**Can We Turn People Into Pain Pumps? On the Rationality of Future Bias and Strong Risk Aversion**

Abstract

Future-bias is the preference, all else being equal, for negatively valenced events be located in the past rather than the future, and positively valenced ones to be located in the future rather than the past. Strong risk aversion is the preference to pay some cost to mitigate the badness of the worst outcome. People who are both strongly risk averse and future-biased can face a series of choices that will guarantee them more pain, for no compensating benefit: they will be pain pumped. Thus, combining these preferences is rationally impermissible. Dougherty (2011) argues that this gives us reason to think that future-bias is rationally impermissible. This argument, and a similar one presented by Greene and Sullivan (2015), relies on the idea that if several preferences are not rationally combinable, then in the absence of some explanation of why they are not combinable, we should conclude that (at least) one of them is rationally impermissible. We take up this question by, inter alia, reflecting on the empirical results of a study we ran that probes people’s risk averse and future-biased preferences. We argue that the data, in conjunction with other considerations, suggests that we should not infer from the irrationality of the combination of these preferences to the irrationality of one of the preferences singly. We also argue that given the descriptive data about these preferences, there is no reason to think that, if one of them is rationally impermissible, then it is future-bias.

**Introduction**

Going back at least as far as Hume (1739) philosophers have thought that all else being equal, most of us prefer to have positively valenced events (positive events) in the future rather than the past, and negatively valenced events (negative events) the past rather than the future. This preference is known as future-bias. Agents who are future-biased are sensitive to an event’s temporal location *per se.* They are not simply sensitive to its temporal location because that location is associated with certain other features, such as the event’s probability, or it intrinsic features, or to various emotional or attitudes states that tend to accompany an event’s location. They are sensitive to its temporal location even holding fixed all these factors. So, for instance, if the sole reason an agent prefers positive events to be in their future, rather than their past, is because that agent gets a lot of utility from *anticipating* future events, and none from anticipating past events, then that agent is not future-biased since she is not sensitive to the temporal location of the event, but rather, to whether or not it produces various pleasurable anticipatory experiences.

Recent empirical work confirms that, all else being equal, people do prefer positive hedonic events—those that involve pleasure—to be located in the future rather than the past, and prefer negative hedonic events—those that involve pain—to be located in the past rather than the future (Caruso, Gilbert & Wilson 2008; Greene, Latham, Miller & Norton 2021a). It also shows that these preferences extend to non-hedonic events (Greene et al. 2021a) and to events concerning other individuals (Greene et al. 2021a). Moreover, empirical work shows that future-bias is not merely a tiebreaker between equally intrinsically valuable states of affairs, but can lead us to prefer experiencing more pain in the past to less pain in the future (Lee, Hoerl, Burns, Fernandes, O’Connor & McCormack 2020; Greene, Latham, Miller & Norton 2021b) and less pleasure in the future to more pleasure in the past (Lee et al. 2020; Greene, Latham, Miller & Norton forthcoming).

Until relatively recently, it was largely assumed that future-bias is at least rationally permissible (Hedden 2015; Heathwood 2008; Hare 2007, 2008, 2013) and perhaps even obligatory (Prior 1959; Schlesinger 1976; Craig 1999; Pearson 2018). At the very least, it was thought that since future-biased preferences are inert—i.e., that they do not affect behaviour since there are no practical choices influenced by those preferences—the preference is largely harmless (Kauppinen 2018). After all, even if one prefers more pain in the past over less in the future, that preference is not action guiding since one cannot bring it about that one had more pain in the past, over less in the future, since the past is causally inaccessible. This marks a key difference between future-bias and (prospective) near-bias (the preference to have positive events in the nearer future compared to the further future, and to have negative events in the further future compared to the nearer future): it is obvious that we can trade between the near and distant future, but it is not obvious that we can trade between the future and the past (Moller 2002, 77). Indeed, this is one explanation of why philosophers have historically focused attention on arguments against the rationality of near-bias without considering whether such arguments would also count against the rationality of future-bias (Greene & Sullivan 2015, 950–3).

Recently, however, several philosophers have argued that future-bias is not inert when combined with certain other preferences (Dougherty, 2011; 2015; Greene & Sullivan, 2015) and, moreover, that when we combine future-bias with those other preferences we are left worse off in some respect and better off in none. On this basis these authors argue that future-bias is rationally impermissible.

Greene & Sullivan (2015), for instance, focus on the interaction of future-bias and regret-aversion. They argue that this combination of preferences leads to a scheduling problem. For instance, suppose that Annie and Freddie are both excited about chewing on a large kangaroo bone. But they cannot (clearly) do so at the same time. (And let’s suppose that the bone will be intrinsically just as pleasantly chewable whether it is chewed first, or second, and that the probabilities of the chewing events are the same regardless of when they happen). If Freddie is future-biased and chooses to chew first, then he knows he will regret his choice after the chewing is finished, in the sense that he will prefer that he had chosen differently.[[1]](#footnote-2) If he had chosen differently, then he would be like Annie in having future, instead of merely past, pleasurable chewing experiences. Meanwhile, Freddie knows that if he chooses to chew second, he will never regret his choice.[[2]](#footnote-3) If he chooses to chew second, then after his chewing is complete both he and Annie will be in the equivalent situation of having only past chewing experiences, and thus he will not prefer that he had chosen differently. Freddie therefore desires to chew second. Notably, then, Freddie’s being future-biased is not inert: it has an effect on which decision he makes: chew first, or chew second. Moreover, Greene and Sullivan think that Freddie’s scheduling preference is irrational, and hence so is future-bias.

We can make things even worse for Freddie. Suppose that because Freddie prefers to chew second, he will pay a small amount of ‘bone cost’ to do so: he will accept a slightly smaller bone if he chews second rather than first. But then, Freddie’s preference leads him to be worse off, since he will prefer a slightly smaller bone, later, to a slightly larger one, earlier, because he prefers to chew second in order to avoid regret. Hence, the combination of future-bias and regret aversion leads Freddie to choose a slightly smaller bone than he would otherwise have done, and this is a reason to think that the combination of these preferences is not rationally permissible. And, again, Greene and Sullivan conclude that it is future-bias that is the rationally impermissible preference.

In a similar vein, Dougherty (2011) aims to show that people who are both future-biased and strongly risk averse can be turned into pain pumps: making a sequence of choices that will certainly leave them with more pain at some time, no less pain at any time, and no other compensating benefit. Strong risk aversion is to be contrasted with risk aversion, where someone is risk averse if they tend to choose, other things being equal, outcomes with lower levels of uncertainty over higher levels of uncertainty. Someone is strongly risk averse is they are willing to pay some cost in order to decrease the badness of the worst outcome. So, for instance, suppose you are faced with a lottery in which there is a 50% chance of winning $1000 and a 50% chance of winning $100. Now suppose that the lottery will allow you to pay, say, $10 to change the payout of the lottery so that you there is a 50% chance of winning $800 and a 50% chance of winning $300. Your expected payout for each lottery is the same, except that you’d need to pay the $10 to get the second range of payouts. So, overall, you will be $10.00 worse off. You are strongly risk averse if you will pay the $10 (or if you will pay something) to shrink the gap between the least good outcome and the best outcome.

Dougherty concludes that at least one of strong risk aversion or future-bias must be rationally impermissible, and he suggests that it is future-bias.

In Dougherty’s thought experiment, a person knows that they will face either i) a four-hour operation on Monday and a one-hour operation on Thursday, or ii) a three-hour operation on Thursday. They will need to remain awake during the operations and each minute of each operation will feature the same amount of intense pain. Next, the person is offered a ‘help-early’ pill that decreases the length of pain on Thursday by 29 minutes if (i), and increases the length of pain on Thursday by 31 minutes if (ii). A risk-averse person accepts this pill because they are willing to increase their expected pain by one minute to reduce the severity of the worst possibility. Finally, the person wakes on Wednesday with temporary amnesia. They are offered a ‘help-late’ pill that increases the length of pain on Thursday by 30 minutes if (i), and decreases the length of pain on Thursday by 30 minutes if (ii). A future-biased risk-averse person will accept this pill because it reduces the severity of the worst outcome of future pain for free: their risk is reduced while their expected future pain remains the same. But by taking both pills the person guarantees themself one more minute of pain on Thursday, for no compensating benefit.

Very roughly speaking both Dougherty’s (2011) and Greene and Sullivan’s (2015) arguments take the same form. They aim to show that when conjoined with some other preference, future-bias leads to bad outcomes. This gives us reason to think that the combination of these preferences is rationally impermissible. Since we have no reason to suspect that the preference with which future-bias is conjoined is rationally impermissible, and since the combination is rationally impermissible, we should conclude that future-bias is rationally impermissible.

We can schematically represent both Dougherty’s and Greene and Sullivans’ argument as follows where P2 is future-bias and P1 is either regret aversion or strong risk aversion.

1. The combination of preferences P1 and P2 is rationally impermissible if it leads to the agent who combines those preferences being worse off in some respect, and no better off in any respect.
2. If each of P1 and P2 is rationally permissible, then the conjunction of P1 and P2 is rationally permissible unless there is strong intuitive purchase on the mechanism that rules out their combination (**Additivity**).
3. There is no strong intuitive purchase on the mechanism that rules out combining P1 and P2 (**Intuitive Purchase**)
4. Therefore, at least one of P1 or P2 is rationally impermissible (modus tollens from 2 and 3 together with 4.
5. Since P1 is rationally permissible, P2 is rationally impermissible.

Before we begin, a note of clarification. There are different ways you might think that the rationality, or not, of preferences is grounded. You might think that it is fairly weak notion, grounded in whether a preference, or a package of preferences, when chosen by actual agents, tends to leave those agents better off or worse off. Call this the view that rationality is grounded in revealed outcomes. By contrast, you might think that the rationality of preferences (or packages thereof) is grounded in something we can learn *a priori* by reflecting on scenarios or thought experiments in which agents who combine certain preferences are shown to be better off, or worse off. On this latter view, whether or not there are any agents who do combine these preferences in this manner is irrelevant: what matters regarding the rational status of the preference, or package of preferences, is that were certain preferences to be combined in a certain manner, they would lead the agents to be worse off. Call this the view that rationality is grounded in a priori reflective coherence. While many or even all all should agree that there are some cases like this (inconsistent preferences, for example), we take it that Sullivan (2018), Greene and Sullivan (2015) and Dougherty (2011, 2015) take this latter view to apply to a larger range of prefences. So difference will be between those who think that thinking about the way packages fail to combine well provides evidence for some relatively intrinsic irrationality of certain preferences, and those who think it just shows what packages do not combine well, and making further non-Humean inferences is not very fruitful.

Much of what we say in this paper takes no stand on which of these is the right view of rationality of preferences. Our discussion of Additivity and Intuitive Purchase, for instance, does not hang on which of these is right. Thus nothing we say there presupposes a view of rationality to which our interlocutors do not subscribe. The same is not, however, true when it comes to the use of the empirical data we collected. Here, we will make use of some empirical facts about actual agents’ behaviour, to suggest that there is no powerful reason to hold that future-bias is irrational. This does not mean that we suppose that either Sullivan, Greene and Sullivan, or Dougherty, imagine that their arguments are sensitive to the behaviour of actual agents. None of them is predicting how people behave; rather, they are showing that there are *a priori* relations between certain preferences such that we know packages of these preferences will lead to poor results. They then argue that if we can be sure that certain preferences are rational (which seems to imply that it’s always rationally acceptable to have them) we should blame the other preference in a poor package for the defects of that package. So, our use of empirical data may not hold much sway with these authors.

We, on the other hand, have a more deflationary attitude to the rationality of preferences which accepts something like the idea that rationality is grounded in revealed outcomes. Given that attitude, our empirical results bear on a related question: is having a certain preference harmful, and in what conditions is it harmful? Most preferences will get us into trouble in certain circumstances (when they form intransitive packages, and there are Dutch bookkeepers around; when they are combined in the ways the current authors we mention describe; when they are not co-satisfiable and so on). But that being said we can still ask “Does a certain preference typically cause trouble? What is the expectation that someone who has that preference will be worse off for it?” So in this paper we do not dispute the a priori findings of these authors: but we explore what we should say about the acceptability of future-bias on a more deflationary, or at least a different, understanding of rationality.

Onward. Notice that (2), which we are calling Additivity, is necessary for either argument to go through. For of course, it does not follow from the fact that P1 and P2 are, when combined, rationally impermissible, that either one of them is singly rationally impermissible. It could be that it is only the combination that is impermissible. This is often the case. It is rationally permissible to prefer roo mince to watermelon, and it is rationally permissible to prefer watermelon to chocolate cake, and it is rationally permissible to prefer chocolate cake to roo mince. But having that set of preferences is rationally impermissible, since having intransitive preferences can lead to one being worse off in some respect, and better off in none. So, the following principle is obviously false: If the combination of P1 and P2 is rationally impermissible, then at least one of P1 or P2 is rationally impermissible.

If that principle were true, then, conditional on Dougherty, and Greene and Sullivan showing that the other preference (regret aversion or risk aversion) are indeed permissible, it would follow that future-bias must be rationally impermissible. But that principle is false. So, in (2) we offer a principle very like one to which Dougherty gestures, according to which we should think that if a combination of preferences is impermissible, that at least one of the preferences being conjoined is impermissible unless we have a clear intuitive idea of why combining them is impermissible, such as, the thought goes, we do in the case of intransitive preferences.

In order for these arguments to be persuasive, we need to be convinced not only that the combination of these preferences is rationally impermissible (and let’s suppose we are) but also (a) that additivity is true, (b) we have no intuitive idea of why combining them would be rationally impermissible, and hence we have reason to think given additivity that at least one of them is rationally impermissible and (c) that if one of the pair is rationally impermissible, it is future-bias.

Thus, a key question is the status of additivity: perhaps each of future-bias and risk aversion is indeed severally rationally permissible, and the combination not, despite there being no strong intuitive purchase on the mechanism that rules out their combination.

Before we turn to consider additivity in more detail, it’s worth noting that one might hold the general view that individual preferences cannot be evaluated as rational or irrational, but *only* combinations of preferences. If that were right, then the arguments above would straightforwardly fail, since they aim to show that because some combination is irrational, some individual preference must be. Interestingly, Sullivan (2018 p.85) gestures towards something like this view when she writes:

One lesson from the case is that we should be able to criticize whole packages of preferences, even when they are disconnected from particular choices. Moreover, while attitudes about the past may never be action‐guiding considered individually, they can be action‐guiding in conjunction with your other beliefs and preferences. And to the extent that we think we can measure well‐being across time (as the Success principle supposes) rather than only at a time, we must take past discounting attitudes as subject to rational criticism

In fact, as we read Sullivan she is not claiming that we can only ever evaluate packages of preferences, and not individual ones, but rather, that in the case in which preferences are disconnected from choices and outcomes, we can nevertheless evaluate them as part of larger packages. Regardless though, her view is clearly that insofar as one finds some combination of preferences to be irrational, one can come to have reason to conclude that it is one member of that package that is the problem: that it is that particular preference that is irrational. That, after all, is precisely how the scheduling argument proceeds. We evaluate the combination of no-regret and future-bias and notice that the combination is problematic, and, from there, we are supposed to go on to conclude that it is the future-biased preference that is the problem in this combination. We are invited to conclude that the no-regret principle is obviously rationally acceptable, and therefore that future-bias must be irrational, given that the package is problematic.

It’s worth, though, saying something about the status of a principle like additivity. It’s tempting to think of this as a principle which might or might not be true of preferences which are rationally permissible, given a completely common understanding of what it is for a preference to be rationally permissible. But that is unlikely to be the case. Something like additivity might actually be a part, or not, of different conceptions of what it is to be a rationally acceptable preference. Indeed, surely it *is*  part of the conception that, say, Dougherty is operating with: what it is for him, in part, for a preference to be rationally acceptable, is for it to combine harmlessly with all other preferences except where there are known systematic issues like intransitivity.

We, on the other hand, think that for a preference to be rationally impermissible it must be that an agent can *at least typically[[3]](#footnote-4)* be criticised for having that preference. We say ‘typically’ rather than always because there are of course extreme circumstances in which one might not be rationally criticised for adopting even intransitive preferences in the presence of a Dutch bookmaker - such as when you are offered a huge payout for doing so which exceeds the possible losses to the Dutch bookmaker in that context. But we do not see the absence of risk-aversion as a special manipulation like that. So, we cannot see why we should criticise an agent who is, for example, future-biased but not risk-averse or *vice versa*. So perhaps the difference between us and Dougherty is whether a preference is rationally impermissible (if individual preferences are) only if it *always* gets us into trouble – i.e., every package of preferences of which it is a part is a worse package for the presence of that presence, as we tend to believe, or whether a preference is impermissible if it *ever* gets us into trouble modulo some to our eyes somewhat arbitrary exceptions, as Dougherty tends to believe.

We do not claim that this paper provides a strong argument (perhaps choosing between these conceptions of rational impermissibility is a conceptual engineering project) for adopting our view: but we do think drawing attention to the role of additivity and how it makes more sense on one conception, and not much on the other, is an important beginning.

The strongest conclusion to our paper, contra Sullivan (2018) and indeed contra Dougherty (2011) Greene and Sullivan (2015) which we won’t defend in detail, is that while we can indeed criticize certain combinations of preferences, it’s the combinations *themselves* that are the problem, not the individual preferences themselves, and, indeed, that individual preferences cannot be rationally evaluated. If that’s right, then advice to give about preferences is “avoid these combinations”. Of course if there are preferences which have huge benefits (e.g., in allowing many other preferences to be satisfied if one has them) and these preferences clash with preferences of a second kind which in most contexts do nothing useful, but get you into trouble combined the first kind, we might give the advice “try not to form preferences of the second kind ever, lest you get into the habit and coming them with the first kind which you will often have”. But such advice seems weaker that saying that preferences of the second kind are “irrational”. It would still be fine to form preferences of the second kind when and only when the first kind are absent. And, we should add, it’s far from clear that even anything as weak as this is true if we were to think of future-bias as a preference of the first kind.

Our aim in what follows, however, is not to directly argue for the claim that it is only combinations of preferences that can be rationally evaluated. Rather, we aim to take up the question of whether we should accept the arguments given above, for the irrationality of future bias, and, in particular, we aim to evaluate the status of Additivity.

Thus, in this paper we take up three questions, several of which are informed by empirical work that we undertook.

The first question we aim to answer will help us in evaluating Additivity. This question is: do people exhibit a certain combination of preferences: risk aversion and future-bias.[[4]](#footnote-5) In particular, we want to determine whether in fact people combine these preferences freely, or whether they tend to be anti-correlated. If the preferences are anti-correlated this could provide reason to think that Additivity is false. Why so? Well, it would suggest that people tacitly recognise that the combination of these preferences is rationally impermissible (since they do not combine them). If there are people who are future-biased, and others who are strongly risk averse, but not people who are both, then this gives us *some* grounds to think that people tacitly judge each of these preferences to be, singly, rationally permissible. Compare: if we find that lots of people prefer A to B, and lots prefer B to C, and lots prefer C to A, but that very few, if any, prefer A to B, B to C, and C to A, then this would give us some sort of defeasible reason to think that those people recognise that this *combination* of preferences is irrational.

Of course, the fact that people display such behaviour is only defeasible evidence for them having a tacit judgement of that kind. It might be that there are evolutionary pressures (say) to not combine these preferences which explains why one does not find that combination, even though it would not be right to say that any individual tacitly judges them to be rationally impermissible. For instance, imagine a case in which even after explaining why intransitive preferences lead to bad outcomes, such people seem baffled, and even explicitly judge that intransitive preferences are perfectly rational, even though as a matter of fact none of those people have such preferences. It would be odd to say that they tacitly judge the combination to be irrational. So even if we were to find that these preferences are anti-correlated, this would not show that people do tacitly make this judgement. But it would provide some weak evidence for the idea, and would certainly make it worth further investigation.[[5]](#footnote-6)

One reading of what we are doing here, then, is trying to empirically establish that future-bias is as a matter of fact rationally permissible. That would be a stronger claim that we have in mind; indeed it’s not clear to us that we should be in the business of asking if individual preferences are rationally permissible, rather than the combinations of preferences, as we discuss above. But there is a weaker thing that such empirical evidence can aid us with: offering advice as to whether, in general, a certain preference is likely to have bad consequences for an agent. Offering situated (or general where the situations are common) advice about what particular preferences to have. Someone who prefers apples to oranges and oranges to pears, might be advised not to get into a situation which would make them prefer pears to apples, for example—at least in an environment where there are Dutch Fruiterers around prepared to exploit this. But this is to say nothing about the individual preferences in general, and, even more controversially, it's to say nothing about transitivity except in exploitative environments.

So, in addition to the interesting discovery that there are perhaps tacit judgements that certain combinations of preferences are impermissible, the data could also weakly suggest something else: that people who are future-biased need not worry too much because that preferences is unlikely to be combined with other preferences in a problematic way.

It’s worth, however drawing an interesting distinction here. The data might suggest that the *recognition* that you are future-biased should not concern you, if it is relatively unlikely to be combined in a problematic way with, at least, a relevant kind of risk aversion. Even if that is right though, it would not follow that the data could be used to generate general counterfactual advice of the kind “if you were to become future-biased it would have no ill consequences”. If you are already risk averse, or perhaps have some other preference like the one called weak forecasting in Sullivan 2018 (a preference which we are far from sure is always good to have let alone required) which combines badly with future=bias, then you should not adopt it - at least not without ridding yourself of the preferences that combine with it badly[[6]](#footnote-7). But we think the evidential information by itself would be important: if the data goes a certain way, it tells us that future-bias does no harm; due to relatively systematic ways in which future-bias fails to combine problematically, in ways that suggest there may be a kind of tacit understanding of what combinations would be problematic. And, on a strong view about the rationality of individual preferences which we are tempted by but do not defend in detail here, that’s about as much as you can expect.

The second question we aim to answer is this. Suppose that Additivity is true, and also that people have no intuitive purchase on why combining the relevant preferences is not rationally permissible. Then it seems that (at least) one of future-bias or strong risk aversion is rationally impermissible. The question is which. In order to answer this question, we can ask and answer another: are people who exhibit future-bias worse off (on average) than those who exhibit risk aversion or vice versa? Answering this question might help to decide which of these preferences is rationally impermissible assuming that one is. This idea is motivated by the thought that, at least in part, the rational evaluation of preferences, or combinations thereof, is grounded in facts regarding whether those preferences tend to make one better, or worse, off. On this sort of view of rationality, what *makes* a preference, or combination thereof, rationally impermissible, is that having that preference tends to make one worse off.[[7]](#footnote-8) If we have a simple view like this, then if it turns out that those who are future-biased tend to be worse off than those who are strongly risk averse, then this would give us a reason to think that if one of these preferences is rationally impermissible, then it is being future-biased not being strongly risk averse. Now, one might think that what matters is not whether as a matter of mere happenstance a preference happens to make one worse off, but whether it tends to make one worse off in some modally robust manner. If one takes this view, then even if future-biased people are indeed worse off than strongly risk averse people, it does not follow that it is the former rather than the latter that is rationally impermissible (assuming one is): for perhaps it is a merely contingent feature of our world that the future-biased are worse off than the strongly risk averse. Still, we take it that before we can even consider the question of whether a particular contingent pattern of being worse-off is modally robust, we need to know what that pattern is. We aim to ascertain this.

Our third and final question is this: how prevalent is the combination of these preferences? Now, determining the prevalence of this combination will not, in itself, shed light on the rational permissibility of either of the preferences taken singly. Even if almost no one *in fact* combines these preferences, if the arguments just articulated succeed, then they show that one of the preferences singly (i.e. future-bias) is rationally impermissible, and that is sufficient for those arguments to be important and interesting since lots of people are in fact future-biased. Nevertheless, we take it that it is of interest whether, if future-bias is rationally impermissible, exhibiting that bias is likely to lead to those who have it in fact being made worse off by combining that preference with strong risk aversion. We can think of the question as this: are future-biased people irrational but lucky (because they do not combine future-bias with strong risk aversion) or are they in a position in which they can easily be turned into pain pumps? Answering this question tells us nothing about whether future-bias is rationally permissible, but it does tell us whether, if future-bias is rationally impermissible it is largely (at least in this respect) harmlessly so or not.

In what follows we examine people’s judgements in a version of Dougherty’s (2011) thought experiment. In doing so we test hypotheses. First, we hypothesized that on Monday, most people would prefer to be on the late series rather than the early series. This is, we predicted that most people would aim to minimize total future surgery time, and will prefer future surgeries totalling 3 hours, rather than 5 hours.

H1: On Monday, most people will prefer to be on the late series.

Next, we predicted that on Sunday, most people would choose to take the yellow pill. That is, we predicted that most people would be strongly risk averse. We made this prediction based on prior research that shows that people show risk aversion of this kind.[[8]](#footnote-9)

H2: On Sunday, most people will be strongly risk averse.

We also predicted that on Wednesday, most people would choose to take the blue pill. That is, we predicted that most people would be future-biased. We made this prediction based on prior research that shows that people are future-biased.[[9]](#footnote-10)

H3: On Wednesday, most people will be future-biased.

Though current data about the prevalence of future-bias and strong risk aversion do not tell us whether to expect people to be both risk averse and future-biased, we decided to predict that most people would be *both* risk averse and future-biased.

H4:  There will be a positive association between people being strongly risk averse and being future-biased. Most people will be both strongly risk-averse and future-biased.

We also had no real basis on which to make a prediction about the connection between risk aversion and future-bias. However, since we predicted that most people would show both preferences, it was natural to predict that there would be a positive association between being risk averse and being future-biased (rather than an anti-correlation).

These hypotheses were preregistered at https://osf.io/7cdyq[[10]](#footnote-11).

We begin, in section 2, by outlining our methodology and results, before in Section 3 considering the upshot of those results for discussions regarding the normative status of future-bias.

**2. Methodology and Results**

**2.1 Methodology**

2.1.1 Participants

100 people participated in the study. Due to the complexity of the situations described in this experiment we recruited undergraduate students in philosophy rather than paid online participants (who could be expected to spend less time reading and thinking about the vignettes). 25 participants were excluded for failing to answer all the comprehension check questions correctly. The remaining sample was composed of 75 participants (35 female, 7 trans/non-binary, 4 prefer not to answer; aged 17-60 mean age 21.33 (SD = 5.78)). Ethics approval for these studies was obtained from the [blanked] Human Research Ethics Committee. Informed consent was obtained from all participants prior to testing. The survey was conducted online using Qualtrics.

2.1.2 Materials and Procedure

All participants see the same series of three screens and answer the same comprehension and probe questions. The first screen introduces two series of operations, the early series and the late series, and participants are told that there is a 50% chance that they are on the early series, and 50% chance that they are on the late series. They are then asked which series they prefer to be on. Since the early series involves a total of 5 hours of surgery, and the late series a total of 3 hours of surgery, we predict that people will, from the standpoint of Monday, prefer that they are on the late series.

 On the first screen they see the following:



They then see the following four comprehension questions:

“If you have the late series, how many hours of surgery will you undergo?”

And given the option of (a) 3 (b) 5 (b) 6

“If you have the early series, how many hours of surgery will you undergo?”

And given the option of (a) 3 (b) 5 (b) 6

“If you have the early series, you will have surgery on how many days?”

And given the option of (a) 3 (b) 5 (b) 6

“If you have the late series, you will have surgery on how many days?”

And given the option of (a) 1 (b) 2 (c) 3

Participants are then asked:

“Which do you prefer?”

And given the option of: (a) the early series or (b) the late series .

Participants then see a new screen which reads as follows:



After having seen this vignette, participants then respond to three comprehension statements to which they can choose (a) true or (b) false.

“The yellow pill decreases the length of the surgery you will have, regardless of whether you are on the early of the late series.”

“The yellow pill makes things better for you if you are on the early series, and worse for you if you are on the late series.”

“The yellow pill makes things worse for you if you are on the early series, and better for you if you are on the late series.”

They are then asked:

“Do you take the yellow pill?”

And can respond with (a) yes and (b) no.

People who are risk averse are willing to pay something in order to decrease the gap between the best and worst outcomes. If the yellow pill is taken on Sunday, then since instead of facing a 50% chance of 5 hours of painful operation on Thursday, and a 50% chance of 3 hours of painful operation on Thursday, instead one faces face a 50% chance of a 4 hours 31 painful operation on Thursday, and a 50% chance of a 3 hours 31 on Thursday. Taking the yellow pill is not costless, however: by taking the yellow pill someone has 1 minute of additional expected pain compared to if they do not take the pill. Nevertheless, we would expect that people who are risk averse will choose to take the yellow pill.

Participants then move the last screen. They see the following:



Participants then respond to the following comprehension statements and are given the option of responding with (a) true or (b) false.

“The blue pill decreases the length of the surgery you will have, regardless of whether you are on the early of the late series.”

“The blue pill makes things better for you if you are on the early series, and worse for you if you are on the late series.”

“The blue pill makes things worse for you if you are on the early series, and better for you if you are on the late series.”

Participants are then asked:

“Do you take the blue pill?”

And given the option of (a) yes (b) no.

Taking the blue pill does not change the amount of expected pain on Thursday. So, it does not change the amount of expected future pain. But it does narrow the gap between the outcomes on Thursday (and does so regardless of whether or not one took the yellow pill). So, people who are future-biased and risk-averse can be expected to take the blue pill.

If, however, people take both pills then regardless of which course they are on, they end up with 1 extra minute of pain, as can be seen from the table below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Effect of yellow pill | Effect of blue pill | Overall result of taking both pills for length of operation on Thursday. |
| Early course | -29 | 30 | 1 |
| Late course | 31 | -30 | 1 |

Notably, by taking both pills it is not simply that people end up with more pain overall, where some of that pain is distributed, say, in the past (after all, a future-biased person will likely think this is a good outcome). Rather, by taking both pills the person guarantees themself one more minute of pain **on Thursday,** for no compensating benefit at any other time.

**2.2. Results**

Before reporting our data, we will summarise our major findings with respect to our four hypotheses. Our hypotheses were, first, that most people will prefer to be on the late series; second, that most people will be strongly risk averse; third, that most people will be future-biased, and fourth, that there will be an association between being strongly risk averse and future biased, and that most people will be both strongly risk averse and future biased. Only our first hypothesis was vindicated. Nearly everyone preferred to be on the late series (and so minimize the total future surgery time). In contrast, we found no evidence to suggest that most people would take the yellow pill (be strongly risk averse), take the blue pill (be future-biased), or that there is an association between people’s choice to take the yellow pill (or not) and take the blue pill (or not).

In support of hypothesis one, we found that nearly everyone reported that they would prefer to be on the late series (69, 92.0%) than the early series (6, 8.0%; χ2(1, N = 75) = 52.920, p < .001).Table 1 below summarises the descriptive data of participants’ responses to the questions “Do you take the yellow pill?” and “Do you take the blue pill?”

|  |  |
| --- | --- |
|  | Do you take the blue pill? |
| Yes | No |
| Do you take the yellow pill? | Yes | 19 (25.3%) | 18 (24.0%) |
| No | 15 (20.0%) | 23 (30.7%) |

Contrary to hypothesis two, we found no evidence that most people will be strongly risk averse. Instead, people were divided between choosing to take the yellow pill (37, 49.3%) and choosing not to take the yellow pill (38, 50.7%; χ2(1, N = 75) = .013, p = .908).

Contrary to hypothesis three, we found no evidence that most people will be future-biased. People were divided between choosing to take the blue pill (34, 45.3%) and choosing not to take the blue pill (41, 54.7%; χ2(1, N = 75) = .653, p = .419),

We also found no evidence that most people will be both strongly risk averse and future biased (contrary to hypothesis four). In fact, we found no evidence of any association at all between people’s choice to take the yellow pill (or not) and take the blue pill (or not), χ2(1, N = 75) = 1.067, p = .302). People’s choices to take each pill (or not) were independent of each other.

3. Discussion

3.1 Our three questions

Let’s begin by answering our three questions. Our third question was: how prevalent is the combination of future-bias and strong aversion?

Neither of our hypotheses about the percentage of people who would be risk averse or future-biased was supported. Rather than finding that a majority would be risk averse, and a majority future-biased, instead we found that half were risk averse, and half future-biased.

Unsurprisingly, then, we did not find that a majority of people were *both* risk averse and future-biased. Rather, we found that around a quarter of people were both risk averse and future-biased. This is an interesting result. It shows, on the one hand, that *most* people do not have the problematic *combination* of preferences. Still, it is notable that ~25% of people do have this combination of preferences, and so roughly a quarter of the population is disposed to be pain pumped, given the right series of choices. This is an important result. Even though this is a minority of people, it’s not an insubstantial minority. So, given that the at the very least, the combination of these preferences is rationally impermissible, this shows that this irrationality shows up in around a quarter of the population we tested. Moreover, it tells us that if, as Dougherty and Greene and Sullivan argue, future-bias is rationally impermissible, then it is not *harmlessly* impermissible. For roughly half of the people who have that preference also have a preference (strong risk aversion) that can lead them to be pain pumped.

Our second question asked: are people who exhibit future-bias worse off (on average) than those who exhibit risk aversion or *vice versa*? If they are, then on the assumption that Additivity is true, this *could* be a reason to think that it is future-bias rather than strong risk aversion that is rationally impermissible. In fact, though, when we look at the population of risk averse people, they do not do better, on average, than the population of future-biased people. Roughly the same proportion of risk averse people get into trouble by also being future-biased, as future-biased people who get into trouble by also being strongly risk averse. So, even if one thinks there is some connection (constitutive or evidential) between a preference tending to make one worse off and being rationally impermissible[[11]](#footnote-12) the data does not support the conclusion that *if* one of future-bias or strong risk aversion is rationally impermissible *then* we have reason to think it’s the former rather than the latter because the former tends to make us worse off than the latter.

Our first question was: do people freely combine strongly risk averse and future-biased preferences? For, one might think, if these preferences are anti-correlated, this *might* give us some reason to think that Additivity is false. We did not find this. However, we also did not find that they were anti-correlated. We found no association between the two preferences. That is why we find that about 25% of the population have both sets of preferences, given that 50% are risk averse and 50% future biased. So, our results here do not provide support for the idea that people tacitly recognize that combining these preferences is problematic. As such, it does not provide support for an argument against Additivity that proceeds by arguing that people’s tendency to be future-biased, and to be strongly risk averse, but not to be both, gives us some reason to think that it is the combination of these preferences that is rationally impermissible, rather than either of the singly.

What light then, does our empirical evidence shed? Well, it *fails* to provide us with a reason to think that if one of future-bias or strong risk aversion is rationally impermissible, then we have reason to think it is the former over the latter. It also *fails* to give us a reason to think that Additivity is false. But of course, the fact that this evidence does not suggest that *if* one of the preferences is singly rationally impermissible, it is future-bias not strong risk aversion, does noy tells us which of the two we should take to be rationally impermissible, assuming one of them is. Moreover, the fact that this evidence does not give us a reason to think that additivity is *false*, does not mean that we have a positive reason to think it is *true*. So it does not mean that we should assume that since the combination of these preferences is rationally impermissible, that at least one of them is singly. At this point, our empirical evidence can be of no further help. But that does not mean that we cannot make progress with these questions.

In what follows we will begin by supposing that one accepts Additivity, and hence that one supposes that at least one of future-bias and strong risk aversion must be rationally impermissible. We will then ask whether there are *other* reasons, not arising from empirical considerations but instead arising from the status of strong risk aversion, for concluding that it is future-bias that is rationally impermissible.

3.2 The status of strong risk aversion

Dougherty argues for strong risk aversion by claiming that if it were rationally impermissible, it would mean that people are irrational to buy insurance from companies that expect to make a profit, and that this is deeply counterintuitive. There are two reasons to doubt the analogy between strong risk aversion when it comes to money (as in the case of insurance) and strong risk aversion when it comes to pain, where it is the latter that is relevant to Dougherty’s thought experiment. One reason to doubt the analogy is that as Greene and Sullivan point out, Dougherty’s thought experiment requires agents to be strongly risk averse in relation to pain, where the pain experiences do not exhibit increasing marginal disutility. Strong risk aversion with respect to money is standardly justified by the increasing marginal disutility of monetary losses. Thus, the rational permissibility of Dougherty’s form of strong risk- aversion cannot be established in the same way (Greene & Sullivan 2015, 955–6).

There is also a second argument to this conclusion which does not rely on the distinction between monetary value and utility. According to this argument one can rationally buy insurance without engaging in strong risk aversion. So even if it is rational to buy insurance, it does not follow that it is rational to be strongly risk averse (about money or experiences). The idea it this: buying insurance is only uncontroversially rational in cases where we are insuring against unrecoverable losses. But consider the case of unrecoverable losses. It might look as though insuring about unrecoverable loss is justified by strong risk aversion. But in fact, the long-term effect of an unrecoverable loss will often, or perhaps always, be to destroy earning capacity, and in such cases the real dollar loss will be so high that purchasing insurance straightforwardly maximises expected monetary value.

In all, then, it is not obvious that if one of strong risk aversion and future-bias is rationally impermissible, then it is the former rather than the latter.

Perhaps, though, there are other, independent, reasons to doubt the rational permissibility of future-bias. We turn to consider that next.

3.3 Independent reasons to doubt the rationality of future-bias

If we know that the combination of future-bias together with strong risk aversion is irrational, then if we also have some independent evidence that future-bias is irrational, then we should be more tempted to conclude that it is future-bias rather than strong risk aversion that is rationally impermissible. Are there such independent reasons? Maybe so.

We know that preference reflection is false for people who are future-biased. Preference reflection is the principle that one is rationally required to now prefer A to B if one believes that one will later prefer A to B (Hedden, 2015, 60). This principle is analogous to the principle of reflection for beliefs, first introduced by van Fraassen (1984). Reflection for beliefs says that you should defer to the beliefs that you anticipate having in the future: that is, if you’re going to come to believe Q (or have credence z in Q) then you should believe that Q (or have credence z in Q) now.

Agents for whom preference reflection is false can be turned into money pumps. Suppose that Annie’s default dinner meal is beef. Now suppose that as of today (Monday) Annie would prefer kangaroo to beef for dinner in two-day’s time (Wednesday). Suppose she thinks it is worth $1.00 to make sure that she gets the kangaroo not the beef on Wednesday. Now suppose Annie also believes that tomorrow, she will prefer that on Wednesday she will have beef rather than kangaroo. Tomorrow, she will pay $1.00 to get the beef not the roo on Wednesday. Then Annie will end up paying $2.00 to gain nothing.

Having said that, preference reflection is controversial.[[12]](#footnote-13) Hedden denies the principle, as does Harman (2009). Hedden offers several arguments against preference reflection, some stronger than others. First, he argues that if true, what you should prefer now depends on facts about personal identity, and that this is troublesome. (Consider fission cases in which it is unclear which, if any, of the post fission products is you). As Hedden notes though, this is a relatively weak reason to discard the principle since it can easily be amended to, for instance, appeal to facts about the preferences of future stages that are I-related to one. We set this kind of worry aside. There are, however, better reasons to worry about preference reflection. Hedden thinks that the principle must be false because it leads to what he calls a bootstrapping problem. This problem arises when you believe that what you will later prefer depends on what you will do now. Here is the idea. Suppose Annie is deliberating about whether to walk to the swamp, or to walk to the headland. She believes that whichever she does she will be glad she did that walk and not the other. Preference reflection, then, says that which walk she ought to prefer now, depends on which walk she believes she will take. If Annie finds herself believing that she will walk to the swamp, then she will believe that she will later prefer the swamp walk to the headland walk, and so by preference reflection she should now prefer the swamp walk to the headland walk. But equally, if she believes she will end up walking to the headland, then she should believe that she will prefer the headland walk to the swamp walk, and she should prefer it now. But then, worried Hedden, preference reflection seems to endorse the following bootstrapping reasoning: I believe I will walk to the swap, so I ought to walk to the swamp” and vice versa. And, Hedden thinks, this reasoning is irrational since you know that whichever walk you take, you will be glad that you did that walk. We could call this the positive bootstrapping problem.

There’s also a related negative bootstrapping problem. Suppose Annie believes that whichever walk she does, she will wish she had done the other walk. So, if she now believes that she will walk to the headland, she ought to believe that she will later prefer the swamp walk to the headland walk, and so by preferences reflection she ought to now prefer the swamp walk to the headland walk. So the principle of reflection endorses reasoning such as “I believe I will go to the headland, and so I ought to go to the swamp”. And, Hedden thinks, reasoning is bad.

Hedden worries that preference reflection is arbitrarily asymmetric. It privileges the preferences of our future selves over those of others, and over those of our past selves.

Consider the self/other case first. We do not, and should not, accept a principle according to which we should defer to the preferences that we believe *others* have. As Parfit (p 187) notes, we might think that something like preference reflection gives us a reason to not be near-biased, because we predict that our later selves will regret our having these preferences. If we spend the money now, rather than save it for later, we predict that our later selves, who are less well off, will regret that we preferred to spend rather than save. And, the thought goes, this gives us a reason, now, not to have this preference. But, Parfit says, we would not accept that kind of reasoning when it comes to the preferences of others. Suppose that Jeremy is self-interested: he is biased towards himself. Suppose that so is everyone else. As a consequence of this, Jeremy does less well because other people act in accordance with their self-biased preferences. Jeremy regrets that other people have these preferences. But, Parfit argues, it does not seem that this gives Jeremy reason to regret that he is self-biased. Hedden concludes that there is something objectionably arbitrary about a principle that applies in the case of deference to one’s own selves, but in the case of others. But the self/other asymmetry is not the one asymmetry. Preference reflection says that we should defer to the preferences of our later selves, not our earlier selves. Parfit (p 157) thinks that a similar principle that deferred to the preferences of our earlier selves would be obviously false. He argues that the fact that his younger self preferred to write romantic poetry to gardening gives his current self no reason at all to have that preference.

Finally, Harman (2009) denies preference reflection because, she thinks, the parallel between beliefs and desires does not hold. Reflection for beliefs is true (if it is) because beliefs always have the same goal: the truth. Given that, a future belief held in a position that is epistemically as good as, or better than, one’s current position, should, it can be argued, be treated as the belief of an expert or authority. But, Harman thinks, what is reasonable to prefer can change over time because who we love changes over time as does their nature. Here’s the idea. It can be unreasonable for me to prefer that I suffer a terrible life-changing experience later this year, from which I would emerge with a particular set of life lessons and character traits, even though, after I have gone through that experience (assuming I in fact will), it will then be reasonable for me to prefer that those things have happened. This shows that as identity changes, what it’s rationally permissible to desire changes too. This is meant to explain why a teenager who has chosen to conceive will later be reasonable in preferring that her child exists, even though it would be better (in all the ways she should care about at the time she chooses) if she waits to conceive late (and hence, argues Harman, prior to conception the teenager should prefer to wait). This is all to say that the fact that an agent can predict that they will come to prefer a certain state of affairs (the child existing, for instance) does not give her reason to now prefer to bring that child into existence.

This argument will be persuasive to the extent that preference reflection seems correct, and for what it’s worth we side with Hedden and Harman here.

Are there other reasons to suppose that it is future-bias rather than strong risk aversion that is rationally impermissible? Perhaps so.

3.4 The argument from multiple pairs

This argument proceeds by observing that future-bias is not only rationally incompatible with strong risk aversion, it’s also, as Greene and Sullivan argue, rationally incompatible with regret aversion. So we have two pairs of preferences whose combination is rationally impermissible. The one shared member of each pair is future-bias. And that gives us some reason to think that if one of these preferences is, singly, irrational, then it is future-bias. For if it’s not future-bias that is irrational, then it must be both risk aversion and regret aversion.

Of course, if regret aversion is independently thought to be a rationally impermissible then this argument would not apply. And if, like us, you do not think that strong risk aversion is rationally permissible, then the irrationality of one of the pairs is already explained, which weakens this thought. Still, if one were undecided regarding which of future-bias and strong risk aversion is rationally impermissible then an argument of this form could be persuasive if one accepts additivity. So, it’s time to return to re-consider additivity.

3.5 The Status of Additivity and Intuitive Purchase

Dougherty’s argument only goes through if additivity is true and so is intuitive purchase. Though we don’t accept additivity (it’s actually hard to find positive arguments for it) we haven’t argued at length directly against it here. But additivity is irrelevant if intuitive purchase is false. And intuitive purchase seems to be an odd principle. Sure, we have a nice mathematical explanation of why certain preferences which are severally rationally permissible are jointly rationally impermissible in the case of, say, transitivity. This is supposed to be the paradigm of intuitive purchase. But it’s unclear that this is a widely, transparently understood principle in the community at large. Certainly, philosophers and economists understand it - because they have seen the simple arguments that show why intransitive combinations of preferences are problematic. But it was worth publishing these arguments. One way to think about Dougherty’s paper is precisely that he offers us just such an argument which shows that jointly, strong risk aversion and future-bias are rationally impermissible. Is this argument any harder to understand than Dutch books arguments, or any less intuitive? We think not. They are *similarly intuitive.* If that is right, then you will draw one of two conclusions with regard to the argument against future-bias offered by Dougherty. If you think Dutch book arguments and the argument for the combined impermissibility of future-bias and strong risk aversion are both ones with respect to which we *have* intuitive purchase, then you should think that in these cases additivity doesn’t apply because we do have intuitive purchase. In that case Dougherty’s argument against future-bias does not go through. On the other hand, if you think Dutch book arguments and the argument for the combined impermissibility of future-bias and strong risk aversion are both ones with respect to which we *lack* intuitive purchase, then you should conclude that additivity is false, since clearly we can have reason to think that there are preferences which are, *merely* *jointly* rationally impermissible even though we have no intuitive purchase on why this is so. In this case, too, Dougherty’s argument against future-bias does not go through.

3.6 Limitations of the Study

At this point it is worth reflecting on some general limitations of this research. First, the vignettes that participants were required to understand are complex and require quite a bit of thought to follow throughout the task. To maximize comprehension, we included several reminders of what each choice involved: that is, we always reminded participants of how much pain would be associated with a particular choice, before they had to report their choice. We also included attention and comprehension checks designed to exclude those participants who did *not* comprehend the vignettes or follow the task. This has enabled us to achieve a sample composed of people who understood and followed the vignettes throughout the task. It is important to note that while we had to exclude 25% of our sample for failing one of the attention or comprehension checks, this number of exclusions is not especially high and is lower than some other studies run on undergraduate philosophy students. For instance, while investigation people’s personal identity judgments, Nichols and Bruno (2010) found that 33% and 44% of students failed to adequately understand a vignette using a single comprehension check.

One might still worry, however, that even those who did pass comprehension did not fully understand the scenario they were considering, or, alternatively, that even if they did, the complexity of the scenario may have affected their choices. One reason to suppose this might be so is that we found somewhat lower levels of future-bias than are typically found in simpler studies. One thing to note, here, is that the amount of future-bias reported varies significantly across different study design and indeed within a study, across different vignettes. For instance, Latham, Miller, Norton and Tarsney (2020) found that between ~40% and ~83 of people reported future-biased preferences when it came to negative hedonic states of affairs. When we look to other studies, we also find a spread of results in this regard. Greene, Latham, Miller and Norton (2021) found that ~85% of participants were future biased when it came to negative hedonic states of affairs such as pains. Latham Miller and Norton (2023) found ~70%, Latham, Miller, Tarsney and Tierney (2021) found ~70%, Latham, Oh, Miller, Shpall and Yu (forthcoming) found ~63% while Greene, Latham, Miller and Norton (2022) found 62%. Our results in this study are a little lower then most of these, though they will within the range reported in Latham, Miller, Norton and Tarsney. This suggests that the complexity of the setup may in fact have reduced future-biased preferences a little. This is unsurprising: in previous studies participants are typically asked whether they prefer N units of pain (say) in the future, of N units in the past. In our study we ask people about taking a pill (or not), and even though we draw attention to the upshots of taking the pill, it is likely that these were still less salient than in previous studies. Nevertheless, given that our results here do fall within the range of already published results in this regard, we think that this is some cause for optimism that people who passed comprehension did by and large understand the set up and did respond appropriately.

A final worry one might have is that perhaps the remaining sample is not representative. Perhaps people that pass all the comprehension questions are more thoughtful, reflective, or intelligent, than those who did not. If that is so, then you might expect such people to be less inclined to combine preferences in in irrational manner. Hence, had we found this combination of preferences to be anti-correlated in our sample, one might worry that, though interesting, this finding does not generalize to the population at large. As things stand, however, while such issues should be borne in mind in what follows, we see no good reason to think that our sample is (especially) unrepresentative.

4. Conclusion

Our aim has been to evaluate a recent argument against its permissibility by (a) conducing empirical research and reflecting on what this tells us about the principles employed by the argument and (b) by independently evaluating those principles. As it happens, we do not think the argument is persuasive in showing us that we should think that one of future-bias and strong risk aversion is rationally impermissible just because the *combination* is (we don’t think there is good reason to accept additivity). Moreover, even if there *were* such reason, we are not yet convinced that the guilty party is future-bias. None of this, of course, shows that future-bias is rationally permissible. It just shows that this particular argument does not give us reason to think that it is not.

References

Briggs, R. (2010). ‘Decision-Theoretic Paradoxes as Voting Paradoxes.’ Philosophical Review 119: 1–30.

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.

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

Ehrlich, I., and Becker, G. S. (1972). Market Insurance, Self-Insurance, and Self-Protection. *Journal of Political Economy* 80(4): 623-648.

Fletcher, G. (2021). *Dear Prudence. The nature and normativity of prudential discourse*. OUP.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings.* Cambridge, MA: Harvard University Press.

Gibbard, Allan. 1999. “Morality as Consistency in Living: Korsgaard’s Kantian Lectures.”

*Ethics* 110: 140–164.

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

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

Greene, P., Latham, A. J., Miller, K., & Norton, J. (2022). “On Preferring that Overall, Things are Worse.” Philosophy and Phenomenological Research. <https://doi.org/10.1111/phpr.12819>

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

Greene, P, Latham, A. J. Miller, K., and J Norton (2022). “Why are People So Darn Past-Biased?”. In Temporal Asymmetries in Philosophy and Psychology. Edited by C Hoerl, T McCormack, and A Fernandes. OUP.

Hare, C. (2007). ‘Self-bias, Time-Bias, and the Metaphysics of Self and Time.’ Journal of Philosophy 104: 350–73.

Hare, C. (2008). ‘A Puzzle About Other-Directed Time-Bias.’ Australasian Journal of Philosophy 88(2): 269–277.

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.

Harman, E. (2009). “I’ll be glad I did it” reasoning and the significance of future desires. Philosophical Perspectives 23, 177-199.

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.

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.

Latham, A. J., Miller, K and Norton, J. (2023). “Against a Normative Asymmetry between Near- and Future-bias” *Synthese* 10.1007/s11229-023-04045-1

Latham, A. J., Miller, K, J. Norton, J. and Tarsney, C. (2020) “Future Bias in Action” *Synthese*. DOI: 10.1007/s11229-020-02791-0 (2021) 198(12), 11327-11349

1. Latham, A. J. Oh, J., Miller, K. Shpall, S. Yu, W., (forthcoming) “Exploring Arbitrariness Objections to Time-Biases” *Journal of the American Philosophical Association.*

Latham, A. J., Miller, K. Tarsney, C. and Tierney, H. (2021). “Belief in Robust Temporal Passage (Probably) Does Not Explain Future-Bias.” Philosophical Studies. 10.1007/s11098-021-01748-4

Latham, A. J. Miller, K., Tarsney, C and Tierney, H. (2022). “Robust Passage Phenomenology Probably Does Not Explain Future-Bias” Synthese

Lee, R., Hoerl, C., Burns, P., Fernandes, A. S., O’Connor, P. A., and McCormack, T. (2020) Pain in the Past and Pleasure in the Future: The Development of Past–Future Preferences for Hedonic Goods. Cognitive Science, 44, e12887. <https://doi.org/10.1111/cogs.12887>

McGuire, M., Pratt, J., and Zeckhauser, R. (1991). Paying to Improve Your Chances: Gambling or Insurance? *Journal of Risk and Uncertainty* 4: 329-338.

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

Moller, D (2002). ‘Parfit on Pains, Pleasures, and the Time of their Occurrence’. Canadian Journal of Philosophy 32(1): 67–82.

Nichols, S. & Bruno, M. (2010). “Intuitions about personal identity: An empirical study”, *Philosophical Psychology*, 23:3, 293-312

Pannequin, F., Corcos, A., and Montmaequette, C. (2020). Are insurance and self-insurance substitutes? An experimental approach. *Journal of Economic Behavior & Organization* 180: 797-811.

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.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap

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

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

Van Fraassen, Bas. (1984). Belief and the Will. *The Journal of Philosophy* 81: 235–256.

1. Greene & Sullivan use ‘regret’ to refer to the preference that one had done otherwise (2015, 957–8). [↑](#footnote-ref-2)
2. Freddie is not near-biased. [↑](#footnote-ref-3)
3. [↑](#footnote-ref-4)
4. In the setup it is assumed that if both are possessed, they are simultaneously; over time all that changes is the perceived external situation, not the attitudes. [↑](#footnote-ref-5)
5. Of course, it’s also the case that sometimes even when people do judge that a combination of preferences is rationally impermissible, they still combine them. In the case of diachronic Dutch books, for instance, some agents might accept a future bet that they know will lead to a sure loss, because they think that the previous bets are ‘over and done with’ and their future-directed credences regarding the bet have not changed. (With thanks to BLINDED for pointing this out). So it certainly does not follow from the fact that people do combine rationally impermissible preferences, that they do not see (explicitly or otherwise) that they are not, jointly, rationally impermissible. [↑](#footnote-ref-6)
6. On a standard causal reading, the data might suggest that backtracking counterfactuals might tell you that if you adopted the presence you would be fine because you wouldn’t have the clashing ones! [↑](#footnote-ref-7)
7. According to other views of rationality, the fact that having certain preferences tends to make one worse off is not constitutive of their irrationality, but rather, it simply *evidence* for it: what makes those preferences irrational is some deeper feature of them, some internal incoherence in the preference itself. See Briggs (2010) for a view of this kind. Even of one takes this view, however, one might still think that the fact that certain preferences do make one worse off is evidence of their irrationality, and so one might still care whether being future-biased makes one worse off than being strongly risk averse. [↑](#footnote-ref-8)
8. See economics and experimental economics literature on insurance and prevention efforts (e.g., Ehrlich and Becker 1972; McGuire, Pratt & Zeckhauser 1991; Pannequin, Corcos & Montmarquette 2020). [↑](#footnote-ref-9)
9. See Caruso, Gilbert & Wilson 2008; Lee, Hoerl, Burns, Fernandes, O’Connor & McCormack 2020; Greene, Latham, Miller & Norton 2021a, 2021b, forthcoming. [↑](#footnote-ref-10)
10. This link is disabled during the refereeing process to avoid de-anonymisation. [↑](#footnote-ref-11)
11. Even if one thinks that the ‘tending’ here needs to be read in a modally robust manner. [↑](#footnote-ref-12)
12. Hedden (2015) and Harman (2009) both deny the principle. [↑](#footnote-ref-13)