 50/50 split. So the heuristic version of the mitigation hypothesis instead predicts the following:

In the first-person negative non-hedonic condition:

(H5a) the mean response will be significantly below 4, and

(H5b) the split between past-biased and non-past-biased participants will not significantly differ from a 50/50 split.

In the third-person negative non-hedonic condition:

(H6a) the mean response will be significantly below 4, and

(H6b) the split between past-biased and non-past-biased participants will not significantly differ from a 50/50 split.

By contrast, the squirrelling version of the mitigation hypothesis predicts that we will simply replicate the Greene et al (2021) results, since mitigation (via squirrelling) can play no role. Hence the squireling version of the hypothesis predicts that:

In the first-person negative non-hedonic condition:

(H5a*) the mean response will be significantly above 4, and

(H5b*) the split between future-biased and non-future-biased participants will not significantly differ from a 50/50 split.

In the third-person negative non-hedonic condition:

(H6a*) the mean response will be significantly above 4, and

(H6b*) the split between future-biased and non-future-biased participants will not significantly differ from a 50/50 split.

In §2 we outline the methodology, analyses and results of the study we ran to test these predictions. In §3 we discuss the upshots of our findings for the simulation and mitigation hypotheses.

2. Experimental Design and Results

2.1 Method

2.1.1 Participants

421 people participated in the study. Participants were U.S. residents, recruited and tested online using Amazon Mechanical Turk, and compensated $0.50 for approximately 5 minutes of their time. 80 participants had to be excluded for failing to follow task instructions. This means that they failed to answer the questions (48), or failed an attentional check question (32). The remaining sample was composed of 341 participants (aged 19-70; 119 female; Mean age 30.36 (SD = 8.66)). Ethics approval for this study 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.

2.1.2 Materials and Procedure

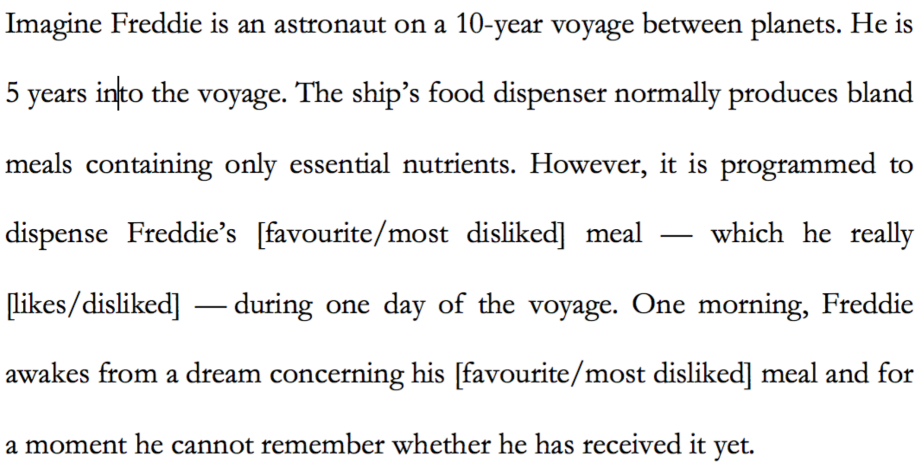
Participants were randomly assigned to one of six conditions. These six conditions reflect some combinations of valence (positive or negative), kind of event (hedonic or non-hedonic) and perspective (first-person or third-person). Two combinations were not tested in the current study, namely the first-person positive hedonic and first-person negative hedonic conditions tested by Greene et al. (forthcoming).

We presented participants with Greene et al.'s (2021) vignettes,[9] but they used Greene et

[9] It is notable that the vignettes are contrived in certain ways in order to avoid certain confounds. In particular, Greene et al (2021) located the relevant party (first- or third-person) on a spacecraft that is out of contact with earth, in order to avoid the possible confound that participants might take themselves to be able to intervene on whether the relevant future event occurs or not. If participants suppose they can prevent the event if it is in the future, then their reported preferences regarding the temporal location of the event might not concern whether the event is past or future, but instead whether the event occurs or does not occur.

al.'s (forthcoming) P/F sliding scale to report their preferences.

Participants in the third-person hedonic conditions received the following vignette in either its positive or negative form:



Imagine Freddie is an astronaut on a 10-year voyage between planets. He is 5 years into the voyage. The ship's food dispenser normally produces bland meals containing only essential nutrients. However, it is programmed to dispense Freddie's [favourite/most disliked] meal — which he really [likes/disliked] — during one day of the voyage. One morning, Freddie awakes from a dream concerning his [favourite/most disliked] meal and for a moment he cannot remember whether he has received it yet.

In each condition, participants were presented with a Likert scale that ran from 1 (I would *strongly* prefer to learn that Freddie's [favourite]/[most disliked] meal was dispensed yesterday, and will not be dispensed tomorrow) to 7 (I would *strongly* prefer to learn that Freddie's [favourite]/[most disliked] meal will be dispensed tomorrow, and was not dispensed yesterday) at the other end of the scale via 4 'I have no preference between these two options' in the middle of the scale. Participants then moved the slider[10] to wherever on the scale they took to reflect their preference.

Participants in the first-person positive non-hedonic condition read a vignette describing the receipt of a community service prize. Participants in the first-person negative non-hedonic conditions read a vignette describing having embarrassing photographs released.

These participants read a version of the following vignette:

[10] In all conditions, the orientation of the slider was randomised, so that which end was a preference for the event to occur in the future and which a preference for the event to occur in the past was randomly varied between participants.

> Imagine you are an astronaut on a 10-year voyage from Earth to set up a colony on a new planet. It is a one-way mission, and there is no way you can return to Earth. You are 5 years into the voyage. Just before you left, you learned that [your home-town mayor plans to award you an important community service prize]/[someone plans to release embarrassing photos of you] at some time during the 10-year period in which you are traveling]. You do not know when they will [award the prize]/[release the photos], and it is not possible to communicate with Earth during the trip, or even once you have arrived on the new planet. You find yourself wondering whether [the prize has been awarded]/[photos have been released] yet.

Participants in the third-person positive and negative non-hedonic conditions read a version of the vignette above, amended into third-person form replacing 'I' with 'Freddie' in all cases (and changing personal pronouns where appropriate).

Participants in the third-person positive non-hedonic condition were presented with a Likert scale that ran from 1 'I would *strongly* prefer to learn that the important community service prize was awarded to Freddie yesterday, and will not be awarded tomorrow' through to 7 'I would *strongly* prefer to learn that that the important community service prize will be awarded to Freddie tomorrow, and was not awarded yesterday' via 4 'I have no preference between these two options' in the middle of the scale. Participants then moved the slider to wherever on the scale they took to reflect their preference. Participants in the third-person negative non-hedonic condition were presented with equivalent statements about Freddie's embarrassing photos, while participants in the first-person conditions were presented with equivalent statements about their own embarrassing photos/community service prize.

In all six conditions participants were then asked to indicate their level of confidence in their previous judgement.[11] After having done so, participants answered a comprehension question: *"In this vignette, you were asked to imagine that [you were]/[Freddie is]…"* to which they could answer (1) an astronaut; (2) a dog; (3) a builder or (4) the home-town mayor. Participants who did not choose (1) were excluded.

2.1.3 Analyses

In order to compare the results between conditions, we re-coded participants' responses in such a way that a response of 5, 6, or 7 reflects future-bias (a preference that the event in question be future if it is positive and past if it is negative) in all conditions. Likewise, a response of 1, 2, or 3 reflects past-bias (a preference that the event in question be future if it is negative and past if it is positive) in all conditions. A response of 4 indicates time-neutrality.

In order to test for future-bias we first ran separate one-sample t-tests to test whether the mean preference significantly differs from 4 in each condition. If the mean is significantly above 4, then overall people might be future-biased, and if the mean is significantly below 4 then overall people might be past-biased. A non-significant t-test result might indicate time-neutrality. (We say *might*, here, since it is consistent with a mean response greater than 4 that half our sample is future-biased and half past-biased, yet the future-biased participants report their preference with greater strength (responses of 6 or 7) than do the past-biased participants (who might report this preference with responses of 3). Such a pattern of responses would not vindicate the claim that people are future-biased overall. *Mutatis mutandis* for a mean of 4, which might reflect a large population of time-neutral participants, or two equal populations of future-biased and past-biased participants.)

[11] Participants were quite confident in their judgements regarding their preferences (M = 5.49, SD = 1.18). Results of a one-way ANOVA showed that there was no significant difference in level of confidence between conditions (p = .687).

In those conditions in which the mean response suggests that people are future-biased, we then combined the proportion of people who were past-biased with those who were time-neutral—we will call these people *non-future-biased*—and in conditions in which the mean response suggests that people are past-biased, we combined the proportion of people who were future-biased with those who were time-neutral—we call these people *non-past-biased.* We then ran separate one-sample χ^2-tests to test whether in those conditions where there was mean future-bias, the *majority of people* responded in a future-biased manner, and whether in those where there was mean past-bias, the *majority of people* responded in a past-biased manner.

Finally, we performed exploratory analyses comparing future-bias across conditions. We did this with an analysis of variance (ANOVA) with between-subjects variables of valence (negative, positive), event (hedonic, non-hedonic), and position (first-person, third-person). In order to perform this analysis, we pooled the results from our six conditions with results provided by Greene et al. (forthcoming) for the first-person positive hedonic condition and first-person negative hedonic condition.

2.2 Results

Overall, we found mean future-bias in all the positively valenced conditions. However, in all these conditions, the split between future-biased and non-future-biased participants did not significantly differ from a 50/50 split. Likewise, we found mean past-bias in all the negatively valenced conditions. However, once again, in all these conditions, the split between past-biased and non-past-biased participants did not significantly differ from a 50/50 split.[12]

[12] In what follows there were also no significant main effects of age and gender, nor were there any significant interaction effects with age and gender. Further, the inclusion of age and gender as factors has no effect on the results that we report in this paper.

Thus, our hypotheses fared as follows.

We hypothesised that in the third-person positive hedonic condition, (H1a) the mean response will be significantly above 4, and (H1b) a significant majority of participants will be future-biased. Our results support (H1a), but not (H1b). Instead, the split between future-biased and non-future-biased participants does not significant differ from a 50/50 split.

We hypothesised that in the third-person negative hedonic condition, (H2a) the mean response will be significantly below 4, and (H2b) a significant majority of participants will be past-biased. Our results support (H2a), but not (H2b). Instead, the split between past-biased and non-past-biased participants does not significantly differ from a 50/50 split.

We hypothesised that in the first- and third-person positive non-hedonic conditions, the mean response will not differ significantly from 4 ((H3a) and (H4a)), and a significant majority of participants will be non-future-biased ((H3b) and (H4b)). Our results support none of these four hypotheses. In each condition, the mean response was significantly above 4, and the split between future-biased and non-future-biased participants does not significantly differ from a 50/50 split.

Finally, we explored the hypothesis that if the squirrelling version of the mitigation hypothesis is correct, then in the first- and third- person negative non-hedonic conditions, the mean response will be significantly above 4 ((H5a*) and (H6a*)), and there will be an even split between future-biased and non-future-biased participants ((H5b*) and (H6b*)). By contrast, if the heuristic version of the mitigation hypothesis is correct, then in the first- and third-person negative non-hedonic conditions, the mean response will be significantly below 4 ((H5a) and (H6a)), and there will be an even split between past-biased and non-past-biased participants ((H5b) and (H6b)).

We find that in both conditions the mean response is significantly below 4, and there is an even split between past-biased and non-past-biased participants. That is, our results support the heuristic over the squirrelling version of the mitigation hypothesis.

We present our more detailed results below. Table 1 summarises the descriptive data from the experiment. After the reverse-coding described in §2.1.3, the 'FB' column represents the proportion of participants who reported future-biased preferences (5, 6 or 7) and the 'PB' column represents the proportion of participants who reported past-biased preferences (1, 2 or 3). The 'N' column represents the proportion of people who reported time-neutrality (4).

Table 1. Descriptive data from all conditions.

Condition	%FB	%PB	%N	Mean	SD	*t-value*	*p*-value
Condition 1: Third-Person Positive Hedonic (N = 52)	61.5	7.7	30.8	4.94	1.31	5.208	<.001
Condition 2: Third-Person Negative Hedonic (N = 52)	13.5	61.5	25.0	3.25	1.37	-3.947	<.001
Condition 3: First-Person Positive Non-Hedonic (N = 60)	43.3	11.7	45.0	4.52	1.20	3.335	.001
Condition 4: Third-Person Positive Non-Hedonic (N = 61)	60.6	11.5	27.9	4.85	1.39	4.795	<.001
Condition 5: First-Person Negative Non-Hedonic (N = 58)	20.7	51.7	27.6	3.45	1.40	-2.993	.004
Condition 6: Third-Person Negative Non-Hedonic (N = 58)	20.7	53.4	25.9	3.41	1.46	-3.051	.003

The results of our t-tests—which, recall, show us whether the mean response differs significantly from a value of '4' (time-neutrality)—appear to show that overall, people are future-biased in all positive conditions (conditions 1, 3, and 4) and that overall, people are past-biased in all negative conditions (conditions 2, 5, and 6).

However, the t-test does not tell us whether the *majority* of people in those conditions are past-biased or future-biased: for that we must look to the results of our χ^2-tests, reported in Table 2 below. Recall that in negative conditions, where there was mean past-bias, we grouped together time-neutral and future-biased participants as a non-past-biased group (non-PB). In positive conditions, where there was mean future-bias, we grouped together time-neutral and past-biased participants as a non-future-biased group (non-FB).

Table 2. Results of χ^2-tests.

Positive Conditions (with mean future-bias)	**%FB**	**%non-FB**	χ^2	***p*-value**
Condition 1: Third-Person Positive Hedonic	61.5	38.5	2.769	.096
Condition 3: First-Person Positive Non-Hedonic	43.3	56.7	1.067	.302
Condition 4: Third-Person Positive Non-Hedonic	60.6	39.4	2.770	.096
Negative Conditions (with mean past-bias)	**%PB**	**%non-PB**	χ^2	***p*-value**
Condition 2: Third-Person Negative Hedonic	61.5	38.5	2.769	.096
Condition 5: First-Person Negative Non-Hedonic	51.7	48.3	0.069	.793
Condition 6: Third-Person Negative Non-Hedonic	53.4	46.6	0.276	.599

The results of our χ^2-tests show us whether the proportion of participants responding in a future-biased versus non-future-biased way in the positive conditions, or in a past-biased versus non-past-biased way in the negative conditions, differs significantly from a 50/50 split. In the positive conditions (1, 3 and 4) the split between future-biased and non-future-biased participants does not significantly differ from a 50/50 split. Similarly, in the negative conditions (2, 5, and 6) the split between past-biased and non-past-biased participants does not significantly differ from a 50/50 split.

In order to compare future-bias across conditions we tested future-bias with a 2x2x2 between-subjects ANOVA. Recall that this was an exploratory analysis. In order to perform this analysis we included the results from Greene et al.'s (forthcoming) first-person positive and negative hedonic conditions, which we did not test in the current experiment. The result of this test found a significant main effect of valence (negative/positive) ($F(1, 434) = 140.625$, $p < .001$). We also observed a significant two-way interaction between event (hedonic/non-hedonic) and valence ($F(1, 434) = 4.815$, $p = .029$). Crucially, we did not observe a significant main effect of position (first-person, third-person), or any significant interaction effect involving position.

The main effect of valence showed that future-bias was significantly higher in positive conditions (*M* = 4.85, *SD* = 1.37) than in negative conditions (*M* = 3.31, *SD* = 1.37).

Simple effects tests using a Bonferroni correction were carried out on the two-way interaction between event type and valence. First, for hedonic conditions, future-bias was significantly higher in positive conditions (*M* = 5.02, *SD* = 1.36) than in negative conditions (*M* = 3.20, *SD* = 1.36; *p* < .001). Second, similarly for non-hedonic conditions, future-bias was significantly higher in positive conditions (*M* = 4.69, *SD* = 1.36) than in negative conditions (*M* = 3.43, *SD* = 1.36; *p* < .001). Third, there was no evidence that valence (negative/positive) had an affect on the relation between event type (hedonic/non-hedonic) and future bias.

3. Discussion

Consider the simulation hypothesis first. This hypothesis predicts that, given the kind of details about the preference target we provide in our vignettes, we will find the same pattern of results across first-person and third-person conditions. If a majority of people (or the mean response) exhibited one sort of bias in a first-person condition, then we should expect the same result in the analogous third-person condition.

Our results, in combination with those of Greene et al. (forthcoming) provide strong support for this hypothesis. In every positive condition, the mean response was future-biased, and in every negative condition, the mean response was past-biased. This was so regardless of whether participants were reporting a first- or third-person preference.

When we look to the majorities, we see a first/third person symmetry in the non-hedonic conditions. In the first- and third-person positive non-hedonic conditions we found the same result: the split between future-biased and non-future-biased participants was not significantly different from a 50/50 split. Likewise, in the first- and third-person negative non-hedonic conditions we found the same result: the split between past-biased and non-past-biased participants was not significantly different from a 50/50 split.

This symmetry is less clear when we compare our third-person hedonic conditions with Greene et al.'s (forthcoming) first-person hedonic conditions. In the first-person positive hedonic condition, a majority of participants reported a future-biased preference, but in the third-person positive hedonic condition the split between future-biased and non-future-biased participants was not significantly different from a 50/50 split. Likewise, in the first-person negative condition a majority of participants reported a past-biased preference, but in the third-person negative hedonic condition the split between past-biased and non-past-biased participants was not significantly different from a 50/50 split.

However, that the mitigation hypothesis is having an effect on people's preferences in the third-person negative hedonic condition—participants are shifting from future-bias to past-bias—lends support to the simulation hypothesis here. For we would only expect mitigation-style reasoning to have this effect if people were simulating the preferences of the preference target.

Most importantly, we found no effect of perspective (first or third-person) or any interaction involving perspective. This supports the idea that where simulation of the preference target is possible, people are inclined to have the same sort of time-biased third-person preferences as they do first-person preferences. That is, participants were inclined to report preferences for the preference target by simulating their own preferences.

It seems plausible that if participants were given even more information about the preferences of the preference target, and indeed, were given information that the preference target's preferences are unlike their own preferences, then they would form somewhat different third-person preferences from their own first-person preferences. As it stands, however, we have no data on this matter. Useful follow-up work would involve presenting participants with vignettes in which the preferences of the preference target are specified, and can be taken to be different from the preferences of the participants themselves.

Our prediction is that if participants know that the preferences of the preference target are time-biased, their preferences on behalf of that preference target will also exhibit time-bias, and will be so regardless of whether their own preferences are time-neutral or time-biased. By contrast, if participants know that the preferences of the preference target are time-neutral, then they will exhibit time-neutral preferences on behalf of the preference target, and will do so regardless of the nature of their own preferences.

Next, consider the mitigation hypothesis. Both versions of the mitigation hypothesis predict that in the third-person negative hedonic condition we will find mean past-bias when using the P/F slider (in contrast to the future-bias Greene et al. (2021) found when people use the A/D slider). This is what we found. The two versions of the mitigation hypothesis, however, make different predictions regarding the first- and third-person negative non-hedonic conditions. The squirrelling version predicts that we will not see a shift towards past-bias because there is no possibility of squirrelling, while the heuristic version of the hypothesis predicts that we will see a shift towards past-bias in these conditions. We found that in all negative conditions, the mean response was past-biased. This supports the heuristic version of the mitigation hypothesis over the squirrelling version.

More support for the heuristic version of the mitigation hypothesis stems from comparing our descriptive statistics and the results of our χ^2-tests to those of Greene et al. (2021). Recall that their methodology resulted in a split between future-biased and non-future-biased participants in the first- and third-person negative non-hedonic conditions that did not significantly differ from 50/50. When we asked participants to respond using the P/F slider we instead found that in these conditions the split between *past-biased* and non-past-biased participants did not significantly differ from 50/50.

This suggests that there is a shift from future-bias towards past-bias in these conditions, which is confirmed by comparing our descriptive statistics to those of Greene et al. In the first-person negative non-hedonic condition, Greene et al. found that 51.4% of participants were future-biased, and 28.2% were past-biased. As reported in Table 1, in the equivalent condition in the present study, 51.7% of participants were past-biased, and 20.7% were future-biased. While the proportion of time-neutral participants is quite consistent across the two studies, Greene et al. found roughly twice as many future-biased as past-biased participants, while we found roughly twice as many past-biased as future-biased participants. We see the very same pattern looking at the analogous third-person conditions across the two studies.[13]

Similarly, Greene et al.'s (2021) methodology resulted in a significant majority of participants reporting future-bias in the third-person negative hedonic condition. While our hypothesis that we would see a significant majority of participants reporting past-bias in this condition was not supported, we instead saw that the split between past-biased and non-past-biased participants did not significantly differ from 50/50.

This, too, suggests that there is a shift from future-bias towards past-bias in the third-person negative hedonic condition, which is once again confirmed by comparing our descriptive statistics to those of Greene et al. (2021). In this condition, Greene et al. found that 61.5% of participants were future-biased, and 22.9% were past-biased. As reported in

[13] We don't see a shift to a significant majority of participants being past-biased in large part because there is a substantial portion of the population who are time-neutral (and who therefore count as non-past-biased). It is nevertheless the case that the percentage of past-biased, as compared with future-biased, participants, is shifting from future-biased towards past-biased.

Table 1, in the equivalent condition in the present study, 61.5% of participants were past-biased, and 13.5% were future-biased. Thus, although we do not see a significant majority of past-biased participants, comparison across the two studies confirms that the change of methodology elicits past-bias from a far larger proportion of participants.[14]

In all, then, when participants respond using a P/F slider, there is a shift towards past-bias in all negative conditions. This is what the heuristic (but not the squirrelling) version of the mitigation hypothesis predicts. Indeed, we found no significant difference between the conditions in which only the heuristic version of the hypothesis predicts past-bias and those in which both versions predict past-bias, which suggests that squirrelling is playing little to no role in explaining the surprising observation of past-bias across the negative conditions when participants respond using the P/F slider.

This provides evidence in favour of the heuristic version of the mitigation hypothesis. Of course, the extant research does not rule out that something else entirely might explain the reported data. It merely shows that of the two explanations on offer—the heuristic and squirrelling versions of the mitigation hypothesis—the heuristic version is best supported.

With that said, there are several reasons one might be hesitant to accept this explanation. First, one might worry that there is little independent reason to suppose that the small change in methodology that results from using the P/F slider rather than the A/D slider will produce the sort of *massive* difference in results that we find. Given this, one might worry that the results of our study are not robust, and instead are merely some kind of unimportant or uninteresting methodological artefact. Second, one might worry that even

[14] Indeed, although we do not see a significant majority of past-biased participants in this condition, the matchup between our results and those of Greene et al. (2021) is striking. Whereas they found that 61.5% of participants were future-biased and 38.5% were non-future-biased, we found that 61.5% of participants were past-biased and 38.5% were non-past-biased. The reason that their study found a *significant* majority to be future-biased, yet we did not find a significant majority to be past-biased, is that they had a larger sample size and thus more power.

if the results are robust, there are aspects of the explanation that require further elaboration if it is to do the work required. Third, one might worry that other recent empirical work undermines the explanation. Let's consider these in turn.

First, let's consider the robustness of the study. While it is surprising that such a small methodological change produces such different results, similar phenomena have been observed in nearby domains. For instance, one of the most well-known demonstrations of a framing effect comes from Tversky and Kahneman's (1981) disease problem. In this problem people are asked to choose between two programs designed to combat a mysterious illness which is expected to kill 600 people. In the positive framing condition, people are asked to choose between program A, in which 200 people will be saved, and program B, in which there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. Meanwhile, in the negative framing condition, people are asked to choose between program C, in which 400 people will die, and program D, in which there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. Importantly, each of the 4 program options has the same expected value, and programs A and C, and, B and D, respectively, are equivalent in terms of their likelihoods. Yet, despite this equivalence, most people choose program A in the positive framing condition, and program D in the negative framing condition.

Another well-known demonstration of small methodological changes producing significant alterations to the results comes from the Wason selection task (Wason, 1968). In this task you are presented a series of 4 double-sided cards and asked which ones you would need to turn over in order to confirm that the cards are consistent with a certain rule. So for example, imagine that the cards have a shape on one side and a colour on the other, and you are presented with a 'square card', 'circle card', 'green card', and 'blue card' and asked which ones you would need to turn over in order to confirm that they were consistent with the rule: if a card has a circle on one side, then it has the colour green on the other side. The correct answer is to turn the 'circle card' and the 'blue card', however, most people chose to turn the 'circle card' and 'green card'.

However, if people are presented with that same task under the description of an everyday setting, then they reliably get it right (Cosmides, 1989; Gigerenzer and Hug, 1992). So for example, imagine the cards have an age on one side and a drink on the other, and you are presented with a 'beer card', 'lemonade card,' '24-year-old card', and '17-year-old card' and asked to confirm that they were consistent with the rule: if a person drinks alcohol, then they must be over the age of 18 years old. Then people reliably choose the correct option of the 'beer card' and '17-year-old card'.

So there is evidence that small methodological differences in framing can have a large effect on participant responses.

The second worry concerns whether enough has been said about the heuristic version of the hypothesis to show how it can do the relevant explanatory work. In particular, one might think that it is not enough that participants come to see the future as uncertain, and to notice that the negative event *may* be mitigated if it is in the future. In addition, participants must suppose that uncertainty surrounding the event is more likely to lead to the event having *less,* rather than *more,* disutility. This is a reasonable assumption given the squirrelling version of the hypothesis: for then, participants have a strategy by which they can mitigate the future badness of the event. But we found past-bias in conditions in which participants cannot have in mind some particular strategy by which this mitigation can be achieved.

Rather, according to the heuristic version of the hypothesis, participants must be supposing that it is possible that the event will be mitigated if it occurs in the future, in a way that it clearly cannot be mitigated if it occurred in the past. This requires the further assumption that future uncertainties will not instead exacerbate the badness of the event. Exacerbation reasoning would tend to favour placing the negative event in the past rather than the future. The defender of the heuristic version of the mitigation hypothesis needs an account of why future uncertainty tends to lead people to suppose that future mitigation might occur, rather than to suppose that future exacerbation might occur.

We noted in the introduction that supporters of the squirrelling version of the mitigation hypothesis might hypothesise that placing events in the future implies forewarning, whereas placing events in the past does not. Thus, since our activities tend to focus on mitigating future problems and not exacerbating them, when people know about problems ahead of time they are potentially in a position to mitigate them. Now, the squirreling version of the hypothesis is one on which individuals consciously look for ways to mitigate, while the heuristic version of the hypothesis is not. Nevertheless, we suggest that the defender of the heuristic version of the mitigation hypothesis can in fact say something similar. If it is plausible that in general, placing events in the future implies that one can be forewarned (and hence forearmed), while placing them in the past does not, then we might expect a general heuristic to arise, which associates placing negative events in the future with the possibility of mitigation, and placing them in the past with no possibility of mitigation. Crucially, this heuristic might explain such preferences even when in fact there is no (practically) possible mitigation.

Third, and most worryingly, there is concern that other empirical results might cut against the heuristic version of the mitigation hypothesis. Latham, Miller, Norton and Tarsney (2020) ran a study in which (some) participants were explicitly told that their choice regarding whether to accept a future electric shock would retrocausally affect whether they had just received 10,000 shocks (if they accept the shock), or had instead just received

10,100 shocks (if they do not accept the shock). This study involved both explicit choice and inequality of outcomes: choosing not to accept one additional future shock causes it to be the case that there were 99 additional shocks in the past.

As predicted, Latham et al. found that future-bias was significantly diminished when people were making choices instead of reporting preferences: people were more inclined to accept the additional shock in choice conditions. This shift towards time-neutrality is predicted by the mitigation hypothesis. Interestingly though, Latham et al. found that even when participants were explicitly making a choice—and thus thinking agentively in the way that the mitigation hypothesis suggests is prompted by the P/F slider—there was not mean past-bias. Instead there was mean time-neutrality, and people were evenly divided between thinking that they would *refuse* the future shock, and thinking either that they would be equally likely to make either choice or that they would choose to accept the future shock. What we do not see, then, in the Latham et al. experiment, is a majority of participants showing past-biased preferences when thinking agentively about unequal outcomes.

This is puzzling. Suppose the shift to past-biased preferences due to use of the P/F slider is the result of people being led to think more agentively (to take themselves to be choosing) and to suppose that the future event might be less bad because somehow it has been mitigated (conditions of inequality). Then we would expect an experiment in which people are explicitly told that they have a choice, and that this will result in an inequality, would report at least as great a shift towards past-bias. Based on the Latham et al. results, however, that is not so. This suggests that either the mitigation hypothesis, even in its heuristic guise, is false, or at the very least that it is not the full explanation for the past-bias that we see when using the P/F slider.

There are, however, some notable differences between the two methodologies that could explain why we see a smaller shift away from future-bias in the Latham et al. experiment.

In that experiment, participants are told that they have been shocked at least 10,000 times in the (near) past—so many times, in fact, that they have lost count. The decision to accept a further future shock, then, will make it the case that they received 10,000, rather than 10,100 shocks, in the past. By contrast, if in the current experiment participants are being led to suppose themselves to have a choice about where on the timeline the negative event is located (past or future), then they are taking themselves to be able to choose whether the entirety of the salient negative event is located in the past, versus in the future, and not merely whether there is some additional disutility which can be avoided in the past, by locating some disutility in the future.

This difference might have an effect because of considerations of diminishing marginal disutility. Perhaps once one has already been shocked 10,000 times, and has lost count of said shocks, the additional disutility of an extra 100 past shocks is insignificant. If that were so, then choosing to accept the future shock does not really result in diminution of past disutility. We suspect that if people were told that they can choose to accept a single future voluntary shock which will make it the case that they were not shocked 100 times in the past, or they can decline the future shock and make it the case that they were shocked 100 times in the past, then people will be more past-biased than they are found to be in in the Latham et al. study.

Nevertheless, this is speculation. Even if we are correct about the explanation for the relative lack of past-bias in the Latham et al. study, overall these findings suggest that the heuristic version of the mitigation hypothesis may not be the full story. Still, we think that in the absence of any other hypotheses about the connection between the P/F slider and past-bias, and in the presence of the evidence we do have, there are reasons to endorse the heuristic version of the mitigation hypothesis.

If this is so, then our results, in concert with earlier work, have substantial implications for philosophical theorising about time preferences. If the simulation hypothesis is correct, then philosophical arguments against the rationality of future-bias that appeal to a lack of future-bias in third-person preferences may not be of value.[15] That's because insofar as there *is* any difference between third-person and first-person preferences it will be the result of one of two things.

First, as in Caruso et al. (2008), it may be the result of an inability to simulate the preferences of the preference target because too little is known about that target. In that case, though, it's unclear why we would think that people's third-person preferences are of much interest. It might be argued that the conditions under which we have preferences regarding what happens to nameless, faceless, and completely unknown individuals are the sort of highly abstract and unnatural conditions that are of limited value in telling us anything useful about our actual third-person preferences. More specifically, it seems plausible that evolutionary mechanisms will generate in us the capacity to simulate the preferences of those around us (albeit imperfectly), and that it will be an advantage to do this well in the sort of social environments in which we find ourselves.[16] The fact that in the absence of any information that provides useful cues to this mechanism, we respond by reporting more time-neutral third-person preferences, might be seen to tell us very little about our third-person preferences, assuming those preferences are, in general, formed through simulation when possible.

Of course, some caution is required here. Our study, too, involves somewhat unnatural conditions, in the form of a science-fictional spacefaring setting. This setting was introduced by Greene et al. (2021) in order to control for certain confounds, and has thus

[15] Such arguments appear in Brink (2011: 378–9), Greene and Sullivan (2015: 968), and Dougherty (2015: 3).

[16] See, for example, Premack and Woodruff (1978), Brüne and Brüne-Cohrs (2006), and de Waal (2008).

been used in follow-up work. However, this opens up questions about the generalisability of the results of this corpus of work. Moving beyond this limitation would be a valuable direction for future research in this area, attempting to control for these confounds in different ways, such that the scenarios in the vignettes are more familiar, or, at least, are unfamiliar in different ways. That would allow us to determine whether these contrivances are playing any role in determining participants' responses. Follow-up studies that aim to evaluate the simulation hypothesis would also do well to include a control condition that removes the information about the third-person and their preferences.

The second way in which first- and third-person preferences can come apart, given the truth of the simulation hypothesis, is if people have enough information about the preference target to attribute to that target different preferences than their own. But in this situation, it's hard to see why we would think that the disparity between first- and third-person preferences gives us any reason to think that time-neutrality is the rational view. That's because we have no reason to think that the preferences we simulate on behalf of the preference target, in this situation, will be time-neutral: all we know is that they will be different from those of the simulator.

Of course, none of this shows that future-bias rational. Rather, it suggests that given the truth of the simulation hypothesis, arguments against the rationality of future-bias which appeal to some asymmetry between first- and third-person preferences do not do the work required. As mentioned in Section 1, both supporters and detractors of future-bias assume that it is restricted to first-person preferences, and detractors have appealed to this assumption in motivating the rejection of future-bias on the grounds that if it is isolated to first-person preferences then it is ad hoc or arbitrary. In some instances, this argument connects with evolutionary debunking accounts of future-bias (Horwich, 1987, 196–8; Maclaurin and Dyke, 2002; Suhler and Callender, 2012). Greene and Sullivan (2015, Section V), for example, argue that third-person preferences are time-neutral because the agent has 'emotional distance' from the experiences in question, and is thus unlikely to be susceptible to the evolved emotions that result in time-biased preferences. But if there is no robust first-person/third-person asymmetry in future-biased preferences, then these arguments need to be abandoned or amended.

Turning now to the mitigation hypothesis, the support that these results give to the heuristic version of the hypothesis tends to undermine a widespread philosophical assumption, namely that we can relatively easily tease apart 'pure' from 'impure' time-biased preferences. A pure hedonic time-biased preference is a preference for a hedonic event to be at one point in time instead of another, "merely because of when it occurs in time" (Lowry and Peterson, 2011, 490). For example, a pure negative past-biased preference would be a preference for a negative event to be future merely because it would occur in the future and not the past. By contrast, if someone prefers that a negative event be future rather than past because its being so will, say, increase total utility, then this is not a pure past-biased preference.

Philosophers debating the rational permissibility of future-bias typically focus on thought experiments involving pure future-biased preferences, in which the *only* consideration relevant to the agent is an event's futurity/pastness. Some recent philosophical work has attempted to argue that it is all but impossible to disentangle pure from impure preferences (Callender, forthcoming) or that insofar as we could do so, pure time preferences are not the sorts of things we should care about (Latham, Miller, and Norton, forthcoming; Greene, forthcoming).[17]

[17] Greene (forthcoming) suggests that when philosophers debate the rational permissibility of future-bias they should ask "is futureness a rationally permissible ground for a preference" and not "are preferences that are solely grounded in futureness rationally permissible?" Thus, according to Greene, philosophers need not restrict themselves to pure future-biased preferences in debating the rational permissibility of future-bias.

If the mitigation hypothesis is correct, then this is further grist to the mill of these arguments. What we see in this study and in Greene et al. (forthcoming) is that changing the methodology from one that employs an A/D slider, to one that employs a P/F slider, makes a surprising difference to the preferences that people report. If the heuristic version of the hypothesis is correct, then the presence of the P/F slider generates a more agentive perspective, which in turn generates a more time-neutral perspective—one that aims to maximise utility overall—and we see an impure past-biased preference because we suppose it possible that the future is mitigable, while the past is not.

If this is true, then to isolate pure future-biased preferences, we need to suppress people's tendency to think agentively. But one might wonder about the usefulness of thinking about pure future-biased preferences and the norms that govern them if these are the preferences we have only when we refrain from thinking agentively.

While clearly we can (and do) have preferences concerning states of affairs we take to be outside of our control, one of the major roles that preferences play in both actual practice and philosophical theorising is as input into deliberation. But deliberation requires taking an agentive perspective. In order to deliberate about whether to *p*, or not-*p*, we must take *p* or not-*p* as open: we cannot deliberate about that which we take to be fixed and immutable.[18] But insofar as we see ourselves as agents with respect to *p*, and hence as being able to deliberate about whether to *p*, this will make our *p*-wise future- or past-biased preferences impure. So, when preferences serve as input to deliberations, the very fact of their doing so undermines their purity. If that is right then, pure future-biased or past-biased preferences would not be of primary philosophical (or psychological or economic) interest.

[18] See Price (1997).

4. Conclusion

Our study provides empirical support for both the simulation hypothesis and heuristic version of the mitigation hypothesis. The simulation hypothesis is supported by symmetries between our first- and third-person results. These results also suggest that arguments against the rationality of hedonic future-bias on the basis of an alleged first/third person asymmetry are ill-founded. The heuristic version of the mitigation hypothesis is supported by our finding a shift towards past-bias when compared to Greene et al. (2021) in all negative conditions. These results consolidate Greene et al.'s (forthcoming) finding of first-person negative hedonic past-bias, and suggest that the previously neglected phenomenon of past-bias is in need of more philosophical attention.

References

Brink, D. O. (2011). "Prospects for Temporal Neutrality". in Craig Callender (ed). T*he Oxford Handbook of Philosophy of Time.* Oxford: Oxford University Press.

Brüne, M. and Brüne-Cohrs, U. (2006) Theory of mind—evolution, ontogeny, brain mechanisms and psychopathology. *Neuroscience and Biobehavioral Reviews* 30: 437-455

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*, 49–57.

Callender, C. (forthcoming). 'The Normative Standard for Future Discounting' *Australasian Philosophical Review.*

Caruso, E., Gilbert, D. T., and T. D. Wilson (2008). "A Wrinkle in Time: Asymmetric Valuation of Past and Future Events". *Psychological Science* 19(8): 796-801

De Waal, F.B.M (2008). Putting the Altruism Back into Altruism: The Evolution of Empathy. *Annual Review of Psychology* 59: 279-300

Dougherty, T. (2015). "Future-Bias and Practical Reason". *Philosophers' Imprint* 15.

Goldman, A.I. (2006). *Simulating minds: the philosophy, psychology, and neuroscience of mindreading.* New York: Oxford University Press.

Greene, P. and M Sullivan, (2015). "Against Time-bias" *Ethics* (125)5: 947-970.

Greene, P., Latham, A. J., Miller, K., & Norton, J (2021). 'Hedonic and Non-hedonic Bias Towards the Future.' *Australasian Journal of Philosophy*, 99:1, 148-163. DOI: 10.1080/00048402.2019.1703017

Greene, P., Latham, A.J, Miller, K., and Norton, J. (forthcoming). 'Why are people so darn past-biased?' In *Temporal Asymmetries in Philosophy and Psychology*, edited by Christoph Hoerl, Teresa McCormack and Alison Fernandes. Oxford: Oxford University Press.

Hare, C. (2007). "Self-Bias, Time-Bias, and the Metaphysics of the Self and Time". J*ournal of Philosophy* 104(7): 350-373.

Hare, C. (2008). "A Puzzle about Other-Directed Time-Bias". *Australasian Journal of Philosophy* 86(2): 269-277.

Hare, C. (2013). "Time—The Emotional Asymmetry". In *A Companion to the Philosophy of Time*, edited by Adrian Bardon and Heather Dyke, Wiley Blackwell, pp. 507-520.

Heathwood, C. (2008). "Fitting Attitudes and Welfare" *Oxford Studies in Metaethics* 3:47-73.

Hedden, B. (2015). *Reasons Without Persons: Rationality, Identity, and Time.* Oxford: Oxford University Press.

Hilton, D. J., & Slugoski, B. R. (1986). "Knowledge-based causal attribution: The abnormal conditions focus model." *Psychological Review*, 93(1), 75–88.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11):587-612.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2):136.

Knobe, J. (2009). Folk judgments of causation. *Studies in History and Philosophy of Science Part A*, *40*(2), 238-242.

Latham, A. J., Miller, K, J. Norton, J. and Tarsney, C. (2020) "Future-bias in Action" *Synthese*. DOI: 10.1007/s11229-020-02791-0

Latham, A.J., Miller, K., and Norton J. (forthcoming). 'Pure and Impure Time Preferences' *Australasian Philosophical Review.*

Lowry, R., and Peterson, M. (2011). *Pure Time Preference*, Pacific Philosophical Quarterly 92:490–508.

Parfit, D. (1984). *Reasons and Persons.* Oxford: Oxford University Press.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences* 1(4): 515-526

Price, H. (1997). Time's arrow & Archimedes' point: new directions for the physics of time. Oxford: Oxford University Press.

Prior, A. N. (1959). "Thank Goodness That's Over". *Philosophy* 34(128): 12-17.

Szpunar, K. K., Spreng, R. N., & Schacter, D. L. (2016) 'Toward a Taxonomy of Future Thinking.' In K. Michaelian, S. B. Klein & K. K. Szpunar (eds.) *Seeing the Future.* Oxford: Oxford University Press.