

# WHY FUTURE-BIAS ISN'T RATIONALLY EVALUABLE

Callie K. Phillips

**Abstract:** Future-bias is preferring some lesser future good to a greater past good because it is in the future, or preferring some greater past pain to some lesser future pain because it is in the past. Most of us think that this bias is rational. I argue that no agents have future-biased preferences that are rationally evaluable—that is, evaluable as rational or irrational. Given certain plausible assumptions about rational evaluability, either we must find a new conception of future-bias that avoids the difficulties I raise, or we must conclude that future-biased preferences are not subject to rational evaluation.

Future-bias is typically thought to be rational, often *obviously* rational.<sup>1</sup> Parfit (1984) offers this now familiar example.

I am in some hospital to have some kind of surgery. Since this is completely safe, and always successful, I have no fears about the effects. The surgery may be brief, or it may instead take a long time. Because I have to cooperate with the surgeon, I cannot have anesthetics. I have had this surgery once before, and I can remember how painful it is. Under a new policy, because the operation is so painful, patients are now afterwards made to forget it. Some drug removes their memories of the last few hours.

I have just woken up. I cannot remember going to sleep. I ask my nurse if it has been decided when my operation is to be, and how long it must take. She says that she knows the facts about both me and another patient, but that she cannot remember which facts apply to whom. She can tell me only that the following is true. I may be the patient who had his operation yesterday. In that case, my operation was the longest ever performed, lasting ten hours. I may instead be the patient who is to have a short operation later today. It is either true that I did suffer for ten hours, or true that I shall suffer for one hour.

---

<sup>1</sup> Others have called this ‘bias against the past’ or ‘past discounting.’

I ask the nurse to find out which is true. While she is away, it is clear to me which I prefer to be true. If I learn that the first is true, I shall be greatly relieved. (165–166)

It seems rational to prefer to be in the first situation in such cases despite the fact that this preference involves preferring to be in a situation that has a greater total amount of pain for you in the overall course of your life. In fact, many philosophers take it to be obvious that this preference is rational and without need of defense. [Prior \(1959\)](#), [Craig \(1999\)](#), [Zimmerman \(2005\)](#), [Heathwood \(2008\)](#), [Hare \(2007, 2009\)](#), and [Kauppinen \(2018\)](#) all claim that this bias is at least rationally permissible, if not rationally required. However, a handful of philosophers—including [Moller \(2002\)](#), [Brink \(2011\)](#), [Sullivan and Greene \(2015\)](#), [Dougherty \(2015\)](#)—raise difficulties for defending the rationality of this kind of bias. In this article I present an argument that would mean that all these authors have got it wrong. Future-biased preferences are not even rationally evaluable—that is, evaluable as rational *or* irrational.

In determining whether an agent is future-biased, I will make use of the definitions offered by [Sullivan and Greene \(2015\)](#).

An agent *S* is [future-biased] with respect to pleasure iff for two exclusive experiences,  $E_1$  and  $E_2$ , where  $E_1$  is at least as pleasurable as  $E_2$ , *S* prefers  $E_2$  because it is a present or future pleasure rather than a past one. . . .

An agent *S* is [future-biased] with respect to pain iff for two exclusive experiences,  $E_1$  and  $E_2$ , where  $E_2$  is at most as painful as  $E_1$ , *S* prefers  $E_1$  because it is a past pain rather than a present or future one. (949)

I will call agents “future-biased” if they are future-biased with respect to pleasure, pain, or both. I will say that a preference is a future-biased preference if it is a preference that meets the description on the right-hand side of one of those biconditionals. In this article I argue that we must reject a plausible assumption about rational evaluability, find a new conception of future-bias that avoids the difficulties I raise, or conclude that future-biased preferences are not subject to rational evaluation.

The conclusion that there are no rationally evaluable future-biased preferences is surprising and counterintuitive, but it has a substantial theoretical advantage that makes this neglected view worthy of our attention. One who is willing to accept the conclusion that future-biased preferences are not rationally evaluable will be freed from the burden to explain the widely assumed *irrationality* of another very common sort of time-bias—namely, near-bias. How to discharge this burden has eluded philosophers since Parfit first made the case for the tension between these biases in *Reasons and Persons* (1984). Near-bias, roughly, is preferring less pleasurable experiences because they are nearer in temporal proximity, or, conversely,

preferring more painful experiences to less painful ones because those more painful experiences are further in temporal proximity.

A popular putative example of near-bias is found in the well-known Stanford marshmallow experiment. This study sought to measure the self-control of children through an experiment in which some children were told they could eat one marshmallow now or wait fifteen minutes and receive two marshmallows. The kids who didn't wait the fifteen minutes to receive the two marshmallows are said to be near-biased—they prefer a lesser good now to a greater good in the future.<sup>2</sup> The claim that near-bias is irrational seems to hardly warrant defense.<sup>3</sup> But Parfit (1984) argues convincingly that it is very difficult to defensibly claim *both* that near-bias is irrational and that future-bias is rational. In brief, the argument is that the way to argue that near-bias is irrational is to claim that a mere difference in when something occurs does not affect its value. But if we make this claim we should also think that future-bias is irrational. The fact that some experience is past shouldn't affect its value for us. Parfit considers possible replies to this argument, but none of them seem to withstand scrutiny. For example, you might think that future-bias is importantly different from near-bias because we can't change the past. If we can't change the past, it is reasonable for us to be less concerned about it.<sup>4</sup> However, we still seem to exhibit future-bias even if the future experience in question cannot be changed. If it were certain that you were going to be tortured later this afternoon, you would still be more concerned about this future painful experience than you would be if it were already in the past. Parfit considers and dismisses other possible defenses of an asymmetry in the rationality of near-bias and future-bias. In my view, these arguments have yet to be challenged successfully. Thus, it's important to find a way to explain or

---

<sup>2</sup> Mischel et al. 1972. It is worth noting that there is now doubt that the children in the study were near-biased in the sense that interests us (i.e., there's doubt that they had preferred the temporally nearer marshmallow eating experience simply because it was temporally nearer). Delaying gratification, waiting to receive two marshmallows had a lot to do with economic and social disadvantages (Watts et al. 2018). It seems children who didn't wait, often from resource poor environments, doubted that there would be marshmallows in the future, or at least assigned it lower probability. If a later outcome is lower probability than the nearer outcome, it will often be rational to choose the nearer. This highlights one difficulty with empirical investigation of near-bias and future-bias. It is difficult to create conditions that would clearly demonstrate a preference for one experience over the other based simply on the temporal location of those events (near or far, past or future) since it is difficult to create conditions where the likelihood of those experiences is the same.

<sup>3</sup> For philosophers and the casual observer at least. The irrationality of near-bias is more controversial among economists and psychologists who are more likely to hold that only structural constraints on preferences are relevant for whether preferences are rational.

<sup>4</sup> This strategy for explaining how future-bias is rational while near-bias is irrational has been pursued by Suhler and Callender (2012) and Dyke and Maclaurin (2002). Latham et al. (2021) investigate empirically the reasons for future-biased preferences and find that the belief that the past is practically irrelevant plays a role, but also find that even if agents supposed that they could causally affect the past, they still exhibit future-bias.

avoid the asymmetry of the rationality of these time-biases. If one accepts my argument that future-biased preferences are not rationally evaluable, there is no asymmetry to explain; this major obstacle for defending the irrationality of near-bias falls to the wayside. In [Section 6](#), I discuss this theoretical advantage and the implications of my argument in connection to near-bias in more depth.

The argument of this article shows that the position that there are no rationally evaluable future-biased preferences can be motivated and is more attractive than it initially appears. Even if the best response to my argument is to reject one of the premises, we'll learn that we must eschew plausible claims about preferences and rational evaluability—a lesson that has gone unappreciated.

## 1 Representation and Rational Criticism

Our definitions of future-bias say that being future-biased is a matter of having certain preferences. So if we criticize the rationality of future-bias, we criticize having certain preferences. Whether it makes sense to criticize particular preferences is controversial. For some, one's preferences are rationally suspect only if they collectively violate some logical or structural constraints on preferences. These constraints rule against preferences that would lead an agent to be Dutch-booked, for example.<sup>5</sup> Views like this tend to be accompanied by a behavior-based approach to preference ascription, and they are typically espoused by those working in decision theory and the theory of rational choice. On these views, preferences are ascribed to an agent on the basis of behavior, choices a third-party observer might describe them as having made. Below I put forward some claims about rational evaluability and argue that those who think future-bias is rationally evaluable are committed to them. By adding some cognitive conditions for the rational evaluability of preferences, these claims align with a *non*-behavior-based approach to preference ascription implicit in some of our ways of speaking about preferences. These approaches need not be exclusive, however. We can acknowledge that there are two senses in which we have preferences. There is a behavioral sense that doesn't come with any built-in cognitive conditions on having preferences, and another sense in which preferences are conscious mental states with certain representational contents that are linked up with beliefs and desires in certain ways. In the course of setting out my claims about the rational evaluability of preferences and their implications for future-bias, it will become clear that someone who thinks future-bias is rationally evaluable must think of such future-biased preferences in the second sense. *Prima facie*, it will seem

---

<sup>5</sup> An agent is Dutch-booked by having intransitive preferences. For example, they prefer *A* to *B*, *B* to *C*, and *C* to *A*. If they are willing to pay a small amount to trade in *A* for *C*, and *C* for *B*, and then *B* for *A*, they can be treated as a "money pump."

that the same must be true for near-bias as well, but as we will see later on, near-bias need not be understood in terms of the second, or non-behavioral, sense in which we have preferences. Let us now turn to rational evaluability. Following this discussion, I'll lay out the argument for the conclusion that there are no rationally evaluable future-biased preferences.

My first claim is that if some preference is rationally evaluable, it should in principle be possible to reflect on the alternatives that the preference concerns (the objects of the preference) and then determine their relative subjective value for us on that basis.<sup>6</sup> The agent can then form a preference for the alternative with greater subjective value. This does not need to involve any kind of decision procedure, or any determination of the subjective or objective probabilities of outcomes. We are just talking about determining a preference, regardless of what is likely to occur or whether any of our actions could bring about one or the other. I do not claim that we reflect on, or that we ought to reflect on, the alternatives in every case, or even in any case. I also do not claim that we must be able to assign a precise subjective value to each alternative independent of one another. Subjective value may be approximate and relative to an alternative. My claim is simply the following.

**Rational Evaluability Claim (REC):** If a preference between some alternatives is rationally evaluable, then we must be able, in principle, to reflect on those alternatives and determine the relative subjective value of the alternatives on the basis of that reflection.

I think this claim is highly plausible on its face, but let me expand on what I take it to involve and make it more precise. What does it take to have reflected on some alternatives, and determine their relative subjective value on the basis of that reflection? At a minimum, we must be able to say that the agent has mental states representing those alternatives, and that she represents them in a way that would allow her to determine their relative subjective value on the basis of that representation. In order to do that, I think an agent must imagine, or cognitively model, the alternatives in a way that captures the features potentially relevant to their subjective value for the agent. Differently put, the representations must disclose features of the alternatives that are the *motivational basis* for a preference between them. For example, for Parfit to be able to reflect on which surgery alternative he prefers to obtain in the case described above, he should imagine the relevant details of both scenarios. This need not be pictorial imagining—it is enough that the content of the representation includes, for the first alternative, that he's had a painful, ten-hour surgery. To represent his other alternative, the content of the representation should include his having a painful, one-hour

<sup>6</sup> By "subjective value" I mean roughly what others sometimes call 'expected utility.' I use the expression 'subjective value,' however, because I find that 'expected utility' strongly connotes a decision theoretic framework, and folks quickly forget that our focus is on determining preferences between alternative experiences, not preferences between possible actions.

surgery later. It's hard to see how Parfit could assign a subjective value to each alternative on the basis of his representations of them if he did not have representations of the alternatives with at least those contents. Think of this as a sort information condition for adequate representation of the objects of a preference.

For another example, consider a case in which my boss asks me to form a preference between watching a movie or scrubbing the floor. If I formed my preference for the former by imagining going to the zoo versus taking a nap, I clearly did not succeed in reflecting on the alternatives and forming a preference for watching a movie over scrubbing the floor via my representations. Now I'm not making any claims here about when or whether you need to adequately represent the objects of a preference for the preferences you form to be rationally evaluable. I'm just interested in the modest claim that you should be *able* to adequately represent the alternatives, the objects of the preference, in order for you to have a rationally evaluable preference between them. In addition, I claim that the adequate representation that you *should* be able to have requires having representations of the objects of the preference for *A* over *B* that disclose the motivational basis for a preference for *A* over *B*.

You might go further here and say that you should be able to adequately represent the alternatives, the objects of the preference, in order for you to have the preference *at all*, not just for it to be rationally evaluable. Imagine that someone asks you to form a preference for either sitting on a round square or standing on a round triangle. This is a case in which some would say you're not *able* to represent the objects of the preference. If that's right, you cannot imagine sitting or standing on impossible objects in a way that would allow you to determine the subjective value of those experiences. If you believe that you cannot represent impossible objects, then even if you grant that in *some* sense you can form a preference between sitting on a round square or standing on a round triangle and express it, it's not appropriate to say that the preference in and of itself is rational or irrational. How could it be? Sure, you've selected one or the other in saying "I prefer to sit on a round square," but you haven't even imagined the relevant alternatives in your mind. Perhaps it's rational or irrational to *attempt* to form that preference. Perhaps a preference between some stand-ins you might be imagining is rational or irrational. Maybe I've always loved the look of triangles, and I've imagined standing on a triangular neon sign with rounded points. But a preference between standing on neon signs of shapes with rounded corners is not a preference between standing on a round triangle or sitting on a round square. In light of this, you might insist that without being able to represent the objects of the preference for *A* over *B*, you really cannot be said to have a preference between *A* and *B at all*. For the purposes of this discussion, I will only make the weaker claim that you cannot be said to have a rationally evaluable preference for *A* over *B*.

in such a situation (again, assuming we cannot represent objects like round squares adequately).<sup>7</sup>

I opt for this weaker claim so as to avoid a confusion about the fact that there are the two senses of having a preference mentioned above: the behavioral and non-behavioral. In the behavioral sense, I think we *can* have a preference between sitting on a round square or standing on a round triangle. We do this by sincerely uttering things like “I prefer sitting on a round square or standing on a round triangle,” or circling the words ‘sitting on a round square’ on a (strange) written questionnaire about one’s preferences. (Never mind that you might not be able to demonstrate such a preference by actually sitting on round squares instead of standing on round triangles.) We *cannot*, however, have a preference between sitting on a round square and standing on a round triangle in the non-behavioral sense of having a preference in which preferences are mental states with certain representational contents that link up with beliefs and desires in certain ways (again, assuming we cannot represent objects like round squares adequately).

Being able to adequately represent the alternatives for a (non-behavioral) preference between *A* and *B* also sometimes calls for representing them in a particular way. When we are considering questions about prudential rationality, we are interested in what an individual prefers for herself. Preferring is a kind of pro-attitude, much like desiring. (Some even consider preferring and desiring to be the same.) Beliefs and pro-attitudes such as desires and preferences are often *de se*—that is, they are beliefs that about *me*, how *I* am, or desires (preferences) about what *I* experience, what happens to *me*. Arguably, when we limit ourselves to prudential rationality, all such preferences will be *de se*. At the very least, it’s plausible that any future-biased preference in the sense we’ve defined above will be *de se*, since it’s a matter of the agent preferring some experience over another for themselves.<sup>8</sup>

A couple of other sorts of distinctions go hand in hand with the *de se/de dicto* distinction. The distinction between imagining “from the inside”

---

<sup>7</sup> There is the question of whether imagining standing or sitting on rounded shapes could be representing the objects of preference *enough*; we of course do not want to require that one be able to represent the objects of the preference perfectly. Fortunately, we need not worry about the difficulty of specifying what exactly is enough for my argument to come. As I will explain further, all that matters is that the agent be able represent objects of the preference enough to motivate preferring one or the other.

<sup>8</sup> Some have written about what are sometimes called “other-regarding” or “third-person” (apparently) future-biased preferences—cases in which we prefer *someone else* have one experience over another because it is past (future) for them. Those preferences are moral preferences, or at least preferences that are outside the scope of prudential rationality, but even with these other-regarding preferences, it is plausible that we do or maybe ought to imagine the alternatives by considering the other person’s point of view in a first-person way “from the inside.” See Parfit 1984, 181, Brink 2011, 378–379, and Dougherty 2015, 3 for discussion of these kinds of preferences.

in a first-person way versus imagining “from the outside” is roughly the distinction between imagining experiencing (or doing) something, versus imagining that something is the case. This distinction lines up fairly neatly with the *de se/de dicto* distinction, which can be described as sort of a first-personal/third-personal distinction. Plausibly the representations involved in potentially future-biased preferences are going to be *de se* contents and involve first-personal representing of what the experiences are like “from the inside.” Recall that future-bias as defined above is a matter of an agent preferring  $A_2 > A_1$  with respect to pleasure or pain because one of these alternatives is past or future for them. The agent’s representations must disclose how pleasurable or painful the experiences are for the agent if they are to disclose the features of the alternatives relevant for determining the relative subjective value of the alternatives of a preference with respect to pleasure or pain. These considerations suggest that adequately representing the objects of a preference with respect to pleasure or pain involves *de se* representing them in a first-person way “from the inside.”<sup>9</sup> These considerations suggest the following.

**Necessary Conditions on Representation (CR):** Relative to a preference  $A_2 > A_1$  with respect to pleasure or pain, an agent  $x$  adequately represents an alternative,  $A_1$ , only if (1)  $x$  represents  $A_1$  so that all of the facts relevant to forming a preference  $A_2 > A_1$  are included in the representation of  $A_1$ , and (2)  $x$  represents the relevant features of  $A_1$  “from the inside.”<sup>10</sup>

We can now make my initial claim about preferences more precise.

**Precise Rational Evaluability Claim (PREC):** If a preference between some alternatives is rationally evaluable, then we must be able to represent those alternatives without violating CR, and determine

---

<sup>9</sup> You might think that even when it comes to forming preferences with respect to pleasurable and painful experience there are times when adequate representation of the alternatives does not require that we be able to consider those alternatives in a first-person way, and that future-biased preferences provide the perfect example: if an experience is in the past surely what that experience is like first-personally is irrelevant. After all, you won’t be reliving that experience; it is a *past* experience. However, I think when we compare possible past experiences (setting aside how they might impact how things are now and who we are) this thought is not compelling. Consider whether you prefer to have had a very painful illness or broken leg during your childhood versus a childhood otherwise the same but absent those experiences. I doubt anyone would prefer the former. Perhaps illness is character building in certain ways, but when we are considering preference with respect to pleasure and pain (not the non-hedonic, such as character), it’s clear that one of the alternatives is more preferable. If that’s right, what experiences are like first-personally must be relevant, even when it comes to preferences involving past experiences.

<sup>10</sup> I’m assuming here that we can think of  $A_1$  in this condition as the “same” alternative even if that alternative would be modeled by a different set of centered possible worlds when evaluated by different agents, or simply by some possible world when referring to the alternative third-personally. I’ll discuss the matter of centered-worlds later.

the relative subjective value of the alternatives on the basis of these representations.

The forgoing discussion shows that PREC is a plausible claim, but it's worth appreciating that those who think we have rationally evaluable future-biased preferences have another reason to accept CR. Recall our definition of future-bias. It tells us that an agent must prefer  $A_2$  to  $A_1$  *because* it is past (future). What is meant by 'because' in these definitions? The natural reading of this definition, and what seems to be intended by authors who have written on the subject of future-bias, is that an agent is future-biased when the *reason* she prefers  $A_2$  to  $A_1$  is that it is past (future). So it seems we should read 'because' in those definitions as concerning reasons.

One important distinction made regarding reasons is the distinction between normative and motivating reasons. While this distinction is usually made in the context of reasons for action, we can easily extend the idea to talk about reasons for preferring. Roughly, normative reasons would be those that justify the preference, and motivating reasons would be those that explain why the agent prefers what she does. Since our definition of future-bias is neutral on whether an experience being past or future is a reason to prefer it, we must be talking about motivating reasons as opposed to normative reasons. Motivating reasons are typically thought of as belief-desire complexes. If this is right, then preferring  $A_2$  to  $A_1$  *because* it is past (future) will involve having a belief (or set of beliefs) that involves a representation of  $A_2$  and  $A_1$ , and represents one of these as future, and one as past. And while it is true that one could have a thin (perhaps dispositional) notion of belief that requires little in the way of representation in order to be attributed with the belief, if it is supposed to contribute to the explanation of the agent's preference between  $A_2$  and  $A_1$ , then there must be some explanatory relation between the alternatives in question, and psychologically accessible content of the beliefs tied up in the agent's motivating reasons. This is because motivating reasons are supposed to *cause* the agent to have that preference. The beliefs involved in the agent's motivating reasons could only be causally efficacious in this way if they have representational content that is likely to be psychologically causally efficacious in forming that preference. Of course, agents could have beliefs that seem to have no reasonable connection in terms of content to other mental phenomena that they in fact cause, even when combined with desires or other mental states. However, in these cases the relationship is better described as a merely causal relationship, as opposed to a motivating relationship. If all of this is right, the agent must represent the alternatives in a way that discloses the motivational basis of a preference between  $A_2$  and  $A_1$  in the spirit of CR in order to have a future-biased preference at all given our definitions of future-bias. It also suggests that whenever we are

talking about future-biased preferences, we are talking about preferences in the non-behavioral sense.

Of course, intuitions will vary on exactly what (and how much) information is relevant to an agent's choice between some alternatives. But if my arguments above are right, philosophers who think that there are future-biased preferences must accept that temporal features of experiences need to be represented in the representations of the alternatives since being future-biased is a matter of preferring some past (future) experience to a future (past) experience *because* it is past (future). This is enough for my argument below.

In [Section 2](#), I will present a future-bias case study and show that an agent cannot form the future biased preference  $A_2 > A_1$  without violating CR. If CR cannot be met for future-biased preferences, we'll have all the ingredients for the argument that there are no rationally evaluable future-biased preferences. This argument can be distilled to the following:

- P1. If you cannot possibly represent  $A_1$  adequately to the preference  $A_2 > A_1$ , then  $A_2 > A_1$  is not rationally evaluable.
- P2. If you are representing the experience in  $A_1$  as present, then you are not representing the experience in  $A_1$  as past.
- P3. If  $A_2 > A_1$  is future-biased with respect to pleasure, then necessarily (if you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing the experience in  $A_1$  as past).
- P4. If  $A_2 > A_1$  is future-biased with respect to pleasure, then necessarily, (if you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing the experience in  $A_1$  as present).

---

C1. If  $A_2 > A_1$  is rationally evaluable, then  $A_2 > A_1$  is not a future-biased preference with respect to pleasure.

C2. If  $A_2 > A_1$  is future-biased with respect to pleasure, then  $A_2 > A_1$  is not rationally evaluable.<sup>11</sup>

P1 is a boiled-down expression of PREC. I take P2 to be a conceptual truth or very nearly one, in no need of defense but it will be addressed further in [Section 2](#) and [Section 3](#). The discussion of CR above may already be enough to convince you of P3 and P4. These premises focus on future-bias with respect to pleasure but a parallel argument can be given for future-bias with respect to pain. In essence, if  $A_1$  is a past experience and  $A_2 > A_1$  is a future-biased preference then P3 says that if you satisfy CR, and so represent  $A_1$  adequately, then you'll represent the experience in question as a *past* experience. This should be uncontroversial given that a future-biased preference is one where a less pleasurable future experience is preferred

---

<sup>11</sup> The necessity operator for premises and conclusions is omitted for simplicity. See the [Appendix](#) for a proof of the validity of this argument.

over a more pleasurable *past* experience (or a more painful past experience is preferred to a less painful future experience).

Similarly, if  $A_1$  is a past experience and  $A_2 > A_1$  is a future-biased preference, then P4 says that if you satisfy CR, and so represent  $A_1$  adequately, then you'll represent the experience in question as a *present* experience. Why think P4 is true? If  $A_2 > A_1$  is a preference that is future-biased with respect to pleasure (pain), then it is a preference motivated (in part) by the fact that the past experience would have been pleasurable (painful) to some extent for you. So to be motivated to form the future-biased preference  $A_2 > A_1$ , you must represent  $A_1$  in a way that allows you to evaluate the relative pleasurable (painful) of the experience for you. Evaluating phenomenal features of an experience such as their pleasure plausibly requires imagining the experience occurring "from the inside" and thus as a *present* experience. So imagining  $A_1$  adequately requires imagining the relevant experience as present. P4 along with P3 will be explored further by looking at the case study in [Section 2](#). I'll then address some possible objections to CR and make some remarks about what my argument means for our understanding of future-bias before closing with a discussion of what it might tell us about near-bias.

## 2 Case Study

**Vacation Case:** A magician approaches you and offers you a choice. He will either grant you a great and exceptionally enjoyable vacation located in your past, or a lesser but still pleasant vacation located in your future.<sup>12</sup>

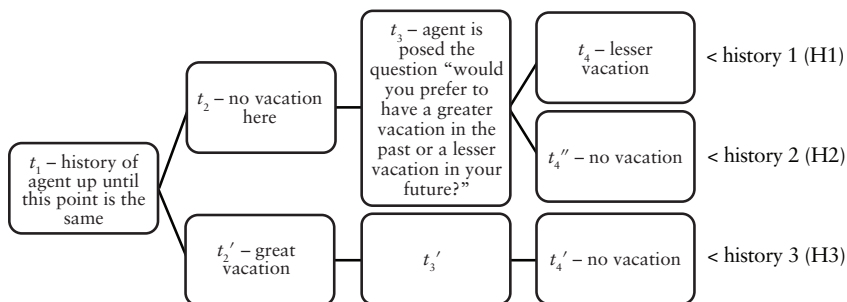
Which do you prefer? Most of us would prefer the future vacation, and for many this seems like a clearly rational preference.<sup>13</sup> It is worth noting that we need not think changing the past is possible in order to assess this type of scenario. We can construct cases like Parfit's surgery case (quoted above) where amnesia rules out that complication, but no changes to the past are involved (and no one needs to have magical power regarding the future). It is already settled what past events have occurred and, in Parfit's surgery case, what events will in fact occur given the agent's situation. Thus,

<sup>12</sup> This is a variation of a scenario described in [Sullivan and Greene 2015](#).

<sup>13</sup> New empirical work on future-bias suggests that people are future-biased with respect to both pain and pleasure (hedonic subjective value) when the future and past events are equal, and also when the alternatives are ten negative experiences in the past versus one negative experience in the future. When it comes to ten favorite meals in the past versus one favorite meal in the future most subjects did not show future-bias by preferring the one future favorite meal ([Greene et al. Forthcoming](#)). This seems to suggest that people are more future-biased with respect to pain or negative experiences than pleasures or positive experiences. However, I think it is important to reevaluate whether these kinds of studies in which subjects are asked to report their preferences between options with a grain of salt in light of my arguments, which show that representing the objects of these preferences adequately may not be possible. I discuss this further in [Section 5](#).

the matter of the necessity of the past is not crucially involved in generating the kinds of cases we want to consider.<sup>14</sup>

For now, let's consider this case in which the agent remembers her actual past, and knows she is located at a particular time, call it  $t_3$ . A branching time diagram will help us to keep in mind the various options as we discuss the case:



How should the agent represent and compare the alternatives for the purpose of assigning a subjective value to them? I will consider three strategies which appear to exhaust the available options.

The first strategy involves comparing the “histories” diagrammed above to one another—in other words, comparing the alternatives as sequences of events without thinking of oneself as located at a particular time in the sequences. With the second strategy, one does not abstract from one’s actual temporal location ( $t_3$ ) but rather attempts to conceive of the alternatives only from one’s actual temporal perspective. The third way is to compare being located at  $t_3$  with being located at  $t_3'$  by imagining having the one temporal perspective and then imagining having the perspective corresponding to the other time. Other apparent strategies will either reduce to these or rely on some controversial view about the metaphysics of time. Let us consider each of these strategies in turn. I’ll first informally say why they’re insufficient for representing the alternatives in a way that meets CR. In [Section 3](#), I’ll return to why the agent will fail to meet CR on each strategy given the formal objects we might use to model representational contents.

Our first strategy is for the agent to decide whether she would prefer a greater vacation in the past to a lesser vacation in the future by comparing history 1 (H1) to history 3 (H3) (refer to the diagram above) considered as histories, or sequences of life events. This involves bracketing, or disregarding, her actual temporal perspective (and any other temporal perspective) in evaluating each alternative.

When we reason this way, the preferable alternative (for the normal person who likes great vacations) is H3. Compared to H1, H3 affords the

<sup>14</sup> We also don’t need to take a stance on the metaphysics of time, though relational (B-theoretic, non-absolute) notions of past and future are used in my arguments. See [footnote 19](#) for more discussion.

greatest utility for the agent's life as a whole. But once she abandons her actual temporal perspective (the perspective of  $t_3$ ) and chooses one history over the other, she seems to no longer be choosing between a great past vacation on the one hand and a mediocre future vacation on the other. This strategy leaves out the pastness and future-ness of the events, a key aspect of the alternatives, and the aspect that makes having a preference between them a potential instance of future-biased preference. Since this strategy leaves out relevant features of the alternatives, it violates CR.

The second strategy is to refrain from abstracting from our actual temporal location at all. If the problem with the first strategy is that we were too quick to abstract from our actual temporal location, then the second strategy may look promising. Unfortunately, this won't work if one of the alternatives we need to represent involves having a different past. If the agent chooses not to stray from her temporal location at  $t_3$  in her representation of the alternatives, then she cannot represent  $t_2'$  as past because while located at  $t_3$ ,  $t_2$  is the past. This is just the obvious point that if we are going to represent the past as being different from how it actually is, then we can't also, as a part of that same representation, represent it as being as it actually is. This does not mean that we can't think of  $t_2'$  as a possible past, but it won't be in a sense that allows us to assess the subjective value of the alternative on the basis of that representation. When we represent alternatives with the aim of assessing how pleasurable or painful the experiences will be for us, we will try to imagine them as actual. With the diagram in view, the point is that if having a different past is possible, and we are to represent what it would be like for it to be *actually our past*—we will be imagining being located on another branch (so to speak) parallel to the one on which we are located. (There is no sense in which one stays located on the same branch while a segment of the branch is replaced. This replacing of some portion of our past is simply represented by the change to a different branch, since we take ourselves to be modeling time and possible histories in a way that respects the fixity of the past, i.e., the view that the past is settled and cannot be changed.)<sup>15</sup> Thus this strategy fails to help us meet the conditions of CR; we can't represent the alternatives without abstracting from our actual temporal location.

One last strategy is available. First, we assess the past vacation alternative from the temporal perspective of  $t_3'$ . Second, we assess the future vacation alternative from the temporal perspective of  $t_3$ , and then compare. *Prima facie*, this seems to be the way an agent *ought* to be representing the alternatives. It resolves the difficulty of the vacations losing their pastness or future-ness with respect to the agent (as with the first strategy), while

<sup>15</sup> I am taking a position on how to represent changing the past, which I won't defend in this article beyond suggesting that if representing having a different past is not represented by the change to a different branch but is instead represented by replacing a segment of the branch, it is just a way of denying that we could represent having a different past at all. Changing the diagram is not a way of representing something with a diagram.

also allowing an agent to represent having a different past (as was missing from the second strategy).

There's a tempting mistake we must avoid. When the agent tries to imagine each alternative from the corresponding temporal perspectives, she might try to do this by just imagining experiencing  $t_3$  or  $t_3'$  in each of these worlds as present. If this is all she does to represent the alternatives, she is just imagining beginning her experiences at  $t_3$  or  $t_3'$  with the memories of the past events of those times. The preference then threatens to transform into a different preference than the one under consideration. It's a preference between a vacation lying in wait to actually be experienced and a mere fuzzy memory. This won't do. The alternatives offered by the magician are between experiencing a great past vacation ( $A_1$ ) and experiencing a mediocre future vacation ( $A_2$ ) —not a fuzzy memory of a great vacation and a mediocre future vacation. You might well have a great vacation in your past even if you can't remember it later.<sup>16</sup> Call this *the memory mistake*. This violates condition (1) of CR.<sup>17</sup>

Being careful to avoid the memory mistake when representing the alternatives is essential if we hope to satisfy CR. But what will that take? The agent must somehow represent having actually experienced the past vacation of the first alternative. On an ordinary conception of times and events, when some event is past of you it was at some time future and present. This suggests that in order to satisfy both conditions of CR and avoid the memory mistake she must imagine that she actually experienced that past event by imagining it as present. When she does this, she has imagined not just being at  $t_3'$  presently, but also being at  $t_2'$  and experiencing  $t_2'$  as present. But when she does this, it starts to seem like she has simply just imagined the relevant sequence of life events, or the history H3. If this is the only way to adequately represent the being located at  $t_3$  or  $t_3'$ , we fail to represent the alternative in question because we have failed to do justice to the fact that a representation of the alternative must involve having a particular temporal perspective with certain events past (future) of the agent. This is true for any pair of alternatives that would feature in a future-biased preference according to our definitions. It's not enough that we imagine having experienced each of the relevant events in the sequence as present. This does not yield a representation of the alternative where the event is past (future), which is essential if the preference is to be a future-biased one

<sup>16</sup> The situation is analogous to Parfit's amnesia case quoted in the introduction.

<sup>17</sup> On some views of prudential rationality, the rational thing to do is to satisfy your present desires. If this sort of view is correct, one might respond that the memory mistake is no mistake at all. One could argue that only present desires are relevant to assessing what alternative it is rational to prefer. Thus, the only significance the past vacation experience has for assessing the alternatives is its impact on my present desires as a memory. To respond in this way is to lose sight of the target. If we hold this kind of view, it is true that it probably won't matter to the outcome of a deliberative process that the memory mistake is committed. Nonetheless, one still has not successfully represented the alternatives in a way that satisfies both conditions of CR.

according to our definitions. The trouble is that adequately representing the alternatives in accordance with CR demands imagining having experienced the relevant past events as present, while also imagining that those events are past and that some other time is the present. A coherent representation cannot represent a particular experience as both past (future) and present.

If all of this is right, it's not possible to represent the alternatives of the Vacation Case in a way that satisfies CR. In [Section 3](#), I'll argue for my conclusion more directly by helping myself to a possible worlds framework for modeling *de se* representational content. In this setting it is most transparent that this predicament is unavoidable. However, readers who are satisfied with the more intuitively presented reasoning of this section may skip it.

### 3 Alternatives and Centered Possible Worlds

The representation of the alternative of having a great past vacation ( $A_1$ ) clearly must represent the agent as having a great past vacation if it is to include the relevant facts and meet (1) of CR. Her representation should also be *de se*. A helpful way to model *de se* contents, first introduced by David Lewis (1979), is with centered-worlds.<sup>18</sup> For our purposes, a centered-world is an ordered pair of a world and a center, where the center is an ordered pair of an individual and a time. (It may help to think of the centered-world as the content of the representation.) In the centered-world (or set of centered-worlds) for  $A_1$ , the agent is at the time  $t_3'$ , and has the property <being such that I experienced a great vacation in the past>. This constitutes the basic representing of the "relevant facts" required by (1) of CR.

How exactly does one represent having that property? When one actually has a property like the property <being such that I experienced a great vacation in the past>, it is usually accompanied by the belief that one has this property. And normally having true beliefs about one's past is due, at least in part, to one's *remembering* the past experience. But  $A_1$  cannot be represented by the set of centered-worlds where the agent at a time ( $t_3'$ ) has a certain belief and memory of a great vacation. Representing  $A_1$  as a matter of having a certain memory gets the alternative wrong—it's perfectly consistent with  $A_1$ , as specified by the magician, that at  $t_3'$  in these centered-worlds you have the property <being such that I experienced a great vacation in the past>, though you don't remember it. So to represent  $A_1$  adequately I must imagine having this property without simply imagining having a memory of this vacation. (This is the memory mistake discussed

<sup>18</sup> I work with Lewis's way of modeling *de se* contents here because it nicely illustrates the difficulty of meeting CR. There are other ways to model *de se* contents. Some will still make precise the difficulty for meeting the CR. Perhaps others will not. Nonetheless, the intuitive problem sketched in [Section 2](#) will remain for anyone sympathetic to PREC and the idea that future-bias is rationally evaluable.

earlier.) As a result, imagining having some memory of a great vacation isn't going to help us meet (1) of CR. We must try a different tack.

First, notice that properties attributing past experiences to someone are relational properties. It is easy to imagine "from the inside" having intrinsic properties such as being six feet tall or even (temporally) present relational properties like standing next to the president. Properties that involve relations to times other than the centered-on time of a centered world cannot be represented first-personally in this way. This is because centered-worlds "zero-in" on a particular time—a particular temporal vantage point—at a possible world. And we noted that the representations of alternatives must incorporate this perspectival component not only to retain the relevant pastness or futureness of the events associated with each alternative, but also to capture the fact that these are *de se* representations—possibilities for *you*, if you are the agent being posed a question like the one from our case study. There is, then, *prima facie* a puzzle about how we are to represent being the bearer of these past-experience properties if one of the relevant relata (in this case an experience) is not located at the time specified by the set of centered-worlds under consideration.

With this in mind, another strategy for attempting to meet both (1) and (2) of CR is to have the agent first represent experiencing the great vacation as present, and then fast-forward, so to speak, to the time the centered-world is centered on. But just stating this option immediately reveals the problem with the suggestion. Imagining experiencing that great vacation as present picks out a *different* centered-world from the one where the great vacation is past; it is one where the agent is located at a different time. This centered-world shares the same possible world as an element but represents a distinct possibility because of the difference in times. (We might write the first as  $\langle w_1, t_2 \rangle$  and the second as  $\langle w_1, t_3 \rangle$ .) Another way to put it, to echo our discussion with the first and third strategies discussed in [Section 2](#), is that the "being past" component of the property is lost if you merely switch between these two representations, each representing some different experience as present. This in turn keeps us from meeting (1) of CR, since future-biased preferences always concern past (future) experiences.

One might be tempted to insist that past-experience properties need no special treatment, and that there is no special way to represent them other than attributing them to your counterfactual self. The agent does not need to represent *experiencing* the past experience in order to fully represent having the property of having some past experience: if I can represent myself to myself, in the actual world, as having the properties that relate to past experiences that I in fact have, there should be no special difficulty for counterfactual representations.

If PREC is true, this response doesn't work. Properties that attribute past experiences or future experiences to an agent must be representable in such a way that an agent may evaluate their subjective value. There can be no evaluating the significance of the property's contribution without

representing the defining aspects of the property. The property in question involves a relation, so the relata must be included in my representation if I am to represent it.

There are two ways that we can understand the past-vacation-experience relatum. We can either take “pastness” to be built into this relatum, perhaps as an intrinsic property, or we can think that ‘past’ describes the relation and the relatum is simply the vacation experience. Only the latter understanding is consistent with the relational notions of past and future we have been using in this discussion.<sup>19</sup> The property <having experienced a great vacation in the past> attributes to the agent a relation to a time (a time when the great vacation occurs) that is past of her. It does not attribute any temporal properties to the vacation or vacation experience itself. With this in mind, we see that the great vacation is a distinct element of the property. Since the property is to be self-attributed in a counterfactual *de se* content, and experiences are properly represented first-personally and “from the inside,” the agent will have to imagine the vacation experience as a present experience in her representation of this alternative if she is to comply with CR. But again, doing this means that the experience is not represented as past in her representation of the alternative. We might put it this way: the agent can merely succeed in imagining a centered-world where “the center” is just an individual, but not where it is an individual and time pair. But it is the fact that centered-worlds include an individual and time *pair* that makes them suitable for representing contents that crucially involve temporal perspective (as is needed for future-bias). We have now reached the impasse seen in our first look at the case study—we seem unable to represent the alternatives in accordance with CR.

So far we have only been discussing  $A_1$ , the great past vacation alternative.  $A_2$ , the mediocre future vacation alternative, involves a relation to another time too. Ultimately, I think the same considerations apply there as well, but I suspect that it is less intuitively compelling because our intuitions about the openness of the future interfere.

Two conclusions are available given the preceding discussion. Either (a), an agent may in some behavioral sense show a preference between the alternatives, but since she does not meet CR and PREC is true, it cannot be that she has a rationally evaluable future-biased preference; or (b) reject

<sup>19</sup> More explicitly, we have been using B-theoretic eternalist notions. I have assumed that an event is past (future) for an agent,  $x$ , if it is past (future) relative to  $x$  at some temporal location. A-theorists and B-theorists alike can allow use of these relational notions of past for  $x$  and future for  $x$ . A-theorists should allow these notions even if they think there are also (perhaps more fundamental) A-properties. It is important to see that we need not take a stand about the metaphysics of time in order to make use of the relational notions of past and future I use in this article. If conclusive arguments in favor of some versions of the A-theory were discovered, one may have the means to resist the arguments here. However, I think we should remain skeptical about the prospects of discovering such arguments. Without successful, independent arguments favoring A-theoretic views about the metaphysics of time, we should not appeal to these views in assessing the rational evaluability of time-biased preferences.

CR (and therefore PREC). In the next section, I will consider some possible reasons for rejecting CR.

#### 4 Objections to the Necessary Conditions on Representation

There are various objections one may have to CR. To start, one might object that there will be cases where we may want to leave out certain experiential aspects when representing an alternative. For example, suppose I am a long-time cigarette smoker but have recently quit. Not knowing that I have quit, a friend offers me a cigarette. Two alternatives present themselves: smoking the cigarette or not smoking the cigarette. When making my decision it may be appropriate for me to choose *not* to imagine how pleasurable it will be to experience smoking that cigarette. Instead, I will try to imagine how the other alternative involves me continuing my commitment to refrain from smoking, having fewer cravings in the future, etc.

These kinds of cases do not pose a problem for the account. Recall that PREC requires only that we are *able* to represent the alternatives in accordance with CR in order for a preference to be rationally evaluable, not that we must always do so. Recall also that CR concerns preferences with respect to pain or pleasure. When making a decision about how to act it will often be appropriate (maybe rationally required) that we bring to bear other considerations beyond what the experience will be like for us qualitatively or perhaps even exclude them. The former smoker's decision, for example, may be one sensitive to non-hedonic values or downstream effects.

Another worry is that taking these conditions seriously will require that we play out every detail of the state of affairs represented by that alternative in extreme detail in order to really represent all the relevant facts. If this isn't required, it is hard to see where to draw the line. We may have to decide this on a case-by-case basis, or perhaps it will be in part settled by some other rationality requirements. There is obviously no way to represent every detail of some alternative that could be described as relevant.

One way to partially allay this worry is to bear in mind that these alternatives will often be counterfactual states of affairs such that when we specify a set of alternatives, the specification includes the salient way in which the counterfactual state of affairs will be different from the actual world. Given that the set of worlds or centered-worlds that are picked out by that alternative are various, we need only represent the nearest one—that is, the one that is most like the actual world, differing only in respect to the specified features. So there will be a great deal that need not be considered since it will overlap with the actual world, or would not make a quantifiable difference to one's assessment of the subjective value of the alternative.

Although it is unlikely that there is any feasible way to eliminate the imprecision involved, this should not be seen as grounds for dismissing the

arguments of [Section 3](#) and [Section 4](#). As I argued in [Section 1](#), the features of the alternatives the arguments turn on are clearly relevant in virtue of what it is to have a future-biased preference at all.

## 5 Exhibiting Future-Bias

If I am right about the Vacation Case, the same difficulties will arise for any future-biased preference, and P3 and P4 of the argument presented at the end of [Section 1](#) are true. In Parfit's surgery case, it is evident that the same representational challenge arises. Even though the agent will not remember experiencing the ten-hour surgery, she must nonetheless represent the scenario as one where she in fact experiences ten hours of pain in her past. The past experience of a painful surgery is a salient feature of the alternative that must be included in order to represent that alternative successfully.

But why, if my argument is correct, does it seem so obvious that we are future-biased? I think there are a number of possible reasons. First, it's possible that we have implicitly (and mistakenly) thought that either first or the third strategies sufficed for representing the alternatives. (It seems unlikely to me that anyone ever thought the second strategy would work. I discuss the second strategy just to cover the logical space.) We are not in the habit of thinking very carefully about how we represent alternatives.

Second, as a possible sociological explanation, we should observe that much philosophical discussion of preferences has occurred in the context of decision theory and discussions of economics, where, as mentioned earlier, preferences are assigned to agents on the basis of their behavior, including verbal reports of having various preferences. There are several recent studies on the extent to which people exhibit future-bias.<sup>20</sup> They all proceed by collecting responses to questions about what people prefer for themselves among some alternatives. It might have been appropriate to say we have future-biased preferences in the behavioral sense of having a preference, if it weren't for the sort of argument given above that this conception of preferences is not compatible with the standard conception of future-bias as preferring some experience to another *because* it is past (future). Future-bias so understood calls for the non-behavioral sense of having a preference on which having a preference requires cognitively representing the objects of the preference. Agents reporting that they have some particular preference doesn't tell us whether they've represented the objects of the preference in question adequately for a future-biased preference. The distinction between these two ways of thinking about preferences seems to have gone unappreciated. There is evidence that we use both the behavioral and the

---

<sup>20</sup> See [Caruso et al. 2008](#) as well as the recent studies by [Greene et al. \(2021\)](#), [Greene et al. \(Forthcoming\)](#), and [Latham et al. \(2021\)](#).

non-behavioral understandings of preferences in ordinary discourse. There is not ordinarily any reason to tease them apart.

It's worth emphasizing that the best response to the argument I've given for the view that there are no rationally evaluable future-biased preferences may well be to reject PREC or CR. One might reasonably judge that the cost of rejecting the apparently obvious claim that we have rationally evaluable future-biased preferences is too high, however initially plausible PREC and CR may have seemed. But it's worth keeping in mind that denying that there are rationally evaluable future-biased preferences has theoretical advantages like the one discussed in the introduction. We would no longer have to reconcile the (apparent) rationality of future-bias with the (apparent) irrationality of near-bias, a problem that has yet to find an adequate solution. On this note, let us finally turn to how we should understand near-bias if we accept my argument for the conclusion that there are no rationally evaluable future-biased preferences.

## 6 Near-Bias

In *Reasons and Persons*, Parfit argues that all the ways of defending or arguing against the rationality of future-bias also apply in a similar way to near-bias. As a result, he thought there was no way to consistently hold that future-bias is rational while maintaining that near-bias is irrational. I have offered an argument for the conclusion that there cannot be any rationally evaluable future-biased preferences, but will it also follow from similar reasoning that there cannot be any near-biased preferences that are rationally evaluable? The answer is probably yes if we define it much like we've defined future-bias. A definition of near-bias along those lines would be this.

An agent *S* is near-biased iff for two candidate future experiences,  $E_1$  and  $E_2$ , where  $E_1$  and  $E_2$  are equally probable and  $E_2$  confers at least as much pleasure as  $E_1$ , *S* prefers  $E_1$  *because* it is closer in temporal proximity.

Whether it is possible to represent the alternatives in a way that satisfies CR on this definition will depend on what is involved with representing an experience as closer or farther in temporal proximity to another. But even if an argument parallel to the argument against the rational evaluability of future-biased preferences can be offered for near-biased preferences, it does not mean we must conclude that we don't procrastinate or make choices that postpone the inevitable to our own detriment. It shows only that there are no rationally evaluable near-biased preferences when near-bias is so defined. Unlike future-bias, near-bias is amenable to being defined in terms of our actions.

If we think that actions can exhibit near-bias in normal cases, and that acting to bring about an experience does not entail that the agent

adequately represents the alternatives (by the lights of CR) prior to acting, the following account is available.

An agent  $S$  is near-biased iff for two alternatives open to  $S$ ,  $E_1$  and  $E_2$ , where  $E_2$  is at least as pleasurable as  $E_1$ ,  $S$  acts to bring about  $E_1$  *because* it is closer in temporal proximity.

This definition circumvents the problems with representation that I have discussed in this article. (Of course, one may have other complaints about it depending on one's views about choice.) A definition parallel to this one cannot be given for future-bias, because there will not always be two *open* alternatives unless we reject the view that it is not possible to change the past. Granting that it is not possible to change the past, we can accept the argument against the rational evaluability of future-biased preferences and maintain not only that we have near-biased preferences, but that these preferences are irrational, as long as one holds a view whereby acting to bring something about can be rational or irrational, and accept a definition of near-bias akin to the one proposed here.

It's worth appreciating that this would not be an *ad hoc* move. Unlike future-biased preferences, it's possible for near-biased preferences to be action guiding, and they can be shown to violate structural constraints on preferences. We can explain the irrationality of near-bias by pointing out that a near-biased agent can be Dutch-booked—the near-biased agent will act in ways that make them worse off. No arguments of this sort apply to future-bias preferences.<sup>21</sup>

Moreover, understanding near-bias in terms of action comports better with psychology and neuroscience research. The literature suggests that near-biased choices are a function of interrelated neural systems. One is a system that is responsible for representing the subjective value of future events (Kable and Glimcher 2007). The second is responsible for the ability to delay gratifications through cognitive control (Makwana et al. 2013). The third is a system involved in episodic memory (Peters and Büchel 2010). If these apparently near-biased choices result from weak self-control, or limited ability to imagine or simulate future events, or beliefs about the probabilities of events, then what behaviorally speaking looks like near-biased preferences are not near-biased in the sense that philosophers often have in mind, and which we have been discussing here: preferring a nearer experience over a more distant experience, all other things equal, *because it is temporally nearer*. However, subjects in these studies do seem near-biased

<sup>21</sup> Dougherty (2011) argues that future-biased agents can be turned into “pain pumps” in cases where agents are also risk averse in a certain way, and thus that future-biased preferences are action guiding as well. However, I think there are reasons to be skeptical about the way risk aversion is applied in these cases. Discussing objections to Dougherty's argument would take me too far afield here. Regardless, arguments like this show only that in when accepting some other principles an agent will choose options that make them worse off. They are not action guiding in the sense that important here: they always feature alternatives that are open to the agent.

in some behavioral sense that can be captured by defining near-bias in terms of actions taken by the agent.<sup>22</sup>

More could be said about near-bias and its relation to future-bias, but we can see that there are promising ways to make sense of the irrationality of near-bias while endorsing PREC and CR. As I've noted, this is a desirable outcome as it allows us to give a principled explanation for why near-bias may be irrational while future-bias is not, something that should encourage us to take seriously the case for the view that there are no rationally evaluable future-biased preferences.

## 7 Appendix

Below is a proof of the validity of the argument at the end of [Section 1](#).

1. Possibly,  $A_2 > A_1$  is rationally evaluable AND future biased with respect to pleasure. [ @ ] (Assumption)
2.  $A_2 > A_1$  is rationally evaluable AND future biased with respect to pleasure. [ $w'$ ] (from 1)
3.  $A_2 > A_1$  is rationally evaluable. [ $w'$ ] (from 2)
4.  $A_2 > A_1$  is future biased with respect to pleasure. [ $w'$ ] (from 2)
5. Necessarily, if  $A_2 > A_1$  is rationally evaluable, then possibly you are representing  $E_1$  adequately to the preference  $A_2 > A_1$ . [ @ ] (P4)
6. If  $A_2 > A_1$  is rationally evaluable, then possibly you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ . [ $w'$ ] (from 5)
7. Possibly, you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ . [ $w'$ ] (from 3, 6)
8. You are representing  $A_1$  adequately to the preference  $A_2 > A_1$ . [ $w''$ ] (from 7)
9. Necessarily, (if  $A_2 > A_1$  is future-biased with respect to pleasure, then necessarily, (if you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as present)). [ @ ] (P1)
10. If  $A_2 > A_1$  is future-biased with respect to pleasure, then necessarily, (if you are representing  $E_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as present). [ $w'$ ] (from 9)
11. Necessarily, if you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as present. [ $w'$ ] (from 4, 10)
12. If you are representing  $E_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as present. [ $w''$ ] (from 11)

---

<sup>22</sup> Thanks to an anonymous referee for this point.

13. You are representing  $A_1$  as present. [ $w''$ ] (from 8, 12)
14. Necessarily, (if  $A_2 > A_1$  is future-biased with respect to pleasure, then necessarily (if you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as past)). [ $@$ ] (P3)
15. If  $A_2 > A_1$  is future-biased with respect to pleasure, then necessarily (if you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as past). [ $w'$ ] (from 14)
16. Necessarily, if you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as past. [ $w'$ ] (from 4, 15)
17. If you are representing  $A_1$  adequately to the preference  $A_2 > A_1$ , then you are representing  $A_1$  as past. [ $w''$ ] (from 16)
18. You are representing  $A_1$  as past. [ $w''$ ] (from 8, 17)
19. Necessarily, if you are representing  $A_1$  as present, then you are not representing  $A_1$  as past. [ $@$ ] (P2)
20. If you are representing  $A_1$  as present, then you are not representing  $A_1$  as past. [ $w''$ ] (from 19)
21. You are not representing  $A_1$  as past. [ $w''$ ] (from 13, 20)
22. Contradiction. (lines 18, 21)
23. Not: Possibly,  $A_2 > A_1$  is rationally evaluable AND future biased with respect to pleasure. [ $@$ ] (reductio)
24. Necessarily, it is not the case that  $A_2 > A_1$  is rationally evaluable AND future-biased with respect to pleasure.
25. Necessarily, if  $A_2 > A_1$  is rationally evaluable, then  $A_2 > A_1$  is not a future-biased preference with respect to pleasure.

Callie K. Phillips  
 College of Charleston  
 E-mail: [calliekp@gmail.com](mailto:calliekp@gmail.com)

#### References:

- Brink, David O. 2011. "Prospects for Temporal Neutrality." In *The Oxford Handbook of Philosophy of Time*, edited by Craig Callender. Oxford: Oxford University Press.
- Caruso, Eugene M., Timothy D. Wilson, and Daniel T. Gilbert. 2008. "A Wrinkle in Time: Asymmetric Valuation of Past and Future Events." *Psychological Science* 19 (8): 796–801. <https://doi.org/10.1111/j.1467-9280.2008.02159.x>.
- Craig, William Lane. 1999. "Tensed Time and Our Differential Experience of the Past and Future." *Southern Journal of Philosophy* 37 (4): 515–537. <https://doi.org/10.1111/j.2041-6962.1999.tb00880.x>.
- Dougherty, Tom. 2011. "On Whether to Prefer Pain to Pass." *Ethics* 121 (3): 521–537. <https://doi.org/10.1086/658896>.

**Acknowledgments** I would like to thank Rebecca Chan, Justin Dealy, Michael Rea, Margaret Schmitt, and Meghan Sullivan for advice and criticism on earlier drafts of this paper. I am also indebted to audiences at the Central Division Meeting of the APA in Denver, 2019, as well as the Time Bias and Future Planning Workshop in Vancouver, Canada, 2015.

- Dougherty, Tom. 2015. "Future-Bias and Practical Reason." *Philosophers' Imprint* 15 (30): 1–16.
- Dyke, Heather and James Maclaurin. 2002. "“Thank Goodness That’s Over”: The Evolutionary Story." *Ratio* 15 (3): 276–292. <https://doi.org/10.1111/1467-9329.00191>.
- Greene, Preston, Andrew J. Latham, Kristie Miller, and James Norton. 2021. "Hedonic and Non-Hedonic Bias toward the Future." *The Australasian Journal of Philosophy* 99 (1): 148–163. <https://doi.org/10.1080/00048402.2019.1703017>.
- Greene, Preston, Andrew J. Latham, Kristie Miller, and James Norton. Forthcoming. "On Preferring That Overall, Things Are Worse: Future-Bias and Unequal Payoffs." *Philosophy and Phenomenological Research*. First Published Online 2021. <https://doi.org/10.1111/phpr.12819>.
- Hare, Caspar. 2007. "Self-Bias, Time-Bias, and the Metaphysics of Self and Time." *Journal of Philosophy* 104 (7): 350–373. <https://doi.org/10.5840/jphil2007104717>.
- Hare, Caspar. 2009. *On Myself, and Other, Less Important Subjects*. Princeton, NJ: Princeton University Press.
- Heathwood, Chris. 2008. "Fitting Attitudes and Welfare." In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, Vol. 3, 47–73. Oxford: Oxford University Press.
- Kable, Joseph W. and Paul W. Glimcher. 2007. "The Neural Correlates of Subjective Value during Intertemporal Choice." *Nature Neuroscience* 10 (12): 1625–1633. <https://doi.org/10.1038/nn2007>.
- Kauppinen, Antti. 2018. "Agency, Experience, and Future Bias." *Thought: A Journal of Philosophy* 7 (4): 237–245. <https://doi.org/10.1002/tht3.392>.
- Latham, Andrew J., Kristie Miller, James Norton, and Christian Tarsney. 2021. "Future Bias in Action: Does the Past Matter More When You Can Affect It?" *Synthese* 198 (12): 11327–11349. <https://doi.org/10.1007/s11229-020-02791-0>.
- Lewis, David. 1979. "Attitudes De Dicto and De Se." *The Philosophical Review* 88 (4): 513–543. <https://doi.org/10.2307/2184843>.
- Makwana, Aidan, Silvia Maier, and Todd Hare. 2013. "Consider This: The Effects of Recommended Choice Feature Consideration on Brain Activity and Behavior During Consumption and Interpersonal Decisions." *European Advances in Consumer Research* 10: 40–43.
- Mischel, Walter, Ebbe B. Ebbesen, and Antonette Raskoff Zeiss. 1972. "Cognitive and Attentional Mechanisms in Delay of Gratification." *Journal of Personality and Social Psychology* 21 (2): 204–218. <https://doi.org/10.1037/h0032198>.
- Moller, Dan. 2002. "Parfit on Pains, Pleasures, and the Time of Their Occurrence." *Canadian Journal of Philosophy* 32 (1): 67–82. <https://doi.org/10.1080/00455091.2002.10716511>.
- Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
- Peters, Jan and Christian Büchel. 2010. "Episodic Future Thinking Reduces Reward Delay Discounting through an Enhancement of Prefrontal-Mediotemporal Interactions." *Neuron* 66 (1): 138–148. <https://doi.org/10.1016/j.neuron.2010.03.026>.
- Prior, Arthur. 1959. "Thank Goodness That’s Over." *Philosophy* 34 (128): 12–17. <https://doi.org/10.1017/S0031819100029685>.
- Suhler, Christopher and Craig Callender. 2012. "Thank Goodness That Argument Is Over: Explaining the Temporal Value Asymmetry." *Philosophers' Imprint* 12 (15): 1–16.
- Sullivan, Meghan and Preston Greene. 2015. "Against Time Bias." *Ethics* 125 (4): 947–970. <https://doi.org/10.1086/680910>.
- Watts, Tyler W., Greg J. Duncan, and Haonan Quan. 2018. "Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links between Early Delay of Gratification and Later Outcomes." *Psychological Science* 29 (7): 1159–1177. <https://doi.org/10.1177/0956797618761661>.
- Zimmerman, Dean W. 2005. "The A-Theory of Time, the B-Theory of Time, and ‘Taking Tense Seriously’." *Dialectica* 59 (4): 401–457. <https://doi.org/10.1111/j.1746-8361.2005.01041.x>.