**Future bias in action: Does the past matter more when you can affect it?**

**Abstract**

Philosophers have long noted, and empirical psychology has lately confirmed, that most people are ‘biased toward the future’: we prefer to have positive experiences in the future, and negative experiences in the past. At least two explanations have been offered for this bias: (i) belief in *temporal passage* (or related theses in temporal metaphysics) and (ii) the *practical irrelevance* of the past resulting from our inability to influence past events. We set out to test the latter explanation. In a large survey (*n* = 1462) we find that participants exhibit significantly less future bias when asked to consider scenarios where they can affect their own past experiences. This supports the ‘practical irrelevance’ explanation of future bias. It also suggests that future bias is not an inflexible preference hardwired by evolution, but results from a more general disposition to ‘accept the things we cannot change’. However, participants still exhibited substantial future bias in scenarios in which they could affect the past, leaving room for complementary explanations. Beyond the main finding, our results also indicate that future bias is *stake-sensitive* (i.e., that at least some people *discount* past experience rather than disregarding it entirely) and that participants endorse the normative correctness of their future-biased preferences and choices. In combination, these results shed light on philosophical debates over the rationality of future bias, suggesting that it may be a rational (reasons-responsive) response to empirical realities rather than a brute, arational disposition.

**1. Introduction**

Think of some experience you deeply enjoy, like eating your favourite meal. Imagine that you just had this experience yesterday. Now, imagine instead that you are going to have this experience tomorrow. Which of these thoughts do you find more appealing?

On the other hand, think of some unpleasant experience, like a painful dental procedure. Imagine that you just had this experience yesterday. Now, imagine instead that you are going to have this experience tomorrow. Which of *these* thoughts do you find more appealing?

If you are like most people, your answers to these questions exhibit a *bias toward the future:* all else being equal, you prefer that pleasant experiences be located in your future rather than your past, and that unpleasant experiences be located in your past rather than your future. More generally, you attach greater evaluative weight to pleasant and unpleasant experiences when they are in the future rather than the past.

Future bias has been pointed out and discussed by philosophers going back at least to Hume (1739). But it has only more recently begun to receive sustained philosophical attention, spurred by influential discussions in Prior (1959) (who highlighted the time-asymmetric character of evaluative attitudes like *relief*) and Parfit (1984) (who drew connections between future bias, temporal discounting, and egoistic bias). Recent contributions to this literature have considered the rationality of future bias (Hare 2007, 2013; Dougherty 2011; Greene & Sullivan 2015; Sullivan 2018; Dorsey 2018; Brink 2011), its psychological underpinnings (Maclaurin & Dyke 2002; Suhler & Callender 2012), and its connection with temporal metaphysics (Yehezkel 2013; Pearson 2018).

Of late, there is a small but growing empirical literature devoted specifically to investigating future bias. That work has offered support for some ideas in the philosophical literature while calling others into question. For instance, Caruso, Gilbert and Wilson (2008) find indirect evidence of a general pattern of future bias, via judgements about appropriate levels of compensation for equivalent events in the past and future. Notably, they find these time-asymmetric attitudes only in a first-person setting (when participants are asked to consider their own past or future experiences), and not in a third-person setting (when considering the experiences of others), in line with the suggestion by Parfit and others that future bias is a distinctively first-personal phenomenon.

Similarly, Greene, Latham, Miller and Norton (2020) found that a majority of participants were future-biased regarding hedonic events—events involving direct pleasures and pains—in a first-person setting, but that there was no such majority regarding non-hedonic events—events that are not directly experienced, or whose experience is not tied to the time of the event. By contrast to Caruso et al, however, Greene et al. found that a majority of participants were future-biased regarding third-person hedonic events.[[1]](#footnote-2)

This growing corpus of empirical research has begun to map the contours of our future-biased preferences.[[2]](#footnote-3) But a question that remains largely empirically unexplored is the *explanation* for this pattern of future bias. The main aim of this study is to begin filling that gap.

Philosophers have suggested at least two explanations for future bias: (i) (implicit) belief in *temporal passage* (or related theses in temporal metaphysics) and (ii) the *practical irrelevance* of the past, resulting from our inability to influence past events. Our aim in this study is to assess the plausibility of the latter hypothesis. In §2, we briefly survey the literature on both hypotheses. There, we distinguish two versions of the ‘practical irrelevance’ hypothesis. According to the first, future bias is an adaptation, hard-wired inflexibly into our affective systems. According to the second, it results from a more general disposition to affectively discount practically irrelevant events, and is mediated by our belief that we cannot affect the past. We extrapolate predictions from each of these hypotheses and endorse the predictions associated with the second version of the practical irrelevance hypothesis.

In §3 we describe our methodology and analyses, and present our results. While our results do not perfectly match our predictions, we nonetheless find support for the second (cognitively mediated) version of the practical irrelevance hypothesis. But we also find a significant level of residual future bias that may not be explained by that hypothesis, and which therefore leaves room for complementary explanations (including hardwired future bias or implicit beliefs about temporal metaphysics).

In §4, we discuss the philosophical implications of these results. In particular, our results may inform the debate over the *rationality* of future bias, suggesting that this bias is at least partially rational in the sense of being ‘reasons-responsive’: when we have practical reason to care about the past, we are less strongly biased toward the future.

**2. The literature and our hypotheses**

As already noted, the philosophical literature has offered two lines of explanation for future bias. The first is what we will call the *temporal metaphysics hypothesis*. According to this hypothesis, we are biased toward the future because of our intuitive/pre-theoretical belief in *temporal passage* or other related metaphysical theses about time.[[3]](#footnote-4) Both the idea of temporal passage and its connection with our attitudes toward past and future experiences are hard to articulate except in metaphors, but these metaphors have deep appeal: it seems to us (or so many philosophers have thought) that we are ‘moving through time’, from the past toward the future, and therefore that our past experiences lie ‘behind us’ while our future experiences lie ‘ahead of us’, in a way that makes it only natural to care more about the future than the past.[[4]](#footnote-5) In turn, it has been thought that these ‘seemings’ either inform, or are informed by, some naïve, or implicit, theory of time that represents an asymmetry between past and future (Callender 2017; Latham, Miller & Norton 2019).

The temporal metaphysics hypothesis is particularly well-represented in the literature inspired by Prior (1959), who argues that verbal expressions of time-asymmetric attitudes like relief require an irreducibly tensed semantics, which in turn implies (the speaker’s belief in) a metaphysics of irreducibly tensed facts or properties. Prior’s argument has been widely criticised (most famously by Mellor (1981, 1983) and MacBeath (1983)), but remains influential. (For a recent defence of Prior’s conclusion against the Mellor/MacBeath reply, see Pearson (2018).) Others have argued for a link between future bias and belief in temporal passage more directly, without going via semantics—for instance, Schlesinger (1976) and Craig (1999).

But though many philosophers have felt an intuitive connection between future bias and our sense of/belief in temporal passage, this connection has proven very hard to elucidate: there is not yet any satisfactory story as to *why* belief in temporal passage should cause us, or make it rational for us, to care more about the future than the past, that goes beyond the level of metaphor and intuition.[[5]](#footnote-6) Thus, many philosophers have felt the need for a less metaphysical explanation.

The second explanatory hypothesis that has been widely discussed in the philosophical literature is what we will call the *practical irrelevance hypothesis*. According to this hypothesis, we attach less evaluative weight to past events because there is nothing 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 do. This idea was perhaps first suggested by Hume, who writes that the greater effect of future events than past events on the will ‘is easily accounted for[:] As none of our actions can alter the past, ‘tis not strange it shou'd never determine the will’ (Hume, 1739, sec. 2.3.7.6).[[6]](#footnote-7) More recently, Kauppinen (2018) has argued that our future-biased preferences are rationally justified (and, he seems to think, explained) by the fact that they have no effect on our choices.[[7]](#footnote-8) Our inability to affect the past also underlies a popular evolutionary explanation for future bias suggested by Parfit (1984, p. 186) and Horwich (1987, pp. 194-196) and developed by Maclaurin & Dyke (2002) and Suhler & Callender (2012).

The observation that the past is inaccessible and hence practically irrelevant actually suggests two distinct explanations for future bias. The first is a direct evolutionary explanation: because our ancestors could gain reproductive advantage by trying to influence future events, but not by trying to influence past events, we evolved cognitive and/or affective dispositions that make a basic distinction between future and past events—and, in particular, that attach much more affective weight to future pleasures and pains than to past pleasures and pains. Let’s call this the *narrow adaptation hypothesis*.

Both Maclaurin & Dyke (2002) and Suhler & Callender (2012) seem to endorse this kind of explanation. According to the narrow adaptation hypothesis, our future-biased preferences are not only a specific adaptation, but are an adaptation that is relatively hard-wired, specified in function, and inflexible.[[8]](#footnote-9) In particular, the class of adaptations we have in mind are relatively developmentally inflexible, in such a way that they tend not to be modified by learning. Moreover, if this explanation is correct, then even though future bias *evolved* because the past is causally inaccessible, we should not expect that our tendency to discount past experiences *is mediated by the belief that* the past is causally inaccessible. By analogy, our gustatory aversion to grass evolved (one assumes) because grass generally has little if any nutritional value for human beings; but our distaste for grass is not mediated by the *belief* that grass is un-nutritious: it just tastes bad to us, and so we don’t eat it.

A second hypothesis, by contrast, holds that future bias is a result/application of a more general mechanism that (whether or not it is an adaption itself) is less specified in function and relatively more flexible. In particular, on this view, the mechanism is one that is sensitive to beliefs (which are sensitive to learning) about which events are causally accessible. The mechanism affectively or evaluatively[[9]](#footnote-10) discounts events that we believe to be causally inaccessible or otherwise practically irrelevant. Let’s call this the *cognitive mediation hypothesis*.

On this view, although the mechanism by which we attach affective weight to past or future events is mediated by our beliefs about their causal accessibility, these beliefs need not be explicit or conscious. But this hypothesis suggests that our devaluation of past events is at least to some extent flexible. If, in some exceptional circumstance, I believe that I have some influence over past events, then I will attach more affective weight to those events. (And, conversely, if I believe that there is nothing I can do to bring about or prevent some future event, then I will attach less affective weight to that event.)

Although the connection between future bias and the practical irrelevance of the past is widely noted and discussed in the philosophical literature, the distinction between these two very different hypotheses has not been widely recognised. But some philosophers have taken sides, at least implicitly, on the central question of whether our bias toward the future is mediated by our belief that the past is causally inaccessible and hence practically irrelevant.

Dorsey (2018), for instance, claims that it is not. Discussing Parfit’s famous case in which an agent does not know whether she has had a very painful surgery yesterday, or will have a moderately painful surgery later today, Dorsey writes that our inability to affect the past ‘makes little difference. One might simply imagine that it is, in fact, possible to change (causally speaking, that is) the past—by some sort of time-travel machine or the hotline to God. Few would opt for the choice to take the less painful surgery today in order to correct the ten-hour surgery yesterday’ (p. 1910). Kauppinen (2018), on the other hand, holds that future-biased preferences are rational only when they do not influence our choices, and seems to think that we in fact exhibit future-biased preferences only when those preferences are practically inert. So although he does not address the question directly, his view strongly suggests that future bias is cognitively mediated by the belief that the past is practically irrelevant.

The first aim of our study is to test the practical irrelevance hypothesis. The mechanism by which its chief competitor, the temporal metaphysics hypothesis, would bring about future bias, has not been well articulated. So it is not entirely straightforward to identify predictions that distinguish these two hypotheses and allow us to adjudicate between them. Nevertheless, it seems plausible that if we devalue the past because we believe that, in some hard-to-articulate sense, past events are ‘receding’ from us while future events are ‘approaching’ (as a result of past-to-future temporal passage), then we should equally devalue past events that we can retrocausally affect (so long as the mechanism of retrocausation does not place those events in our own experiential future, which in our vignettes it does not).

By contrast, the practical irrelevance hypothesis (or at least, the cognitively mediated version of that hypothesis) predicts that future bias will diminish in scenarios where the past is practically relevant—in particular, in scenarios where we can *affect* the past. Hence, the discovery that future bias can be mitigated in situations where we can affect the past would lend support to the practical irrelevance hypothesis over the temporal metaphysics hypothesis.

In addition, the practical irrelevance hypothesis seems to predict that future bias will be mitigated in cases where the past and future outcomes at stake are *unequal*. That is, if the practical irrelevance hypothesis is true, we would expect people to assign *some* weight to past disutilities, while assigning *greater* weight to future disutilities. So, for instance, they will generally prefer a given quantity of past suffering to the same quantity of future suffering, but will prefer a small quantity of future suffering to a large enough quantity of past suffering. We call this the *past discounting hypothesis*. (To avoid a potential confusion: This hypothesis asserts that we discount the past as a whole relative to the present/future, but does not assert—or deny—that we discount the *further* past relative to the *more immediate* past.) We call the rival hypothesisthat people will uniformly prefer any amount of past suffering (no matter how large) to any amount of future suffering (no matter how small) the *past indifference hypothesis*.

We think both versions of the practical irrelevance hypothesis will predict the past discounting hypothesis (with the qualification, as we explain below, that a strong enough version of the cognitive mediation hypothesis will predict that, under some circumstances, people will not even *discount* the past but will be fully time-neutral). First, although the narrow adaptation hypothesis says that our time-asymmetric dispositions are relatively inflexible, it does not predict that we are inflexibly *absolutely* future-biased such that past utilities and disutilities play no role at all in determining our preferences and choices. After all, in general, an organism will do better if it prefers more total lifetime utility to less, and so will do better if it prefers small future disutilities to large past disutilities. Further, affective responses to past events might serve various adaptive purposes even if they can’t directly guide our choices in the way that attitudes toward future events do. For instance, a persistent negative affective response to a painful past experience might help an organism *learn* to avoid similar painful experiences in future. Finally, although we might be adapted to attach no affective weight to *episodic memories* of past hedonic experience, it is hard to see how evolutionary mechanisms alone could make us indifferent to past experience in general: the *belief* that some experience (which I do not remember) happened to me in the past and the belief that some experience will happen to me in the future plausibly are supported by the same general-purpose capacity for semantic representation, in which case it is difficult to see how we could be hard-wired to experience strong affective responses to the former sort of belief but no response at all to the latter.

The cognitive mediation hypothesis predicts that in conditions under which people can affect the past, the past indifference hypothesis will be false. This means that it predicts that in those conditions people will either be time-neutral, or they will discount the past (as per the past discounting hypothesis) but they will not be indifferent to the past. That is because our unconcern with the past, according to the cognitive mediation hypothesis, results from the belief that the past is causally inaccessible. In the absence of that belief, we should show concern for the past.

If the only source of future bias is cognitively mediated beliefs about the practical irrelevance of the past, and if people are able to fully take on board or *internalise* that the past is relevant when it is stipulated to be so, we should predict that when people are asked to assume that their choices can affect the past, they will be time-neutral, giving the same weight to past and future experiences. Call this the *strong* version of the cognitive mediation hypothesis.

If, however, there are other sources of future bias, then we should predict that although we will show some concern for the past when it is made causally accessible, we will still not show as much concern for the past as we show for the future. In other words, we should predict the past discounting hypothesis. We should also predict the past discounting hypothesis if we suspect that people will only partially internalise the stipulation that they can affect the past. After a lifetime of engaging with situations where we believe *with certainty* that the past is causally inaccessible, it would not be surprising if, even when it is stipulated that the past is causally accessible, we do not fully internalise this stipulation, and continue to discount the past to some extent. Thus, what we call the *weak* cognitive mediation hypothesis predicts only the past discounting hypothesis, and not time-neutrality.

By contrast, the temporal metaphysics hypothesis suggests an absolute, all-or-nothing distinction between past and future. Metaphysical properties like *being past* or *receding from the present* do not come in degrees.[[10]](#footnote-11) If one thinks, for instance, that it is rational to attach evaluative weight to future events because they are *approaching the present* and therefore an appropriate object of some attitude like *anticipation*, this feature will be entirely absent from past events, to which we therefore have no reason to attach evaluative weight. (Parfit (1984, pp. 181-2) makes a version of this argument.) It therefore seems reasonable to treat evidence in favour of the past discounting hypothesis over the past indifference hypothesis as evidence against the temporal metaphysics hypothesis.[[11]](#footnote-12)

We had no experimental evidence upon which to choose between the temporal metaphysics and the practical irrelevance hypotheses. Nevertheless, since no account of the former appeals to evolutionary and psychological mechanisms, and so no real mechanism by which beliefs about temporal metaphysics could lead to future bias has been articulated, we predicted that evidence would support the practical irrelevance hypothesis. Hence we predicted that future bias would be mitigated in situations where past and future outcomes are unequal.

We also aim to determine which version of the practical irrelevance hypothesis is best supported. To do so, we distinguish a number of dimensions along which the two versions of that hypothesis can be expected to yield different predictions.

First, the narrow adaptation hypothesis predicts that future bias cannot be mitigated in cases where the past is causally accessible. Hence the narrow adaptation hypothesis predicts that there will be no significant difference between choice and preference situations (situations in which one can and cannot affect the past, respectively), since the same inflexible mechanism is at work in both cases. By contrast, the cognitive mediation hypothesis predicts that future bias can be substantially mitigated in choice situations, since it sees future bias as deriving from a general sensitivity of our affective states and evaluative attitudes to the causal accessibility and manipulability of particular events/features of the world. Hence, the cognitive mediation hypothesis predicts that people are less likely to exhibit future bias when making choices about causally accessible past events than when reporting preferences about causally inaccessible past events.

Second, the narrow adaptation hypothesis seems to predict a divergence between prospective and normative judgements, while the cognitive mediation hypothesis predicts no such divergence. *Prospective* judgements are judgements about what we *would* choose or prefer: for instance, that we *would* choose less suffering in the future over more suffering in the past. *Normative* judgements are judgements about what it would be better to choose or prefer: for instance, that it would be better to choose less suffering in the future over more suffering in the past.

The narrow adaptation hypothesis holds that future bias is relatively functionally specific and inflexible. If this is the correct explanation of our future-biased preferences and choices, there is no particular reason why we should endorse them as normatively correct, rather than treating them as a simple matter of preference.[[12]](#footnote-13) The cognitive mediation hypothesis, on the other hand, suggests that prospective and normative judgements should go hand in hand: we discount the past *for a reason* (because it is practically irrelevant), and that reason plays a cognitive role in shaping our attitudes. This gives us a natural basis from which to intuit that our attitudes are normatively correct, and not just arbitrary matters of taste.

In sum, then, we think that:

(a) the practical irrelevance hypothesis predicts (as per the past discounting hypothesis) that people will exhibit significantly less future bias when past disutilities are greater than future disutilities (i.e. in unequal outcome conditions), than when past and future disutilities are equal (i.e. in equal outcome conditions), while the temporal metaphysics hypothesis predicts that there will be no such difference (as per the past indifference hypothesis).

(b) the cognitive mediation hypothesis predicts that people will be significantly less future-biased in choice conditions than in preference conditions, while the temporal metaphysics hypothesis and the narrow adaptation hypothesis both predict that there will be no difference between these conditions.

(c) the narrow adaptation hypothesis predicts that people will more strongly exhibit future bias when asked for a prospective judgement (about what they *would* choose or prefer) than when asked for a normative judgement (about what they *should* choose or prefer), while the cognitive mediation hypothesis predicts that there will be no significant difference between people’s degree of future bias in normative vs. prospective judgements.

As noted, we predicted that our findings would support the practical irrelevance hypothesis over the temporal metaphysics hypothesis. We had, however, little basis for choosing between competing versions of the practical irrelevance hypothesis. On the basis of introspection on our own judgements, we ultimately decided that the cognitive mediation hypothesis was more plausible, and so we predicted findings consistent with that hypothesis: that people will be significantly less future-biased in choice conditions than in preference conditions, and that there will be no significant differences between normative and prospective judgements.

**3. Experimental Design and Results**

**3.1 Method**

*3.1.1 Participants*

1462 people participated in the study. Participants were U.S. residents, recruited and tested online using Amazon Mechanical Turk (MTurk), and compensated $0.50 for approximately 5 minutes of their time. Given recent worries about the quality of data collected through MTurk, concerning both the quality of human responders and the presence of bots, we adopted a number of quality control measures.[[13]](#footnote-14)

First, we used only those MTurk participants who have a task approval rate of at least 95% and who have had their tasks approved at least 1000 times. That means that all our participants had already successfully completed at least 1000 other studies, and received at least a 95% approval rating on these tasks, a standard that can be expected to eliminate most bots.

Second, our study included both task instructions and attentional checks that doubled as comprehension checks. We excluded participants who failed either to follow instructions or to correctly answer an attentional check/comprehension question. In total, 719 participants were excluded for either failing to answer the questions (142) or failing the attentional check/comprehension question (577). The remaining sample was composed of 743 participants (aged 20-76; 237 female, 3 prefer not to answer). Mean age 34.13 (SD = 9.99). Ethics approval was obtained from the University of Sydney 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*

We developed a single base vignette that could be minimally modified for each of the conditions.[[14]](#footnote-15) The following is the base vignette seen by participants in the choice conditions. The differences between the equal and unequal choice conditions are indicated in square brackets.

**Vignette: Choice Conditions**

You have been abducted by an advanced alien species, the Predictors. The Predictors have decided, out of scientific curiosity, to subject you to a lengthy series of painful electric shocks. You try to count the shocks, but you quickly lose track. This continues for several days.

Finally, the Predictors are ready to release you. Feeling regretful for the suffering they have caused, they will repair any physical harms of the experiment, and give you a pill that will reverse any traumatic psychological effects (though without erasing any of your memories). Thus, you will suffer no further physical or psychological harm from the experiment after you are released.

But before releasing you, the Predictors give you a choice: You may choose to voluntarily undergo one final electric shock before you are released. The Predictors inform you that they had already predicted your response to this offer before the experiment began. If they predicted that you WOULD choose to undergo the final shock, then they shocked you exactly 10,000 times since your kidnapping. If they predicted that you WOULD NOT choose to undergo the final shock, they shocked you [10,100 / 10,001] times. As their name suggests, the Predictors are extremely good at “predicting” human behavior: In fact, they have developed the technology for time travel, and upon observing your choice, one of the Predictors will travel a week into the past and report your choice so that the Predictors can shock you the appropriate number of times. [Thus, if you choose to undergo the final shock, you will leave the Predictors’ spaceship having been shocked 10,001 times in total, while if you do not, you will leave having been shocked 10,100 times. So if you choose to undergo the further shock you will receive a total of 10,001 shocks, and you choose not to undergo the further shock you will receive a total of 10, 100 shocks: an extra 99 shocks. / Thus, whatever choice you make, you will leave the Predictors’ spaceship having been shocked 10,001 times in total. Either you will have been shocked 10,000 times in the past, and you will be shocked one more time, in the future, just before they release you, or you will have been shocked 10,001 times in the past, and you will receive no further shocks in the future. So in either case, you will receive 10,001 shocks in total.]

In the equal choice conditions participants read the version in which, whatever choice they make about the final (future) shock, they will leave the Predictors’ spaceship having been shocked 10,001 times in total. The choice, then, is whether the 10,001st shock is in the future or the past. In the unequal choice conditions, participants read the version in which, if they choose to undergo the final (future) shock, they will leave the Predictors’ spaceship having been shocked 10,001 times in total, while if they do not, they will leave having been shocked 10,100 times, all before the time of their choice.

After having read either the equal or unequal choice vignette, participants in the normative choice conditions were presented with a 7-point Likert scale running from ‘The better choice is to accept the voluntary shock’ via ‘Accepting and refusing are equally good choices’ to ‘The better choice is to refuse the voluntary shock’. The orientation of the scale was randomised in all conditions. Participants in the prospective choice conditions were presented with a 7-point Likert scale running from ‘I would definitely choose to accept the voluntary shock’ via ‘I would be equally likely to accept or refuse’ to ‘I would definitely choose to refuse the voluntary shock’.

Participants in the preference conditions saw a version of the following vignette:

**Vignette: Preference Conditions**

You have been abducted by an advanced alien species, the Predictors. The Predictors have decided, out of scientific curiosity, to subject you to a lengthy series of painful electric shocks. You try to count the shocks, but you quickly lose track. This continues for several days.

Finally, the Predictors are almost ready to release you. Feeling regretful for the suffering they have caused, they will repair any physical harms of the experiment, and give you a pill that will reverse any traumatic psychological effects (though without erasing any of your memories). Thus, you will suffer no further physical or psychological harm from the experiment after you are released.

As they prepare to release you, the Predictors mention that before your abduction, you were randomly assigned to one of two schedules of shocks. If you are on Schedule 1, then you were shocked 10,000 times in the last several days, and before your release you will be given one final shock. If you are on Schedule 2, then you were shocked [10,100 / 10,001] times in the last several days, but you will not receive any additional shocks before your release. [Thus, if you are on Schedule 1 you will leave the Predictors’ spaceship having been shocked 10,001 times, while if you are on Schedule 2, you will leave having been shocked 10,100 times. That’s because if you are on Schedule 1, you will have been shocked 10,000 times in the past, and you will be given another shock in the future before you leave. If you are on Schedule 2, you will have received 10,100 shocks in the past, but you will not receive any shocks in the future. / Thus, whichever schedule you are on, you will leave the Predictors’ spaceship having been shocked 10,001 times. Either you are on Schedule 1, in which case you were shocked 10,000 times in the past, and will be given one more shock in the future, before your release, or you are on Schedule 2, and you were shocked 10,001 times in the past, and will not be shocked again in the future.]

Participants in the equal preference conditions saw the version of the vignette in which, if they are on Schedule 1, they have been shocked 10,000 times and will receive one further (future) shock, and if they are on Schedule 2, they have been shocked 10,001 times and will receive no further (future) shocks. Thus, whichever Schedule they are on they receive 10,001 shocks overall. Participants in the unequal preference conditions likewise saw a vignette in which, if they are on Schedule 1, they have been shocked 10,000 times and will receive one further (future) shock. However, if they are on Schedule 2, they have been shocked 10,100 times and will receive no further (future) shocks. Thus, if they are on Schedule 1 they will leave the spaceship having been shocked 10,001 times—but they will receive a future shock—and if they are on Schedule 2 they will leave the spaceship having been shocked 10,100 times, all of which have already occurred before the time at which they report their preference.

After having read either the equal or unequal preference vignette, participants in the normative preference conditions were presented with a 7-point Likert scale running from ‘It is definitely better to be on Schedule 1, and not be on Schedule 2’ via ‘Neither schedule is better than the other’ to ‘It is definitely better to be on Schedule 2, and not be on Schedule 1’. Participants in the prospective preference conditions were presented with a 7-point Likert scale running from ‘I would definitely hope to be on Schedule 1, and not be on Schedule 2’ via ‘I would not have a preference to be on either schedule’ to ‘I would definitely hope to be on Schedule 2, and not be on Schedule 1’.

Participants were randomly assigned to one of eight conditions. These eight conditions reflected all possible combinations of evaluation (prospective or normative), attitude (preference or choice) and outcome (equal or unequal).

After recording their responses, participants proceeded to a new screen, where they answered one of two attention checks/comprehension questions. Participants in the choice conditions were asked ‘*In this vignette you were asked to imagine that…*’ and had to select one of the following options:

If you DON’T accept an extra shock you will receive 10,001 shocks in total

If you DON’T accept an extra shock you will receive 10,100 shocks in total

If you DON’T accept an extra shock you will receive NO shocks in total

If you DON’T accept an extra shock you will receive 10,000 extra shocks

Participants in the preference conditions were asked ‘*In this vignette, on which Schedule will you receive the most shocks overall…*’ and had to select one of the following options:

Schedule 1

Schedule 2

Schedule 3

None of the above - You will receive the same number of shocks on each Schedule.

Participants who answered the question incorrectly were excluded.

**3.2. Results**

Before describing our results, we will start by summarising how our results address each of our hypotheses.

First, we predicted that if the practical irrelevance hypothesis is correct, then we should see a significant reduction in future bias in unequal outcome conditions relative to equal outcome conditions (as per the past discounting hypothesis). Conversely, we predicted that if the temporal metaphysics hypothesis is correct, then we should see no difference in future bias between unequal outcome conditions and equal outcome conditions (as per the past indifference hypothesis). Our results vindicate the practical irrelevance hypothesis: we observe significantly reduced future bias in unequal outcome conditions relative to equal outcome conditions.

Second, we predicted that if the cognitive mediation hypothesis is correct, then we should see significantly less future bias in choice conditions relative to preference conditions. Conversely, we predicted that if the temporal metaphysics hypothesis or the narrow adaptation hypothesis is correct, then we should see no significant difference between these conditions. Our results vindicated the cognitive mediation hypothesis: we observe significantly reduced future bias in choice conditions relative to preference conditions.

Third, we predicted that if the narrow adaptation hypothesis is correct, then we should see a significant difference between normative and prospective conditions. Conversely, if the cognitive mediation hypothesis is correct, then we should see no significant difference between normative and prospective conditions. Here, our results were mixed. We observed a significant interaction effect between normative/prospective and equal/unequal, with results in the unequal outcome conditions supporting the narrow adaptation hypothesis and results in the equal outcome conditions supporting the cognitive mediation hypothesis.

Table 1 below summarises the descriptive results from our experiment. The ‘FB’ column represents the proportion of participants who respond in a future-biased manner. That is, they would choose *not* to receive a future shock, or would prefer to be on the schedule *without* the future shock. These responses were coded as 5, 6 or 7. The ‘non-FB’ column represents the proportion of participants who respond in a non-future-biased manner. That is, they would choose to receive a future shock, or would prefer to be on the schedule with the future shock. These responses were coded as 1, 2 or 3. The ‘4’ column represents the proportion of participants who chose the middle value on the Likert scale, indicating that they would be equally likely to choose either option, or would have no preference between the two schedules, or that the two choices/schedules are equally good. Table 1 also presents the results of the t-tests and one-way 𝜒2-tests for each condition.

*Table 1. Descriptive data from all conditions.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Condition**  | **%FB** | **%non-FB** | **%4** | **Mean** | **SD** | **t-test** | ***p*-value** | ***X2*** | ***p*-value** |
| **Condition 1:**Normative/Choice/Equal | 64.9 | 24.5 | 10.6 | 4.78 | 1.93 | 3.902 | <.001 | 8.340 | .004 |
| **Condition 2:**Normative/Choice/Unequal | 40.2 | 53.3 | 6.5 | 3.49 | 2.28 | -2.154 | .034 | .391 | .532 |
| **Condition 3:**Normative/Preference/Equal | 83.7 | 4.7 | 11.6 | 5.66 | 1.34 | 11.485 | <.001 | 39.116 | <.001 |
| **Condition 4:**Normative/Preference/Unequal | 66.3 | 28.6 | 5.1 | 4.48 | 2.18 | 2.179 | .032 | 10.449 | .001 |
| **Condition 5:**Prospective/Choice/Equal | 60.2 | 35.9 | 3.9 | 4.42 | 2.12 | 1.997 | .048 | 4.282 | .039 |
| **Condition 6:**Prospective/Choice/Unequal | 60.0 | 36.7 | 3.3 | 4.31 | 2.13 | 1.387 | .169 | 2.844 | .092 |
| **Condition 7:**Prospective/Preference/Equal | 76.2 | 8.8 | 15.0 | 5.43 | 1.61 | 7.902 | <.001 | 22.050 | <.001 |
| **Condition 8:**Prospective/Preference/Unequal | 69.0 | 26.0 | 5.0 | 4.73 | 1.98 | 3.689 | <.001 | 14.440 | <.001 |

Our t-test results show that the mean level of future-biased responding is significantly above 4 in all conditions except for Condition 2 (normative/choice/unequal) and Condition 6 (prospective/choice/unequal). Thus, in the majority of conditions we tested we find that, overall, participants respond in a future-biased manner. However, in Condition 2 we find that on average, people think it is better to choose the future shock if it means that they would have received fewer shocks overall. In Condition 6 we find a mean that does not differ significantly from 4. However, this does not indicate that the majority of participants are responding with 4. Indeed, only 3.3% of participants gave this response (indicating that they would be equally likely to accept or refuse the future shock). Likewise, just because a mean is significantly above, or below, 4, does not mean that the majority of participants are responding in a future-biased or non-future-biased manner. Thus, it is informative to consider the t-test results in conjunction with the results of our one-way 𝜒2-tests.

Our one-way 𝜒2-tests show that a significant majority of people respond with a 5, 6, or 7, in all conditions except for Condition 2 (normative/choice/unequal) and Condition 6 (prospective/choice/unequal). Thus, in the majority of conditions we tested we find that the majority of participants respond in a future-biased manner. However, in Condition 2 we find that people are evenly divided between thinking that it is better to choose the future shock if it means that they would have received fewer shocks overall, and thinking either that the two choices are equally good or that it is better to refuse the future shock. Conversely, in Condition 6 we find that people are 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 would choose to accept the future shock.

To compare people’s future bias across conditions we used a 2x2x2 between-subjects ANOVA. The results found a significant main effect of Choice/Preference *F*(1, 735) = 32.057, *p* < .001, and Equal/Unequal *F*(1, 735) = 31.457, *p* < .001. We also observed a significant two-way interaction between Normative/Prospective and Equal/Unequal *F*(1, 735) = 8.188, *p* = .004.

The main effect of Choice/Preference showed that people’s future bias was significantly stronger in preference conditions (M = 5.07, SD = 1.98) than in choice conditions (M = 4.25, SD = 1.98).

The main effect of Equal/Unequal showed that people’s future bias was significantly stronger in equal outcome conditions (M = 5.07, SD = 1.98) than in unequal outcome conditions (M = 4.25, SD = 2.02).

Simple effects tests using a Bonferroni correction were carried out on the two-way interaction between Normative/Prospective and Equal/Unequal. The effect is shown in Figure 1 below. Firstly, in normative conditions, people’s future bias was significantly higher for equal outcomes (M = 5.22, SD = 1.99) than for unequal outcomes (M = 3.98, SD = 1.98; *p* < .001). Conversely, in prospective conditions, there was no significant difference in future bias between equal outcomes (M = 4.92, SD = 2.00) and unequal outcomes (M = 4.52, SD = 1.98; *p* = .052).

Secondly, for equal outcomes, there was no significant difference in future bias between normative and prospective conditions (*p* = .154). However, for unequal outcomes, we found significantly higher future bias in prospective conditions relative to normative conditions (*p* = .009).

*Figure 1.* Two-way interaction effect between Normative/Prospective and Equal/Unequal. Error bars represent 95% confidence intervals.

**4. Discussion**

Our results mostly—but not entirely—confirm our predictions from §2 and support the cognitive mediation hypothesis over both the narrow adaptation hypothesis and the temporal metaphysics hypothesis. Before diving into the details of these results and their philosophical implications, it will be useful to highlight and concretely describe our central result, namely, the reduction of future bias in conditions where participants can choose between a smaller pain in the future and a larger pain in the past. These are the two choice/unequal conditions (conditions 2 & 6), the only conditions in which we did *not* see mean (and majority) future bias.

In these conditions, as a participant, you are asked to imagine that you have been given a choice whether to accept a final shock, and that your choice will retrocausally determine how many shocks you received in the past (which you can remember only imprecisely). Specifically, the Predictors will report your choice to themselves in the past, and use that report to determine how many shocks to give you. Accepting one future shock will thereby prevent 100 past shocks.

In these conditions, both the temporal metaphysics hypothesis and the narrow adaptation hypothesis seem to predict that participants should remain just as future-biased as they are in ordinary situations where they cannot affect the past. The past shocks are still *in the past* (‘receding from the present’, ‘over and done with’, etc), and presumably have whatever metaphysical properties the temporal metaphysics hypothesis might take to explain our indifference to past events.[[15]](#footnote-16) Likewise, the narrow adaptation hypothesis predicts that indifference to past events is simply a ‘hard-wired’ preference that does not depend on one’s beliefs about the details of particular situations. Thus neither of these hypotheses predicts that future bias will be mitigated in the choice/unequal conditions.

The cognitive mediation hypothesis, on the other hand, says that we are generally indifferent to the past because we care less about things that we believe we cannot affect, and we are usually all-but-certain that we cannot affect the past. On this hypothesis, imagining yourself in a situation where you *can* affect the past should make you less future-biased (perhaps to the point of complete time-neutrality). Given that the past stakes are substantially higher than the future stakes in these conditions (100 shocks vs. 1 shock), the attenuation of future bias proposed by the cognitive mediation hypothesis can explain why some participants think they would, or that it is better to, accept the future shock, even when in ordinary (non-retrocausal) situations, they would prefer 100 past shocks to one future shock. Thus, it seems that the cognitive mediation hypothesis is uniquely able to explain the fact that, on average, people think it is better to accept one future shock in order to prevent 100 past shocks, and more generally, the reduction of future bias in the choice/unequal conditions. Our results in these conditions show that, in line with the cognitive mediation hypothesis, the past *does* matter more when you can affect it.

More broadly, our results contain three key findings. First, as predicted, we confirmed the past discounting hypothesis and rejected the past indifference hypothesis: participants were less future-biased when the past stakes were larger than the future stakes. Participants more strongly agreed that they would choose/prefer to receive a future shock in conditions where this would result in receiving 10,001 shocks instead of 10,100 shocks in total, than in conditions where they would receive 10,001 shocks either way. This supports the practical irrelevance hypothesis over the temporal metaphysics hypothesis. Since both versions of the practical irrelevance hypothesis predicted this result, however, it does not support one version over the other.

Second, as predicted, we found a statistically significant difference in future bias between choice conditions and preference conditions. In particular, we found that participants more strongly exhibited future bias in preference conditions than in choice conditions. This finding supports the cognitive mediation hypothesis over the temporal metaphysics hypothesis and the narrow adaptation hypothesis.[[16]](#footnote-17)

Third, however, where we had predicted that there would be no statistically significant difference between degrees of future bias in normative compared to prospective conditions, we instead found a mixed result. In the equal outcome conditions (comparing one past shock to one future shock), we found no difference between participants’ prospective judgments and their normative judgments. But in the unequal outcome conditions (comparing 100 past shocks to one future shock), we did find a difference, with participants significantly more likely to report that they *would* exhibit future bias in their choices or preferences (i.e., that they would choose/prefer more pain in the past rather than less pain in the future) than to endorse those choices/preferences as normatively correct. This casts some doubt on the cognitive mediation hypothesis, which (at least naïvely) predicts that we are future-biased only insofar as we have evident reason to be (namely, the causal inaccessibility of the past), and that when the past is made causally accessible, our tendency toward future bias and our normative endorsement of future bias should diminish in equal measure.

On the whole, we think that our results support the weak version of the cognitive mediation hypothesis. First, we see clear evidence of cognitive mediation in the finding that participants are less future-biased when the past is made causally accessible (i.e., in choice conditions versus preference conditions). But, contra what the strong version of the cognitive mediation hypothesis would suggest, we do not find that participants in choice conditions uniformly become time-neutral, and therefore uniformly report that they would/ought to choose the future shock in the choice/unequal conditions, and that they would/ought to be indifferent in the choice/equal conditions.

As we noted, this could be because we are asking participants to imagine a very unusual situation in which they can retrocausally influence the past, and they are unable completely to internalise this stipulation and eliminate their deeply ingrained habit of future bias. If so, then we should conclude that the cognitive mediation hypothesis is fully vindicated. Alternatively, we might take these results to suggest that future bias is only *partially* cognitively mediated. On this interpretation, the fact that we find residual future bias even in choice conditions—that is, we find past discounting rather than time-neutrality—is taken to show that the cognitive mediation hypothesis is not a complete explanation, and that something else explains the residual future bias.

Second, there is the unanticipated mixed finding regarding our final hypothesis. As an explanation for this finding, we speculate that in the unequal outcome conditions (and especially in the choice/unequal conditions), there are a significant number of participants who can see the normative appeal of temporal neutrality, while continuing to feel an affective bias toward the future and hence to prospect that their choices/preferences would exhibit future bias. This affective resistance to reasoning that favours temporal neutrality is another way in which future bias is imperfectly cognitively mediated. We do not see a statistically significant normative/prospective gap in the equal outcome conditions, according to this explanation, because here there is no real normative argument for temporal neutrality for participants to feel pulled by. Since the past and future stakes are equally balanced, there is nothing wrong—even from a temporally neutral point of view—with favouring the future.

In summary, then, our first result supports the practical irrelevance hypothesis over the temporal metaphysics hypothesis, our second result supports the cognitive mediation hypothesis over both the temporal metaphysics hypothesis and the narrow adaptation hypothesis, while our third result is equivocal, partly supporting the cognitive mediation hypothesis and partly supporting the narrow adaptation hypothesis. On the whole, however, we take our results to provide cautious support for the cognitive mediation hypothesis, while also leaving some things unexplained (namely, residual future bias in choice conditions and the normative/prospective gap in unequal outcome conditions). These remaining explananda suggest the presence of other, complementary sources of future bias, perhaps including those posited by the temporal metaphysics or narrow adaptation hypotheses (a point which we elaborate below).

Our results have at least three notable implications for the broader literature on future bias. First, by supporting the practical irrelevance hypothesis, our results weaken the potential link between future bias and beliefs about temporal metaphysics. As we noted earlier, if we devalue the past because we believe that past events are receding from us while future events are approaching, then we should equally devalue past events that we can affect.[[17]](#footnote-18)

Significantly, although we found that future bias diminishes in conditions where the past is causally accessible—something we would not expect to find if the temporal metaphysics hypothesis were true—we also found that participants still exhibited a substantial degree of future bias even in the conditions where they could affect the past. In all, while this strongly suggests that the temporal metaphysics hypothesis is not the whole story, it also leaves room to appeal to something beyond cognitively mediated practical irrelevance to explain future bias. It could be, then, that the temporal metaphysics hypothesis is part of this explanatory story. Further empirical work would be needed to test whether this is so.

Indeed, there are two avenues of exploration that suggest themselves given these findings. First, we have treated the temporal metaphysics hypothesis as a single hypothesis. But in fact, we can disentangle at least two different versions of that hypothesis. First, there is the version we have focused on in this paper, according to which people’s beliefs about the structure of the temporal dimension (namely, that it is dynamical) play a role in explaining future bias. Call this the *temporal metaphysics beliefs hypothesis*. Alternatively, one might think that it is not people’s *beliefs* about temporal metaphysics, but rather their *experience* or *phenomenology* of time (e.g., a phenomenology as of time passing, or as of the future drawing closer and the past receding), which play that explanatory role. Call this the *temporal phenomenology hypothesis*. Follow-up studies could aim to assess and adjudicate between these two hypotheses, by determining whether the strength of future bias correlates with, on the one hand, people’s tacit beliefs about temporal metaphysics, and the other hand, their temporal phenomenology. If there are no such correlations, then this would suggest that residual future bias is not the result of either tacit metaphysical beliefs or temporal phenomenology, but of something else entirely.

Second, while our results support the practical irrelevance hypothesis, they count against the narrow adaptation hypothesis, and hence against the sort of direct evolutionary explanation for future bias that has been popular in the recent literature.[[18]](#footnote-19) Once again, however, our results leave a degree of unexplained future bias, as well as the unexplained normative/prospective gap in our unequal outcome conditions, which might be accounted for by a more functionally specific, developmentally inflexible mechanism that discounts the past. Still, the fact that making the past causally accessible mitigates our future bias indicates that, at best, the simple evolutionary explanation is not the whole story, and, at worst, may not be any of the story.

Third, our results are relevant to the philosophical debate over the *rationality* of future bias.[[19]](#footnote-20) A general policy of attaching less affective weight to events and experiences that one cannot affect, and that cannot count comparatively for or against one’s practical options, seems more rationally justifiable than a policy of weighting events purely on the basis of their temporal locations, regardless of their causal accessibility (at least insofar as affective states are subject to rational appraisal). Furthermore, the fact that we care about past events more when we have some practical *reason* to care about them (i.e., some reason to give them weight in practical deliberation) indicates that our bias toward the future is rational in the sense of being ‘reasons-responsive’. This contrasts both with explanations based on beliefs about temporal metaphysics (which, even if empirically correct, do not provide any obvious rational justification for future bias—see footnote 5) and with the view that future bias is simply an ‘arational’ preference that is outside the scope of rational assessment (as, for instance, one might think that gustatory preferences are).

Of course, none of this shows that future bias is in fact rational, especially given that there is some amount of future bias that may not be accounted for by our beliefs about the causal accessibility of the past. Nevertheless, our results are suggestive. And they show where more research is needed. In addition to studies that test whether the temporal metaphysics hypothesis plays any role at all in explaining future bias, it might be useful to run a ‘time-reversed’ version of the present study, focusing on cases in which agents cannot affect their future experiences.[[20]](#footnote-21) If the cognitive mediation hypothesis is correct, we would expect future bias to be mitigated in situations where we *cannot* affect the *future*, just as it is in situations where we *can* affect the *past*. Finally, it would be profitable to run studies in which the ratios of past and future stakes are allowed to vary more than they do in our study (which only considers 1:1 and 100:1 ratios), or in which participants can freely specify how much past gain would be required to compensate for a given future loss, in order to more precisely measure the degree to which we discount the past (and, perhaps, quantify the reduction of this discount factor in cases where we can affect the past, or cannot affect the future).

Such studies could both help to confirm (or deny) the evidence we found for the cognitive mediation hypothesis, and explain some of the findings that we have left unexplained, like the residual future bias we exhibit even under the supposition that the past is causally accessible, and the partial gap between our prospective and normative judgments about past/future trade-offs. We will then be better positioned both to describe our overall pattern of attitudes toward past and future experiences and to decide whether that pattern is rational, irrational, or arational.[[21]](#footnote-22)

**References**

Ahler, D., Roush, C., & Sood, G. The micro-task market for lemons: data quality on Amazon’s Mechanical Turk. Unpublished manuscript, 22 January 2020. URL: <http://gsood.com/research/papers/turk.pdf>

Ainslie, G., & Haslam, N. (1992). Self-control. *Choice over time*, *177*, 209.

Akerlof, G. A. (1991). Procrastination and obedience. *The American Economic Review*, *81*(2), 1-19.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, *4*(1): 29-34.

Bickel, W. K., Yi, R., Kowell, B. P., and Gatchalian, K. M (2008). ‘Cigarette smokers discount past and future rewards symmetrically and more than controls: Is discounting a measure of impulsivity’. *Drug and Alcohol Dependence* 96: 256–262.

Breslin, P. A. S. (2013). An Evolutionary Perspective on Food and Human Taste. *Current Biology*, 23(9): 409-418.

Brink, D. O. (2011). ‘Prospects for Temporal Neutrality’. in Craig Callender (ed.) T*he Oxford Handbook of Philosophy of Time.* Oxford: Oxford University Press.

Callender, C. (2017). *What Makes Time Special?* Oxford 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.

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

Dorsey, D. (2018). ‘Prudence and Past Selves.’ *Philosophical Studies* 175(8):1901-1925.

Dougherty, T. (2011). On whether to prefer pain to pass. *Ethics* 121: 521–37.

Maclaurin, J. & Dyke, H. (2002). ‘‘Thank goodness that’s over’: the evolutionary story.’ *Ratio* 15:276–92.

Frederick, S., Loewenstein, G., & O’Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of economic literature*, *40*(2), 351-401.

Garrett, B. (2018). ‘Experience and Time.’ *Acta Analytica* 33 (4):427-430.

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

Greene, P., Latham, A. J., Miller, K and Norton, J. (2020). ‘Hedonic and non-hedonic bias towards the future’. *Australasian Journal of Philosophy* <https://doi.org/10.1080/00048402.2019.1703017>

Guo, T., Ji, L. J., Spina, R., & Zhang, Z. (2012). Culture, temporal focus, and values of the past and the future. *Personality and Social Psychology Bulletin*, *38*(8), 1030-1040.

Grossi, G. (2017). Hardwiring: Innateness in the age of the brain. *Biology & Philosophy*, 32(6), 1047-1082.

Hare, C. (2007). ‘Self-bias, time-bias, and the metaphysics of self and time.’ *Journal of Philosophy* 104: 350–73.

Hare, C. (2013). ‘Time – the emotional asymmetry.’ In *A Companion to the Philosophy of Time*, ed. A. Bardon and H. Dyke, 507–20. Hoboken, NJ: Wiley-Blackwell.

Hausman, J. A. (1979). Individual discount rates and the purchase and utilization of energy-using durables. *The Bell Journal of Economics*, 33-54.

Holcombe, A., Latham, A. J., Miller, K., & Norton, J. (ms). ‘The Rationality of Near Bias towards both Future and Past Events’.

Horwich, P. (1987). *Asymmetries in Time: Problems in the Philosophy of Science*. Cambridge, MA: MIT Press.

Hume, D. 2000 (1739). *A Treatise of Human Nature*. Oxford: Oxford University Press.

Kauppinen, A. (2018). ‘Agency, Experience, and Future Bias.’ *Thought: A Journal of Philosophy* 7(4):237-245.

Krebs, J. (2009). The gourmet ape: evolution and human food preferences. *The American Journal of Clinical Nutrition*, 90(03): 707-711.

Latham, A. J., Miller, K., and Norton, J. (2019). ‘Is our naïve theory of time dynamical?’ *Synthese*. DOI: 10.1007/s11229-019-02340-4[2]

Latham, A. J., Miller, K., and Norton, J.(forthcoming). ‘An Empirical Investigation of Purported Passage Phenomenology’ *Journal of Philosophy*. <https://philpapers.org/rec/LATAEI>

Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., & Waldman, I. D. (2015). ‘Fifty psychological and psychiatric terms to avoid: A list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases.’ *Frontiers in Psychology, 6*:

MacBeath, M. (1983). ‘Mellor’s Emeritus headache.’ *Ratio*, 25:81–88.

Mellor, D. H. (1981). ‘Thank goodness that’s over.’ *Ratio*, 23:20–30.

Mellor, D. H. (1983). ‘MacBeath’s soluble aspirin.’ *Ratio*, 25:89–92.

Molouki, S., Hardisty, D. J., & Caruso, E. M. (2019). The sign effect in past and future discounting. *Psychological science*, *30*(12), 1674-1695.

Norenzayan, A., Hansen, I. G., & Cady, J. (2008). An angry volcano? Reminders of death and anthropomorphizing nature. *Social Cognition*, 26(2): 190-197.

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

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.

Schlesinger, G. (1976). ‘The stillness of time and philosophical equanimity.’ *Philosophical Studies* 30:145–59.

Suhler, C. and C. Callender. (2012). ‘Thank goodness that argument is over: explaining the temporal value asymmetry.’ *Philosophers’ Imprint* 12:1–16.

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

Takeuchi, K. (2011). Non-parametric test of time consistency: Present bias and future bias. *Games and Economic Behavior*, *71*(2), 456-478.

Tarsney, C. (2017). ‘Thank goodness that’s Newcomb.’ *Analysis* 77(4):750-759.

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economics letters*, *8*(3), 201-207.

Yehezkel, G. (2014). ‘Theories of Time and the Asymmetry in Human Attitudes.’ *Ratio* 27(1):68-83.

Yi, R., Gatchalian, K. M., & Bickel, W. K. (2006). Discounting of past outcomes. *Experimental and clinical psychopharmacology*, *14*(3), 311.

1. We should note that there is a very large literature on other kinds of time biases. For example, psychologists and behavioural economists have had much to say about *future* *near bias* (the preference to have positive events in one’s near future rather than one’s far future, and negative events in one’s far future rather than one’s near future). Rather confusingly, in psychology the term ‘future bias’ has sometimes been used to refer to future near bias (See Takeuchi, 2011). But this is not what we mean by this term.

The literature on future near bias has focussed on why, and the rate at which, people ‘temporally discount’ (i.e. show future near bias) (Frederick, Loewenstein & O’Donoghue, 2002, p. 377; Holcombe, Latham, Miller & Norton, ms). While early experiments suggest what Ainslie & Haslam (1992, p. 59) call a ‘pervasive devaluation of the future’ (see also Thaler, 1981, Hausman, 1979, and Akerlof, 1991), a more recent meta-analysis (Frederick et al., 2002) describes ‘tremendous variability’ in empirical findings of the rate at which people discount the future.

With few exceptions, this literature has ignored *past* near bias (the preference for positive events in one’s near past rather than one’s far past, and negative events in one’s far past rather than one’s near past). Rare exceptions include Yi, Gatchalian & Bickel (2006), Bickel, Yi, Kowell & Gatchalian (2008), Molouki, Hardisty, & Caruso (2019) and Holcombe et al. (ms). Holcombe et al. report the surprising finding that some people are what they call ‘far-biased’: they prefer positive events in the far past over the near past, and negative events in the near past over the far past.

In addition, there are large literatures that attempt to describe various patterns in our inter-temporal preferences that exhibit themselves in, for instance, voting behaviour, saving behaviours, and other ways in which we plan for the future. Notably, however, none of this research speaks directly to the phenomenon of future bias, since it typically focuses exclusively on our preferences regarding future events, and sets aside our preferences regarding past events. [↑](#footnote-ref-2)
2. Though we should be careful not to assume the universality of results we report here, given the cross-cultural differences found by Guo, Ji, Spina & Zhang (2012). [↑](#footnote-ref-3)
3. It is worth emphasising that the temporal metaphysics hypothesis only claims that our (tacit) *belief* in a dynamical theory of temporal metaphysics explains our future-biased preferences. The success or failure of this hypothesis has no direct implications regarding the *true* metaphysics of time. Perhaps the world is non-dynamical, but we falsely believe it to be dynamical, and this false belief explains our future bias. Or perhaps the world is dynamical, and we believe it to be dynamical, but this belief is not the source of our future bias. [↑](#footnote-ref-4)
4. Recent empirical work by Latham, Miller & Norton (forthcoming) suggests that, surprisingly, non-philosophers only weakly agree that things seem this way. [↑](#footnote-ref-5)
5. For instance, Yehezkel writes that ‘the failure to offer any substantial justification for the asymmetry in our attitudes based on the flow of time stems from the inability to offer any non-trivial account of the flow of time. It is difficult to see what difference is made by the claim that ‘future events are moving closer to reality,’ given that all that is meant by this claim is that ‘in the future, future events will be closer to the present.’ This is a mere truism, as evident by the analogous claim, regarding the past, according to which ‘in the past, past events were closer to the present.’ The attempt to justify the asymmetry between past and future based on the flow of time *per se* thus seems to collapse into triviality.’ (Yehezkel 2013, pp. 6-7). Similar sceptical remarks are made by Parfit (1984, p. 178) and Hare (2013, pp. 510-11), among others. [↑](#footnote-ref-6)
6. Though this passage is often quoted to associate Hume with the practical irrelevance explanation for future bias, this is probably a mistake: Hume is here talking specifically about effects *on the will*, and the next sentence reads: ‘But *with respect to the passions*, the question [of what explains ‘the superior effects of the same distance in futurity above that in the past’] is yet entire, and well worth the examining’ (Hume 1739, 2.3.7.6; emphasis added). In trying to account for the past-future asymmetry with respect to the passions, Hume entertains a number of hypotheses, including a version of the temporal metaphysics hypothesis (2.3.7.9), but does not seem to take the practical irrelevance of the past as an explanation for its weaker effect on the passions. [↑](#footnote-ref-7)
7. Kauppinen does not claim that our past-directed preferences are *always* practically inert in the relevant sense. But he holds that when a future-biased preference would influence the agent’s choices, or would contradict an earlier preference on which she has already based a choice, future bias is rationally impermissible, *and moreover* is no longer psychologically typical. [↑](#footnote-ref-8)
8. There is a great deal of controversy regarding whether there is a useful or interesting concept of being hard-wired or innate (see Grossi (2017) for an overview), given that with the exception of inborn reflexes, very few psychological capacities are entirely inflexible (see Lilienfeld, Sauvigné, Lynn, Cautin, Latzman & Waldman 2015). For our purposes, all that matters is that there is a category of adaptations that are relatively more inflexible and specific than others, which are significantly more flexible. [↑](#footnote-ref-9)
9. We will talk of both affective and evaluative discounting in what follows. We suppose that both our affective states and evaluative judgments exhibit a bias toward the future and that our affective future bias plays some role in explaining our evaluative future bias. Which of these we focus on in the context of a hypothesis depends on which role we take our affective states to be playing. For instance, if the narrow adaptation hypothesis is true, then the hard-wired mechanism responsible for future bias probably works via our affective states, so we focus on these. On the other hand, if we’re asking about the rationality of future bias, our evaluative judgments are more obviously subject to rational appraisal, and so we focus on these. [↑](#footnote-ref-10)
10. Of course, there are A-theoretic properties that do come in degrees, like *distance from the present*. But most philosophers do not think that these gradable properties have normative significance (as evidenced by the fact that few philosophers endorse any form of pure time preference among future events). [↑](#footnote-ref-11)
11. As on the question of cognitive mediation, the philosophical literature seems to be divided on whether we merely discount past pleasures or pains, or are entirely indifferent to them. Hume (1739) treats the past-future asymmetry as a matter of degree (our representations of future events are more vivid than our representations of future events), and so pretty clearly thinks that the past is merely affectively discounted. This ‘discounting’ view is also suggested by Suhler & Callender (2012, p. 5) and endorsed by Yehezkel (2014, p. 11). On the other hand, Parfit (1984) writes that: ‘I do in fact regard my past suffering with complete indifference. I believe that, in this respect, most other people are like me’ (Parfit 1984, p. 173) (Parfit notes, however, that he knows some people who ‘find knowledge of their past pains mildly distressing.’) Dorsey (2018) takes a similar line, suggesting that the level of past pain that could outweigh a given unit of future pain is ‘indeterminately high’ and that it is appropriate to prefer the past pain ‘for virtually any imaginable level of painfulness’ (Dorsey 2018, p. 1915). And Sullivan (2018, p. 58) claims that ‘we assign no value to a merely past painful experience or pleasurable experience.’

 Among proponents of the temporal metaphysics hypothesis, Schlesinger (1976) seems to endorse complete indifference to past hedonic experience, arguing that we regard past experiences with an ‘equanimity’ which, if we accept the B-theory of time, should extend to future events as well. Other proponents of the temporal metaphysics hypothesis, like Prior (1959), Craig (1999), and Pearson (2018), do not address the question directly. To our knowledge, no one in the literature has endorsed both the temporal metaphysics hypothesis and the view that we merely discount past hedonic experience. [↑](#footnote-ref-12)
12. There is a range of cases that illustrate this idea. For instance, we know that evolutionary pressures in our past, when food was much more scarce, have given us all relatively inflexible preferences for food that is high in fat and carbohydrates. Of course, this is not entirely inflexible, but it is relatively inflexible. Clearly, however, we don’t assume that these preferences are normatively the right ones: in our current environment, they often lead to obesity and early death (e.g., Breslin 2013; Krebs 2009). Likewise, we know that we have a relatively inflexible disposition to see agency in the world. Since detecting agency was very important in our evolutionary history, we evolved mechanisms that are very sensitive to cues that there is an agent present. Those mechanisms are in fact over-sensitive, and so signal the presence of agents where there are none, including signalling the presence of agents in natural phenomena such as volcanoes, crop variation, and so on. But we can have this disposition to see agency in natural phenomena without endorsing it as normatively correct (e.g., Barrett 2000; Norenzayan, Hansen, & Cady, 2008). [↑](#footnote-ref-13)
13. See Ahler, Roush & Soud (ms) for a discussion of some of the problems associated with collecting data using MTurk and the prevalence thereof. [↑](#footnote-ref-14)
14. These vignettes are based on a thought experiment described in Tarsney (2017). [↑](#footnote-ref-15)
15. The implications of the temporal metaphysics hypothesis are somewhat complicated by the possibility of backward time travel. When an agent supposes that she herself will travel into the past, she may well believe that some experiences in the ‘physical past’ are in her ‘experiential future’, and therefore take the same attitudes toward them that she would toward future experiences. But, although our vignettes involve backward time travel, they don’t involve the agent (the character whose perspective subjects are asked to adopt) traveling into the past, and so we should still expect the temporal metaphysics hypothesis to straightforwardly predict indifference to all past experiences. [↑](#footnote-ref-16)
16. It also contradicts Dorsey’s claim that causal access to the past ‘makes little difference’ to our time-asymmetric attitudes (Dorsey 2018, p. 1910). On the other hand, the finding of residual future bias in choice conditions contradicts Kauppinen’s overly bold claim that ‘our intuitions are temporally neutral when it comes to active choices regarding pleasure and pain’ (Kauppinen 2018, p. 243). [↑](#footnote-ref-17)
17. It is important to be clear that these results speak only to the role that (implicit or explicit) *beliefs* about temporal metaphysics play in our asymmetric attitudes toward past and future, and do not tell us anything about the *true* metaphysics of time. In particular, even if we were to fully reject the temporal metaphysics hypothesis, this would not justify the conclusion that dynamical theories of time are false. There are many reasons to be drawn to dynamical theories of time, beyond thinking that they might explain or justify our future-biased preferences. (Indeed, one could still hold that facts about the metaphysics of time *justify* future bias, without claiming that those facts, or our beliefs about them, play a predominant role in *explaining* that bias.) So, unsurprisingly, these results give us little reason to have one, rather than another, view about temporal metaphysics itself. [↑](#footnote-ref-18)
18. Exegesis is challenging here, since most discussions of potential evolutionary explanations for future bias in the extant literature do not distinguish between a narrow adaptation for indifference to past experience and a more general adaptation that has the cognitively mediated downstream effect of indifference to past experience. The remarks on future bias in Horwich (1987) and Maclaurin & Dyke (2002), for instance, are most naturally read as positing a narrow adaptation only insofar as their explanations do not mention any more general mechanism or broader adaptation. Suhler & Callender (2012) offer a very intricate evolutionary explanation of future bias, but their story remains open to multiple interpretations at key points. On the one hand, they emphasise that ‘our explanation does not make the implausible claim that there evolved some wholly new, dedicated mechanism for valuing past and future’ (p. 13). But in context, they seem to be simply pointing out that the selective forces that favoured concern for the future and indifference to the past operated on organisms with already-developed, general-purposes capacities for representation (of past and future events) and affective response. The core of their explanation for future bias (p. 12) seems to be simply that, with these cognitive capacities in place, evolution preferred organisms who had strong affective responses to representations of future pleasures and pains, but not to representations of past pleasures and pains. [↑](#footnote-ref-19)
19. Philosophers who see our future-biased attitudes as rational (at least when they concern the agent’s own pleasant and unpleasant experiences) include Prior (1959), Kauppinen (2018), Dorsey (2018), and Pearson (2018). Those who regard future bias as irrational include Dougherty (2011), Brink (2011), Greene and Sullivan (2015), Sullivan (2018), and Garrett (2018). [↑](#footnote-ref-20)
20. Thanks to an anonymous referee for this excellent suggestion. [↑](#footnote-ref-21)
21. James Norton would like to thank the Icelandic Centre for Research (grant 195617-051). Andrew J. Latham would like to thank the Ngāi Tai Ki Tāmaki Tribal Trust for their support. [↑](#footnote-ref-22)