Time Biases and Rationality

Wen Yu

The University of Sydney

Faculty of Arts and Social Sciences

Department of Philosophy

April 2025

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Wen Yu

Acknowledgements of Funding

This research reported in this thesis was supported by the award of a Faculty of Arts and Social Sciences Postgraduate Research Scholarship (SC3602).

Attribution Statement

Part of Chapter 1 is reproduced from my paper "On the rational evaluability of future-bias" (2024) published in *Inquiry: An Interdisciplinary Journal of Philosophy.* There are also cited works in experimental philosophy that I have co-authored.

Abstract

This dissertation offers a *partial* defense of the rationality of future-bias and near-bias.

Future-bias is the psychological phenomenon of preferring good things to be future rather than past, and bad things to be past rather than future. Near-bias, on the other hand, is preferring good things to be in the nearer future/past, and bad things to be in the further past/future.

An initial puzzle arises as future-bias often strikes us as sensible and rational, but near-bias is commonly deemed as criticizable and irrational. But both kinds of preferences manifest a sensitivity to mere differences in temporal locations, and thereby would seem to share the same rational status.

There are mainly three positions that have been taken in response to this puzzle. Time-neutralism is the view that both kinds of time-biases are irrational. Permissivism is the opposite view that both kinds of time-biases are at least sometimes rational. There is also the hybrid position, according to which future-bias is rational but near-bias is irrational. Many have justified the hybrid position by pointing out that the difference between pastness and futurity, unlike mere differences in temporal distances, is deep and normatively significant.

My own understanding of time-biases, however, has led me to a place barely occupied. I contend that time-biases had better not be thought of as being sensitive to temporal locations in *metaphysical time* – a conception that most occupants of the above three positions have presumed. Time-biases under the metaphysical-time conception do not seem to be the kind of preferences we actually have; nor are they worth defending.

Inspired by Derek Parfit's (1984) discussion on personal identity and egocentric concern, I reconceptualize time-biases as a kind of preferences that are (roughly) sensitive to what we may call *personal time*. Personal-time-biases seem to be what we actually have. And they are justifiable in terms of intra-personal bonds of concern – the inter-personal counterparts of which are also present in the special relationships that call for moral partiality (such as friendship). Somewhat surprisingly, however, the resulting picture of intra-personal partiality seems to fly in the face of

our pervasive pro-future-bias intuition. Indeed, it seems to indicate a hybrid position in the opposite direction – one that's near-bias-friendly but future-bias-unfriendly.

The other major theme of the thesis is ground clearing. Given the plethora of literature on time-biases from different domains of studies – including philosophy, psychology, and economics – there is the danger of talking past each other. Even among philosophers, different presumptions have been made regarding what time-biases are and how to go about evaluating their rational status. These background presumptions invite more general questions regarding the nature of preferences, the manifold of "rationality", the apparatuses of mental representation, and so on.

So, it's important to tease out the (potential) disagreements over those questions and see how they might affect approaches to the rationality of time-biases. Despite having my own commitments, I have also attempted to – with little prejudice – lay out the points of disagreement and how they interrelate. To some extent, I am drawing a map for theory choice concerning time-biases. The reader may approach with their own antecedent presumptions and see where they lead. They may end up in a position they find uncomfortable, in which case they may either accept the end result or question their starting point – or, of course, question the map itself.

Chapter Summary

Chapter 1 introduces the phenomenon of future-bias and near-bias, as well as the different senses in which they may be called rational or irrational. I pay special attention to how time-biases – and preferences in general – are understood differently in different domains of inquiry. For the purpose of the upcoming chapters, however, I settle on a conception of time-biases as a kind of preferences *qua* evaluative mental states. This mentalist conception serves "thicker" notions of rationality that philosophers in the literature of time-biases are mainly interested in.

Chapter 2 examines the purported "normative standard" of near-bias – the standard of exponential discounting – and its relation to future-bias. I argue that the normative standard is unmotivated. I also argue that it's unfruitful to criticize future-bias on the grounds of structural rationality.

Chapter 3 starts by examining the rationality of time-biases given the metaphysical-time conception, with an emphasis on future-bias. Although there are arguably some deep metaphysical

asymmetries between the past and the future, I contend that these asymmetries are not reason-giving. Given the metaphysical-time conception, future-bias is as arbitrary as near-bias. But I also bring good news for friends of time-biases: Our time-biases do not track metaphysical time. Instead, they track what we may call personal time. And personal-time-biases may not be that arbitrary.

Chapter 4 assumes the personal-time conception and plays defense against an argument against time-biases which hinges on the claim that time-biased agents fail to exhibit a principal concern for their lifetime wellbeing. I argue that the argument is not axiologically neutral: It rests on an assumption about lifetime wellbeing that subjectivists of wellbeing can reasonably reject. The chapter also touches on some other issues concerning wellbeing and time – for instance, how future-directed and past-directed desires may or may not be relevant to wellbeing.

Chapter 5 offers a positive justification for time-biases (given the personal-time conception). I argue that insofar as we can reasonably be partial to our dearest and nearest, a similar kind of intra-personal partiality should also be reasonable. This is because the grounds for inter-personal partiality in special relationships also obtain intra-personally (that is, obtain among different person-stages of single individuals). There is a somewhat surprising upshot, however: The resulting picture of intra-personal partiality, though having a near-biased shape, doesn't seem to have a future-biased shape. Our pro-future-bias intuition is not vindicated.

Acknowledgements

I am indebted to a large number of people. First and foremost, my supervisor Kristie Miller, without whom this thesis simply cannot exist.

Thanks are also due to Anthony Bigg, Patrick Dawson, Brigitte Everett, Jordan Lee-Tory, Rongting Li, Rasmus Pedersen, Asher Shang, Shira Yechimovitz, an anonymous referee from *Inquiry*, the audiences at the conference *Personal Persistence and Personal Ontology* and *the 8th & 9th Annual Conference of the International Association for the Philosophy of Time*. Their valuable comments have been formative of my arguments.

Finally, my deepest gratitude to my loving and supportive parents, and all the brilliant music I have been listening to when composing this thesis.

Content

1. Introduction: What Time-Biases Are and How to Evaluate Them

This thesis concerns the rational status of time-biases. By time-biases I mean *future-bias* and *near-bias*. To be future-biased, speaking roughly, is to have greater concern for things that lie in our future over those that have receded into the past. Parallelly, to be near-biased is to have greater concern for things that are temporally nearer over those that are more distant. Future-bias and near-bias are cousin attitudes that are sensitive to differences in temporal locations.

A lot has been written on the rationality of time-biases. Philosophers have reached more consensus about near-bias than with future-bias. Near-bias is commonly regarded as irrational. Henry Sidgwick's sentiment is still widely shared by contemporary philosophers:

> The mere difference of priority and posteriority in time is not a reasonable ground for having more regard to the consciousness of one moment than to that of another. The form in which it practically presents itself to most men is "that a smaller present good is not to be preferred to a greater future good". (Sidgwick 1907: 380-81)

Our everyday practice seems to corroborate Sidgwick's verdict. We are near-biased creatures. But we also sense that there is something off about this. We criticize people for indulging their present desires for alcohol and food at the expense of future health. We urge each other and ourselves to moderate expenses and save for the future. When I binge on Netflix late at night, knowing that I'll be groggy and unproductive tomorrow, I feel like a shitty person. Near-bias appears to be some sort of myopia of the mind, a rational defect, or a failure to take care of oneself.

When it comes to future-bias, philosophers are more divided. On the one hand, there seems good reason to think that future-bias is as irrational as near-bias. The difference between future and past, for one thing, seems likewise an instance of what Sidgwick calls "mere difference of priority and posteriority in time". A fully rational person, it seems, should have equal concern for all temporal parts of her life. But intuitions in favor of future-bias are recalcitrant. Past ordeals are over and done with, we often say, so that they no longer matter (as much). Indeed, we criticize people for not being able to leave the past behind. One who agonizes over past sufferings as much as

impending sufferings appears more pitiful than admirable. Savoring on past success as much as aspiring for future success seems more foolish than prudent.

Some apologetics of future-bias have attempted to vindicate these intuitive judgements. Insofar as they regard near-bias as irrational, they will argue that the parity between near-bias and future-bias is merely apparent, and that the difference between being past and future is of special significance. Still others in favor of future-bias are more pessimistic about the business of justifying it. It's been suggested that future-bias is a "brute fact about rationality" (Heathwood 2008: 61). Some even suspect that future-bias is beyond the veil of rationality, being something "that help[s] to define the type of rational creature we are rather than something that is itself up for rational assessment" (Scheffler 2021: 102).

In all, regarding the rationality of time-biases, there are three general positions that one may want to take. The first is *hybridism*. Being "commonsensical", it's the view that future-bias is rational but near-bias is irrational.[1] And there is *time-neutralism*, according to which near-bias and future-bias alike are irrational. Finally, *permissivism* is the view that both kinds of time-biases are rational.

I came to this topic deeply attracted to hybridism. This thesis, however, does not turn out to be a defense of hybridism. At the end of the day, I offer no conclusive verdict regarding which position should be taken. In particular, I find no good reason for believing in hybridism except for our entrenched intuitions. It seems an unstable position susceptible to encroachment from both time-neutralism and permissivism. But nor do I end up with a clear-cut case for time-neutralism or permissivism.

All this may sound disappointing to the reader. But not all is in vain. I take it that I have advanced the debate on several grounds. To anticipate a few, I rebut a recent argument to the effect that future-bias cannot be evaluated as rational or irrational (§1.4.2). I argue that one of the standard arguments for time-neutralism – *the Argument from Wellbeing* – is on far shakier grounds than it may appear (Ch.4). I also join some other philosophers in arguing that hybridists are ill-advised to build their case on the metaphysics of time (Ch.3). Finally, I suggest that a good case (though

[1] Following the terminologies of Latham, Miller, Oh, Shpall, and Yu (2024). Of course, there is also the barely occupied "hybridist" position according to which near-bias is rational but future-bias is irrational – more on which in Ch.5. I use the term here in the more popular sense.

not a perfect one) can be made for permissivism if one is willing to take on a revisionary conception of what time-biases are (Ch.5).

Part of what I am doing, then, is ground-clearing. As we shall see, the question "Are time-biases rational?" is multiply ambiguous. When philosophers approach this matter, they do not always have in mind the same conception of rationality – or even of what time-biases are. Sometimes they make it clear where they are coming from. But sometimes these antecedent presumptions are hidden or lost in translation. It would be beneficial, then, to tease out these presumptions – without prejudice – and gain clarity about where the disagreements and agreements really lie. And lots of non-trivial conditional claims shall emerge from such ground-clearing. So, I am to some extent drawing a map for theory choice. The reader may approach with their own antecedent presumptions and see where they lead. They may end up in a position they find uncomfortable, in which case they may either accept the end result or question their starting point – or, of course, question the map itself.

I now turn to the first pressing question: What *are* time-biases? (§1.1-1.3) I don't think there is *a* correct answer. But some answers are better suited than others for our purpose. I shall primarily focus on time-biases as preferences, where preferences are understood as comparative evaluative mental states. I begin with future-bias (§1.1-1.2) and then turn to near-bias (§1.3). By taking on board this conception of time-biases, we can then move on to the question: What does it take for time-biases to be rational or irrational? (§1.4)

1.1. Future-Bias

The phenomenon of future-bias, with its apparent rationality, is most famously illustrated by the following case by Derek Parfit:

> *My Future and Past Operations*: 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 co-operate with the surgeon, I cannot have anaesthetics. 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. 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*. (1984: 165-66, emphasis mine)

Parfit *prefers* that he is the patient who has had a ten-hour long operation rather than the patient who has a one-hour long operation awaiting. In other words, Parfit prefers a greater amount of pain in the past over a lesser amount of pain in the future.

I take this to be a paradigmatic case of future-bias. Much of the subsequent literature on future-bias revolves around this case. It tells a great deal about future-bias. First and foremost, future-bias is taken to be a kind of *preference*. Second, it's a kind of preference that is *sensitive to* the pastness and futurity of the alternatives. Third, it invites the reader to realize the existence and prevalence of future-bias. Parfit predicts that if you were in his predicament, you'd have the same preference. Finally, it also jogs our intuition that future-bias is rational. Most of us would find Parfit's preference perfectly sensible. Indeed, it may seem ludicrous to *not* have the preference.

The first two points speak directly to the question "What is future-bias?" I shall, for the purpose of this thesis, understand future-bias as a kind of *preference* that is past/future-sensitive. This is not a trivial claim as it may appear. As things stand, there are other phenomena that are sometimes lumped together with future-bias but are potentially distinct – namely, *temporal relief* and *the valuation asymmetry*. (Similarly for near-bias.) I prefer to pull these phenomena apart. And to do this, I shall take on some presumptions regarding what preferences are.

On my preferred conception of preferences, preferences are *mental states* that are *evaluative* and *comparative*.[1] It will become clearer what all this means after comparing future-bias with temporal relief and the valuation asymmetry. But let me put my cards on the table. First and foremost, I take preferences to be mental states with representational content. This *mentalist* conception is to be contrasted with the *behaviourist* conception which reduces preferences to choice behaviors or mere dispositions to choose. Secondly, by saying that they are evaluative and comparative, I mean that the formal object of preferences is betterness (or comparative value relations more generally) – in the sense that the formal object of beliefs is truth.[2] Preferences necessarily represent two (or more) alternatives. And to prefer *A* to *B* is to represent *A* as better than *B* (or more valuable than *B*) in some relevant respects.[3] I should also add (if not already implied by what's said) that preferences are not merely "urges and impulses, cravings and passions and itches" (Pettit 2006: 137). Preferences may be partially underwritten by impulses and itches, but one can prefer despite impulses and itches to the opposite.[4] Finally, I take that one normally has privileged introspective access to her own preferences. As such, self-report of preferences is reliable (though not infallible) evidence of preferences. (This is not to say choice behaviours and emotional reactions, etc., are not good evidence. See next paragraph.)

As a methodological sidenote, I do not intend this to be a conceptual analysis of our ordinary concept – though I think all the above is not in great tension with our ordinary conception. Nor do I claim this to be the only correct conception of preferences. It might be the case that there are

---

[1] Fehige & Wessels (1998); Pettit (2006); Hansson & Grüne-Yanoff (2022); Hausman (2012); cf. Broome (2006).

[2] This is reminiscent of the "guise of the good" analysis of desires. That is, I take preferences to be under the "guise of the better". That analysis of desires is unpopular, as it seems normal for one to desire something one takes to be bad. (Tenenbaum 2007, Schafer 2013) But it does not seem coherent to say that one prefers *A* to *B* but regards *A* to be worse than *B*. No matter what you think about desires (and how they relate to preferences), the "guise of the better" analysis of preferences seem apt.

[3] Or you might think that the formal objects of preference are normative rather than evaluative properties. Gregory (2021) holds that preferences just *are beliefs* about normative reasons. On that view, to prefer *A* to *B* is to believe that there is more normative reason that *A* obtains than that *B* obtains. For our purposes, we need not decide whether the formal objects of preferences are normative or evaluative. Nor do we have to decide whether preferences just are beliefs. I'm personally in favour of Mulligan's (2015: 176-78) suggestion that preferences are *sui generis* mental states non-reducible to beliefs (or any other kinds of mental states for that matter).

[4] Daniel Friedrich invites us to consider Pablo the pedophile:

> Pablo has pedophile desires but also believes that there is nothing good whatsoever about realizing his pedophile desires; he believes acting out his desires would be a moral abomination, would be prudentially unwise (as he would rot in hell), and wouldn't even give him any momentary pleasure (as he would instantly be struck by fear and shame). (2017: 69)

Here Pablo prefers not to be a pedophile – as he considers it a better state-of-affair if he refrains – despite desires or "itches" to the opposite.

multiple distinct concepts of preferences that are legitimate and useful, capturing something important about us and serving important theoretical roles (Sen 1982; cf. Hausman 2012). However, the mentalist conception explicated above is one that is normally adopted by philosophers who are interested in the rationality of preferences (and especially the rationality of time-biases). It's thus a conception that best serves our purposes.

One way to capture the gist of this conception is that preferences are partially belief-like and partially desire-like. They are belief-like in the sense that they aim to represent the world accurately and are at least to some extent evidence-sensitive. This is evident in our deliberation (Pettit 2006). When I deliberate about what to prefer, I try to figure out what is better, what bears the preponderance of reasons, or what is preferable. Upon discovering the gruesome facts about animal farming, or upon being convinced that animals have rights, I come to prefer a vegetarian lifestyle. On the other hand, preferences are also desire-like in that they have motivational potentials. This is not to say that preferences are defined in terms of dispositions to *act*. I prefer world peace rather than havoc. But I am not disposed to do anything. Also, the motivational profile of preferences is not restricted to actions. They may also motivate certain emotions: Upon learning that world peace has obtained, I feel glad. This open-minded view about how preferences may motivate (or fail to motivate) is especially important considering future-bias. The past is over and done with. Parfit is not motivated at all to make it the case that his future-bias is satisfied.[1] Nonetheless, we should expect his preference to motivate *something* – for instance, feeling nervous when checking with the nurse which patient he is.

So, preferences are mental states that represent comparative evaluations. We can now give a more precise definition of future-bias. Slightly modifying the definition by Preston Greene and Meghan Sullivan (2015: 498-99):

> **Future-bias:** An agent *A* is *positively* future-biased iff for two exclusive states-of-affairs S_1 and S_2, where S_1 is at least as valuable as S_2, *A* prefers S_2 over S_1 *at least in part because* S_2 is future but S_1 is past; An agent *A* is *negatively* future-biased iff for

[1] Of course, you might say that if Parfit (believed that he) had the choice, he would be motivated to make it the case that he already suffered – and likewise, that if I could bring out world peace with a finger click, I would do it. These hypotheticals may be true – so long as they are not taken to be *definitions* of preferences (cf. Binmore 1994, Guala 2019) – but they do not help to understand ordinary instances of future-bias.

> two exclusive states-of-affairs S_1 and S_2, where S_1 is at least as disvaluable as S_2, *A* prefers S_1 over S_2 *at least in part because* S_1 is past but S_2 is future.

Insofar as Parfit's self-report is accurate, he has a future-biased preference. That is, Parfit is (negatively) future-biased by preferring the state-of-affairs of his having had a ten-hour operation in the past over the state-of-affairs of his having a one-hour operation in the future.

Some clarifications about the definition. First, I take alternatives of preferences to be states-of-affairs that have temporal locations as constituents (which may or may not obtain) that can serve as representational content and are also bearers of values. Second, the saying "at least as (dis)valuable" in the definition should be read as concerning value comparisons *prior to* factoring temporal considerations – what we may call *unadulterated* value-assignments. One way to understand that the past operation is "at least as disvaluable as" the future one is to think of how Parfit would prefer if the alternatives were temporally co-located: He would prefer the shorter operation. And this just means that he regards the longer operation as *unadulteratedly* more disvaluable.[1]

Thirdly, it's very important to understand that Parfit prefers the past operation (at least in part) *because* it's past. Parfit's case is fairly contrived so as to make "all else equal". But one might prefer the past operation for a variety of other reasons if all else were not equal. For instance, one might be "future-biased" simply because she has an important job interview later today she cannot afford to miss. Or one might be "future-biased" simply because she thinks that the earlier the operation is done, the better the curative effects. We should call such agents *merely apparently* future-biased. They base their preferences not on the pastness and futurity of the alternatives *per se*, but other considerations merely contingently associated with temporal locations. (It's a far less interesting question if such preferences are rational.) Rather, when we talk about future-bias (and near-bias), we mean *genuine* future-bias that is at least in part based upon considerations of the alternatives

[1] What if Parfit is masochist and therefore prefers lesser future pain over greater past pain (assuming that what counts as pain is attitude-independent)? I'd say that he is positively future-biased since he positively values pain. On an alternative reading of the definition, unadulterated (dis)value is considered objective. On this reading (and the assumption that pain is objectively bad), Parfit the masochist is not future-biased. He is not negatively future-biased because he prefers what's of objective disvalue to be future; and he is not positively future-biased because pain is objectively bad. But intuitively, Parfit the masochist fits the bill of future-bias. It thus seems that unadulterated (dis)value should be understood as subjective.

being past or future.[1] And let's say that a future-biased preference is *pure* if and only if is solely and completely based upon considerations of pastness and futurity (in addition to considerations of unadulterated value, of course). Our everyday future-biased preferences, even if genuine, are unlikely to be pure. And perhaps pure time-biases are not the kind of thing we can actually have, even when reacting to carefully contrived scenarios. (More on this shortly.) That said, when we approach the matter of rationality, it is often more convenient to abstract away elements of impurity and focus on pure time-biases.

Fourthly, it's likewise important that Parfit so prefers because the past operation is *past*. Future-bias is *perspectival*. It's issued from a perspective from which the longer operation is over and done with and the shorter operation awaits to be experienced. Parfit does not merely prefer a longer operation (say) on Wednesday over a shorter one on Thursday. (He doesn't have any fixation on calendar dates as such.) Nor does he merely prefer an *earlier* operation to a *later* one. (He wouldn't prefer the longer operation when it's future though earlier). Rather, the *pastness* and *futurity* of the alternatives are essential to how his future-biased preference is represented.

Talking about pastness and futurity is a risky business given the debates in the metaphysics of time. I shall discuss the (ir)relevance of the metaphysics of time to the rationality of future-bias in Ch.3. But the perspectival-ness of future-bias, in the sense I am concerned here, does not discriminate between the A-theory and the B-theory. Both sides can agree that we can and often do have representational attitudes that are perspectival. The A-theorists may want to analyse such perspectival-ness in terms of representing the alternatives as having irreducible monadic A-properties (Prior 1959). The B-theorists may want to give a token-reflexive analysis of the representational content in terms of B-relations – an analysis with essential reference to *the indexical present* (cf. Mellor 1998). I remain neutral here regarding how exactly we represent perspectivally.[2] I am merely saying that future-bias is not preference concerning calendar date, or preference for something to be earlier or later.[3]

---

[1] It might be helpful to follow Lowry and Peterson (2011) and talk about "grounds" for preferences, i.e., "the set of considerations which lead to your having the preference, and which would justify the preference, were they considerations of the right kind" (2011: 493). We may distinguish a *partial* ground for a preference from a *complete* ground. For a pure time-bias, the difference in temporal locations constitute the complete ground. For a genuine but impure time-bias, the difference serves as only a partial ground.

[2] I argue in Ch.3 that time-biases are sensitive to personal-time rather than metaphysical-time. This subtle distinction does not matter for my present purpose.

[3] Earlier I said that I take alternatives of preferences to be states-of-affairs with temporal locations as constituents. This needs to be read loosely given the B-theory of time. The A-theorists say that such

This roughly concludes what I take future-bias to be. They are preferences understood as comparative evaluative mental states. More specifically, they are preferences over past and future alternatives, in virtue of their being past and future. Being future-biased, Parfit represents the past operation as less disvaluable relative to the future operation because of their being past and future respectively. In saying all this, I am non-committal about how we come to be future-biased (in an evolutionary or psychophysiological sense), though some hypotheses will be brought up when relevant.

1.2. Future-Bias, Temporal Relief, and the Valuation Asymmetry

Future-bias is a kind of past/future-asymmetric attitude. There are two other kinds of past/future-asymmetric attitudes [1] that look similar to future-bias and are often lumped together in philosophical discussions – namely, *temporal relief* and *the temporal valuation asymmetry* (*the TVA* for short). Following some recent publications[2], I caution against a unified treatment. We should not *presume*, to say the least, that these three phenomena are of the same nature and (therefore) share the same rational status. This is not to say that these phenomena are unrelated. They likely share certain psychological and evolutionary underpinnings. But that doesn't entail that they are one and the same or share the same rational status. For sure, evaluating the rational status of future-bias might well inform what we should make of the other two phenomena (or *vice versa*). It's just that it's rather unclear what the relation should be. In what follows, I compare future-bias with temporal relief and the TVA – but not merely for setting the latter two aside. The comparison also helps to clarify what future-bias is and how we should approach its rational status.

1.2.1. Temporal Relief

temporal locations are A-properties and that they can go directly into representational content. The B-theorists deny that states-of-affairs are "tensed". But they would nonetheless allow perspectival representations of such "tenseless" states-of-affairs. The more general point is just the Fregean insight that representational contents are more fine-grained than states-of-affairs.

[1] There are lots of other attitudes that may be called past/future-asymmetric. For example, we fear death (future non-existence) but do not regret not having been born earlier (past non-existence). Whether this is rational may have something to do with the rational status of future-bias (Brueckner & Fischer 1986). But it's a topic too involved to address here.

[2] Bacharach (2022), Hoerl (2015, 2022), Caruso, Latham & Miller (2024).

Temporal relief is the phenomenon that we feel relieved when and *only when* a negative experience is over. It's past/future-asymmetric: We are relieved when the negative experience is past but not when it's future (or present).

The phenomenon gained philosophical prominence due to Arthur Prior's (1959) "Thank goodness that's over" argument for the A-theory of time. Imagine that Prior had a root canal on June 15, 1954 and was relieved when the surgery was over:

> *Thank Goodness That's Over:* One says, e.g., "Thank goodness that's over!", and not only is this, when said, quite clear without any date appended, but it says something which it is impossible that any use of a tenseless copula with a date should convey. It certainly doesn't mean the same as, e.g., "Thank goodness the date of the conclusion of that thing is Friday, June 15, 1954", even if it be said then. (Nor, for that matter, does it mean "Thank goodness the conclusion of that thing is contemporaneous with this utterance". Why should anyone thank goodness for that?) (Prior 1959: 17)

Prior's point is that temporal relief only makes sense if it has A-theoretic content (or turns on a belief that has A-theoretic content). He then deduces that the A-theory is the correct metaphysics of time. The argument has been challenged by the B-theorists on various grounds.[1] Some B-theorists claim that A-theoretic content is inessential to temporal relief. *Perspectival* content – analysed in B-theoretic terms in conjunction with indexicality – suffices. Prior is relieved, they'd say, that the root canal is earlier than this very token of relief. Some other B-theorists concede that temporal relief has A-theoretic content but reject that this implies the existence of A-properties. Some A-theorists find these B-theoretic responses wanting.[2] Still others offer evolutionary explanations of temporal relief orthogonal to the metaphysics of time.[3]

This is not the place to adjudicate this debate (though I will discuss how the A-/B-theories relate to *future-bias* in Ch.3). What matters here is this: What is temporal relief? How does it relate to future-bias?

[1] MacBeath (1983), Mellor (1981, 1998).
[2] Zimmerman (2005), Pearson (2018).
[3] Maclaurin & Dyke (2002), Suhler & Callender (2012), Bacharach (2022), Hoerl (2022).

Undeniably, temporal relief is a kind of emotion. And it's an emotion that "turns specifically on the fact that a painful or otherwise unpleasant episode has actually taken place but has now ended" (Hoerl 2015: 220).[1]

This doesn't entail, however, that temporal relief is not subject to rational evaluation. (I'm at this stage using "rational" loosely.) While some might think that temporal relief is some "brute" emotion subject to only causal explanation but not justification[2], it's a more popular view that temporal relief has representational content and can be judged as "correct" or "fitting". And it does seem that temporal relief is intimately related to future-bias. Parfit himself talks about feeling relieved (1984: 66) when being told that his operation is over and done with; and we tend to judge his relief as fitting.[3] Not to mention that the two case studies read very similarly.

Unsurprisingly, temporal relief and future-bias tend to be lumped together in the literature.[4] It's a popular presumption that they are the same phenomenon, having the same rational status. And Prior is sometimes interpreted as arguing from the rationality of *future-bias* to the A-theory. But on the face of it, an emotion is not a preference. Preferences do not necessarily carry phenomenal qualities. We can deliberate over preferences but hardly over emotions. (Pettit 2006: 137-38) Emotions may be evaluative, but they are not comparative. Preferences have no valences (Mulligan 2015: 177), but temporal relief is a positive emotion (Hoerl 2022: 217).

What underlies the tendency of blending temporal relief with future-bias, I surmise, is this *satisfaction thesis*: Temporal relief is caused by, and an expression of, a future-biased preference coming to be satisfied.[5] That is, Prior has always had a future-biased preference for the root canal

[1] Temporal relief is distinct from what we may call counterfactual relief – the kind of relief we feel when things turn out not so bad as expected. Prior can be always in full knowledge of how things shall unfold for him (thus no counterfactual relief) but nonetheless have temporal relief.
[2] Mellor (1981); cf. Mellor (1998).
[3] Though note that Parfit's relief seems sitting in between temporal relief and counterfactual relief, or a combination of both. More on this soon.
[4] Craig (2000), Tarsney (2017), Greene & Sullivan (2015), Fernandes (2021), Maclaurin & Dyke (2002), Suhler & Callender (2012).
[5] Also pointed out by Bacharach (2022: 257). He then argues against the satisfaction thesis in detail. One reason is that if that were true, we should expect positive emotion to incrementally build up in Prior during the course of the root canal – as his future-bias is being gradually satisfied – rather than a "sudden flood of relief" at its conclusion (ibid.). I am unsure about this objection. It might be that Prior prefers the root canal to be "completely" past so that the preference cannot be satisfied "bits by bits". Also, it's plausible that Prior prefers the root canal to be past rather than *present* (as a limiting case of future-bias) so that he gets no satisfaction until its final conclusion.

to be past rather than future (or present), which only comes to be satisfied when it's over. And it's unsurprising that satisfaction of preferences should lead to some positive emotion.

There are reasons to doubt the satisfaction thesis. To begin with, it seems a fair assumption that temporal relief is a universal, or near universal, phenomenon. It's doubtful, however, whether future-bias is universal. Even restricted to negative hedonic experiences with equal payoff (e.g., equal amount of past pain *versus* future pain), empirical findings suggest that future-bias is far from universal (Green et al. 2021a, 2021b, 2022a). Or imagine some time-neutralist philosopher claiming that she's not future-biased at all. We should not expect her to lack temporal relief even if there's no reason to doubt her lack of future-bias. Even if there are people who lack temporal relief, that would seem more like some psychophysiological abnormality rather than due to a lack of future-bias.

And all this coheres nicely with a cluster of recent proposals[1] that temporal relief is less "sophisticated" than Prior presumes. On the proposal I favor (Bacharach 2022), temporal relief is rooted in experiences. Rather than being mediated by beliefs or preferences, it's directedly caused by the cessation of negative experiences. It's a kind of "deactivating" emotion that turns on the very fact that some negative stimuli have ceased to be active, and thereby necessarily contemporaneous with the cessation of the experience. Thus, it's also necessarily contemporaneous with the belief that the experience has ended, if there is such a belief. But temporal relief is not *mediated by* the belief. Rather, the cessation of the experience is the common cause of the relief and the belief. This means that future-bias is incidental to temporal relief. All that needs for Prior to be temporally relieved is that he has been undergoing a painful experience.[2] It's unnecessary for him to have mentally represented the temporal location of the experience either prior to or during the experience.

The details of the explanation may be negotiable. Nonetheless, the above considerations taken together suffice to make the satisfaction thesis suspicious. But let's suppose it's true that temporal relief necessarily turns on the satisfaction of future-bias. Still, a unified treatment regarding

[1] Bacharach (2022), Hoerl (2015, 2022); cf. Maclaurin & Dyke (2002), Suhler & Callender (2012).
[2] Nonetheless, it does not mean that temporal relief is some "brute" emotion that cannot be "justified" or "made intelligible". Bacharach (2022) and Hoerl (2022) hypothesize that temporal relief endows evolutionary advantage to sophisticated creatures capable of anticipating the future. The anticipation of temporal relief motivates us to put ourselves through negative experiences with positive outcomes (such as a root canal).

rational status seems presumptuous. Generally, preferences and emotions may be subject to different standards of evaluation. What makes temporal relief "correct", "fitting", or "rational" is a complicated matter that I cannot fully address here. But one thing you may wonder is why we should be *relieved* – rather than feeling some other positive emotion – when future-bias comes to be satisfied (Hoerl 2015: 221-22). More importantly, it seems apt to say that relief can be fitting when the preference is satisfied *regardless of* whether the preference itself is rational or not. At least, it doesn't seem incoherent to say that although I shouldn't be future-biased in the first place, it's nonetheless appropriate for me to feel relieved when this preference is satisfied. Rationality-wise, the relation between the two phenomena is flimsy to say the least.

That being said, the phenomenon of temporal relief may gesture to an *error theory* of future-bias. This is what I mean: I have taken future-bias to be a kind of comparative evaluative attitude that has perspectival representational content. And I have been assuming that we tend to be future-biased in this sense (though perhaps never purely). This assumption is based on philosophers' self-reports and empirical studies.[1] However, it might be suggested that future-bias as we have observed is just a special kind of temporal relief understood along the line of Bacharach (2022) explained above.[2] That is, if temporal relief directly turns on the cessation of negative experiences and needn't involve comparative representation of values, and if future-bias is just a kind of temporal relief, then future-bias cannot be the kind of preference I've been assuming. Hence the error theory: Our future-bias is *merely apparent*.

Parfit's case helps to illustrate the point. The case has gone a long way to making "all else equal". So do the empirical studies that follow suit. But there is a confounding factor that's not eliminated and perhaps in principle ineliminable. That is the *anticipation* of future experiences – which is in itself a kind of experience that can be pleasant or unpleasant.[3] When anticipating the potential future operation, Parfit is now in an unpleasant experiential state: He is fearful and anxious. It seems plausible, then, that his future-bias is in part based on an aversion to such a state of unpleasant anticipation he is *currently* in. If told that the operation is over and done with, he'd no longer anticipate it anxiously. All this so far does not amount to an error theory. It merely suggests

[1] Empirical studies suggest that at least in first-person hedonic scenarios (of which Parfit's case is an instance), future-bias is prevalent and robust (Greene et al. 2021a, 2021b, 2022a, Lee et al. 2020, Lee et al 2022) .I say more about other scenarios in Ch.3.
[2] Bacharach (2022), Hoerl (2022), and Caruso, Latham & Miller (2024) suggest the error theory.
[3] That anticipation (and retrospection) confers "external utilities" is well-observed in the literature on near-bias: Loewenstein (1987), Van Boven & Ashworth (2007), Harris (2010), Hardisty & Weber (2020).

that future-biased preferences in Parfitian scenarios may be more impure than they appear. But taking a step further, one might think that the unpleasantness of anticipation is *all* that's driving Parfit's (self-report of) future-bias. That's the error theory: Parfit's anticipation is parallel to Prior's root canal. Parfit comes to be relieved because he's no longer anticipating, just as Prior comes to be relieved because he's no longer in pain. And just as we might mistake Prior's relief for future-bias (as with the satisfaction thesis), we might mistake Parfit's relief for future-bias. For all we know, we are never *genuinely* future-biased.

I see the attraction of the error theory. I don't know how to (dis)prove it. But there is some *prima facie* reason against it. To start, I acknowledge that anticipation no doubt poses an empirical problem. When empirically testing future-bias, we'd like to know to what extent our future-bias is genuine. We don't want the unpleasantness of anticipation to systematically distort our data. To this end, David Brink suggests modifying the Parfitian scenario "so that it involves administration of a drug that blocks anticipation of future pain, much as the doctors induce amnesia to block recollection of the pain of the operation" (2011: 379). But that scenario verges on incoherence. It seems that in order for one to form a future-biased preference, one must be able to anticipate the future event in the first place. That is, anticipation seems necessary for adequately representing the alternatives. (More on adequate representation in §1.5.) So, we would be at a loss responding to Brink's scenario. More generally, if adequate representation must involve anticipation, it might be impossible to empirically isolate the pure from the impure. Thus, we should be cautious when making claims about how strongly we are (genuinely) future-biased.

The error theory, however, is a much more sweeping claim. And there are reasons against it. First, there is some empirical evidence suggesting that we are genuinely future-biased. It turns out that our preferences are sensitive to ratios of unadulterated (dis)value.[1] To illustrate, one who prefers 2 hours of past pain over 1 hour of future pain may no longer be future-biased when past pain increases to 10 hours. Such sensitivity is at least *prima facie* evidence that we make comparative evaluations about the past and future alternatives themselves.[2] Third-person data aside,

[1] Greene et al. (2021a, 2021b, 2022a), Lee et al. (2020), Lee et al. (2022).

[2] The error theory may strike again here: Such sensitivity is merely apparent – it can be all accounted by the (un)pleasantness of imagining the past alternative which is proportionate to the badness of the past alternative. I don't have a conclusive response. It's indeed plausible that imaginative retrospection plays a role – just as experiential memory plays a role (Lee et al. 2022). But I think that there is even lesser reason to think retrospection is all that's going on given that retrospection tends to be emotionally less intense than anticipation. (Van Boven & Ashworth 2007, Caruso et al. 2008)

introspection deserves some trust. The error theory says that we are systematically mistaken about what our "future-bias" is about. We think that we prefer pain to be past rather than future but all we in fact care about is how unpleasant our present anticipation is. That sounds possible but presumptuous. Philosophers who have studied this matter in depth have offered justification for future-bias. And the justification points to some features concerning pastness and futurity. [1] (It's a separate matter whether the justification is any good.) It's certainly true that we can be sometimes confused about our own mental states. But such a systematic delusion still seems implausible.

All that being said, for all we know, the error theory is true. Even if it's true, however, it does not mean that the rationality of future-bias becomes an irrelevant topic. Maybe we are not in fact future-biased. But perhaps we *should* be. Perhaps if certain metaphysical theories of time are true (Ch.3) or certain theories of personal identity are true (Ch.5), then we should be future-biased. Or perhaps we shouldn't. The normative question is important regardless.

A quick recap: The relation between future-bias and temporal relief may be much more tenuous than many have thought. The satisfaction thesis seems false. Even if it's true, settling the rational status of future-bias does not directly tell us what to make of temporal relief (or *vice versa*.) It's an interesting question how much temporal relief feeds into future-bias, but there is no compelling reason to believe in the error theory.

1.2.2. The Valuation Asymmetry

We now turn to the TVA (the valuation asymmetry)[2], the other phenomenon frequently lumped together with future-bias.[3] First recorded by Caruso, Gilbert and Wilson (2008), the TVA is the phenomenon that people accord greater value (in the form of monetary rewards and the like) to future events compared to equidistant past events that are otherwise identical. In one experiment, participants were invited to imagine that they had just returned from the vacation home of a friend *or* will visit the vacation home. They were then asked to pick a bottle of wine from a varying price range they see fit as a thank-you gift (which the friend will receive after the vacation in either case).

[1] There is also empirical evidence that folk justification for future-bias resembles philosophical justification (Latham, Li, Miller & Yu manuscript).
[2] Empirical studies of the TVA include Caruso et al. (2008), Burns et al. (2019), Halabi et al. (2022), Caruso, Latham & Miller (2024), Latham, Li, Miller & Yu (manuscript).
[3] Caruso et al. (2008), Suhler & Callender (2012), Tarsney (2017), Fernandes (2021) identify the TVA with future-bias or at least identify them as manifestations of the same mechanism.

It turned out that on a population level, the future vacation deserved a wine 37% more expensive compared to the past vacation.

There are two important features of the original Caruso et al. studies. The first is that the TVA was manifest (and robust) in between-person analyses (as above) but disappeared in within-person analyses.[1] Comparing their choices of wine when the participants had been confronted with *both* the past and the future vacations, their choices were in alignment. They take this pattern to suggest that "participants valued the future event more than the past event, but considered this asymmetry *irrational*" (2008: 798). The second is that the TVA was found to be fully mediated by an *affective asymmetry* in imagining future and past events. This suggests that we value future events more *because* we experience more affective arousal when imagining the future than the past.

On a phenomenological level, the TVA and future-bias are distinct. But it's easy to see why they tend to be identified. For a start, it seems that they can be both straightforwardly modelled as *temporal discounting*. In standard decision theories, preferences are represented by utility functions. The future-discounting model for near-bias (Ch.2) is one such instance: Near-bias is a matter of discounting the future relative to the present in the manner that the more distant future, the more (dis)value discounted. Similarly, one might want to represent future-bias with discounting functions. One might want to think of the disvalue accorded to the past operation by Parfit (as of now) as a function of three ingredients: (a) the disvalue that would be accorded if it were present; (b) how distant in the past the operation was; and (c) a discounting parameter. (Sullivan 2018: 77-82) Similarly for the disvalue accorded to the future operation in terms of a function of future-discounting. Then we can think of Parfit as discounting the past more than the future – indeed, to the extent that past pain with much greater unadulterated disvalue turns out to have lesser discounted disvalue.[2] The TVA can also be represented by temporal discounting. Insofar as monetary value is a stable and reliable indicator of other sorts of value, the TVA can be

[1] To be accurate, Caruso et al. (2008) did not do between- and within-person designs separately. They confronted participants with both scenarios in counterbalanced order and did two kinds of analyses.
[2] Sullivan (2022b), Suhler & Callender (2012). Notice that on this model, prospective near-bias, retrospective near-bias, and future-bias are represented at one go. Value peaks at the present (being undiscounted) and decreases in both directions. But this may not be an accurate representation of our actual pattern of preferences. This model assumes that our time-discounting is context-independent. But it's not difficult to imagine someone with the following pattern of preferences: On the one hand, she does not care about past pain at all (that is, she absolutely discounts the past) if *compared with* future pain. On the other hand, she cares equally about pain in the near and in the distant past – and she cares *a lot*. Thus, one might want to instead isolate future-bias from near-bias and represent future-bias as a matter of discounting the past (directly) *relative to the future*.

seen as a matter of discounting the past to a greater extent than the equidistant future. And as we are only concerned about comparing the past with the future here, we may simply say that both phenomena are a matter of discounting the past *relative to the future – past-discounting* for short.[1]

Moreover, the affective asymmetry found to mediate the TVA has also been thought to underpin past-discounting. According to this popular cluster of hypotheses, the past is discounted more relative to the future because the affective asymmetry has evolved as an advantageous heuristic. The full story is involved, but here is the gist.[2] The future is in general much more controllable than the past. Thus, it would be advantageous for us to be much more motivated to care about and pursue future goals rather than to cry over spilled milk. Thus, an affective asymmetry – the tendency to experience greater affective arousal when entertaining the future relative to the past – may have evolved as a heuristic for valuing attitudes. That is, we tend to take affective arousal as an indicator of value and therefore accord greater value to the future relative to the past.[3] But the heuristic is, after all, a heuristic. It overgeneralizes to scenarios where control is irrelevant, as in Parfit's future-bias and the TVA. Taken together, it seems that the TVA and future-bias are different manifestations of temporal discounting, resulting from a "misfiring" heuristic. Call this *the counterpart thesis*: The TVA and future-bias are counterpart manifestations of the very same mechanism.

But I think that the relation between the TVA and future-bias may not be so neat. Let us begin with a puzzle. As mentioned earlier, Caruso et al. (2008) find the TVA to be manifest in between-person analyses but absent in within-person analyses. They take this to indicate that the TVA is regarded as *irrational* (even by those who display the TVA). Regarding future-bias, however, philosophers predict entrenched and widespread intuition that it is rational. Such an asymmetry in *perceived rationality*, if real, would put great pressure on the counterpart thesis. If the TVA and future-bias are manifestation variants of the same underlying mechanism, we should presumably expect their perceived rational statuses to align. Indeed, there is empirical evidence of such an

[1] See also Caruso, Latham & Miller (2024: 6). They point out that if we *define* the TVA as a matter of temporal discounting, then it may turn out a conceptual truth that those who have the TVA are future-biased. I opt for a more "superficial" definition of the TVA here (following Hoerl 2022).
[2] The direction asymmetry (that the future feels closer than the equidistant past), the uncertainty asymmetry (that we know less about the future), and perhaps some other asymmetries may also feed into the affective asymmetry or more directly lead to past-discounting. See Maclaurin & Dyke (2002), Suhler & Callender (2012), Fernandes (2022); cf. Bacharach (2022), Caruso, Latham & Miller (2024).
[3] This relates to what I said about anticipation and retrospection earlier. The affective asymmetry is in effect an asymmetry between anticipation and retrospection. There I suggest that apart from explaining the "pure" bit of future-bias, the affective asymmetry can also be responsible for the "impure" bit.

asymmetry. Latham, Li, Miller and Yu (manuscript) *directly* probed perceived rational statuses. We found that participants had a strong tendency to judge future-bias to be rational (indeed, not only permissible but also *obligatory*) but also a strong tendency to judge the TVA to be irrational.

What should we make of this? Given the inevitable impurity of actual future-bias discussed earlier, it might be suggested that the counterpart thesis is only to some extent true, but in a way that's consistent with the asymmetry in perceived rationality. Recall temporal relief. I've hypothesized that present states of anticipation may always to some extent drive (apparent) future-bias. Assuming temporal relief is not a matter of temporal discounting, then, our actual future-bias – as probed by Parfitian scenarios – may always contain a considerable dose of impurity. That is, the counterpart thesis may be in part true of our actual future-bias, which can be seen as a blend of temporal discounting and temporal relief.[1] This is a quite reasonable hypothesis. But it's also a potential symmetry-breaker of the perceived rational statuses of future-bias and the TVA. As we know, our intuitions are often vague. They lack specified content that's introspectively transparent. (McGahhey & Van Leeuwen 2018) And insofar as future-bias is an amalgamation of past-discounting and temporal relief, it might be that our intuition that future-bias is rational latches onto not past-discounting but temporal relief. [2] In other words, the perceived rationality asymmetry may be in fact between temporal relief and the TVA, rather than between (pure) future-bias and the TVA.[3]

Here is a different explanation of the perceived rationality asymmetry that likewise preserves the counterpart thesis *to some extent*. This time it's our intuition regarding the TVA rather than future-bias that misleads. Consider an analogy. Lots of us have egocentric preferences (never mind if we should). I prefer misfortune to befall some stranger rather than myself. I am an other-discounter, as it were. But suppose that I'm asked a question about adequate compensation. Someone is an innocent victim of a car accident. Despite my egocentric preference, I'm inclined to judge it as *inappropriate* to accord lesser monetary compensation should the victim be some stranger rather than myself. My judgement about compensation does not reflect my other-discounting because questions about compensation, reward, etc. "invoke a host of moral intuitions about fairness and desert" (Sullivan 2022a: 107) and perhaps also reflect ingrained social and legal conventions.

[1] See also Bacharach (2022).
[2] It's much less plausible that the TVA is to a substantive extent underwritten by temporal relief.
[3] This does not directly address why we (presumably) also judge *positive* future-bias to be rational. But perhaps a similar story can be told about the pleasantness of positive anticipation.

Perhaps something similar is going on with the TVA (in within-person analyses and direct normative judgements). Perhaps, despite that we do regard past-discounting as rational, we do not regard it morally fair or socially appropriate to show gratitude with an inferior wine just because the past vacation is of lesser value.

Either way, the perceived rationality asymmetry does not throw out the counterpart thesis. It merely indicates that actual future-bias is considerably impure (which is unsurprising) or that the TVA can be masked by moral and social considerations. There is some plausibility in both explanations. But there are other reasons against the counterpart thesis. In particular, we have good reason to suspect that the underlying mechanisms of the TVA and future-bias are much more variegated and diverse – and that they don't overlap much. If so, then they are not counterpart manifestations in any substantive sense and moreover, it would be presumptuous to think that they are rationally on par (even when they are pure and undistracted by extraneous considerations).

There is firstly a purely empirical observation. Contra Caruso et al. (2008), subsequent studies failed to observe a robust TVA (in within-person analyses) and moreover, failed to see the TVA being mediated by the affective asymmetry.[1] Empirical findings of future-bias, however, have been largely consistent.[2] There are multiple possibilities regarding what's going on here. In light of the above points about moral and social considerations and in preservation of the counterpart thesis, one might suggest that the TVA is often masked or distorted in empirical settings. There are just too many factors that might weigh into considerations of monetary reward or compensation.[3] But it doesn't mean that the TVA isn't there, or that it's not in fact driven by the affective asymmetry.

We might need more empirical data to see if this hypothesis is plausible. But I also suspect that *future-bias* far from being fully explained by the affective asymmetry (in conjunction with temporal relief). Let's begin by noticing that future-bias is, so to speak, a within-person phenomenon. Recall Caruso et al.'s (2008) inference from the lack of within-person TVA to its

[1] Burns et al. (2019), Halabi et al. (2022), Caruso, Latham & Miller (2024), Latham, Li, Miller & Yu (manuscript). Burns et al. (2019) also find developmental disassociation among the TVA, future-bias, and the affective asymmetry.
[2] Greene et al. (2021a, 2021b, 2022a), Greene, Holcombe, Latham, Miller & Norton (2021), Lee et al. (2020), Lee et al. (2022), Miller et al. (2022).
[3] More in this regard, Sullivan (2022a: 107) suggests that "prudential rationality would seem to dictate asking for the maximum compensation rate one is capable of negotiating in any of these cases – you should not dwell on the suffering of the overtime but rather your boss's capacity to pay".

perceived irrationality. Their thought seems to be that since the affective asymmetry is a quick-and-dirty heuristic vulnerable to misfiring, we are unreflectively dispositioned to exhibit the TVA. But this disposition can be counteracted on reflection (as shown in within-person analyses). Prompted by the juxtaposition of both the past and the future scenarios that are otherwise identical, we come to be under the (perceived) rational pressure to override what the affective asymmetry suggests.[1] While this may be inaccurate about the TVA given what's just said about moral and social considerations, I take their general point: Overt comparison of alternatives tends to lead to more deliberate and considered judgements (though it's a further question if the more deliberate must be more rational). This suggests that future-bias, as a kind of comparative evaluation, is probably more deliberate and considered than merely taking at face value an evolved heuristic. It's thus plausible that, especially in "concocted" scenarios where one can clearly see that "all else are equal", our future-bias is (largely) underwritten by considerations other than the affective asymmetry. There is some empirical evidence that in future-bias, we are not just letting the heuristic overgeneralize unhinged. Latham et al. (2020) and Greene et al. (2022c) found that future-bias is sensitive to the (perceived) controllability of the past and the future. Also, Latham, Li, Miller and Yu (manuscript) presented participants with philosophical arguments for and against future-bias and then probed their normative judgements. Participants generally agreed with those arguments for future-bias. This indicates that the perceived rationality of future-bias survives reflection, which in turn suggests that future-bias is not merely a product of the affective asymmetry. (Not to mention philosophers who have painstakingly reflected on this matter and arrived at the considered judgement that future-bias is rational.) These empirical findings echo the more general point that preferences are not merely "urges and impulses, cravings and passions and itches" (Pettit 2006: 137). These things can feed into preferences. One may prefer to satisfy her cravings. But preferences are not just slaves of passions. We can and often do deliberate over preferences – aiming at the formal object of comparative value.

To be clear, this is not to say that the affective asymmetry is irrelevant to future-bias. It's just unlikely the full story. Although I don't have a full story to tell about future-bias (or the TVA)[2], the

[1] See also Latham, Li, Miller & Yu (manuscript) and Fernandes (2021: 4007).

[2] The existing evidence regarding future-bias is not very informative. That future-bias is (a bit) less pronounced in third-person and non-hedonic conditions suggests that the affective asymmetry may play *some* role (as first-person and hedonic states-of-affairs tend to evoke stronger emotional responses and therefore magnify the affective asymmetry). On the other hand, there is evidence that future-bias does *not* rest on beliefs in the A-theory of time or the phenomenology of robust passage (Latham et al. 2022). But a positive explanation would be more forthcoming, I surmise, if I'm correct that future-bias tracks not metaphysical-time but personal-time (Ch.3).

take-home message is that it's rather unclear how future-bias relates to the TVA. We have just seen reasons to question the counterpart thesis (most prominently the asymmetry in perceived rationality). For all we know, the underlying mechanisms of the TVA and future-bias may not considerably overlap. And it would be unwarranted and premature to presume that they share the same rational status.

1.3. Near-Bias

In addition to pastness and futurity, our preferences also seem to be sensitive to temporal distances. That is, we tend to be near-biased. Near-bias can be future-directed (prospective) or past-directed (retrospective). To be *prospectively* near-biased, roughly, is to prefer good things to be sooner rather than later and bad things later rather than sooner. Prospective near-bias has been extensively studied in psychology, sociology, and economics[1] – usually under the label "future-discounting" or "delay-discounting". To be *retrospectively* near-biased, on the other hand, is to prefer good things to be in the nearer past and bad things to be in the more distant past. This kind of near-bias has gathered relatively little attention.[2]

For the purpose of this thesis, I define near-bias in a similar manner as future-bias. I take near-bias to be a kind of comparative evaluation that can motivate, but not necessarily manifests in, choice behaviors. More precisely, I take prospective near-bias to be:

> **Prospective Near-bias:** An agent A is *positively prospectively* near-biased iff for two exclusive *future* states-of-affairs S_1 and S_2, where S_1 is at least as valuable as S_2, A prefers S_2 over S_1 at least in part because S_2 is *in the nearer future than* S_1; An agent A is *negatively prospectively* near-biased iff for two exclusive *future* states-of-affairs S_1 and S_2, where S_1 is at least as disvaluable as S_2, A prefers S_1 over S_2 at least in part because S_1 is *in the more distant future than* S_2.

Retrospective near-bias receives a similar definition.

[1] For overviews see Read (2004), Frederick et al. (2002), Soman (2005), Grüne-Yanoff (2015).
[2] But see Yi et al. (2006), Molouki et al. (2019), Greene, Holcombe, Latham, Miller & Norton (2021).

Just like future-bias, near-bias is essentially *perspectival*. (Hedden 2015: 53; cf. Callender 2022) When I prefer the root canal to be later rather than *sooner*, I do not merely prefer the root canal to be later rather than *earlier*. If that were what I prefer, you would expect me to also prefer the root canal to be in the *nearer past*, and also prefer it to be *future* rather than *past*. But that cannot be what I prefer, insofar as I am also, expectedly, retrospectively near-biased and future-biased. Near-bias (prospective or retrospective) is therefore perspectival: A near-biased agent represents the temporal profiles of her alternatives in terms of *directed* temporal *distances* relative to the *present* – for example, as in the nearer future (in the case of prospective near-bias). Again, I remain neutral regarding how exactly such perspectival-ness is or should be represented – in terms of A-properties or B-relations.

We also need to distinguish between merely apparent, impure, and pure near-bias. I may prefer the root canal to be scheduled next week rather than tomorrow morning *solely* because I have a speech tomorrow afternoon and I'd rather not ruin it with a swollen mouth. Here my near-bias is *merely apparent*. Or I might be genuinely but *impurely* near-biased. I might have that preference partially because of the upcoming speech and partially because of the difference in temporal distance. Or I might have a *pure* near-bias: All I care about is near-future *versus* distant-future.[1]

Suppose that my merely apparent near-bias is perfectly rational. And suppose that my pure near-bias is irrational (as lots of philosophers would say). What should we say about my impure near-bias? Well, you might think it's *to some extent* rational (if graded rationality makes sense) because it's a mixture of rationality and irrationality, as it were. Or you might think it's plainly irrational. We can proceed without deciding on this issue by just focusing on *pure* near-bias.[2] The same goes for future-bias.

[1] These distinctions can be easily captured by utility functions. Considerations such as a swollen mouth go into unadulterated utilities. Uncertainty goes into the "expected" part of expected utilities. Considerations of temporal distances *per se* are captured by the distance parameter in conjunction with the discounting parameter of future-discounting functions. But be careful that since the literature tends to focus exclusively on prospective near-bias, the futurity of the temporal locations is often supressed in models of future-discounting.

[2] Another way to proceed is to focus on not preferences but *grounds* for preferences. (Lowry & Peterson 2011, Greene 2021; cf. Callender 2021b) A ground for a preference is some consideration that motivates the preference. In my impure near-bias, swollen mouth constitutes a *partial* ground and temporal distances constitute another *partial* ground. The question to focus on is whether temporal distances constitute *a* legitimate ground for near-bias – however impure or pure.

By adopting this *mentalist* definition of near-bias, I'm probably departing from how most economists and psychologists think of near-bias. Instead of conceptualizing preferences as evaluative mental states with representational content, lots of economists and psychologists focus on choice behaviours. In fact, they tend to take choice behaviours to be somehow constitutive of preferences. In its crudest form, preferences are *identified with* (actual) choice behaviours. My preference for a marinara over a margarita just *is* the choice for marinara over margarita that I've made. This *behaviourist* conception[1] goes hand-in-hand with their (almost) exclusive focus on prospective near-bias.

There are general objections to behaviourism and I'm pretty convinced by them.[2] But this is not a debate to be had here. We may grant that behaviourism is a legitimate conception. The problem is that it's unhelpful here. Given crude behaviourism, there is no future-bias or retrospective near-bias. But we want to talk about them. There are softened versions of behaviourism that may allow us to ascribe near-bias and retrospective near-bias. For instance, Kenneth Binmore (1994) identifies preferences with hypothetical choices and Francesco Guala (2019) identifies preferences with dispositions to choose.[3] So, if Parfit would choose the past operation if he were given the choice, or if he is dispositioned to choose the past operation, Parfit may be future-biased. But even if such ascriptions of preferences make sense[4], our evaluation of time-biases would be unduly restricted. To anticipate, some philosophers think preferences can be *correct or incorrect* in the sense that beliefs can be true or false. It's a matter of accurately *representing* the world. Mentalism can of course allow for such evaluation. But hypothetical choices or dispositions are not correct or incorrect in this sense. This is not to say that mentalism commits one to such evaluation. You might be a "consequentialist" and think that *all* that matters is the consequences of choice behaviours. Mentalism can accommodate this as well: As preferences motivate choice behaviours, they are criticisable on consequentialist grounds when they result in choice behaviours.

---

[1] Behaviourism is not one unified position but a cluster of interrelated thoughts. (Pettit 2006, Dietrich & List 2016, Hausman 2012) In addition to "ontological" behaviourism just mentioned, there is also "methodological" behaviourism, which roughly says that choice behaviours are the only *evidence* of preferences. Methodological behaviourism does not imply that there is no future-bias or retrospective near-bias. It's just that we lack evidence of them. Methodological behaviourism is also false, I'm inclined to think. I take that we have introspective access to preferences and that self-reports provide (defeasible) evidence.

[2] Pettit (2006), Dietrich & List (2016), Hausman (2012).

[3] These two versions may be the same if dispositions are analysed in terms of hypothetical choices.

[4] I am unsure. Does the hypothetical choice have to be effective in the sense of making it case that the past is a certain way, or does a mere declaration of choice suffice?

We will take a closer look at different approaches to rational evaluation shortly (§1.4.1). For now, a brief overview of the empirical findings on future-discounting – with a few cautions in light of the diverging conceptions of preferences.

For a starter, there is a substantive divergence between what economists take to be "the normative standard" of future-discounting and how people actually discount the future. According to the normative standard, one discounts exponentially and retains a constant discounting rate over time.[1] Being an exponential discounter, one's comparative evaluation of future alternatives is sensitive to only the temporal distance *between* the alternatives (and also unadulterated value). She doesn't care about how much these alternatives are delayed from *the present*. As such, insofar as the exponential discounter does not change her discount rate over time, her preference does not change over time. She is "diachronically consistent".[2] But we are not exponential discounters. Our actual future-discounting behaviours are more accurately modelled by hyperbolic discounting functions.[3] With positive states-of-affairs, hyperbolic discounting is characterized by increasing impatience as the alternatives approach in time. A hyperbolic discounter tends to be diachronically inconsistent. She may find it worthwhile to wait for an extra day in exchange for a larger reward when both rewards are distant, but later "reverses" her preference when the rewards become more immediate. This is because she is sensitive to how distant the alternatives are relative to *the present* over and above the temporal distance between the alternatives.

I have much more to say about the normative standard of future-discounting in Ch.2. But the point here is that if we are not behaviourists, we need not assume that all reversals in *choice behaviours* are reversals in *preferences*. This is because *counter-preference* choice behaviours are possible for mentalists who take preferences to be comparative evaluations.[4] Given mentalism, a certain choice behaviour for *A* rather than *B* does not necessarily imply a corresponding preference for *A* over *B*. In particular, I take there to be a genuine difference between *succumbing to temptation* and *acting*

[1] Strotz (1955), Frederick et al. (2002); cf. Callender (2021a, 2022), Hedden (2015: 50-55).
[2] There are combinations of changing discounting-rates and non-exponential discounting that can honour diachronic consistency. But exponential discounting is the only non-gerrymandered solution. (Rasmussen 2008, Hedden 2015: 50-55, Callender 2022) Also notice that the normative standard here is better read as having incorporated perspectival-ness, for it would otherwise prescribe – among other things – inflating the past for exponential discounters with a non-zero discounting rate. More on the normative standard in Ch.2.
[3] Frederick et al. (2002), Loewenstein & O'Donoghue (2002).
[4] They are certainly impossible given crude behaviourism. (Thaler 1980, Ostapiuk 2022)

upon changed preferences.[1] There is a genuine difference in motivation, for example, between the visceral urge for a second drink writ large and a considered change of mind that a second drink is worthwhile after all. This point is echoed by some economists dissatisfied with the standard model of future-discounting. They point out that some cross-temporal reversal in choice behaviours[2] are results of "visceral influences such as hunger, sexual desire, physical pain, cravings" heightened by temporal immediacy which shall not be blended into functions of future-discounting (Frederick et al. 2002: 372). Alternative models to hyperbolic discounting have been proposed to account for such reversals in choice behaviours.

To clarify, the point here is *not* that we are, after all, *not* hyperbolic discounters (let alone that we are exponential discounters). The point is that we should not *presume* choice behaviours to be always accurately reflecting preferences.[3] This is worth keeping in mind since the empirical studies on future-discounting tend to take a behaviourist approach.

Apart from this potential dissociation between choice behaviours and preferences, there is another reason to caution against reading off near-biased preferences directly from empirically established discounting functions. As several authors have observed[4], the discounting functions calculated from observed choice behaviours may in fact – to a worrisome extent – reflect considerations or motivations other than sensitivity to temporal distances. Part of the problem lies in that some researchers are just not interested in isolating the "pure" from the "impure". But even when it comes to studies that have painstakingly attempted to isolate temporal locations from other factors[5], caution is still warranted. As we have seen from the previous discussions on future-bias, it might be the case that some confounding factors (e.g., external utilities conferred by states of anticipation) just cannot be controlled away. Indeed, empirical studies on near-bias have yielded widely divergent discounting rates, which suggest a consistent failure to isolate pure near-bias.[6] Then there is also the deeper worry: For all we know, we are never genuinely near-biased.[7] This is the error theory of near-bias, motivated by similar considerations as the error theory of future-bias.

---

[1] See also Pettigrew (2020: 165-67), Thoma (2018). Cf. Hausman (2012) who has a mentalist conception but take preferences to incorporate the totality of motivational forces.
[2] One may think they are not genuine choice behaviours because they are in some sense "unfree". (Some think weakness of will or akrasia is conceptually impossible.) Regardless, the point is that they tend to be *deemed as* choice behaviours that reflect preferences.
[3] Self-reports are likewise only defeasible evidence, as I have suggested about future-bias.
[4] Frederick et al. (2002), Read (2004), Callender (2021a, 2022), Greene (2021), Latham et al. (2021).
[5] Loewenstein (1987), van Boven & Ashworth (2007), Hardisty & Weber (2020).
[6] See Frederick et al. (2002: 377-94) for a detailed discussion of methodological difficulties.
[7] See also Callender (2021a), Latham et al. (2021).

Space forbids an adequate evaluation, but I should confess that I find the error theory more attractive here than there. When I reflect upon my apparent near-biases, it seems that other sorts of considerations – uncertainty, instrumental values, and so on – are always salient. I can further excuse myself if succumbing to temptation is distinct from temporary change of preference. But perhaps I'm deluding myself. Perhaps all this is *post hoc* rationalization. I don't have a conclusive answer. In any event, if the error theory is true (though it's a big "if"), this may be a welcome revelation. If genuine near-bias is irrational (as many would say), it would be good news that we not systematically making mistakes.

1.4. Rationality and Adequate Representation

When philosophers and authors from with varied interests talk about time-biases being rational or irrational, they may have in mind a variety of senses of rationality. Here I offer a brief (and slightly biased) overview of the different dimensions of rational evaluation one might take when approaching time-biases (§1.4.1). Distinguishing these different senses of "rationality" is important for the purpose of the following chapters – namely, evaluating the rational status of time-biases. Then I turn to a recent argument (Phillips 2021) that threatens to bankrupt much of our project (§1.4.2). It's argued that there cannot be rationally evaluable future-bias, given a mentalist conception of preferences. I show how the argument fails.

1.4.1. Rationality: Reasons, Oughts, and So Forth

For the rest of the thesis, I shall assume a mentalist conception of preferences unless indicated otherwise. Preferences are considered as representational mental states with the formal object of betterness or value comparison. For Parfit to prefer the past operation over the future one is for him to represent the former alternative as comparatively better. Preferences are partially belief-like and partially desire-like in the sense of being evidence-sensitive but also motivating. By virtue of having the future-biased preference, Parfit is motivated to (say) check with the nurse and shall consider it good news that he's already suffered. If Parfit were convinced by the time-neutralist arguments, he might not so prefer.

Given this mentalist conception, how might one go about criticising Parfit's preference? Let us begin with a distinction between *rationality* and *correctness*. As the terms are usually used in the

literature, rationality is an *internal* evaluation sensitive to the mental states (especially epistemic states) of the subjects, whereas correctness is *external*. In other words, rationality is subjective whereas correctness is objective. (We may alternatively speak of subjective rationality *versus* objective rationality. But that's a bit cumbersome.) A glass of petrol is cleverly disguised as gin and tonic. You (reasonably) believe that it's gin and tonic and take a sip. (Williams 1979) You are being *rational* by taking a sip; so the story goes. But it's *incorrect* for you to take a sip. In other words, you *ought not* to take a sip.[1] Correctness and rationality are connected. You ought not to take a sip because the *normative reason* out there counts against taking a sip: Petrol is by all accounts bad for you. But you are being rational because your action is correctly responding to your *apparent reason*. You have a (reasonable) belief that it's gin and tonic. What you believe, if it were true, *would* be a normative reason for you to take a sip. Gin and tonic is really good for you (let's assume). Your action is rational because it would be correct if what you (reasonably) believe were in fact true.[2]

Somewhat confusingly, when philosophers talk about the "rational status" of time-biases, they often have in mind correctness instead of rationality. Their terminologies are somewhat deviant.[3] And I apologize for having perpetuated this confusion. In the opening paragraphs, time-neutralism, hybridism, and permissivism were defined in terms of "rationality". But as I understand time-neutralism, it is the position that we *ought not* to be time-biased. That is, time-neutralism concerns normative reasons for or against time-biases – independent from what we happen to believe (about time). The focus on correctness is most transparent in the context of the metaphysics of time (Ch.3). For example, the claim that future-bias would be "irrational" if the B-theory is true is obviously saying that if the B-theory is true, the preponderance of normative reason "out there" would count against future-bias. It does not concern if we believe in the B-theory or not. But the carelessness is understandable in this context. As I suggested, it remains

[1] I tie "oughts" with correctness, though some might prefer distinguishing two kinds of oughts (subjective and objective). (cf. Schroeder 2018: 290).
[2] A few clarifications: (1) The term "apparent reason" comes from Parfit (2011). Some might prefer the label "subjective reason". (2) One can *have* an apparent reason *R* for *A* and also perform *A* without being motivated by *R*. In this case, *R* is not a *motivating reason*. Her *A-ing* may fail to be rational for failing to be correctly *responding to R*. (3) The characterization of apparent reason here is crude and unsatisfactory (Vogelstein 2012), though it suffices for our purposes. Among other things, it's up to debate whether your belief that it's gin and tonic needs to be *reasonable* (whatever that means) in order for your taking a sip to be rational. Suppose it turns out that our future-bias rests upon A-theoretic beliefs that are unreasonable. It would then be up to debate if our future-bias is rational.
[3] Errol Lord observes that "nearly all views about rationality in metaethics and a prominent group of views about rationality in epistemology" make the common assumption that "rationality solely depends on internal features of the agent" (2018: 5).

rather unclear why we are future-biased. The inquiry into (subjective and internal) rationality thus does not seem a very fruitful pursuit. Hence the lack of careful distinction.

It's not that we can say nothing about rationality. Given the connection between correctness and rationality, by answering the question of correctness, we also get a bunch of conditional answers to the question of rationality. For example, suppose that future-bias would be correct if the A-theory is true, then I would be rationally future-biased *if* my future-bias is based upon a (reasonable) A-theoretic belief – regardless of whether the A-theory is in fact true. And – as a sidenote – we need to be careful about what counts as apparent reason when it comes to preferences. Preferences are evaluative attitudes that are partially belief-like. Some might even say that they just *are* beliefs about betterness (11, fn.3). In any event, preferences are very intimately tied to *normative beliefs*. Thus, we need to exclude normative beliefs from the apparent reasons to which rationality is sensitive. Suppose that I believe the A-theory to be true and also believe that it provides normative reason for future-bias. We do not want to say that I'm *ipso facto* rationally future-biased because my future-bias is based on (or just is) this normative belief. Rather, whether I am rational should depend on whether the A-theory is *in fact* reason-providing. To avoid bootstrapping, apparent reasons should be restricted to non-normative beliefs (e.g., that the A-theory is true).

In subsequent chapters, when what's at issue is correctness and normative reasons, I will make this clear. But I shall sometimes use "rationality" and its cognates as an umbrella term for whatever evaluation one might apply to time-biases.

The above correctness/rationality distinction is drawn within *substantive rationality*. For my purpose, I understand substantive rationality as concerning *individual* attitudes. But we are sometimes also interested in how one's preferences hang together – with each other or with other kinds of attitudes. Criticizing *combinations* (or patterns) of attitudes is a matter of *structural rationality*.[1] One prominent instance of structural-rationality criticism in the context of time-

[1] The precise distinction between substantive/structural rationality is notoriously tricky, especially if one allows requirements of structural rationality to take narrow scopes (cf. Broome 2013). Here I'm merely making an impressionist characterization. There is an intuitive distinction between having intransitive preferences and (say) preferring to drink petrol knowing it's petrol – though the latter may also be said to constitute a pattern of attitudes that do not "cohere".

There is an involved debate regarding how structural rationality and substantive rationality are related. (Kiesewetter & Worsnip 2023) Some think that all requirements of structural rationality are

biases is the "preference reversal" argument that hyperbolic discounting is irrational (Ch.2). It roughly says hyperbolic discounting is irrational because it leads to a diachronically inconsistent pattern of preferences. The charge of irrationality is not zooming in on a particular preference, whether in terms of correctness or rationality. Rather, it is a certain pattern of preferences over time that's considered rationally objectionable, and that pattern can be identified without looking into the specific content of the individual preferences.[1]

Then there is the distinction between *state-given* and *object-given* reasons (Parfit 2001). This distinction cuts across correctness/rationality as well as substantive/structural rationality. An eccentric millionaire offers you a $100 reward for desiring to drink petrol. Is there a (normative) reason for you to so desire? We might say that you have a state-given reason for desiring to drink petrol but lack any object-given reason. In this case, being in the state of *having such a desire* is good for you. But there is nothing desirable about the *object* of the desire, namely, drinking petrol. (There is a sense in which state-given reasons are highly context-sensitive but object-given reasons are not so.)

Debates over time-biases are mainly concerned about object-given reasons. When debating whether future-bias is correct, what's commonly at issue is the value of *past and future pain*, instead of whatever accidental harms or benefits may follow from one's *being in a future-biased state*. That said, some philosophers are also interested in state-given reasons. (Parfit 1984: 170-77, Scheffler 2021: 90-93). Insofar as being future-biased and not ruminating on the past can facilitate a more positive outlook on life, one might think, we'd be better off if future-biased. I submit that there might be state-given reasons for and against time-biases. And it's a genuine question if state-given reasons should factor in all-things-considered oughts. But state-given reasons are highly contingent and difficult to entertain from the armchair. In what follows, I shall set aside state-given reasons and take the ought of correctness to be completely determined by object-given (normative) reasons.

Recall the hypothesis that future-bias is a product of the affective asymmetry. It says that it's advantageous for us to have the evolved heuristic to discount the past relative to the future.

---

reducible to norms of substantive rationality. Some take the opposite direction of reduction. Still others regard them as somewhat independent. I'm not taking a stand on this issue.

[1] Structural requirements can be diachronic or synchronic. (cf. Hedden 2015) The alleged constraint evoked here is diachronic.

Although this is often taken as offering an evolutionary debunking argument *against* future-bias, one might want to turn this into a state-given justification *for* future-bias. For sure, the heuristic sometimes misfires. But – as the thought goes – we are better off by being in the state of *being dispositioned to* (or having the general tendency to) discount the past. That may be true; though I'm skeptical.[1] In any event, that's a separate question from object-given reasons.

To brief recap, we've made three distinctions that help to map out different dimensions of "rationality": (1) Correctness *versus* rationality; (2) The structural *versus* the substantive; and (3) Object-given *versus* state-given reasons. In the remainder of the thesis, I assess three prominent arguments against time-biases: *The Argument from Preference Reversal* (Ch.2), *the Argument from Arbitrariness* (Ch.3), and *the Argument from Wellbeing* (Ch.4). The first one concerns structural rationality. The other two are (mainly) about correctness.

Two things before concluding this section. I have not yet said anything about the rationality of choice behaviours. Here I make a minimal assumption: Given that a certain choice behaviour is motivated by a certain preference, the rational status of the choice is somewhat dependent on that of the preference. By being purely near-biased, I choose to suffer tremendously in the distant future rather than to suffer mildly immediately. Insofar as my near-bias is irrational, I'd say that my choice is irrational *because* it's motivated by an irrational preference. In rational evaluation, preferences are primary and choices are secondary. But this is not to say that there is always one-to-one mapping. For instance, some think that although cyclical preferences are in themselves rationally kosher, it would be irrational to act upon these preferences and make a choice cycle (cf. Andreou 2007). Throughout this thesis, I focus primarily on evaluating preferences as mental states but say relatively little about choice behaviours.

Finally, one does not have to consider all these dimensions of rational evaluation as legitimate or applicable to time-biases. One might think that time-biases, as preferences, are not the kind of things that can be *correct* or *incorrect*.[2] For some Humeans, there are no normative reasons that are independent of one's preferences. Some preferences are simply self-justifying. Thus, some of the upcoming discussions may fail to engage with these Humeans from the get-go. Also, one might

[1] Caruso, Latham, Miller & Yu (manuscript) studied how time-biases associate with happiness (as measured by a variety of subjective scales) and found the association to be weak and mixed.
[2] Street (2009), Goldman (2006), Hedden (2015: 150-51); cf. Parfit (2011).

think that there are no requirements of structural rationality.[1] So, the structural criticism of hyperbolic discounting would be irrelevant to some. These are all fundamental and deep disagreements. Rather than taking sides on these issues here, I shall, along the way, draw attention to places of divergence when needed.

1.4.2. Can Future-Bias Be Irrational?

There is, however, an argument to the effect that *future-bias* is beyond the veil of rational evaluation. According to Callie K. Phillips (2021), given mentalism about preferences, future-bias turns out *rationally unevaluable* because the alternatives cannot be adequately represented. An adequate representation would require representing the past/future states-of-affairs as both past/future *and present*. But, as the argument goes, that is impossible.

The argument fails. It rests on a confusion about the temporal perspectives involved in representation. Once the confusion is cleared, it should be obvious that future-bias can (in principle) be adequately represented and thus be rationally evaluated. A more detailed response is given in my (2024).

Let me clarify upfront that the actual target of Phillips's argument may be more circumscribed. She focuses on *first-person hedonic* future-bias. It's unclear how the case can be generalized as advertised. Later I will say something about the other sorts of future-bias (as well as near-bias). For now, let us focus on the first-person hedonic.

Take as a case study Parfit's *My Future and Past Operations* (quoted in §1.1). Let $A_2 > A_1$ stand for Parfit's preference for A_2 (the longer operation in the past) over A_1 (the shorter operation in the future). Phillips argues that Parfit's does not have a rationally evaluable future-bias:

> P1. If you cannot possibly represent A_2 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_2 as *present*, then you are not representing the experience in A_2 as *past*.

[1] cf. Kiesewetter (2017), Lord (2018).

P3. If $A_2 > A_1$ is future-biased with respect to pain, then necessarily (if you are representing A_2 adequately to the preference $A_2 > A_1$, you are representing the experience in A_2 as *past*).

P4. If $A_2 > A_1$ is future-biased with respect to pain, then necessarily, (if you are representing A_2 adequately to the preference $A_2 > A_1$, you are representing the experience in A_2 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 pain.

C2. If $A_2 > A_1$ is future-biased with respect to pain, then $A_2 > A_1$ is not rationally evaluable. (Phillips 2021: 582, adjusted)[1]

The argument is valid. It is also rather general. If sound, then *no* first-person hedonic preference is both future-biased and rationally evaluable. (I will sometimes drop the "first-person hedonic" qualification). By Phillips's own admission, the conclusion is "surprising and counterintuitive" (574). But that is an understatement. Notice that it's the mere fact that *some* alternative(s) is past *or* future that the argument trades on. So, the argument can be generalized to near-bias (cf. 592) – and moreover, to mundane preferences directed at a single past or future temporal location. If that sounds incredible to you, you should suspect some premise(s) to be false.

We should reject P4. Adequate representation in no way requires the past or future alternatives to be represented as *present*. Before digging in, a quick look at the other premises. P2 says that an alternative cannot be represented as both past and present. It's regarded as "a conceptual truth or very nearly one" (582). I agree. P1 draws a conceptual link between adequate representation and rational evaluability. Stated in more detail:

> 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.
>
> Differently put, the representations must disclose features of the alternatives that are the motivational basis for a preference between them. (577)

[1] Phillips introduces a new case study concerns pleasure. The quoted argument is minimally modified in alignment with Parfit's case.

This sounds very plausible. A behaviourist can afford to not care about what goes on in the mind. But we mentalists do care. Indeed, as I take a preference regarding *A* and *B* as a comparative evaluation regarding *A* and *B*, I'm willing to endorse her stronger claim that such reflection and disclosure of motivational basis is necessary "in order for [one] to have the preference *at all*" (578). Phillips gives an example of a preference for sitting on a round square. (ibid.) That's a preference that no one can have, I'm inclined to agree. One just cannot coherently represent the alternative of sitting on a *round square*. One may succeed in conjuring up something "similar" – say, sitting on a stool that's partially curvy and partially rectangular – and find *that* attractive. But that would be a preference with *different* objects.

But what exactly does adequate representation amount to when it comes to *first-person hedonic* future-bias? Why would one think that future-bias is a bit like preferring sitting on a round square? Both P3 and P4 are supposed to follow from 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 *[features] relevant*[1] 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*". (2021, 580, emphasis mine)[2]

As I understand it, (1) of CR applies to all preferences. Granted, there may always be some vagueness regarding what counts as "relevant". But it's clear that with future-bias, the relevant features include the *pastness* and *futurity* of the alternatives. Such perspectival-ness is a defining feature of future-bias, though I'm open-minded, as suggested earlier, regarding whether such perspectival-ness must be represented in terms of A-properties or can be adequately captured by B-relations in a token-reflexive manner. Assuredly, P3 is true.

[1] The original wording is "facts" rather than "features". But I take "features" to be more appropriate since the objects of preferences need not obtain.
[2] Note that P1 merely says that rational evaluability requires the *possibility* of meeting CR. It's not the case that one must actually meet CR in every instance of forming the preference of the relevant type. See Phillips (590) for reasons to opt for the weaker claim.

P4 is supposed to follow from (2) of CR, a condition special to the first-person hedonic.[1] I find (2) quite plausible. However, understood properly, it in no way motivates P4.

To imagine an experience from-the-inside, as I understand it, is to imagine *undergoing* the experience from the perspective of the *experiencer*.[2] I am attempting to form a preference between eating a durian and eating a banana. But I have never tasted or even smelled a durian before. And since (as I know) the taste of durians is rather peculiar and unlike any other food I have tried, I cannot adequately imagine myself eating a durian *from-the-inside*. Since I have no inkling of the *what-it-is-likeness* or *experiential features* of eating a durian (the texture, the aroma, etc.), as (2) suggests, I cannot make a comparative evaluation between the alternatives. I am simply at a loss if eating a durian would be worse or better for me. Of course, I can imagine *from-the-outside* that I am eating a durian. I can mentally picture myself sitting by a dining table, chewing a paste of yellow and creamy stuff. But that is no help. I might form *some* preference on that basis – say, a preference for being an adventurous person willing to try exotic food – but that would not be the hedonic preference in question. (Paul 2014)

The reader may find representation from-the-inside unnecessary. You may think statistics and testimonial evidence about durians would suffice. (cf. Paul 2014) I shall nonetheless grant (2). What I reject is the following inference: "Evaluating phenomenal features of an experience such as their [painfulness] plausibly requires imagining the experience occurring 'from the inside' and *thus* as a *present* experience." (583, emphasis mine)

This may be a tempting inference to make. I am imagining from-the-inside *eating* a durian. So, I must be representing the states-of-affairs as present. How else could I be said to occupy the experiential perspective? But that's confused. Consider episodic memory. Episodic memory can be from-the-inside and rather vivid, as if one is "reliving" the past experience.[3] I am remembering from-the-inside eating a durian. But unless I'm somehow delusional, I am representing the

[1] It seems that (2) of CR is *overdetermined* by the preferences being first-person and being hedonic:

> When we limit ourselves to prudential rationality, all such preferences will be *de se*. (579)
> 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. (583)

Anyhow, I grant that CR is true as a constraint on the first-person hedonic.

[2] I am following the standard understanding of the from-the-inside/-outside distinction. (Williams 1973, Velleman 1996, Ninan 2008)

[3] Episode memory and *anticipatory* imagination are contrasted as two modes of "mental time-travel", the former being past-directed and the latter being future-directed. In Parfit's case, of course, it's imagination that's taking him in both temporal directions.

experience as past and *only as past*. I do not represent it as present. I am not confusing a past durian-eating experience for an *ongoing* durian-eating experience.

To make this point clearer, think of an experiential perspective as consisting of an *individual* and a *temporal location*. Representing from-the-inside is representing from an experiential perspective of a certain individual at a certain temporal location. But imagining *from* a certain perspective does not entail *identifying with* that perspective. This is true of both individuals and temporal locations.

Let us imagine David Bowie living in Cold-War-era West Berlin. He produced "the Berlin Trilogy" there, including the song "Heroes" with lyrics describing a couple kissing by the wall. You may picture – as if from the lens of a movie camera – a pale guy with his newly dyed black hair, tweed vest, and newsboy hat, walking by the Berlin Wall, raising his head and turning his gaze towards a couple kissing by the wall (presumably Iggy Pop and his then-girlfriend). This is not imagining from-the-inside (of Bowie). But you can instead take the experiential perspective of Bowie and imagine what it is like for him to be walking by the wall, hearing an approaching military helicopter, seeing the couple kissing, and being mesmerized by the extraordinary scene. This is imagination from-the-inside. But it represents the experiences as *Bowie's* rather than *yours*. That is, you do not *identify* the experiential perspective as your own perspective.

This is reminiscent of David Velleman's (1996) distinction between *notional subject* and *actual subject*. In my from-the-inside imagination from Bowie's perspective, I am the actual subject who does the imagining, whereas Bowie is the notional subject occupying the centre of my imagined content. I do not identify myself with the notional subject. It is stipulated to be Bowie, not myself. Nonetheless, the notional subject and the actual subject can coincide in from-the-inside imagination. This happens when you imagine yourself living in Cold-War-era West Berlin, where the notional subject is identified with the actual subject.

Apart from a notional individual, an experiential perspective may also consist of a temporal location. We may call it a *notional time*, to be distinguished from the *actual time* at which the episode of imagination takes place.

Notional time can coincide with actual time in representational attitudes. This happens in mundane perceptual experiences. (Though Velleman's distinction is initially only concerned with

imagination.) My visual experience of a car passing is from-the-inside and it represents the event as occurring at the present. This can also happen in imagination. I may imagine from-the-inside some experience as present. I may fantasize about how wonderful it would be if I were on vacation right now. But notional time does not have to coincide with actual time. We can also "mentally time-travel" towards the past and the future by taking different temporal perspectives. And mental time-travel can be from-the-inside. This is just to say that we can, in representation, take the experiential perspective of a notional time distinct from the actual time and thus represent experiences as past or future.[1] In episodic memory, when having an occurrent recollection of eating a durian, the notional time – from the perspective of which I remember from-the-inside – is some past time. Mental time-travel can also take the form of imagination, and it can be past- or future-directed. That is, we can anchor our from-the-inside imagination at a notional time which we take to be past or future. I may imagine from-the-inside enjoying a summer vacation next year. And I may imagine from-the-inside living in Cold-War-era West Berlin.

All this is to say that P4 does not follow from (2) of CR. Representing from-the-inside does not entail representation as present because the notional time need not be identified with the actual time. And we have a positive account of adequate representation so that both clauses of CR are met. Insofar as one can take the experiential perspective of a notional time distinct from the actual time – whether past or future – there is no principled obstacle to forming future-bias that is rationally evaluable.[2]

---

[1] Nor does imagining as present entail imagining from-the-inside. The two distinctions cut across each other, just as the from-the-inside/-outside distinction and the *de se*/non-*de se* distinction cut across each other.

[2] For readers interested in possible world semantics: Phillips (2021: 587-90) notes that uncentred worlds are inadequate for representing Parfit's preference. This is correct because future-bias is perspectival. But she further argues that centred worlds are of no help. Identifying centres as ordered pairs of individuals and times, she reasons that representing the past alternative as past *and* as present would require "switching between" different centres, where a distinctive temporal perspective is lost:

> Imagining experiencing [the ten-hour operation] as present picks out a *different* centred-world from the one where [the operation] is past; it is one where the agent is located at a different time. This centred-world shares the same possible world as an element but represents a distinct possibility because of the difference in times. …[T]he "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…
>
> …[t]he agent can merely succeed in imagining a centred-world where "the centre" is just an individual, but not where it is an individual and time pair. But it is the fact that centred-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). (2021: 588, emphasis in original)

Just to clarify, I have been following a paradigm according to which the "from-the-inside-ness" of representation is *not* a constituent of representational content. (You may think of it as a mode of representation.) And when I talk about "represent as", I am concerned about the content which involves the notional time. So, in memory and in past-directed imagination, the experiences are represented simply as past.[1] But you might dislike this way of carving out content. You might think that there is still a legitimate sense in which the past experience *is* represented *as present*. We can accommodate this. We can say that it is represented as present from-the-perspective-of-the-*notional-time*. And when you're stepping into Bowie's shoes, the experience is represented as mine from-the-perspective-of-Bowie. Likewise, we can say that when being future-biased, Parfit

---

The reasoning is fallacious. It makes a similar mistake as lumping together the notional perspective and the actual perspective. We can adequately represent Parfit's preference insofar as we recognize that there can be two centres serving different functions.

The first thing to notice is that Phillips is assuming that future-bias has B-theoretic content (588, fn.19) – that is, pastness is represented in terms of the earlier-than relation rather than as a monadic property. Representing B-theoretic content is indeed more complicated than A-theoretic content (to which I turn later), but entirely feasible.

The general strategy is to postulate two centres – one corresponding to Parfit's actual temporal location and the other corresponding to the notional temporal location from which he imagines from-the-inside – and let them stand in the earlier-than relation in content. This is not an *ad hoc* move. Imagination is always to some extent constrained by what we believe about the actual world, and objects in the imagined worlds can be attributed with relations relating to actual objects. So, there is the need to make imagined content parasitic on believed content. And we can add centres into the mix. Indeed, there have been detailed proposals in the literature on imagining from someone else's perspective. (Recanati 2007, Ninan 2008)

Without debating over the details, we may follow Ninan (2008) and think of the content of imagination as two-dimensional centred intensions, i.e., functions from centred belief worlds to sets of centred imagined worlds, so as to capture how certain beliefs get 'imported into' imagination. With regard to Parfit's preference, we may firstly identity a belief-centre that comprises his actual temporal location; and secondly, identify a imagine-centre that is anchored at the time at which the past operation takes place. Then, make it part of the imagined content that the time of the imagine-centre stands in the earlier-than relation to the time of the belief-centre. It looks like this:

{<(*w, i, t*), (*w', i, t'*)>: *i* is undergoing a long and painful operation at *t'* in *w'* and *t'* is later than *t*}.

As such, we adequately represent Parfit imagining from-the-inside from a temporal location which he takes to be past. There is no 'switching between' centres because the centres serve different functions and work in concert.

Now, suppose instead that future-bias has A-theoretic content. In this case, there is no need to incorporate a belief-centre, as being past is considered as a monadic property:

{(*w', i, t'*): *i* is undergoing a long and painful operation at *t'* in *w'* and *t'* is past}.

[1] This is a popular way of characterizing third-person from-the-inside imagination: When you imagine *being* someone else (which may sound paradoxical), your personal identity is no part of the representational content. (Lewis 1979, Velleman 1996, Ninan 2008) Now, one might instead distinguish different types of content such that your identity *is* part of the content. For example, Perry (2010) suggests that you are a constituent of the "back-up" content though not of the "official" content. No matter. The important thing is that there is some distinction to be made, not how exactly it is made. Perry may say that you back-up-represent the experiences as yours but official-represent the experiences as Bowie's. Still, no incoherence would follow.

represents the past operation as *past* from-the-perspective-of the-*actual-time* and also as *present* from-the-perspective-of-the-*notional-time*. If you prefer, then, there is some sense in which P3 is true and some sense in which P4 is true. Still, there is no sense in which P3 and P4 are both true. In other words, nothing incoherent with presenting future-bias would follow from this way of talking about content more "fine-grainedly".

I have so far argued that there is no *principled* obstacle to forming rationally evaluable (first person-hedonic) future-biased preferences. The temporal features of future-bias can be taken care of by distinguishing notional time from actual time. That said, Phillips's argument does raise an interesting issue. If you share the general sentiment that adequate representation is no easy task, you might suspect that we are not doing a good job.

Our ability of perspective-taking is limited. It is rather difficult to imagine what it would be like from experiential perspectives that are unfamiliar. Temporal distance, however, tends to bring on *psychological distance*. It may then seem that some preferences involving temporally distant alternatives evade adequate representation.

The worry is more evident with near-bias.[1] Lots of our near-biases concern the far-distant future, the experiential perspective from which may be rather alien. Think of smokers who indulge in their immediate desires and incur a huge risk of getting cancer in their later life. We tend to think of such imprudence as plainly irrational. But one might think what is really going on is *not* a considered (though defective) evaluative attitude based on adequate representation, but a failure in perspective-taking due to experiential poverty. It sounds plausible that they are just *incapable* of representing the future adequately and according a subjective value to the prospect of getting cancer.

Imagining from-the-inside having cancer is not merely to imagine lots of coughs and pills and infusions. It is a "lived experience" infused with emotions, values, and outlooks on life that are "transformative" of who we are (Paul 2014.) And no one can adequately imagine the lived

[1] Phillips herself is ambivalent about the rational evaluability of near-bias. She thinks whether her argument generalizes to near-bias depends on "what is involved with representing an experience as closer or farther in temporal proximity to another" (2021: 592). But near-bias is also perspectival. Futurity is a defining feature of prospective near-bias and pastness is a defining future of retrospective near-bias. So, a parallel argument can be run against near-bias. The argument, of course, would be unsound because the counterpart of P4 would be false.

experience of having cancer, it might be thought, unless one has already experienced it (or something similar). The epistemic predicament of the smoker is similar to one who lacks the phenomenal concept of pain (Loar 1990) and therefore cannot really prefer pain to be past. Bluntly put, instead of regarding the smoker as irrationally or incorrectly discounting the disvalue of having cancer, one might say that having cancer is just not an alternative they can adequately represent. Her preference is rather like an attempt to prefer sitting on a round square that's doomed to fail to "hit the target".

Whether that's the correct thing to say, of course, depends on the one hand, how demanding adequate representation should be, and on the other, what's actually going on in our minds. Now, recall that (2) of CR is supposed to be a constraint on first-person hedonic preferences. What should we say about third-person or non-hedonic time-biases? Does the requirement of representation from-the-inside generalize? Here are some rough thoughts. As it seems to me, (2) should also be a constraint on the *third-person hedonic* (but perhaps in a less stringent manner). Suppose that I've never experienced pain, it doesn't seem that I can prefer that Parfit has endured the longer operation. Suppose instead that I am a normal person and (as I know) painful experiences do not vary *much* interpersonally, it seems that I should be able to imagine from-the-inside what's it like for Parfit to be in pain and therefore form a rationally evaluable third-person preference. Notice that psychological distance cuts across the first-/third-person distinction and comes in degrees. So, one might think that as a general matter, representation from-the-inside may be adequate to greater and lesser extents depending on psychological distance. Rational evaluability may therefore also come in degrees.

Then consider the *first-person non-hedonic*. I'm inclined to think that (2) of CR may generalize to some extent here because the *experiential* may go beyond the (merely) hedonic. Having cancer, for example, is a lived experience not reducible to pain and pleasure, but nonetheless seems to require imagination from-the-inside. As already suggested, reading some research papers on cancer and casually observing some patients may not suffice for adequate representation. Of course, there are non-hedonic events that are non-experiential. Say you are supposed to form a preference between being betrayed nine times in the past *without ever knowing* and being betrayed one time in the future *without ever knowing* (Brueckner & Fischer 1986: 216). In this case, it would be misguided to imagine from-the-inside what it's like to be betrayed – the indignation, the loss of self-esteem, etc. – and form a preference on that basis. (The difficult question remains of what it *does* take to

adequately represent alternatives that are completely non-experiential. I say something about this in Ch.3.)

When forming preferences, adequate representation of the alternatives is indeed a risky and difficult business. It can go wrong in all sorts of ways – some are no fault of our own, but some are. Nonetheless, there is no principled obstacle to representing adequately the temporal features of time-biases.[1] That is, time-biases can be rationally evaluated as comparative evaluations that are representational. And it is the task of the following chapters to go about making these evaluations.

[1] This relates back to the error theory. It remains possible (though improbable I think) that all our actual future-biases are merely apparently because we constantly fail to adequately represent the past and future alternatives but are instead merely concerned about the *present* states of retrospection and anticipation. See also Phillips (2021: 585-85, 87-88) on "the memory mistake".

2. Time-Biases and Preference Reversal

This chapter focuses on structural rationality. In particular, I question what we may call "the economist view" or "the normative standard" of near-bias as well as an analogous position on future-bias. The normative standard says roughly that a near-biased agent is not being irrational insofar as her discounting function does not lead to preference reversal over time – i.e., preferring *A* over *B* at an earlier time but preferring *B* over *A* at a later time. The normative standard thus recommends exponential discounting of the future (with constant discount rates) but condemns non-exponential discounting. Although economists have said little about future-bias, it's pointed out by some philosophers that parallel reasoning can be applied to future-bias. Future-biased agents are such that they typically undergo preference reversal. So, if it's correct that preference reversal is "a cardinal sin" (Callender 2021a: 233) that no rational agent should commit, then future-bias is a rational mistake.

Call these arguments *the Arguments from Preference Reversal*.[1] They aim to establish that non-exponential near-bias and future-bias are irrational on the grounds of structural rationality. Despite the initial appeal, I find these arguments wanting. Most importantly, it has not been made clear what "preference reversal" consists in and why it's bad. I introduce *the Arguments from Preference Reversal* in §2.1. In §2.2, I examine a popular thought that preference reversal is structurally irrational because it leads to *exploitation*. But the exploitation perspective is strained because there is no clear and necessary connection between irrationality and exploitation. Better to leave exploitation behind and motivate diachronic constraints on preferences on independent grounds. And to do so, we need to understand more carefully the notion of preference reversal. But as I argue in §2.3, we see no good, non-question-begging way to specify a diachronic constraint that would support the conclusions (given a better understanding of preference reversal). To be clear, none of this is to prove that the economists are wrong about near-bias or that future-bias is rationally kosher. It's just that I see no convincing argument from structural rationality to those ends. Finally, in §2.4, I briefly consider a family of arguments against future-bias that appeal to the

[1] Such a reasoning can be found in Strotz (1955), Roelofsma & Read (2000), Greene & Sullivan (2015: 950), Grüne-Yanoff (2021), Smartt (2022); cf. Callender (2021a, 2022)

practical upshots of future-bias working in concert with certain other attitudes (e.g., regret-aversion). I find these arguments also unconvincing.

2.1. The Economic View and the Normative Standard

There is the view that structural rationality exhausts all there is to rationality. On this view, individual attitudes cannot be criticised as substantively irrational or incorrect, but patterns of attitudes can be criticised as irrational. Though a minority view among philosophers, it's popular among economists. When it comes to time-biases, this exclusive focus on structural rationality amounts to an exclusive focus on diachronic consistency – and more specifically, an antipathy towards *preference reversal*, "a cardinal sin" in decision theory (Callender 2021a: 233).

Thus, it has virtually become a truism that "the normative standard" of *near-bias* is *exponential discounting* (Callender 2021a). For exponentially discounting is the only (non-gerrymandered) way of future-discounting that avoids preference reversal. Unfortunately, we do not actually meet this normative standard. We tend to be hyperbolic discounters and therefore tend to undergo preference reversal over time.

I'll turn to an example shortly but just to be precise here, exponential discounting is neither sufficient nor necessary for avoiding preference reversal. One needs to discount exponentially *and* keep her discount rate constant over time. And a non-exponential discounter may avoid preference reversal if the way her discount rates change over time is some certain function of calendar time.[1] But to have one's discount rates dependent on calendar time, arguably, is objectionably strange. (And we are not such non-exponential discounters.) In other words, exponential discounting is "the only non-gerrymandered type of discounting" (Hedden 2015: 52) that is guaranteed to avoid preference reversal and therefore the ideal way to avoid preference reversal. I shall grant that *the Argument from Preference Reversal*, if successful, succeeds in driving a normative wedge between exponential discounting with a constant rate and non-exponential discounting with gerrymandered rates despite that they both avoid preference reversal. And in what follows, for simplicity, I shall sometimes speak *as if* the normative standard concerns

[1] Callender (2022: 129-30), Hedden (2015: 52-53).

exponential *versus* non-exponential discounting (and not also how discount rates evolve over time).

An exponential discounter is such that "the proportionate difference between how much [they] care about one time and how much [they] care about some other time depends only on how far apart those points in time are"[1] (Hedden 2015: 52). In other words, for an exponential discounter, "[t]here will always be the same proportionate difference in how much this person cares about two future events." (Parfit 1984: 160) Suppose that I prefer 1 marshmallow in-365-days over 2 marshmallows in-366-days. And this is because I care about only the *temporal distance between the alternatives*, and I discount the value of marshmallows by 60% per day. An extra day of delay makes the marshmallows 60% less valuable, so to speak. Fast forwarding to one year later, so long as I keep my discount rate constant, I shall likewise prefer 1 marshmallow immediately over 2 marshmallows tomorrow. Despite that one entire year has passed, the distance between the alternatives has not changed. Being an exponential discounter with a constant rate, I'm guaranteed to stick to my earlier preference.

We tend *not* to be exponential discounters. Our discounting functions tend to be *hyperbolic* or quasi-hyperbolic[2], manifested in diminishing patience when good things approach in time and constantly putting off unpleasant tasks. Suppose that I now prefer 2 marshmallows in-366-days over 1 marshmallow in-365-days. From my present perspective, it's worth my while to wait for one extra day to secure one extra marshmallow. But as a hyperbolic discounter, fast forwarding to one year later, one day's wait no longer seems worthwhile. I thus *reverse* my previous preference by preferring 1 marshmallow immediately.[3] But preference reversal is *structurally irrational*. So, since hyperbolic discounting leads to preference reversal, it's structurally irrational.

It's worth emphasizing that the charge of irrationality here is a matter of structural rationality – and moreover, a matter of *diachronic* structural rationality. The normative standard of exponential discounting does not say that it's irrational to be *near-biased*. It's perfectly fine to be near-biased, insofar as you discount exponentially (and also keep constant your discount rate). This reflects the

[1] Presumably, she is also sensitive to pastness and futurity. This point will be important in §2.3.
[2] Grüne-Yanoff (2015), Loewenstein & O'Donoghue (2002), Sullivan (2022), Callender (2021a).
[3] Depending on one's particular discount rates, a hyperbolic discounter may not exhibit preference reversal in this *particular* example. But the point is that as long as one discounts hyperbolically (and more generally non-exponentially), preference reversal is unavoidable at certain times, in certain situations (if one were to form the relevant preferences).

"preferences are preferences" presumption of rational decision theory. It has no business in poking into individual preferences and criticising them as incorrect or irrational.

Although economists are mostly interested in only prospective near-bias, it has been observed by many that a similar argument can be raised against future-bias, since a typical future-biased agent also undergoes preference reversal.[1] (One might also generalize the normative standard to retrospective near-bias for obvious reasons. Much of what I will say about prospective near-bias and future-bias can be read as responding to this generalization as well.)

Consider again Parfit's *My Future and Past Operations*. Suppose that the longer operation is scheduled on Wednesday and the shorter operation is scheduled two days later, on Friday. It's now Monday so that both alternatives are future. Parfit prefers to undergo the shorter operation on Friday. This is a typical and presumably very sensible preference for him to have. Parfit is not extremely *far*-biased (i.e., drastically *inflating* the distant-future) and reasonably so. It's now Thursday and Parfit wakes up in a fog. Being future-biased, he prefers having had the longer surgery on Wednesday already. This pattern of preference reversal is similar to that of a non-exponential discounter and is likewise rationally objectionable; or so the argument goes.

One immediate question is whether this could establish that *future-bias* is irrational. Since the charge is one of structural rationality, at least on the face of it, it at best establishes that Parfit's diachronic *pattern* of preferences is irrational. That is, it only says that it's irrational for Parfit to have a certain preference at an earlier time *and then* a reversed preference at a later time. It does not say that the latter (reversed and future-biased) preference is irrational.

This is a delicate question that raises tricky issues about the nature of structural rationality. I can only make some brief remarks here. For one, it might be the case that the opponents of future-bias are perfectly satisfied with the "weaker" conclusion that Parfit should not have this pattern of preferences. This conclusion, though perhaps non-ideal to some, is still far from trivial.

[1] Brink (2011), Dougherty (2011), Greene & Sullivan (2015), Hedden (2015: §5.2), Kauppinen (2018), Ahmed (2018), Scheffler (2021). Brink remarks on Parfit's case:

> [Future-bias] also makes the preference unstable. When I view both procedures prospectively or retrospectively, I have the temporally neutral preference to minimize suffering. It is only when the greater suffering is past and the smaller suffering lies in the future that I display the temporally biased preference for the greater past pain. (2011: 377)

But perhaps one can go further. Here is an obvious observation: Insofar as a future-biased agent undergoes preference reversal, there seems to be some *rational asymmetry* between her earlier and later preferences. That is, the fact of preference reversal casts in an unfavourable light her later future-biased preference but not her earlier preference. One way to pursue this intuition is to adopt a *narrow-scope* constraint on preference reversal.[1] A narrow-scope constraint would be roughly in the form of: *If* you prefer *A* over *B* at t_1, *then it's irrational* for you to prefer *B* over *A* at t_2 (for any $t_1 < t_2$). A *wide-scope* constraint, in contrast, would be roughly like: *It's irrational* for you to [prefer *A* over *B* at t_1 *and* prefer *B* over *A* at t_2]. For our purpose, the most important difference is that the narrow-scope formulation, but not the wide-scope formulation, reflects a *rational asymmetry*. The narrow-scope constraint lays the blame on the later but not the earlier preference, so to speak, whereas the wide-scope constraint is indifferent.[2] That is, the narrow-scoper says that so long as you are already in the "antecedent state" of $A > B$ at t_1, it's irrational for you to $B > A$ at t_2. Instead, you should stick to your earlier preference.

This is not to adjudicate the general debate between wide- and narrow-scoping. The suggestion is merely that a narrow-scope formulation seems more attractive here because it delivers a good sense in which it is *future-bias* that is to blame.[3] Importantly, the narrow-scoper is *not* saying that future-bias is thus *substantively irrational* (let alone incorrect). A narrow-scope constraint remains a structural constraint: Rational criticism is "triggered" only if one satisfies the antecedent state. So, the narrow-scoper is *not* looking at Parfit's later future-biased preference, taken in isolation, and saying that *it* is irrational (by virtue of being a future-biased preference). (For the moment I am taking on the economists' assumption that structural rationality exhausts rationality. But something about substantive rationality *may* follow if you believe in substantive rationality and also that substantive rationality is intimately related to structural rationality. I turn to that when concluding this chapter.)

---

[1] For an overview of the narrow-scope *versus* wide-scope debate see Kiesewetter & Worsnip (2023: §2.1).
[2] Another way to spell out the difference is to say that while Parfit can satisfy the wide-scope constraint by revising either his earlier preference or his later preference, he can only satisfy the narrow-scope constraint by revising his later preference (insofar as he is already in the antecedent state). But I am unsure if this way of talking is apt when it comes to diachronic constraints. If "ought implies can" and if we are focusing on the times between t_1 and t_2 (which seem to be the only relevant times), then presumably he ought to satisfy the constraint by not becoming future-biased, regardless of which formulation. This is simply because Parfit cannot change his past preference.
[3] If one narrow-scopes here, then presumably she should also narrow-scope about non-exponential discounting. That is, on pain of being arbitrary, she should also identify a rational asymmetry between my earlier and later marshmallow preferences. I'll let the reader to decide if that's a welcome result.

Regardless, I shall remain neutral on whether the argument aims to establish the "stronger" conclusion that future-bias is to be blamed for structural irrationality. But for convenience, I shall sometimes say things like "future-bias is irrational". In the context of this chapter, this does *not* mean that future-bias is substantively irrational or that *every* future-biased agent is structurally irrational (for one could have other temporal preferences that are atypical and therefore do not undergo preference reversal).

To put my cards on the table, despite the orthodoxy, I don't think there is any plausible and non-*ad hoc* structural constraint that can establish that non-exponential discounting or future-bias is irrational. Just to anticipate the upcoming discussions, it's obviously not true that all instances of preference reversal (intuitively construed) are irrational. But when trying to specify exactly under which conditions preference reversal is irrational, no informative and non-*ad hoc* answer seems forthcoming.

And there is this initial concern. Although the term "preference reversal" is frequently used to characterize the preferences of non-exponential discounters and future-biased agents, one may wonder if it's ever apt. Near- and future-bias are perspectival. (§1.1, §1.4.2) Being future-biased, Parfit's preference on Thursday is *not* one for an operation on Wednesday over an operation on Friday. Rather, it's one for a *past* operation over a *future* one. Similarly for the hyperbolic discounter who cares about how distant the future alternatives are *relative to the present*. But preference reversal, understood intuitively, has this form: $A > B$ at t_1, and then $B > A$ at t_2 (for some $t_1 < t_2$). To say that Parfit undergoes preference reversal, it seems that the content of his preference must be individuated in terms of calendar time.[1] The perspectival-ness is lost. Although this is not yet an argument that Parfit is not structurally irrational, it suggests that much hangs on how preferences are individuated, which I discuss in detail in §2.3. Just for now, let's grant that non-perspectival way of talking. For Parfit's preferences, let A be *having-a-shorter-operation-on-Friday* and let B be *having-a-longer-operation-on-Wednesday*. As such, Parfit undergoes preference reversal in

[1] This echoes Callender's (2022) observation that "tense drops out of" (126) exponential discounting:

> The discounting is only a function of two differences (big reward – small reward) and (late date – soon date). If you discount, it's okay so long as you always do it the same way. "When" you are therefore doesn't matter. (127)

Though the normative standard does not, or at least should not, make near-bias completely non-perspectival. We don't want to say that it's rationally required of an exponential future-discounter (with a non-zero discount rate) to *inflate* the past. More on this point in §2.3.

accordance with the above intuitive schema. The question remains, however: On what grounds is preference reversal irrational?

2.2. Preference Reversal and Exploitation

It's observed that hyperbolic discounters can act in ways that are self-defeating or are subject to exploitation.[1] George Ainslie vividly describes the predicament of a hyperbolic discounter: "Often I'll choose in the opposite direction when I'm close and when I'm distant, which means I'll regularly do things at one time and undo them at another." (2001: 40)[2] Or to take a more contrived example, consider a hyperbolic discounter Mia. Mia got lucky and won a free beach stay in June next year (*B*). To accommodate contingencies and different tastes, a fancy cruise trip in September next year (*C*) is also on the menu. For just a small amount of money, she could get an upgrade to *C*. Fair deal, she thought. She traded *B* for *C*. It's now May next year and four months from *C*. Mia no longer finds it worthwhile to wait for *C*. She'd rather take a break from work immediately even though the beach stay (*B*) is intrinsically less attractive. Lucky for Mia, she can reschedule for *B* for just a small amount of money. Mia reschedules.

But notice what has just happened: With this sequence of choices, Mia ends up with the alternative she started with – namely *B* – and a bit *poorer*. She has paid for something that she could have had for free! Being a hyperbolic discounter, she willingly made a sequence of choices that have rendered her, *from her own perspective*, worse off. (Mia cares about money. But money isn't essential; money can be substituted with anything that the agent happens to value – say, the time and energy spent in rescheduling the trips.) A rational agent would not, it seems, get herself into such an unfortunate situation.

This is reminiscent of *the money-pump argument* against *cyclical* preferences.[3] Consider Maurice who has pairwise preferences that come in a cycle (Broome 1991: 100-05). When comparing visiting Rome (*R*) and mountaineering in the Alps (*M*) he prefers *R* over *M*. When comparing *M*

[1] See for instance Strotz (1955), Ainslie (2001), Calender (2021a), Ahmed (2018).
[2] Though you might think that at least some such behaviours are not instances of preference reversal but succumbing to temptation (§1.3).
[3] Originated in Davidson et al. (1955). Notice that a set of preferences can be intransitive without being cyclical. It's a debated issue whether agents with intransitive but non-cyclical preferences can likewise be money-pumped (cf. Gustafsson 2010).

and staying at home (*H*), he prefers *M* over *H*. When comparing *H* and *R*, however, he prefers *H* over *R*. Maurice's preferences form a cycle: $R > M$; $M > H$; $H > R$. Maurice can be turned into a money-pump. Suppose that he begins with *M* at hand (*M* being his *status quo* alternative). Insofar as he always acts on his pairwise preferences, he will pay some cost to move from *M* to *R*, and then from *R* to *H*, and then from *H* to *M*. He thereby cycles back to his *status quo* alternative but with gratuitous monetary loss. Indeed, for cyclical preferences, the trading cycle can be repeated *ad infinitum* (or till one is bankrupt). Maurice's trading behaviours seem rationally indefensible.

Though this may be obvious, it's worth pointing out that as standardly interpreted, a general rational constraint against cyclical preferences is viable only regarding *synchronic* preferences (or at least when the relevant preferences remain constant in the meantime.) That is, there is no general constraint against changing preferences over time resulting in a *diachronic cycle*. So, time-biased preference reversal is not an instance of cyclical (or symmetric) preferences in the relevant sense.

What makes Maurice relevant for our purpose is that both Maurice and Mia can be money-pumped. More precisely, they both suffer what Brian Hedden calls *diachronic tragedy*:

> A *tragic attitude* is one such that if you have it, there are *possible cases* where you will prefer performing each member of a *tragic sequence* at the time it is available, even though you prefer not to perform the sequence *as a whole*. (2015: 75, emphasis mine)

In Mia's case, the tragic attitude (or more precisely the tragic *combination* of attitudes) is $C > B$ at t_1 and $B > C$ at t_2. The tragic sequence consists of trading *B* for *C* at t_1 and trading *C* for *B* at t_2. At both times, Mia acts in accordance with her pairwise preference at the time. But she performs a tragic sequence. *At both times*, she prefers not performing the tragic sequence as a whole over performing it (given that she cares about money).

According to *the pragmatic interpretation* of the money-pump argument against cyclicity, cyclical preferences are irrational *because* they are a tragic attitude.[1] That is, because cyclical preferences make one vulnerable to diachronic tragedies, they are irrational. This interpretation is pragmatic

[1] McClennen (1990), Schick (1986), Andreou (2007), Rabinowicz (2008), Ahmed (2017). One alternative interpretation is that money pumps are a dramatic devise to *illustrate* the irrationality of cyclical preferences. (Davidson et al. 1955; Gustafsson 2013). On this view, it's not the case that cyclical preferences are irrational *because* they are exploitable.

in the sense that it trades on the pragmatic consequences of the choice behaviors resulting from a tragic attitude. (On a charitable interpretation, the argument goes from *possible* pragmatic consequences to general requirements of rationality. So, Maurice would *not* be off the hook if it happens that no one is around to exploit him.)

It thus seems that we have a parallel argument against preference reversal. In a schematic form:

(1) Preference reversal over time makes one vulnerable to diachronic tragedy.
(2) If an attitude makes one vulnerable to diachronic tragedy, then it is irrational.
(3) Therefore, preference reversal over time is irrational.

You might worry if this argument is irrelevant to future-bias. Presumably, it's impossible to affect the past, and therefore a typical future-biased agent is not vulnerable to diachronic tragedy. Parfit may on Monday pay some money to ensure that he shall have the shorter operation on Friday. When it's Wednesday, he prefers otherwise. But that's it. He might regret having made the earlier trade. But there is no tragic *sequence* he can ever perform.[1] This is not a real issue, however, for the argument aims to establish that *all* instances of preference reversal are irrational. (We may think of the causal irrelevance of the past as a contingent exploitation-blocking factor parallel to the absence of exploiters.) The real issue is this: We don't want to put a general ban on preference reversal. The above argument proves *too much*.[2]

Indeed, this is a rather general problem with pragmatic arguments: they tend to prove too much. It has been argued by many that a variety of tragic attitudes are such that we have independent reasons to judge them as *rational*. That is, diachronic tragedy is very probably not a sufficient condition for irrationality. Thus, there is good reason to reject premise (2) of the above argument.

[1] You might think that there is some structural constraint governing preferences and regret. I turn to this in §2.4. Or you might think that Parfit is nonetheless vulnerable to diachronic tragedy because he may be under the delusion that the past can be affected and therefore cough up some money to make his preference "satisfied". This can be granted. But I don't think exploitability is the real issue.
[2] Parfit may suffer diachronic tragedy after all if he is also *risk-averse*. As Tom Dougherty (2011) shows, if Parfit is future-biased and risk-averse, there is a possible scenario in which he would "take a series of pills that leave [him] with more pain and better off in no respect" (526). Parfit can be pain-pumped, as it were. The problem is that the tragic attitude here is the *combination* of risk-aversion and future-bias. Although Dougherty (2011: 535-36) further suggests that since risk-aversion is rational, it's future-bias that's to blame, Preston and Sullivan (2015: 955-56) object that risk-aversion is irrational in this scenario.

Although the sentiment is widely shared that "the connection between exploitation and irrationality is very tenuous" (Callender 2021a: 244), objections to (2) are variegated. For example, some hold that cyclical preferences are often rational (MacIntosh 2010).[1] It's also pointed out that violation of the reflection principle for preferences leads to diachronic tragedy (Hedden 2015: 83-84) but the reflection principle is not generally true (Harman 2004).[2] Here, rather than replicating all the relevant discussions, I only wish to showcase the implausibility of (2) with an example closest to home – one of preference reversal. (It is thus also a direct counterexample to the conclusion.) However, as soon will be clear, there is an uncharitable reading of the argument that would multiply counterexamples too easily. But once it's made clear what the more charitable reading is, it turns out that the counterexample casts doubt on both (1) and (2).

Consider Maya. Just as Mia the hyperbolic discounter, she undergoes preference reversal with the same preferences and performs the same tragic sequence. That is, she trades *B* (beach stay in June) for *C* (cruise trip in September) at t_1 and then trades *C* back for *B* at t_2. Unlike Mia, however, Maya's preference reversal is not underwritten by hyperbolic discounting. It's rather caused by a change in taste. At t_1, she prefers *C* over *B* because she just doesn't get what's fun about beach stays. By t_2, however, she has developed a fascination for surfing and sunbathing. After exposing herself to these activities, she gradually grows into them. At t_2, then, she prefers *B* and thereby trades back. This is somewhat unfortunate for Maya, but one would be hard-pressed to say that her preference reversal is *irrational*. People grow out of and into preferences all the time. There cannot be a general rational constraint against all such instances.

But this objection has gone a bit too quickly. Though perhaps an obvious point (cf. Pettigrew 2019: 198-201), it's sometimes not made clear that diachronic tragedies indicate irrationality only when they are *predictable* by the would-be victims. In Wlodek Rabinowicz's words, a money-pump argument is effective only if it's "based on the assumption that the agent to be exploited knows at least as much as her would-be exploiter" (2008: 146). Or as Brian Hedden puts it:

> It is one thing to suffer misfortune through no fault of your own, where you couldn't have avoided disaster even if you tried. But if you head down a disastrous course

[1] Rational cyclical preferences may include Maurice's preferences, for it's not an unnatural reading of Maurice's alternatives that their values are not comparable.
[2] The reflection principle roughly says that you should now defer to what you *believe* to be your future preferences (no matter whether you will in fact have those preferences).

> knowing full well that it is disastrous, then it seems there must be something wrong with you. (2015: 75)

Consider your repeated failures in the stock market. Or consider one who has been lured deep down a conniving pyramid scheme. These people may be accused of being foolish or reckless. But they don't have to be irrational in the sense that one who (simultaneously) believes contradictory things is. So, for a pragmatic argument to have any bite, the exploitation setting must be such that the unfortunate agent knows (or is in a position to know) whatever the exploiter knows for the purpose of exploitation. Thus, "diachronic tragedy" in both (1) and (2) of the above argument has to be understood as predicted (or predictable) diachronic tragedy; otherwise (2) would be patently false. When it comes to preference reversal, this means that the agent, at the time of holding her *status quo* alternative, predicts not only the upcoming trading opportunities but also her own *preference reversal*.

It's arguably true that non-exponential discounters and future-biased agents often predict – or at least are in a position to predict – their own preference reversal, for their time-biases are expected to be somewhat stable. So, let's assume that Mia the hyperbolic discounter faces a predicted diachronic tragedy. Let's also stipulate that Maya has foreknowledge about her preference reversal caused by exposure to new experiences (though this is not realistic). The problem, however, is that once it's made clear that Mia and Maya "see what's coming" for them, we do not expect them to suffer diachronic tragedy (McClennen 1990, Schick 1986, Rabinowicz 2008). At least, so long as Mia and Maya are minimally sophisticated agents capable of incorporating their future perspectives into decision plans, they will simply refuse to trade at all.[1]

This observation leads many theorists to conclude that the implication of irrational attitudes from diachronic tragedy is at best indirect. According to Rabinowicz, pragmatic arguments are best understood as making *conditional* recommendations such as: "If you are going to make your decisions in a disunified way, then you'd better satisfy these constraints [on mental states]" (2008: 140). In the same vein, Chrisoula Andreou (2007) takes the money-pump argument against cyclicity to have established only that it's irrational to have cyclical preferences *and* always choose

[1] There is a fool-proof method of backward-induction reasoning to safeguard against diachronic tragedy (Rabinowicz 2008). This is not to say that one has to actually use that method. Put yourself in the shoes of Mia or Maya, it should be obvious what to do.

in accordance with those preferences.[1] (She also takes there to be independent reason that cyclical preferences are sometimes rational.)

All this seems correct to me. On the pragmatic interpretation, it's argued that what's bad about certain attitudes is that they guarantee diachronic tragedies in certain situations. But if one can also avoid diachronic tragedies by being more sophisticated choosers, then it's hard to see what's intrinsically or categorically wrong with these attitudes. Differently put, we can understand pragmatic arguments as directly speaking against certain patterns of choice behaviours – but without direct implications about the underlying attitudes. Thus, there seems to be no coherent reading of the argument that succeeds. For the premises to be true, one has to modify (1) into saying that it's preference reversal in conjunction with myopic choice strategies that leads to diachronic tragedy (and correspondingly modify the antecedent of (2)). But then one does not have the desired conclusion.

However, as one might object, even if it's only directedly established that it's irrational for one to undergo preference reversal *and* always choose in accordance with her preferences[2], shouldn't we blame the former conjunct rather than the latter conjunct? (cf. Andreou 2007) But that cannot be a universal truth. Consider again Maya. I don't see anything rationally objectionable with her preference reversal (even given that she sees everything coming for her). It cannot be rationally required of her to keep her earlier and later preferences constant. At t_1, she hasn't yet *experienced* the fun of surfing and sunbathing (despite having anticipated the *fact* that she will be exposed to new experiences), so she has every right to stick to her preference for the cruise trip. It may even be psychologically impossible for her to defer to her future preference given that she just doesn't get what's fun about sunbathing and surfing as of now. And at t_2, her preference reversal is also perfectly rational given her new experiences.[3] We have every reason, then, to hold that Maya's preferences are fine. It's just that, given her preferences, she had better not be myopic in her choices.

Now, as the reader may have been thinking, there is some obvious difference between Maya and Mia the hyperbolic discounter. Intuitively, Maya undergoes preference reversal *for good reaso*n,

[1] Notice that both Rabinowicz and Andreou assume that preferences are not just choice behaviours.
[2] Rabinowicz (2008) would not even endorse this weaker conclusion. He takes the myopic chooser to be foolish but not *irrational*.
[3] You might think that Maya is irrational because she violates the reflection principle for preferences (56, fn.2). But the principle is widely regarded as false. In particular, it arguably fails to hold when believed preference change is due to new experiences, of which Maya's case is one instance (cf. Harman 2009).

but Mia does not: Her preference reversal is willy-nilly, as it were. I don't deny this. And I'll attempt to spell out this difference in more technical terms in the next section. But the point here is that whatever the difference is, it's not reflected in the above pragmatic argument. And, as it seems transparent to me, the argument cannot reflect that difference however modified: For it trades on the mere fact of preference reversal but is indifferent to what causes preference reversal. Indeed, this indifference to "bad" and "good" kinds of preference reversal may indicate that pragmatic arguments are generally not apt when applied to *diachronic* patterns of attitudes. This is not an unfamiliar point. John Cantwell (2003) argues that although pragmatic arguments work for synchronic attitudes (though this much is also doubtful), they are invalid in diachronic contexts. Preston Greene concurs: "The problem is that dynamic inconsistency is not the same sort of thing, normatively speaking, as synchronic inconsistency. And so, the economist has stretched consistency reasoning past where it can reasonably go." (2021: 263)

In all, we should leave the money-pumps behind. It's ill-advised to argue against non-exponential discounting and future-bias on the grounds that they lead to diachronic tragedy, for the connection between irrationality and exploitation is too tenuous. In particular, there is the need to distinguish "good" cases of preference reversal from "bad" cases, but that is something the money-pumps cannot deliver. To do so, we need to take a closer look at the very notion of preference reversal.

2.3. Ultimate Preferences and Preference Change

Preference reversal can be seen as one type of preference change. Indeed, the most acute type. Carrying over the characterization from above, preference reversal is understood as going from (strictly) preferring *A* over *B* at an earlier time to (strictly) preferring *B* over *A* at a later time. There are other types of preference change that are less acute. For instance, one may evolve from being indifferent between *A* and *B* to having a preference for *A* over *B*, or retain her preference ordering but vary its strength. For the sake of generality, we may consider if there is some plausible constraint on *preference change* over time. Let us begin with the following principle:

Invariance: Rational agents do not change their preferences over time.

Invariance sounds implausible. Consider again Maya who prefers at t_1 cruise trips over beach stays but comes to prefer at t_2 the opposite because she has developed a fascination for surfing and sunbathing in the meantime. Or consider Layla who is at t_1 morally indifferent between omnivorism and veganism but comes to prefer veganism at t_2 because she has in the meantime learned about the appalling reality of animal farming. These instances of preference change seem perfectly rational.

In order not to rule them as irrational, we might want to make a distinction between *ultimate* preferences and *derivative* preferences.[1] Intuitively, although Maya and Layla have changed their preferences in some ordinary sense, it might well be that there is "a bedrock layer" (Hedden 2015: 46) to their preferences that has remained invariant. Maya has always preferred the type of vacation that's more enjoyable. What has changed is what she enjoys (or takes herself to enjoy). Layla has always preferred a type of diet not causing gratuitous animal suffering, all else being equal. What has changed is her empirical belief about animal suffering. In both cases, what they ultimately value does not change, so to speak. It's only their derivative preferences that have changed, due to exposure to new information or new experiences. To simplify the matter, we may consider both instances of preference change as underwritten by *belief change* (about descriptive matters). Although exposure to new experiences may not seem entirely reducible to mere belief change, we may lump them together for our purposes – since exposure to new experiences is not what's at issue with time-biases. We thus have the following principle suggested by Brian Hedden (2015: 47):[2]

> **Utility Conditionalization:** (1) It is rationally required that your *ultimate* preferences (i.e., preferences over *maximally* specific possibilities) do not change over time. (2) It is rationally required that your *derivative* preferences (i.e., preferences over *non-maximally* specific possibilities) change only as your *beliefs* change.[3]

[1] Lewis (1988), Hedden (2015: 45-47), Pettigrew (2019: 19-22).
[2] Hedden thinks that all rational constraints are synchronic but not diachronic. So, he does not endorse Utility Conditionalization. I draw much aspiration from Hedden's discussions, though I do not deny that there may be diachronic constraints.
[3] As Hedden notes, the second clause can be "combined with a variety of views about rational belief changes" (2015: 47) including Bayesian Conditionalization which is analogous to Utility Conditionalization. As we are not concerned about rational belief change here, this minimal formulization suffices.

Utility Conditionalization takes care of Maya and Layla nicely. It's also neat and intuitive. As Hedden notes, it's difficult to conjure up a plausible alternative constraint "other than the vague injunction not to change your preferences too dramatically too often" (2015: 48).

Utility Conditionalization has some potential problems that need not detain us. You may find ultimate preferences psychologically unrealistic. (Do we really have all these fine-grained preferences?) For our purposes, however, the important thing is that in order to rule Mia's and Parfit's preference reversals as irrational, ultimate preferences need to be understood as excluding *perspectival* preferences. That is, maximally specific possibilities should correspond to *uncentred* rather than *centred* worlds.[1] But Utility Conditionalization, thus formulated, seems false.

To see this, let's first assume that ultimate preferences must be non-perspectival. That is, they can be adequately represented by uncentred possible worlds. (Derivative preferences are also non-perspectival – in line with the stipulation that a hyperbolic discounter undergoes preference reversal in the sense of firstly $A > B$ and then $B > A$.) If that's so, then Parfit would fail condition (1) of Utility Conditionalization. This is because Parfit's ultimate preferences can only be fleshed out in terms of calendar dates in order to be non-perspectival. On Monday, for all pairs of worlds that differ only in respect to the scheduling of the operations, he prefers those in which a shorter operation occurs on Friday over those in which a longer operation occurs on Wednesday. However, on Thursday, when Wednesday has become past, Parfit's ultimate preferences are such that for all pairs of worlds that are otherwise identical, Parfit prefers those with a longer operation on Wednesday over those with a shorter operation on Friday. His ultimate preferences have been reversed.

What if ultimate preferences can be perspectival? Consider instead Utility Conditionalization* which takes maximally specific possibilities to be centred worlds and therefore allows ultimate preferences to be *centre-sensitive*. Derivative preferences remain non-perspectival. Accordingly, the beliefs that may rationalize change in derivative preferences can be perspectival. [2] Utility

[1] More on centred-world representations see §1.4.2: 42, fn.2.
[2] On the B-theoretic token-reflexive account of *de nunc* attitudes, the perspectival-ness here may not be strictly speaking a matter of representational content, but (also) a matter of mode of representation. If that's your preferred account of perspectival representation, then preference reversal should not be considered as merely a matter of content.

Conditionalization* sanctions Parfit's preferences – or at least if he is *always* future-biased.[1] We can stipulate that Parfit's ultimate preferences are always such that for all pairs of centred worlds that differ only in the temporal locations of the centres, he prefers those in which pain is past over those in which pain is future. But then the change in his (non-perspectival) derivative preferences are vindicated by the change in his *perspectival* beliefs about his own temporal location in relation to the operations in question.

The very same reasoning goes for the hyperbolic discounter. So, Utility Conditionalization condemns time-biased preference reversal but Utility Conditionalization* does not. The question then becomes: Which principle should be adopted? Frankly, I don't know if Utility Conditionalization* is true. But there is good reason to reject Utility Conditionalization.

It is a mundane observation that our preferences are perspectivally sensitive to *personal identity* in addition to temporal locations. That is, our preferences can be centre-sensitive along both dimensions.[2] Consider a case reminiscent of the classical discussions on *de se* attitudes. A philosophical journal announces that "The Problem of the Essential Indexical" by John Perry and "Attitudes *De Dicto* and *De Se*" by David Lewis are the only two pieces on the shortlist for the Best Philosophical Paper Award of 1979. Like Rudolf Lingens in Perry's tale, Perry temporally forgets about who he is. Reading the journal, Perry is indifferent regarding which paper wins. He reads them and considers them as equally good. Then there comes the sudden realization that *he is Perry*. Acquiring a *de se* belief, he now much prefers that Perry's paper wins.

According to Utility Conditionalization, Perry's preference change is irrational. If ultimate preferences are non-perspectival and must remain invariant, then Perry cannot rationally change his preference regarding which paper wins. By contrast, Utility Conditionalization* can accommodate Perry's preference change. I consider this a compelling reason for Utility Conditionalization* and against Utility Conditionalization. To be clear, the point here is not that egoism is true. Perhaps it's incorrect or substantively irrational for Perry to be biased towards

[1] What if Parfit is at an earlier time time-neutral but becomes future-biased at a later time? This may seem a change in ultimate preferences and therefore irrational by the light of Utility Conditionalization*. But it might well be that Parfit becomes future-biased because of some belief change – say, coming to believe in the A-theory of time. Carving out ultimate preferences in this way may raise some technical issues (as you may take the A-theory to be impossible), but I find it intuitively plausible that such preference change need not be structrually irrational.

[2] These two dimensions become one if you identify centres as person-stages. (Lewis 1979)

himself. That's up for debate. The point is that it's preposterous to rule as irrational Perry's preference change on purely *structural* grounds. But Utility Conditionalization prohibits all preference changes in response to *de se* information. It overstretches the boundaries of plausibility and should be rejected. (If Perry's selfishness somehow clouds our judgements, consider biases towards our nearest and dearest.)

At this stage, one might want to modify the principle to the effect that ultimate preferences can be *de se* but not *de nunc*. Call this Utility Conditionalization**. It would sanction Perry but condemn Parfit. Utility Conditionalization** may be true (but see the next a few paragraphs). But I wonder what could motivate that principle – except for a need to rule as irrational hyperbolic discounters and future-biased agents. This is tantamount to just stipulating that they are irrational. I don't want to overstate my argument here. Nothing can stop one from making such a stipulation. But the dialectic is this. We have started with an intuitive case against preference reversal and thereby against time-biases. It makes it appear as if everyone has a good idea about what preference reversal is and why it's bad. But as things turn out, exploitation is not the real issue, and not all instances of preference reversal are bad. So, we turn to consider more sophisticated principles governing preference change over time (by teasing apart ultimate and derivative preferences.) Then we find that the only plausible candidate is one that's tailor-made to establish the conclusion. But that's hardly an argument. It's more of a crude appeal to intuition.

Now, I'm willing to concede that perhaps the debate just boils down to intuition. But it's very questionable if intuition really counts in favour of Utility Conditionalization**. For one, an exponential discounter doesn't seem much more rational than a hyperbolic discounter. The normative standard poses no restriction on discount rates. An exponential discounter "may care only about the moment" (Callender 2021a: 241). And "even fairly moderate discount rates lead to extreme differences in how much you care about two particular times, when those two times are far enough apart" (Hedden 2015: 53). For sure, the exponential discounter will not undergo preference reversal. But on intuitive grounds, invariance hardly seems a universal virtue. Why should it be more virtuous to reaffirm an extremely near-biased preference at an early time rather than repudiate it? Also, setting preference reversal aside, it's rather unclear why one should only care about the temporal distances between two future events (Hedden 2015: 54). These questions

haven't been properly answered or even considered.[1] Finally, if the argument boils down to an appeal to intuition, then friends of future-bias should remain completely unmoved, given our widespread and entrenched pro-future-bias intuition.

In addition to being *ad hoc*, there is one more problem with Utility Conditionalization** (and Utility Conditionalization). The normative standard of near-bias reprimands non-exponential discounting but recommends exponential discounting. Craig Callender (2022) observes that the difference between the two kinds of discounting can be characterized as "tensed" *versus* "tenseless": The non-exponential discounter cares about where "now" is, but the exponential discounter does not. That's only *partially* true. The normative standard, if it is to be remotely plausible, cannot be *thoroughly non-perspectival*. This is because it had better be restricted to *prospective near-bias*. (Callender 2022: 123-24, Hedden 2015: 53) Suppose instead that it generalizes from the prospective to the retrospective. Then for any exponentially and prospectively near-biased agent with any *non-zero* discounting rate, the normative standard would require her to *inflate* the past relative to the future and moreover, to *inflate* the distant past relative to the near past. That is, in order to be structurally rational, such an agent who discounts the future exponentially has to be past-biased and *retrospectively far*-biased (to the same extent of her prospective near-bias). This completely non-perspectival version of the normative standard does satisfy Utility Conditionalization**. But this completely non-perspectival normative standard is not even remotely plausible. If anything, the normative standard should recommend *discounting* the past exponentially. But that would mean that the normative standard violates Utility Conditionalization**, for the normative standard permits preference change in response to *de nunc* belief change.

In brief, any version of Utility Conditionalization is caught in a dilemma. The principle either allows or condemns *de nunc* preference change. If the former, then it has no bite against non-exponential discounting or future-bias. If the latter, then the resulting "normative standard" would be one that nobody would endorse.

In light of this dilemma and also the above observations that the normative standard doesn't seem intuitively compelling, one gets the impression that the normative standard is an unstable

[1] In Ch.5 I argue that the best justification for near-bias is in terms of intra-person bonds-of-concern. But we have no reason to expect the resulting discounting functions to be exponential.

position.[1] It suggests two more appealing alternatives. On the one hand, there is zero-discounting (or some alternative discounting functions that are well-motivated). On the other, there is the permissivist position that permits near-bias of any form. That brings us to the realm of substantive rationality and correctness, which will occupy the remaining chapters.

The takeaway of this section is that there isn't any well-motivated and plausible constraint on preference change that would render irrational non-exponential discounting and future-bias. And I have drawn attention to the internal tension of the normative standard. I turn to future-bias in the next section where I briefly address a family of recent arguments against future-bias that are neither strictly pragmatic nor (explicitly) relying upon principles like Utility Conditionalization. But I wish to address one potential objection before that. This is the "cheating by redescription" objection which says that I have been illegitimately carving out the content of time-biases.

In order to get a better understanding of preference reversal (and preference change in general), I've appealed to the distinction between ultimate and derivative preferences. I grant that Parfit has undergone preference reversal by locating his preference reversal on the level of derivative preferences. Derivative preferences are assumed to have non-perspectival content. I also claim that this is consistent with his ultimate *perspectival* preferences being invariant. Then I argue that any plausible version of Utility Conditionalization should allow ultimate preferences to be perspectival. Now, here is a quick-and-dirty argument that is probably unconvincing but delivers a similar message. According to my preferred definitions of time-biases, their contents are essentially perspectival. This is to disregard the somewhat artificial ultimate/derivative distinction and opt for a manner of content individuation that's "superficial" but introspectively immanent. By this way of carving out content, Parfit simply *does not* undergo preference reversal. On Monday, he has a preference regarding two *future* states-of-affairs. On Wednesday, he prefers a *past* state-of-affairs over a *future* one. He doesn't reverse his earlier preference. He just loses the old preference and gains a new one concerning different alternatives. And it's difficult to see how that would be structurally irrational. At least not any more irrational than Perry's preference change – one going from an indifference between Lewis winning and *Perry* winning to a preference that *I* win. But that sounds like cheating, you might say: One can redescribe the content of preferences as fine-grained as one likes so as to give off the *mere impression* of rationality. And it might be further

[1] Callender (2021a, 2022) and Greene (2021) agree that the normative standard is unstable. While Callender has sympathy for permissivism, Greene stands by zero-discounting.

objected that my official argument (which makes the ultimate/derivate distinction) also somehow involves cheating by redescription – by illegitimately smuggling perspectival-ness into ultimate preferences and beliefs, perhaps.

A similar worry of redescription has been raised regarding cyclical preferences. Recall Maurice with cyclical preferences over trip plans. But does he really have cyclical preferences and is therefore irrational? Perhaps not. Perhaps his preferences are not merely over trip plans but are over the *choice behaviours* themselves. As John Broome describes it:

> Mountaineering frightens him, so he prefers visiting Rome. Sightseeing bores him, so he prefers staying at home. But to stay at home when he could have gone mountaineering would, he believes, be cowardly. That is why, if he had the choice between staying at home and going mountaineering, he would choose to go mountaineering. (1991: 101)

The relevant alternatives are now fine-grained choice behaviours – e.g., *choosing-to-stay-at-home-when-the-alterative-is-mountaineering*, which Maurice disprefers because *this choice* would reflect his cowardness. Individuating preferences this way, Maurice turns out to not have cyclical preferences. Broome then worries that if one can always fine-grain preferences like this, no one has cyclical preferences. Indeed, all structural constraints would be vacuous. But Broome's worry is unfounded. For sure, one may *redescribe* preferences as one likes. But (unless some really bad version of behaviourism is true) redescriptions *don't make things real*. This point has been made by James Dreier:

> For we might notice that a peculiar set of preferences like Maurice's *could* be rationalized by more finely individuating options, and still ask whether Maurice *does* in fact individuate the options in that finer way (1996: 259-60, emphasis in original).

Which description of Maurice's preferences is accurate depends on, among other things, whether he cares about making decisions that appear cowardly. That is, whether Maurice really has cyclical preferences depends on what he really cares about. The same is true about time-biases. I am not hoping to *make it the case* that Parfit's preferences have perspectival content by mere redescription. What's in fact going on in his mind comes first, and that determines which description is accurate. With genuinely future-biased (and near-biased) agents, it's undoubtedly the case that their

considerations are in fact perspectival. This perspectival-ness goes into the explicit content given the "superficial" but introspectively immanent manner of content individuation. But if you prefer to make an ultimate/derivative distinction and put a non-perspectival restriction on derivative preferences (in line with the received characterization of preference reversal), then perspectival-ness goes into ultimate preferences. However you individuate preferences, perspectival-ness must appear somewhere. That's accurately reflecting what time-biased agents really care about instead of cheating by redescription.

2.4. More on Preferences and Choice Behaviours

While it's long observed that future-biased agents frequently undergo preference reversal, we don't tend to find it problematic – at least not to the extent that hyperbolic discounting is regarded as problematic. What might be responsible for this asymmetry in intuitive judgements? One reason, presumably, is that future-bias seems *practically inert*. While it's not unusual for hyperbolic discounters to counteract their earlier efforts due to preference reversal, such pragmatic repercussions do not follow from future-bias, for one cannot causally affect the past (though one may regret the past). A family of recent arguments, however, challenge the idea that future-bias is practically inert. (Dougherty 2011, 2015, Greene & Sullivan 2015, Sullivan 2018) They point out that future-bias is nonetheless *indirectly* action-guiding. Future-bias can, in concert with certain other attitudes, lead one to act in certain ways. And these practical upshots are, as the argument goes, rationally objectionable. In this regard, future-bias turns out as problematic as hyperbolic discounting.

There is an initial ambiguity in these arguments (Braddon-Mitchell et al. 2023). They can be read as straightforward pragmatic arguments. On this reading, they aim to establish that future-bias is irrational *because* it can lead to unacceptable pragmatic consequences – although these pragmatic consequences are not strictly speaking diachronic tragedies (in Hedden's sense).[1] Alternatively, one may read them as dramatizing and concretizing the inherent irrationality of future-bias. The pragmatic consequences are, on this reading, *evidence* of irrationality (Dougherty 2015). Despite my general misgivings about pragmatic arguments, I shall remain neutral between these two readings. One major problem with these arguments, as I see it, is that they are not very dialectically

[1] Except for Dougherty's (2011) pain-pump, which seems a straightforward diachronic tragedy. (fn.16)

forceful.[1] The allegedly problematic pragmatic consequences are such that defenders of future-bias can legitimately take them to be acceptable. The "pragmatic spin" does not seem to add much, and the arguments more or less boil down to a proclamation that future-bias is irrational.

As these arguments are all structured around elaborate case studies requiring detailed exposition, I can only be brief here. I refer to Dale Dorsey (2016, 2021: 260-79) who does an excellent job painstakingly analysing those cases and pointing out some alluring ambiguities. The upshot is that these arguments either rely on principles of rationality that are suspicious[2], or more or less beg the question against future-bias (or both). My misgivings are in a similar vein. Indeed, I suspect that I'll only be rehashing some of Dorsey's arguments in a less articulate form.

I'll focus on just one such argument to reveal the general problem. Consider what Preston Greene and Meghan Sullivan (2015) call *the scheduling problem*:

> *Fine Dining:* Jack wins a free meal at a fancy French restaurant on Monday morning, and he must schedule the meal for a night sometime in the next week. Given his flexible schedule, every night is equally convenient for him, and there are no other considerations that would make the meal more enjoyable or more likely to occur on one night rather than another. Therefore, Jack schedules the meal for Monday night. As expected, it is an incredibly delicious meal. On Tuesday morning, Jack strongly prefers that his restaurant experience were in the future, rather than the past. And so he regrets scheduling the meal for the previous night. (2015: 959)

Jack is absolutely future-biased and not near-biased at all.[3] Now suppose that Jack is also *regret-averse*, where regret is understood as "a type of preference – namely, preferring that one had done otherwise" (Greene & Sullivan 2015: 957). Regret-aversion, according to Greene and Sullivan, is rational. More precisely, they take the following principle to be true:

---

[1] Dougherty concedes that the arguments are of limited persuasive force:

> I will not insist that if someone reflects on my argument by itself, then she is rationally obliged to accept the conclusion as true. What I do insist on is that encountering this argument should lead anyone to reduce his or her credence in the claim that future-bias is rationally permissible. (2015: 12)

[2] See also Bnefsi (2023), Braddon-Mitchell et al. (2023).

[3] The assumption that Jack is *absolutely* future-biased is convenient though unnecessary. But it's crucial that Jack is not near-biased so that his future-bias is not counterbalanced.

> *Weak No Regrets:* If an agent has full and accurate information about the effects of the options available to her, then it is rationally permissible for her to avoid options she knows she will regret in favor of ones she knows she will never regret. (958)[1]

Jack reasons that he should schedule the dinner on the last day available because that's the only course of action that he'll never regret. This reasoning, according to *Weak No Regrets*, is rational. So, he schedules the dinner on the last day. But this scheduling behaviour, according to Greene and Sullivan, is "absurd" (960).[2]

The argument against future-bias then goes as follows: Jack acts in a rationally indefensible way, which is the result of a combination of regret-aversion and future-bias. And since his regret-aversion is rational (as *Weak No Regrets* is true), it follows that future-bias is irrational.

This argument has been challenged on various grounds.[3] (The move from the attitude-combo being irrational plus regret-aversion being rational to future-bias being irrational is especially suspicious.) The point I'd like to make here, instead, is that the argument more or less *boils down* to an appeal to intuition that future-bias is irrational. As such, defenders of future-bias can honestly remain unmoved.

To put my cards on the table, I don't find Jack's scheduling behaviour absurd at all. Indeed, I fail to see where the opposite intuition could come from – if not an antecedent commitment that future-bias is irrational. As I see it, Jack has reasoned and decided in an honourable way. And this is simply because, I surmise, I cannot shake the intuition that future-bias is rational (or correct).[4] That is, although Greene and Sullivan, as time-neutralists, may sincerely find that Jack postpones the meal "for no good reason" (2015: 960), it's also a perfectly sensible intuition for the opposite

---

[1] It's "weak" because it permits rather than requires avoiding regret. The stronger version faces dire objections. (Harman 2004; cf. Sullivan 2018: 94)

[2] They also consider a variant case in which there is no deadline, where it's unclear whether Jack should *ever* schedule the meal: "If Jack is unlucky enough to live forever, he may find that he never has the meal." (2015: 959) I find this case problematic. Very briefly, I surmise that if Jack is regret-averse, then he'll most regret never having scheduled the meal. So, he won't miss the meal after all. (cf. Bnefsi 2023, Dorsey 2021: 268, fn.46)

[3] Braddon-Mitchell et al. (2023) suggest that it's possible that future-bias and regret-aversion are both rational individually, but the combination is irrational. Dorsey (2016, 2021) argues that *Weak No Regrets* is at best a useful heuristic and should be rejected if future-bias is rational. Bnefsi (2023) argues that once one carefully distinguishes permanent from temporary regret and also considers the degrees of regret, the resulting, more accurate principle does not sanction Jack's scheduling behaviour.

[4] Despite that I take there to be other good arguments against future-bias. (cf. Ch.2)

side to have that Jack acts for *very good reason*. After all, scheduling the meal on the last day avoids his potential future-bias-induced regret.[1] This seems as good a reason as (say) scheduling on the last day because *that* meal would be most delicious and therefore not regrettable.

This is not to say that these arguments are of no worth. I surmise that some defenders of future-bias may find Jack's behaviour odd, and that may make them feel uneasy. But such oddity need not mean outright irrationality. The circumstance is odd – there is absolutely no other consideration that may favour a certain date. Jack is also an odd fellow. He's future-biased but not near-biased at all. And he somehow manages to reason immaculately from his future-bias and regret-aversion to an ideal plan. (This is a very intricate piece of reasoning. Other case studies involve even weirder set-ups.) In light of all these uncommon features of the set-up, it's not immediately clear, to say the least, whether the oddity of his scheduling behaviour should be seen as an indication of irrationality or instead just an unpredicted but innocent upshot of an unlikely circumstance.

It's worth emphasizing that it's in itself an important insight of these arguments that future-bias can be action-guiding. It undermines a popular thought that future-bias is rationally kosher because it cannot lead to bad consequences (cf. Kauppinen 2018).[2] That said, it's a further question if the said consequences are indeed bad, and also a further question what is to blame for those bad consequences. (In the above case, regret-aversion, or the unfortunate combination of regret-aversion and future-bias, is a candidate culprit.)

To briefly summarize, I have argued in this chapter that the normative standard of near-bias is unmotivated. The pragmatic argument against preference reversal is problematic on multiple grounds. And there is no non-*ad hoc* constraint on preference change that vindicates the normative standard. For similar reasons, it's ill-advised to argue against future-bias on the grounds of structural rationality.

In the upcoming chapters, I will largely leave behind structural rationality. As I see it, the most important question is whether future-bias and near-bias are correct preferences – that is, whether they are well supported by normative reasons. We'll be looking at future- and near-biased preferences as individual attitudes and the worldly features that may speak for or against them.

[1] See also Dorsey (2021: 267-68).
[2] Though I do not endorse that rational criticism must be based on consequences.

At this stage, one might wonder how structural rationality, substantive rationality, and correctness all hang together. One might wonder if there are some entailment relations among these dimensions of rational evaluation. These are huge questions that I don't have conclusive answers. But very briefly, with respect to how substantive rationality relates to correctness, as mentioned in §1.4.1: Substantive rationality hinges on apparent reasons. And apparent reasons are (reasonably) believed contents such that, if true, would be normative reasons. Thus, whether our actual time-biases are substantively rational is partly a matter of what would be normative reasons for time-biases, but also partly an empirical matter – for it depends on what beliefs underwrite our time-biases (and the reasonableness of those beliefs). So, although substantive rationality and correctness are intimately related on a theoretical level, I will not speak directly to the substantive rationality of time-biases.

Regarding how structural rationality relates to substantive rationality, the matter is much more contentious.[1] I'm somewhat attracted to reductionism of structural rationality. This is the view that norms of structural rationality are not *sui generis*, but a free lunch derivative of norms of substantive rationality.[2] Suppose that cyclical preferences are irrational. A reductionist would say that this is *because* for any set of cyclical preferences, at least one of the individual preferences is not adequately supported by the subject's apparent reasons.

I'm unsure if reductionism is true. But if you are a reductionist, then it seems that you have more reason to endorse my arguments in §2.2 and §2.3. That is, you'd have independent reason to believe that there isn't a structural constraint against non-exponential discounting or future-bias. Central to reductionism is *the guarantee thesis*:[3] For any instance of structural irrationality, it's guaranteed that at least one individual attitude in the combination is substantively irrational. Parfit's preference reversal is supposed to be structurally irrational. Must he be substantively irrational? No, I surmise, *even if* future-bias is in fact *incorrect*. This is because I don't think future-bias is necessarily incorrect. I can imagine some feature *F*, that if it were to obtain, would make future-bias correct. And I can imagine Parfit reasonably believing in *F* (though falsely) and appropriately basing his future-bias on this belief. That is, his future-bias can be substantively irrational. So can

[1] See Kiesewetter & Worsnip (2023) for an overview.
[2] Kolodny (2007); cf. Worsnip (2022).
[3] The guarantee thesis is necessary but not sufficient for reductionism. The argument below relies on the guarantee thesis but not the whole package of reductionism. Though see Worsnip (2022: 65-90) for a battery of objections against the guarantee thesis.

his earlier preference, more obviously. And that's just to say that Parfit need not be structurally irrational after all.[1]

All of the above, I concede, is contentious. And I don't hang my case on this argument. But I take that the previous arguments have made a decent case that the criticism of time-biases from structural rationality is not on firm grounds.

[1] The same goes for the hyperbolic discounter, although the relevant feature *F* may be weirder.

3. Arbitrariness and the Metaphysics of Time

3.1. Structural Rationality and Correctness

We have in the previous chapter explored an argument against time-biases that centres on preference reversal. The charge of irrationality there is structural. It says that a near-biased agent with certain discounting functions and a typical future-biased agent undergo preference reversal over time, and that *that pattern* of preference reversal is irrational. The latter claim is not as warranted as it may appear. It's unclear what general principle of structural rationality can be evoked to support it – except for an *ad hoc* one that targets these very time-biases.

But even if time-biased agents are *not* structurally irrational, it may still be that their preferences are *incorrect*. Correctness (or objective rationality; §1.4.1) has individual attitudes as their primary targets. It's a matter of how well certain attitudes are supported by normative reasons (as opposed to apparent reasons). Consider one who prefers the world being destroyed over her finger being scratched. This preference of hers coheres with whatever other preferences she has at the time, and is stable over time – and more generally, does not (together with her other attitudes) violate any norm of structural rationality. Still, something is wrong with this preference, most of us would say: She *ought not to* prefer the destruction of the world. This preference is incorrect. It is not supported by normative reasons.

This chapter will primarily focus on the *correctness* of future- and near-biases. That is, I shall be concerned about the normative reasons for and against time-biases. The charge of incorrectness differs from the charge of structural irrationality. The claim that near-bias is incorrect – which many philosophers hold – condemns *all* near-biases. An exponential discounter with a constant but non-zero discount rate, for instance, is regarded as having incorrect preferences. She just ought not to discount the far-future at all – no matter what her discount function is or how it evolves over time. Likewise, to establish that future-bias is incorrect, one need not look at patterns of preferences over time. Past pain just ought not to be discounted relative to future pain, regardless of whether future-biased agents undergo preference reversal (although they usually do).

One prominent argument against time-biases is framed in terms of *arbitrariness*. Future-bias and near-bias are said to be arbitrary preferences. And arbitrariness falls foul of rationality. As Tim Smartt summarizes:[1]

> According to "the non-arbitrariness principle" one shouldn't allow arbitrary features of a good to influence one's preferences about that good. One kind of feature that's often taken to be arbitrary is the temporal position of the good. For example, other things being equal, non-arbitrariness holds that one shouldn't have different preferences about the prospect of getting a free donut today or next week. To do so would introduce an arbitrary feature into one's preferences, making them irrational. (Smartt 2021: 302)

As I see it, the argument from arbitrariness is best understood as concerning *correctness* rather than *rationality* (structural or non-structural). (This is not just a quibble over the word "rationality".) That is, what's at issue is normative reasons for or against time-biases. This question can be pursued without looking into what time-biased agents happen to believe or whether their beliefs are justified. I clarify this matter in §3.2.

To anticipate, the bulk of this chapter shall be focusing on *future-bias*. Smartt is a time-neutralist who thinks negatively of both near-bias and future-bias. Both kinds of biases are considered arbitrary because differences in "the temporal position of the good" are arbitrary. But lots of philosophers think differently. These hybrid theorists hold that although differences in *temporal distance* are arbitrary, differences in being *past or future* are not. Expectedly, they tend to point to some deep *metaphysical asymmetry* between the past and the future to justify future-bias.[2] I investigate this appeal to the metaphysics of time in §3.3. The verdict is negative: It doesn't seem that *any* metaphysical asymmetry can give rise to the desired *normative asymmetry*. The purported metaphysical asymmetries, so to speak, are not arbitrariness-breakers.

In §3.4, I argue for a more general claim that this whole enterprise of seeking justification for future-bias in the metaphysics of time is misguided. Our intuitions about permissible future-bias

[1] See also Parfit (1984: 123-26), Sullivan (2018: 34-47), Greene & Sullivan (2015: 949-53), Kauppinen (2018: 244), Callender (2021b: 315-18), Dorsey (2021: 254-56).
[2] Prior (1959), Parfit (1984), Lowry and Peterson (2011), Craig (1999), Pearson (2018), Hare (2013), Heathwood (2008).

do not track bias towards the *metaphysical-future*. Rather, they track bias towards what we may call the *personal-future*. If any sort of "future-bias" is worth defending, it's the latter sort, i.e., personal-future-bias. But as one's personal-future may not be the metaphysical-future, metaphysical asymmetries cannot be the desired arbitrariness-breaker. The normative status of personal-time-biases shall be explored in the remaining two chapters.

3.2. Reasons for Preferences

The charge that time-biases are arbitrary, as I understand it, is concerned with the *correctness* of time-biases. That is, it's said that since time-biases are arbitrary, one *ought not* to be time-biased. That said, the relevant discussion has not been sufficiently attentive to the distinction between correctness and *rationality*. Nor is the term-of-art "arbitrariness" adequately elucidated. We need to get some confusion out of the way.

In the context of time-biases, the notion of "arbitrary" preferences and their rational failure are often introduced by cases. In *Reasons and Persons*, Parfit introduces the now-famous case *Future-Tuesday-Indifference* to show a particular way in which attitudes are "irrational" – namely, being arbitrary. He then proceeds to examine whether time-biases are likewise arbitrary. More recently, Meghan Sullivan (2018: 34-47) argues against time-biases *via* the principle of "Non-Arbitrariness". She justifies the principle on the grounds that "it offers the best explanation of why we judge a family of cases as paradigmatically irrational" (2018: 37) – including Parfit's case. The principle of Non-Arbitrariness and Parfit's case are as follows:

> *Non-Arbitrariness:* At any given time, a prudentially rational agent's preferences are insensitive to arbitrary differences. (Sullivan 2018: 36)

> *Future-Tuesday-Indifference:* Throughout every Tuesday [a certain hedonist] cares in the normal way about what is happening to him. But he never cares about possible pains or pleasures on a future Tuesday. Thus he would choose a painful operation on the following Tuesday rather than a much less painful operation on the following Wednesday. This choice *would not be the result of any false beliefs*. This man knows that the operation will be much more painful if it is on Tuesday. Nor does he have false beliefs about personal identity. He agrees that it will be just as much him who

> will be suffering on Tuesday. Nor does he have false beliefs about time. He knows that Tuesday is merely part of a conventional calendar, with an arbitrary name taken from a false religion. *Nor has he any other beliefs* that might help to justify his indifference to pain on future Tuesdays. This indifference is a *bare fact*. When he is planning his future, it is simply true that he always prefers the prospect of great suffering on a Tuesday to the mildest pain on any other day. (Parfit 1984: 124, emphasis mine)[1]

Evidently, something is off about the hedonist. But what exactly goes wrong? And how should we understand "insensitive to arbitrary differences" in Non-Arbitrariness?

Parfit links arbitrariness to lack of reasons: "Preferring the worse of two pains, for no reason, is irrational." (1984: 124) And here is Sullivan's diagnosis of the hedonist: "[H]is preferences vary in arbitrary ways… If the hedonist is indifferent between experiencing pain on Wednesdays and Thursdays, he should likewise be indifferent to whether his pain is scheduled for a Tuesday." (2018: 39) (Note that Parfit and Sullivan use "indifference" differently. Sullivan means *not* that the hedonist should not care about future-Tuesday pain, but that he should not care *whether* pain is future-Tuesday or future-Wednesday, etc.)

Parfit's claim is ambiguous. On one reading, the hedonist is irrational because he has no *apparent* reason to be future-Tuesday-indifferent. He is being *subjectively arbitrary*. This reading seems true to the set-up. His future-Tuesday-indifference is "a bare fact", not based on any (however false or unjustified) belief about the significance of future Tuesdays. His attitude doesn't even make sense from his *own (belief-channelled) perspective*. Sullivan's diagnosis appears to be of a similar sort: Since all the dates are on par *from the perspective of* the hedonist, it's irrational to be exceptional about Tuesdays.[2]

I tend to agree that subjective arbitrariness is irrational. (cf. Street 2009, Smith 2009) But subjective arbitrariness does not seem to be what the time-neutralists have (or should have) in mind. Presumably, time-biases are often *not* arbitrary from the subjective perspectives of time-biased

[1] The hedonist is sensitive to differences in calendar time, so this case does not beg the question against time-biases. (Sure, he is *future*-Tuesday-indifferent. But this perspectival feature doesn't seem essential to the case.)

[2] Or perhaps she is concerned about *structural* rationality (as the conditional indicates). A bit later (2018: 39) she cites Tversky and Kahneman's (1981: 453) claim that "rational choices should satisfy some elementary requirement of consistency and coherence".

subjects. This is especially evident with respect to future-bias. Philosophers and ordinary folks alike, tend to think that there is *something* significant about the past/future-asymmetry (even though they may fail to articulate it properly). At any rate, the issue of subjective arbitrariness is largely an empirical issue. And little concerning the correctness of time-biases can be revealed by investigating subjective arbitrariness. So, I shall instead focus on arbitrariness of an *objective* sort. To say that time-biases are objectively arbitrary is to say that they are incorrect (in a specific way) and not supported by normative reasons.

The future-Tuesday-indifferent hedonist is also being objectively arbitrary, insofar as there is really nothing special about future Tuesdays: Future-Tuesday pain isn't less painful, doesn't bring instrumental benefits, and so on.

To get a better handle on time-biases being objectively arbitrary, we may think of normative reasons for preferences as facts about *differences between alternatives*. More precisely, reason-statements can be in this form:

> The fact that alternative A has attribute f and alternative B has attribute g is a reason for S to prefer A over B ($A \neq B$; $f \neq g$).[1]

We might also stipulate that attributes f and g are "along the same dimension" so that they can be properly compared. Suppose that eating pho is very healthy but eating pizza is not healthy. These two alternatives differ along the health dimension. This difference – that pho is healthier than pizza – is a normative reason for me to prefer having pho rather than pizza. But not all differences are reasons. That "pho" contains fewer letters than "pizza", presumably, is not a reason for me to prefer one way or another. This difference is *arbitrary* in the sense of being *non-reason-giving* – or slightly differently put, *normatively irrelevant*.

[1] Notice that given this way of regimenting reasons for preferences, a reason for preferring A over B *just is* a reason *against* preferring B over A. So, be careful not to double-count reasons. Also, the fact that A and B share one and the same attribute f is not by itself a reason for being indifferent between A and B. No such single fact is a reason for indifference; otherwise we would end up with a plethora of reasons for indifference, outweighing any potential reasons for $A > B$ or $B > A$. Instead, we need a "meta-principle" for indifference which roughly states that one ought to be indifferent when reasons for $A > B$ and reasons for $B > A$ are tied. (cf. Lowry & Peterson 2011, my §5.6) Also, I do not take this way of of regimenting reasons to be the only correct way, or as making any commitments about the nature of reasons (e.g., whether reasons are fundamentally comparative).

The claim that time-biases are arbitrary, then, amounts to this: Mere differences in temporal locations are *non-reason-giving*, i.e., *normatively irrelevant*. Defenders of time-biases deny this. A defender of future-bias, for instance, would say that the fact that one painful operation is *past* but the other is *future* is a reason for preferring the former. The difference between being past or future is reason-giving and normatively relevant, that is.

The claim that the past/future-asymmetry (henceforth, PF-asymmetry) is non-reason-giving is a rather strong one; its denial, rather weak. The bare claim that PF-asymmetry generates *some* reason may not suffice to justify all instances of future-bias one wishes to justify. Many defenders of future-bias take PF-asymmetry to be more than a tie-breaker. Future-bias can be permissible, they think, when the payoff is not equal – as in *My Future and Past Operations*. To determine whether future-bias is permissible in that very case, the reason for future-bias has to be weighed against the countervailing reasons, including a difference in the amount of pain. More generally, the normative status of a preference is determined by the *overall balance* of reasons. It could be the case that although PF-asymmetry gives *some* reason for future-bias, Parfit ought not to be future-biased all-things-considered in that scenario with very unequal payoff. Regardless, I am only concerned about whether PF-asymmetry is reason-giving *at all*. If it can be established that it is utterly normatively irrelevant – as I shall attempt in the next section – then there is no question about weight.

Also, I set aside a potential distinction between *requiring* reason and *permitting* reason (cf. §5.6). Some defenders of future-bias think that future-bias is permissible but not required. To justify that position, they might have to appeal to the requiring/permitting distinction and claim that PF-asymmetry generates permitting reason but not requiring reason. (Alternatively, one might think that whenever future-bias is permissible, it's also required. In that case, the distinction between requiring reason and permitting reason is inessential.) The distinction will become important in Ch.5 where I evaluate time-biases under the *personal-time* conception. But here, under the *metaphysical-time* conception, by non-reason-giving I mean not generating *any* reason – requiring or permitting. If my upcoming arguments are convincing, they should convince the reader that PF-asymmetry just has no normative significance at all (in relation to future-bias). The requiring-or-permitting question does not arise.

Two more things to mention before diving into the metaphysics of time: Although I shall talk much more about future-bias than near-bias, I do not take it for granted that near-bias is

impermissible. As we shall see, some metaphysical theories of time *appear to* suggest that differences in temporal distances – in addition to PF-asymmetry – are reason-giving, although the appearance turns out deceiving.

Finally, as it should be clear, I am assuming non-Humeanism about normative reasons. I take there to be normative reasons "out there" independent from what one happens to prefer or value. In particular, the claim that mere differences in temporal locations are non-reason-giving is universal and attitude-independent. The Humeans may demur. Indeed, it seems that a Humean must claim that if one happens to (*really and intrinsically*) care about *past* pain much more than future pain, then she should be *past*-biased. This seems to me a *reductio ad absurdum* of Humeanism, though I cannot engage in this fundamental divide here.

3.3. Time-Biases and the Metaphysics of Time

On the metaphysical-time conception of time-biases, to be future-biased is to be biased towards the metaphysical-future. This metaphysical-time conception is often taken for granted (but see §3.4). Unsurprisingly, much discussion about future-bias refers to theories about the metaphysics of time. These metaphysical theories of time are mainly structured around two questions. The first concerns whether reality is irreducibly *tensed*. There we find the chasm between the A-theory and the B-theory. The second question concerns the *ontological status* of past, present, and future: Are past, present, and future stuff all real? It's been thought that given certain answers to either question (or both questions), there is some PF-asymmetry that's shown to be reason-giving. I argue to the contrary.

3.3.1. Tenseless Reality

I've mentioned the A-theory and the B-theory in Ch.1, but here is a more detailed picture.[1] According to the A-theory of time, reality is irreducibly tensed. *Being past*, *being present*, and *being future* are all monadic and objective properties. They are intrinsic to the states-of-affairs, objects, etc., that instantiate them. Things constantly change in these properties as the objective present is constantly moving. The B-theorists deny the existence of these monadic and objective properties.

[1] See also Craig (2000) and Zimmerman (2005).

Instead, the tenseless relations *being earlier than*, *being simultaneous with*, and *being later than*, etc., are all there is to temporal reality. Time is rather like space in this respect. Just as *being here* is an observer-dependent matter, there is no objectively privileged present. To say something is present is just to say that it's simultaneous with a designated observer.

Prima facie, the B-theory cannot vindicate time-biases whereas the A-theory looks much more promising. The B-theory seems a non-starter because its *ontology* is not tensed. Differently put, there is simply no PF-asymmetry given the B-theory, except for an observer-dependent one. It seems that the B-theory can at best vindicate some tenseless preferences – say, preferring bad things to be *earlier* rather than *later*. But that's neither future-bias nor near-bias.

But this seems too quick. What about the Mellor/MacBeath response to Author Prior's *Thank Goodness That's Over* argument? Don't David Hugh Mellor and Murray MacBeath show that to vindicate perspectival attitudes, reality does not have to be tensed? I have in §1.2.1 cautioned against lumping together temporal relief and future-bias. Nonetheless, the debate over temporal relief appears relevant here, as it concerns how perspectival attitudes relate to temporal reality. So, let's revisit Prior's argument for the A-theory. Imagine that Prior had a root canal on June 15, 1954 and was relieved when the surgery was over:

> *Thank Goodness That's Over:* One says, e.g., "Thank goodness that's over!", and not only is this, when said, quite clear without any date appended, but it says something which it is impossible that any use of a tenseless copula with a date should convey. It certainly doesn't mean the same as, e.g., "Thank goodness the date of the conclusion of that thing is Friday, June 15, 1954", even if it be said then. (Nor, for that matter, does it mean "Thank goodness the conclusion of that thing is contemporaneous with this utterance". Why should anyone thank goodness for that?) (Prior 1959: 17)

Prior's contention is that temporal relief is appropriate (if and) only if the A-theory is true. And since temporal relief is obviously appropriate, the A-theory is true. Although this argument is sometimes interpreted as an argument from *future-bias*[1], I consider that a mistake. As I've argued, the relation between temporal relief and future-bias is tenuous. First, the satisfaction thesis seems

[1] Craig (2000), Tarsney (2017), Greene & Sullivan (2015), Fernandes (2021), Maclaurin & Dyke (2002), Suhler & Callender (2012).

false. That is, temporal relief does not have to be caused by, or express, a future-biased preference coming to be satisfied. Moreover, even if the satisfaction thesis is true, one cannot deduce from temporal relief being appropriate to future-bias being correct, since it might well be true that the appropriateness-condition of relief concerns the satisfaction of the preference (which may be in itself incorrect) but the correctness-condition of preference concerns normative reasons for and against the preference. That said, the debate over temporal relief is instructive for our purposes.

Prior's argument involves an important (suppressed) step from tensed representation to tensed reality. The now-standard B-theoretic response by Mellor (1998; cf.1981) and MacBeath (1983) concedes that tensed *representation* is indeed necessary for making temporal relief appropriate. What's denied is that tensed representation entails tensed reality. These B-theorists concede that the representation content of Prior's relief involves a monadic tensed property: "The root canal is *past*!" But this tensed representation has a tenseless truth-maker. It's true iff the root canal concludes before this very utterance. It has a token-reflexive truth-condition. Temporal relief is thus seen as an instance of the phenomenon of "essential indexicals" (Perry 1979). "Past" in representation functions much like the indexical "I". Such indexicals are cognitively and motivationally significant. But just as it doesn't follow that there is some monadic property I-ness in reality, nor does it follow that reality is tensed.

Whether this response succeeds in the case of temporal relief is up for debate. But suppose that the B-theorists can accommodate irreducibly tensed representation. Will that help with time-biases? No, you might think. Because the B-theoretic *reality* remains tenseless. And normative reasons – as opposed to motivational reasons – are given by what's "out there" instead of our representations thereof. Either I have a normative reason to admire Hesperus (aka. Phosphorus) or I don't. Representing the heavenly entity in different guises may affect my motivation to admire, but the normative reasons are not subject to such vicissitudes. Likewise, though it may be true that I can only come to be future-biased by representing states-of-affairs as past and future, such tensed representations do not generate normative reasons over and beyond what the tenseless reality can provide.[1]

[1] Here I echo Pearson's (2018: 23-30) objection to the Mellor/MacBeath response on the grounds that they confuse motivational reason with normative reason. Prior's relief, according to Pearson, is correct only if it tracks monadic temporal properties in reality as opposed to his temporal representation.

Yet again, this is too quick. Normative reasons, though tenseless given the B-theory, can be perspectival in a more innocuous manner. Consider a form of egoism according to which everyone ought to be utterly self-serving. This is presumably false, but clearly not incoherent. This is because it's conceptually innocent to posit *agent-relative* reasons (Nagel 1974).[1] The egoists are committed to reasons of the following form: The fact that φ promotes the wellbeing of *Nagel* is a reason for *Nagel* to φ. This is a reason for Nagel, but not for anyone else. (Though Parfit may have a parallel reason to serve *himself*.) The B-theorists can likewise allow for *time-relative* reasons. They may say something like: That the longer operation is earlier-than-*Thursday* and the shorter operation is later-than-*Thursday* is a reason for one to, *on Thursday*, prefer the former over the latter. This reason does not apply to (say) Monday, though the tenseless relations are the same when it's Monday. So, it seems that the B-theorists can make sense of normative reasons relative to temporal perspectives.

But are the B-theorists warranted to posit such time-relative reasons? I think not. It's certainly true that the two alternatives stand in different earlier-than/later-than relations to Thursday. But why should this difference be reason-giving? On the B-theory, time is much like space. It would be weird to think that Buridan's ass, facing two identical piles of hay - one to its left and one to its right – has any reason to prefer one over another. It's even weirder to think that if the ass turns around so that what was left becomes right and *vice versa*, its preference should reverse accordingly.

Granted, the B-theorists do not literally spatialize time, so the analogy is imperfect. Importantly, unlike space, time nonetheless has a *direction*. PF-asymmetry, on the B-theory, can be seen as an asymmetry between being *earlier* and being *later* relative to certain perspectives. And it's not as if the occupant of the temporal perspective can simply turn around in time and shift the direction. We can grant such a directioned PF-asymmetry. But is it reason-giving?

This may depend on how the B-theorists understand the direction of time. A B-theorist may take direction to be *content-neutral* in the sense that it is a metaphysical add-on not associated with the happenings *in time*. It neither gives rise to, nor is reducible to, asymmetries in causation, memory, deliberation, knowledge, entropy, etc. If that's how direction is understood – a frivolous add-on – it's difficult to see its normative relevance. Suppose that God declares that space is directed like

[1] How best to regiment agent-relative reason-statements is a fraught issue. (Ridge 2023) Here I opt for a rough formulation that resembles Nagel's. More on this in Ch.5.

that: The direction points to where heaven is, say. Do we then have reason to care about that direction? Does the ass then have reason to prefer the stack (say) closer to heaven? Well, no, unless caring about that direction somehow helps us to get to heaven (by making God happy). This points to an alternative understanding of direction. The B-theorists may instead think that the directioned PF-asymmetry is determined by – or is what determines – the content asymmetries in causation, memory, deliberation, knowledge, entropy, etc.[1] If PF-asymmetry is not content-neutral, then the question becomes whether *those* asymmetries in content are reason-giving.

This question has to be handled carefully. I argue that there is indeed some sense in which those asymmetries may be normatively relevant – but not in the desired sense.

Consider first the entropy asymmetry. At least in our local (temporal and spatial) environments, entropy always increases in time, and that's why we never see medium-rare steaks turning rare, or spilled milk unspilled. Is this asymmetry reason-giving? Well, it's true that we might have reason to prefer medium-rare steaks over rare steaks, or to prefer milk unspilled. But these reasons are all given by the particular differences between the alternatives instead of the entropy asymmetry *per se*. Put differently, it's obviously false that we always have reason to prefer higher-entropy states over lower-entropy or *vice versa*. My preference for medium-rare steaks (with higher entropy) and for unspilled milk (with lower entropy) testifies to that.

Consider then the memory asymmetry: We remember the past but not the future. The future can only be, to some extent, anticipated and predicted. I have discussed memory and anticipation in Ch.1. They can be reason-giving by way of conferring *external utilities*. For instance, one might have reason to be *merely apparently* future-biased because remembering a terrible experience is not as distressing as anticipating the experience. But these external utilities concern different temporal spans than the experience itself and can be acknowledged from an atemporal perspective. Rather than providing reasons for genuine future-bias, they provide mundane reasons for preferring less distressing experiences.

[1] I make no assumption about which of these asymmetries are more fundamental (in case they determine the direction of time). It might be the case that there is a "master asymmetry" (probably the entropy asymmetry) that gives rise to all the other asymmetries. Or perhaps these asymmetries are somewhat independent. At any rate, I proceed by considering each asymmetry in isolation. That leaves open the possibility that these asymmetries, taken together, somehow possess reason-giving force which each of them individually lacks. But that possibility seems very remote to me.

Then, the knowledge asymmetry. Partly due to the memory asymmetry, our knowledge about the future is much patchier and flimsier than our knowledge of the past. The knowledge asymmetry is not a universal one. It's normatively relevant only to the extent that it justifies *probability discounting* – which is only contingently and non-universally associated with temporal locations. Yet again, probability discounting is not genuine future-bias.

Finally, the asymmetries in causation and deliberation. I deal with these two asymmetries together because as I see it, these two asymmetries are only relevant when they point to an asymmetry in *causal control*. (It's hard to see why one should care about the differences between causes and effects that are out of one's own causal reach. And deliberation makes sense only when it's potentially causally effective.) Yet again, causal control seems normatively relevant only to the extent of affecting the probability of events occurring. But probability-discounting is not time-discounting. Moreover, in case Parfit can control his future fate, presumably he has more reason to prefer the *future* operation because that's something he can *prevent* or *mitigate* (setting aside considerations about long-term health benefits).[1] (Even here, he only has reason to be merely apparently rather than genuinely past-biased. Again, probability-discounting is not time-discounting.)

In all, it's safe to conclude that the B-theory cannot vindicate future-bias. Granted, there is a PF-asymmetry on the B-theory – one consisting of how events stand in earlier-than/later-than relations to the indexical "now". But this asymmetry, if stripped away of content, seems utterly normatively irrelevant. On the other hand, when we look into the more palpable asymmetries associated with PF-asymmetry, we see them generating reasons not for genuine future-bias, but reasons concerning features of the alternatives that are only contingently associated with time (such as probabilities and external utilities).

I should also point out that the asymmetries in entropy, causation, etc. are not peculiar to the B-theory. Everyone acknowledges their existence (though they may offer different explanations). Moving onto other theories of time, I shall set aside those asymmetries and consider only the PF-asymmetries distinctive of these theories.

[1] This resonates with the evolutionary explanation of future-bias in terms of the affective asymmetry – which in turn arises from the epistemic asymmetry and the causal asymmetry. I suggested in §1.4.1 that although there might be *state-given* reason for *being dispositioned to be* future-biased (as it's in general survival enhancing), that doesn't make future-bias *correct*.

3.3.2. Tensed Reality and Unreality

Turn now to the A-theory. I happen to think that the A-theory is false, but let's set that aside. The A-theory does seem more promising since it acknowledges an objective and observer-independent PF-asymmetry. In this tensed reality, it's an objective matter which things are past, present, or future. And these tensed facts change as the present marches on. Accordingly, the normative reasons for future-bias on offer are irreducibly tensed. These reasons are constantly being lost and gained. When Monday *was present*, Parfit had no reason to prefer the Wednesday operation over the Friday operation because they were both *future*. But as it *is now* Thursday, Parfit has a reason to prefer the Wednesday operation because it's *past*. Or so the A-theorists would say. On reflection, however, it doesn't seem that the A-theory has any advantage over the B-theory. All it adds to the B-theoretic reality is those observer-independent tensed properties that are in constant flux. PF-asymmetry thus boils down to the fact that "the present advances towards future events, while drawing away from past events" (Yehezkel 2013: 73). That is, past states-of-affairs *were* once present while future states-of-affairs *will be* present. But if that's the end of the explanation, we should be left wondering: Why is there any reason to prefer *was-once-present* pain over *will-be-present* pain? Why not the other way around?[1] Indeed, the purported explanation is no more informative than the mere statement that there is reason to prefer past pain over future pain. Given the A-theory, being past just *means was-once-present,* and being future just *means will-be-present*. As Gal Yehezkel puts pointedly, "[the] attempt to justify the asymmetry between past and future based on the flow of time *per se* thus seems to collapse into triviality" (2013: 73). And by the same token, no reason for near-bias seems forthcoming. It is an utterly uninformative "justification" that there is reason to prefer *will-be-present-in-two-hours* pain over *will-be-present-in-one-hour* pain.

But perhaps we have been looking into the wrong faultline in the metaphysics of time. In addition to whether reality is tensed, there is also the question about the *existence* (or *reality*) of the past, the present, and the future. The distinction between being existent and non-existent does seem a deep one. Although the question of tense and the question of existence are distinct, answers to these two questions can be combined, resulting in more comprehensive temporal metaphysics. Hopefully, some normative asymmetry shall emerge from these combinations.

[1] Also see Gallois (1994: 56), Miller (2021: 43), Yehezkel (2013: 73), Karhu (2023: 878), Hare (2013: 510) for this objection.

The B-theorists invariably endorse *eternalism*: Past, present, and future objects, states-of-affairs, etc. *all exist*.[1] Combining the B-theory and eternalism results in *the block theory*. As a B-theoretic metaphysics, the block theory makes no observer-independent distinction between the past and the future. Moreover, there is no change in what exists; everything exists eternally. Past pain is thus as real as future pain, just as what's to your left is as real as what's to your right. The block theory is as unfriendly to time-biases as you can get. If you agree that the B-theory cannot vindicate future-bias, then you'd certainly agree that adding eternalism is of no help.[2]

The A-theorists, on the other hand, are divided among presentism, the moving spotlight theory, and the growing block theory. *The moving spotlight theory* combines the A-theory with eternalism: Past, present, and future things all exist; but different things get "lit up" as the spotlight (i.e., the present) progressively moves forward. Being eternalist, this theory fares no better than the block theory. We have seen that the A-theory by itself cannot offer a normative asymmetry any more than the B-theory. Nothing remotely reason-giving can emerge after adding eternalism.

Moving on to *presentism*. The presentists hold that *only* the present exists. As the present marches on, the totality of existence is constantly changing. Thus, as of Thursday, pain on Wednesday does not exist. But nor does pain on Friday. *Prima facie*, presentism lacks an PF-asymmetry regarding existence. But perhaps as A-theorists, the presentists can say more: Past pain *once existed* (but no longer will); future pain *will exist*. However, this seems to again collapse into triviality. It's merely stating what presentism means – the same sort of triviality as in the bare A-theoretic claim that past pain was once present but future pain will be present. It remains a mystery why once-existed pain matters less than will-exist pain.[3]

---

[1] For general discussions concerning eternalism and its rivals see Merricks (2006), Craig (2006), and Miller (2013). Note that "existence" here is used in an *unresisted* sense. The question about existence concerns the totality of reality as opposed to existence at times. There is emerging skepticism, however, about the substantiveness of the eternalism/non-eternalism debate. It has been argued (e.g., Sider 2006) that the debate is merely verbal: The rivals are using different languages (in particular, different "existence"-concepts) to talk about the same ontology. If that's true, then it would be unsurprising that the non-eternalists views would fail to vindicate time-biases. Conversely, if the non-eternalists fails to vindicate time-biases, this can be taken as *prima facie* evidence that the debate is non-substantive.

[2] Cockburn (1998) and Hare (2007: 361) suggest that the block theory entails time-neutralism.

[3] Though it may seem that presentism can justify *present*-bias. I doubt anyone would like to defend present-bias *only* (rather than as a limiting case of future-bias or near-bias). That aside, I don't think it can justify present-bias – see below on the shrinking block.

Consider then *the growing block*: The totality of reality progressively grows in size as the present marches on, and the present marks the edge of the block that's succeeded by nothing. The growing block forward-lookingly resembles presentism but backward-lookingly resembles eternalism. The past *exists* but the future *does not* (though *will*) exist. So, there is indeed a deep PF-asymmetry on the growing block – one concerning existence. However, *if anything*, this asymmetry seems to get things backward: "For *prima facie,* it is real events which should concern us, rather than unreal events." (Yehezkel 2013: 73)[1] It makes some intuitive sense to say that we should prefer pain to be non-existent rather than existent. But it makes no sense that we should prefer pain to be existent rather than non-existent. If anything, the growing block seems to vindicate past-bias rather than future-bias.[2] (Not saying that this is indeed true; see below on the shrinking block.)

This brings us to *the shrinking block theory*. The shrinking block is the mirror image of the growing block. On the shrinking block, the future and the present exist but the past doesn't exist: "The present is the constantly eroding edge of the future, which thus shrinks incessantly. There is less and less future as the present proceeds." (Casati & Torrengo 2011: 240) No one seriously defends this view. But Casati and Torrengo (2011) suggest that it's not as absurd as it may seem. And it appears congenial to future-bias (Hare 2013: 511). That is, if our world is a shrinking block, then there is a PF-asymmetry between the non-existent past and the existent future. And this asymmetry seems reason-giving in the right way: There is reason to prefer pain to be past rather than future because this amounts to preferring pain to be non-existent rather than existent. Likewise for pleasure: Future pleasure is existent and therefore preferable.

---

[1] See also Maclaurin & Dyke (2002: 286)

[2] On a growing block, the past and the present, though both existent, may nonetheless differ in important respects. Dean Zimmerman, a presentist, considers a response on behalf of the growing-blockers:

> A headache is only *truly* painful when it is present; yesterday's headache, although it exists, is no longer painful. It has a past-oriented property, *having been painful* – a sort of backwards-looking relation to the property *being painful*. But actually, being painful is a matter of simply having the property itself, not standing in some other relation to it; and that's why it no longer concerns us. (2005: 215, original emphasis)

Similarly, some growing block theorists, in response to the epistemic challenge that a growing block resident cannot know it's *now* now rather than past, embrace the "real but dead past" thesis (Forrest 2004). On this view, past state-of-affairs are not really *happening* because activities can only take place on the edge of the block. Thus, as of now, past pain is (tenselessly) existent but *not really painful*. My past self is much like a zombie.

If past pain is not really painful, then it may seem that one ought to prefer pain to be past rather than present. Even so, this cannot justify future-bias (except for, perhaps, the limiting case of present-bias). The "dead past" growing block is relevantly similar to a presentist world. Past pain was painful and future pain will be painful, but neither *is* painful. Again, there seems to be no normative asymmetry here.

But this is still an unsatisfactory response. The problem is that the existence/non-existence (or real/unreal) distinction along the temporal dimension is a peculiar one the normative significance of which we don't have any intuitive grasp.[1] I can understand the difference between existent *past* pain and non-existent *past* pain and its normative relevance: I have reason to prefer past pain to be non-existent in the sense that it never occurred.[2] I can also understand how the distinction between real coffee and ersatz coffee, or that between real Santa Claus and fictional Santa Claus, can be reason-giving: Real coffee tastes better, and it would be nice if Santa were real. But the PF-asymmetry on the shrinking block is nothing like those distinctions. It's a distinction that cuts across all the like-sounding distinctions we are familiar with, and one that's valid only when marking the past from the future (and the present). To put it differently, the normative significance of future pain being *real* is very obscure due to a lack of sensible *contrast class*. (Dorato 2006, Norton 2015) It's real *not* in the sense of being not ersatz or not counterfeit or not fictional. Indeed, the only contrast class one can point to is "being unreal" *as the past is*.[3] But if that's all the shrinking blockers can say, then the explanation again collapses into triviality: It's just saying that past pain doesn't matter *because it's past*. Merely attaching the label "is unreal" to the past doesn't help; we are still left wondering about the normative relevance of being unreal.

The shrinking blockist explanation, as things stand, remains uninformative and mysterious. I don't want to declare this is where the explanation must unfortunately stop. Perhaps there are things that can be said but haven't been said yet. But I'm pessimistic about a satisfactory answer. Also, past pain *was once existent*. But tenses do not spell magic, as shown by the above discussions about the other A-theoretic metaphysics.

One quick note on near-bias before turning to the last candidate. I've been focusing on different ways to spell out PF-asymmetry. PF-asymmetry doesn't have to be all-or-nothing. Quentin Smith (2002), for example, defends "degree presentism": Present things exist to the fullest extent, and the extent to which past and future things exist are inversely related to their distance from the present. (cf. Parfit 1984: 181) Degree presentism would vindicate near-bias if the shrinking block vindicates

[1] This gestures to the above scepticism (86, fn.1) that the dispute here is merely verbal.
[2] This need not be inconsistent with future-bias (even absolute future-bias) in light of that preferences are essentially comparative. I may regard past pain as irrelevant when compared with future pain but nonetheless consider past pain as relevant when timing is not an issue. There is nothing obviously incoherent about this.
[3] Dorato (2006) and Norton (2015) argue that the growing block and the shrinking block are metaphysically equivalent: They differ only with respect to how the label "is real" is applied, but the label doesn't really mean anything substantive.

future-bias. Moreover, the shrinking blockers can in principle adopt gradable existence for the future and thereby also vindicate near-bias. But as I've argued, it's not enough to have a neat mapping of "formal structures". A satisfactory story is still wanting regarding how normative reasons should emerge from such formal structures.

Finally, there is the *open future versus* the fixed past. The future is full of open possibilities but the past is over and done with. No point in crying over spilled milk. But be more careful so that you don't spill it tomorrow morning. While the open future intuition is widely shared, it proves extremely difficult to pin down its exact content.[1] For some philosophers, it just amounts to the knowledge asymmetry or the asymmetry in causal control, or the thesis that the future doesn't exist but the past exists (i.e., the growing block). Setting aside these suggestions, there aren't many alternative proposals of its content available.[2] While the details differ, these remaining proposals share a common feature: The future, but not the past, is in some sense *underdetermined*[3]. Importantly, the sense of underdeterminacy here had better not be identical with what's at stake in the debate of nomological determinism/indeterminism – for presumably, whether the world is deterministic or not, it goes in both directions. Here is an impressionist way to make concrete the idea of underdeterminacy. Imagine a model of *the branching future*: The temporal structure of the world is like a tree with a trunk and multiple branches (and branches on branches). The present is the end of the trunk after which the branching begins. The multiple future possibilities are represented by the branches. As the present marches on, the branches are trimmed away *except for one* – which becomes part of the trunk (i.e., part of the fixed past).[4] (cf. MacFarlane 2003: 325-26, Barnes & Cameron 2008)

There are genuine questions regarding how to understand such a model. People disagree over whether the branching future entails nomological indeterminism (or determinism!) and in which

[1] See MacFarlane (2003), Barnes & Cameron (2008), and Torre (2011) for overviews.
[2] I also set aside equating the open future with the failure of bivalence, for two reasons. Firstly, as pointed out by Barnes & Cameron (2008) and Torre (2011), one should not tie the open future with a controversial semantic thesis, though the underlying intuition may lend some support to this semantic thesis. Secondly and relatedly, as a semantic thesis, the failure of bivalence tells little about the metaphysics. If anything, it's the underlying metaphysics that's normatively informative.
[3] Barnes & Cameron (2008) would dislike this word. They'd say that whereas the growing block has an underdetermined future, a branching future is *over*determined. This disagreement doesn't matter much for my purpose. I just need a word different from "indeterminacy" so that the view doesn't sound as if it's just nomological indeterminism.
[4] My sense is that this model is A-theory/B-theory-indifferent. The B-theorist may understand the "present" indexically so that the "trimming away" of branches are not objective but observer-dependent.

sense the future is underdetermined. But to give the open future its best shot, let's say that nomological *indeterminism* is true so that the future is genuinely full of open possibilities. (The past *was once* full of possibilities, but only one possibility has been actualized.) Intuitively, this model is different from everything else we've canvassed. The model also seems to force a distinction between *being real* and *being existent*: The future branches exist but are *not yet* real; only one of them *will be* real. And there is indeed a deep PF-asymmetry: The past is a single actualized possibility, but the future is full of open possibilities.

I have tried hard to spell out the open future in an intuitive way. But here is the disappointment. For one, the triviality problem seems to re-emerge: Why does it matter that something is real (already actualized) rather than unreal (to be actualized or trimmed away)? The answer had better not be that reality is the fixed past but unreality is the open future. For another, if anything, the model seems to get things backward (like the growing block). If all these future branches are genuinely open possibilities that are on par, and it's as of now (Wednesday) underdetermined if Parfit is about to undergo the operation on Friday, then it seems his concern for the future operation should be "dispersed" rather than "intensified" relative to his past-directed concern. After all, there is the genuine possibility that he shall be pain-free on some branches not yet trimmed away. (Again, this sounds like probability discounting.) Finally, there is the worry that this model – though *looking* different from the other proposals canvassed above – is not really saying anything distinct. This is indeed a recurring worry in the open future literature. According to some, the asymmetry either collapses into some other asymmetries or is barely coherent. At least, since no informative explanation has been given, there is no good reason to think that the open future justifies future-bias.

I wish to conclude that PF-asymmetry is non-reason-giving. None of the salient ways to specify the asymmetry sounds promising. Although this doesn't prove that no alternative proposal is possible, the dialectics so far should warrant a pessimistic induction. There is little reason to expect some other obscure asymmetries to save the day.

It's a recurring theme that justification for future-bias from PF-asymmetry collapses into triviality. Chris Heathwood acknowledges this, but he also thinks that no non-trivial answer is ever needed: "Why is it preferable for a pain to be over and done with? I'm afraid the only answer may be: it just

is." (2008: 61)[1] Is it legitimate to posit a *brute fact* here? It's as painful to argue against a brute fact as to posit it. It's true that explanation must end somewhere. I can only let the reader decide if it should end here.

There is a worse problem with this brute fact, however. If Heathwood has in mind (as it seems) the brute fact that there is always reason to prefer pain to be in the *metaphysical past*, he would make wrong predictions. Indeed, all attempts to vindicate time-biases by appealing to features of the metaphysics of time, brutish or not, seem misguided to me. It doesn't seem that our time-biases *actually track* metaphysical-time. Nor does it seem that time-biases *ought to track* to metaphysical-time.

3.4. The Irrelevance of the Metaphysics of Time

3.4.1. Normative Variation and the Metaphysics of Time

The discussion so far brings bad news for future-bias. Although there is apparently some deep metaphysical PF-asymmetry, it's hard to see its normative relevance. (Accordingly, there is bad news for near-bias.) There is more bad news: Any proposal that PF-asymmetry is reason-giving (brutish or not), if it proves anything, would prove *too much*. This is the argument from *normative variation*. The rough idea is that if PF-asymmetry is reason-giving, it counts in favor of future-bias *across the board*. But future-bias is not permissible across the board: The normative status of future-bias varies, despite that PF-asymmetry is always there. In particular, it's suggested that future-bias is *impermissible* when it's other-directed and when it's about non-hedonic events. This argument from normative variation, if successful, is powerful. It would show that PF-asymmetry, however specified, *cannot* be what's reason-giving. It would foreclose any potential proposal of that sort.

The argument gets something right. PF-asymmetry would indeed prove too much (but also too little) – this is what I shall argue for in the next section. However, I am not convinced that we can

[1] Caspar Hare concurs: "[I]f there are normative facts concerning future-bias they are brutish facts – facts that philosophy is ill-suited to uncover." (2013: 519)

get that conclusion from normative variation in the above sense. I find reasons for believing in normative variation wanting.

It has long been hypothesized that our actual pattern of future-bias is not uniform. Future-bias seems present under certain conditions but absent under some others. Parfit, for one, suggests that our future-bias exhibits a *first-/third-person asymmetry*: "[T]here is this difference between our attitudes to past suffering in our own lives and in the lives of those we loved." (1984: 184) Many other authors concur.[1] We tend to be future-biased about our own weal and woe but time-neutral when it comes to other people (including loved ones). Or so it's hypothesized.

One other salient asymmetry hypothesized is the *hedonic/non-hedonic asymmetry*: Our future-bias is confined to hedonic states-of-affairs.[2] Consider *My Future and Past Betrayals*: "[E]ither some friends of yours have betrayed you behind your back nine times in the past or some friend will betray you behind your back once in the future." (Brueckner & Fischer 1986: 9) Being betrayed is considered non-hedonic because you may be betrayed without ever knowing it and without suffering any emotional damage from it. Then there is the asymmetry: Probably you would rather have one instance of future betrayal. But the set-up of the case is otherwise identical to *My Future and Past Operations*. More in this regard: If there is equal hedonic payoff (same amount of future and past pain), you'd very probably be future-biased. But when it comes to the same amount of past and future betrayal, it makes sense for you to be indifferent. The same seems true of other kinds of non-hedonic events – disgrace, being cheated upon, academic achievement, etc.

So, there are two purported *descriptive* asymmetries: The first-/third-person asymmetry and the hedonic/non-hedonic asymmetry. If these descriptive asymmetries obtain, then it's just a short step towards *normative variation*: Future-bias, though permissible (or even obligatory) regarding the first-person hedonic, is impermissible regarding the third-person or the non-hedonic. After all, defenders of future-bias take seriously empirical facts about future-bias. Future-bias, as a deeply entrenched feature of human psychology, is considered innocent until proven guilty. But if future-bias is absent under certain conditions, it's only natural to assume that future-bias isn't permissible under these conditions, absent good reason to think otherwise. As Alison Fernandes notes, defenders of future-bias work with the presumption that our actual future-bias "is best

[1] Brink (2011), Hare (2008, 2013), Greene & Sullivan (2015).
[2] Brueckner & Fischer (1986), Brink (2011), Hare (2013), Dougherty (2015).

explained by its being justified" (2021: 4003). Moreover, proponents of the descriptive asymmetries also report aligning intuitions about normative status.[1] Time-neutrality about the third-person and about the non-hedonic, just as future-bias about the first-person hedonic, is what they find intuitively correct. And since normative intuitions are taken as important data for first-person hedonic future-bias being in fact permissible – occasionally endowed with a Moorean status (Heathwood 2008: 61) – it's only natural that the opposite intuitions about the third-person and the hedonic should carry the same weight.

The normative variation thesis, if true, is worrisome. Alison Fernandes (2021) argues from this against an A-theoretic justification of future-bias.[2] But the problem is more general. It would seem that *any* proposal that PF-asymmetry is reason-giving would be self-sabotaging. For instance, on the shrinking block, past betrayals are just as non-existent as past pain, and your past pain is just as non-existent as my past pain. If this asymmetry were reason-giving, it would give us too many reasons.

But there is no compelling reason to believe in these descriptive asymmetries. Nor would normative variation straightforwardly follow from the descriptive asymmetries. I argue for both claims in the rest of the section. (My aim is to challenge the argument from normative variation and is therefore relatively modest. I do not aspire to definitively prove that there are no descriptive asymmetries or that normative variation is false.)

Let us begin with the hedonic/non-hedonic asymmetry. First, philosophers' prediction is not confirmed by empirical studies. Greene et al. (2021a, 2022b) find that although non-hedonic future-bias is a tad less prevalent and pronounced than hedonic future-bias, there isn't an obvious

[1] Parfit (1984: 181-84), Kauppinen (2018), Nguyen (2022), Hare (2008, 2013), Stokes (2016).
[2] Fernandes focuses on the first-/third-person asymmetry. In addition, she argues that the A-theory would only justify absolute future-bias, which is inconsistent with our being non-absolutely future-biased. I cannot engage with that argument in detail but I wish to point out two issues. Firstly, she relies on data of the TVA where "future-bias" is indeed rather weak. But when it comes to *future-bias*, we do seem to be strongly future-biased (though perhaps not absolutely). (cf. §1.2.2) Secondly, the A-theorists may adopt a view on which pastness comes in degrees. (cf. Smith 2002)

asymmetry.[1] Although the existing evidence is rather limited and somewhat mixed, it should at least cast some doubt on the descriptive asymmetry.[2]

But there is a deeper worry. I think that we should not put too much stock in the empirical data – even if they appear to show a robust asymmetry. That's because due to the nature of non-hedonic events, it's unclear what our temporal preferences would be *about* in non-hedonic conditions. Nor is it clear what our normative intuitions are latching onto. Consider a paradigm example of non-hedonic time-neutrality:

> *Past or Future Infidelity:* I learn that my wife plans to be unfaithful to me, for the first time, this week. I very much do not want this to happen. But, being a non-confrontational sort of a fellow, I decide not to try to prevent it from happening. I isolate myself from her – I retreat to Yorkshire and work on some philosophy. Time passes. Has the dread event occurred? I do not know. (Hare 2013: 508)

Caspar Hare finds himself time-neutral: "I would not in any way care about whether her infidelity was in the immediate past or the immediate future." (ibid.) But we need to pause and ask: What's so bad about being cheated upon? And where is such badness temporally located?

One major badness of infidelity is its *instrumental* badness. Infidelity leads to anger, jealousy, quarrels, constant suspicion, a relationship forever tainted, and sometimes even divorce. In comparison, the intrinsic badness of being betrayed pales into insignificance. (In fact, I personally don't see any intrinsic badness in betrayal.) Importantly, these instrumental bads spread out into the *future*. Insofar as Hare cares about these things, his "time-neutrality" makes perfect sense. But this attitude is not so much giving equal weight to past/future infidelities, but is rather (reasonably) expressing pronounced concern for how the spousal relationship goes forward. Dale Dorsey makes the same point about disgrace[3]: "The bulk of the *overall* disvalue of disgraces is surely their

[1] There is also empirical study exploring different types of hedonic goods. Latham, Miller, Oh, Shpall & Yu (2024) find that future-bias is stable across sensations and moods, and also across different kinds of moods.

[2] As Greene et al. (2022b: 160, fn.16) note, since they only use conditions of equal payoff, it may be the case that our non-hedonic future-bias is very weak. Perhaps it only serves as a tie-breaker but shall disappear when the payoff is unequal. On the other hand, hedonic future-bias is shown to outweigh very unequal payoff. (Greene et al. 2021a, 2021b, 2022a)

[3] Brink (2011: 378) imagines a case of huge past disgrace (obscene drunk behaviour!) *versus* small future faux pas.

instrumental disvalue – lost careers, relationships, reputation, and so on… [T]hese bads will occur in the future *relative to the disgrace itself*." (2016: 357, emphasis in original) This observation seems generally true of all non-hedonic events (positive or negative) we care about.

By instrumental (dis)value, I have in mind mainly temporally local events such as (in the case of infidelity) instances of quarrels, feelings of suspicion, etc. But perhaps that's not yet the whole story. The badness of a relationship being forever tainted, you might think, is not reducible to these temporally local instances. (Velleman 1991, Nguyen 2022: 11-12) Rather, it makes a significant sector of your life a *failure*. Or think about achievement. What's so good about publishing a seminal philosophical paper? The citations, subsequent research fundings, etc., are of course important, but presumably, they are not all there is. Its significance also lies in making your career as a philosopher much more successful and meaningful. Such (dis)values do not seem located at some discrete times. They instead cast your *whole life*, or a great portion of your life, under a (dis)favourable light. A slightly different way to think of this is in terms of *signifying* value (Dorsey 2015). The fact that your paper becomes seminal is (hopefully!) evidence that you are virtuous as a philosopher. But the goodness of being virtuous is obviously not confined to the instance of that paper being printed.[1] All this suggests that our apparent time-neutrality about non-hedonic events cannot be taken at face value. This is because our future-biased tendencies, even if in effect, would have been swamped by considerations of other values that are not temporally local.

There may seem an obvious fix. In order to elicit accurate responses, it might be suggested, that all we need to do is to control away these "confounding" values. It needs to be stipulated, for example, that Hare feels no emotional stress thinking of his wife cheating on him, that he doesn't take this to indicate that he is unattractive, that his marriage goes forward as usual as if nothing untoward has happened, so on and so forth. Perhaps Hare would remain time-neutral after all these are stipulated away. But I wonder what that time-neutral preference would be about. Nothing bad (or good) in infidelity is left, I'm tempted to think. If an instance of infidelity is not felt about and leaves no trace whatsoever, I can hardly make myself care about it. Even if there is some residue intrinsic, temporally confined disvalue, it's unclear if our responses would reveal anything. The "purified" scenario would be too dematerialized and too far removed from reality to evoke valid and reliable responses. This point has been made by Anh-Quân Nguyen:

[1] Much more on temporally local and non-local values in relation to wellbeing in the next chapter.

> We need a sufficient level of grounding to thought experiments for our intuitions to latch onto – if idealised too much, we risk that our intuitions do not form reliable responses, even if we think that we can confidently isolate intuitions from each other and form responses. (2022: 23)

The problem of idealization, of course, is a rather general one. We've encountered the problem with hedonic future-bias. (§1.2) There the major issue is with the (dis)value conferred by states of anticipation and retrospection. For Parfit's case, Brink suggests purification by "administration of a drug that blocks anticipation of future pain" (2011: 379). This scenario perplexes imagination. It's unclear how I could come to care about future pain if I'm unable to anticipate it in the first place. Fortunately, this purifying strategy is not needed (unless we need to pin down the exact discount rate). We know that our hedonic future-bias tends to be strong enough to outweigh the unequal payoff *plus* the potential discomfort of anticipation. When it comes to non-hedonic events, however, there is a deeper conundrum. Their (dis)value is predominantly non-intrinsic and temporally non-local. So, a purified scenario would be barely imaginable. (Purely) non-hedonic time-bias or time-neutrality would be a very alien attitude that evades our understanding.

In all, there is no good reason to believe in the descriptive asymmetry. It's not even clear what it would mean for one to be non-hedonically time-neutral. This in turn casts under a very suspicious light our intuitive judgements about its normative status. As the descriptive content of the targeted attitude remains obscure, our normative intuitions towards it should fail to reliably tell for or against anything.

So, normative variation is unmotivated insofar as non-hedonic events are concerned. Turning to the first-/third-person asymmetry, it's hypothesized that we shift to time-neutrality in response to the weal and woe befalling other people. And this shift is regarded as appropriate. Consider this third-person hedonic scenario envisaged by David Brink:

> My daughter undertakes volunteer work in a remote and largely inaccessible part of the world. I receive a message from someone that travelled through her village that she has a terminal disease that has become quite painful and will soon kill her. I am depressed. When I am told later that this was substantially correct but mistaken about the timing so that my daughter has already died, I feel no relief that her pain is behind her. Yet again, this narrows the scope of the bias. (2011: 379)

I share the time-neutralist intuition. I find myself preferring my daughter still alive and in pain. (Though we should be careful not to identify relief with future-bias.) But this case is cleverly deceiving because death is involved. However you think about death, it's not any garden-variety badness like pain. At least, it makes perfect sense to prefer a loved one to be still alive (but destined to die) rather than already dead.[1] When death is removed from the picture, however, I find myself future-biased. I'd rather that my daughter has already suffered (more) in the past.

My intuition may be in the minority (cf. Dorsey 2016: 367). But here is a plausible explanation for conflicting intuitions here. It seems that we are more inclined to be future-biased when being "imaginatively empathetic" with the target subject. Caspar Hare confesses that he'd be time-neutral if his daughter is far away and out of touch, but future-biased if he's by her side. What explains this shift in attitudes, according to Hare, is not so much proximity or communication *per se*. Rather, it's because proximity and communication influence our capacity "to engage, imaginatively, with their present conditions" (2008: 277). When Hare is with his daughter, he finds himself empathetically adopting his daughter's preference, which is, presumably, a future-biased one.

There is experimental evidence that empathy affects our third-person attitudes. Caruso et al. (2008) observe that the temporal valuation asymmetry virtually disappears in third-person conditions. For example, we tend to assign equal compensation to a third party regardless of whether the harm takes place in the past or in the equidistant future. This is in tension with Greene et al.'s (2021a, 2022b) finding that there is no clear first-/third-person asymmetry in future-bias: We are almost as future-biased regarding other people. Although the TVA need not be identified with future-bias[2], there remains something mysterious – namely, that the TVA is first-/third-person asymmetric but future-bias is (almost) symmetric. Greene et al. hypothesize that this discrepancy has something to do with how detailed and "personal" the target subject's predicament is portrayed. In their vignettes, the target subject is a particular astronaut (Freddie) on a special mission, whereas Caruso et al. (2008: 299) talk about a "randomly selected person from the local area". Again, this difference presumably means different levels of empathetic engagement. Greene et al. (2022b: 241) further predict that under empathy-inducing conditions, should one be told that the

[1] Dorsey (2016: 367) finds the case deceiving for slightly different reasons.
[2] Among other things, people judge the TVA to be inappropriate but future-bias to be appropriate. (Latham, Li, Miller & Yu manuscript; §1.2.2)

target subject is *time-neutral*, one might be time-neutral "on behalf of" her even if one has first-person future-bias.

So, the purported descriptive asymmetry is not robust, to say the least. Furthermore, there is a plausible explanation for this: When there is much empathetic engagement, third-person attitudes tend to align with the (presumed) first-person attitudes of the target. What follows? Does all this speak *for* or *against* normative variation? The question seems to boil down to this: Ought one empathetically engage with the target's predicament and defer to her preference? A positive answer would count against normative variation. If first-person future-bias is correct and third-person attitudes should align with first-person attitudes, then it's hard to see how third-person future-bias could be incorrect.

I'm inclined to think that empathetic engagement is a virtue rather than a vice in this context. Firstly, it's undeniable that empathetic engagement is at least sometimes conducive to rather than antithetical to appropriate attitudes (Hare 2008: 276-77). This is certainly true in some contexts of social and moral decisions. Being exposed to the suffering of farm animals in a visceral manner helps to see that their welfare matters. The same seems to hold in the present context: Empathetic engagement is conducive to revealing what really matters.

Secondly and relatedly, it's not even clear to me that from an utterly emotionally detached perspective, we would be forming genuinely *past-/future-neutral* preferences. Recall the discussion on adequate representation in §1.4.2. As Callie Phillips (2021) argues, insofar as hedonic future-bias is concerned, adequate representation of the alternatives involves (1) representing them as respectively past and future and (2) revealing the hedonic values of the alternatives. The same, of course, applies to genuinely *past-/future-neutral* preferences.[1] But it seems that emotional detachment is an obstacle to registering the temporally embedded perspective as well as imagining from-the-inside the painfulness of suffering. That is, a decent amount of empathetic engagement seems necessary for adequately representing the alternatives, without which the resulting attitudes would fail to be our targeted attitudes. Finally, I see no prior reason to think that the correct first-

[1] You might think that a genuine past-/future-neutral preference does not require perspectival representation. A non-perspectival representation will suffice. I find this implausible. I am *not* being non-egoist by being impartial to Wen and Gwen if I don't know that I am Wen.

person and third-person preferences should come apart here. It's not as if Hare's (or Brink's) interests are somehow in conflict with his daughter's.[1]

In all, the argument from normative variation is unsuccessful. It doesn't provide good reason to give up the claim that PF-asymmetry is always reason-giving. But there is some other reason against this claim. It seems obvious to me that our time-biases *do not* track metaphysical-time. Or at least, time-biases *should not* track metaphysical-time. Instead, our time-biases (roughly) track and should track what we may call *personal-time*.[2] Metaphysical-time and personal-time roughly coincide in ordinary life. So, it may evade immediate observation which of them we are (or should be) tracking. But metaphysical-time and personal-time are conceptually distinct and can materially come apart. Metaphysical-time is *gruesome* insofar as correct preferences are concerned. If that much is true, then there is compelling reason that the metaphysical PF-asymmetry is normatively irrelevant. Instead, what's reason-giving is features concerning personal-time, and they give reasons for biases towards the personal-future and the personal-near. (At least that's what defenders of time-biases should say.) This coheres with and further supports the above arguments (§3.3) that none of the specifications of the metaphysical PF-asymmetry makes it reason-giving. It's just misguided to seek justification for future-bias in the metaphysics of time.

3.4.2. Metaphysical-Time and Personal-Time

On the personal-time conception of time-biases, time-biases track and should track (roughly) personal-time rather than metaphysical-time. My argument for the personal-time conception is not new. André Gallois (1994), Alison Fernandes (2022), Todd Karhu (2022), and Kristie Miller (2021, 2022) have advanced similar arguments. And the argument is (uncreatively) intuition-based. But I take the intuitions to be very compelling.

Begin with this *backward time-travel* case adapted from Alison Fernandes (2022: 193):[3]

---

[1] Things become more complicated if we think that future-bias should track personal-time (or more precisely, intra-person bonds of concern). Some defenders (Karhu 2022) of this view take this to imply that we should be third-person time-neutral because other people's life is equally not in our personal-future or personal-past. (cf. Bnefsi 2019, Sung 2024) I suggest in Ch.5 that third-person time-neutrality is compatible but not entailed by this view.

[2] As the reader may have realized, personal-time is not really a temporal dimension. It might then be thought that personal-time-bias is not really *time*-bias. Maybe so. Maybe personal-time-bias doesn't deserve the label "time-bias" – only metaphysical-time-bias do. If the reader is concerned about labels, feel free to read personal-time-bias as a different sort of preference.

[3] Which is in turn inspired by a case from Tarsney (2017).

> *Time-Travel Shocks (I):* You have the misfortune to encounter a philosophical fiend. The fiend offers you the following choice. Either (a) You will experience 5 electric shocks in the next five minutes, or, (b) You will experience 10 electric shocks five hours ago – and the fiend will send you *back in time* by five hours using his time machine so that you can *live through and experience* these past shocks (and immediately bring you back to his departure point afterwards). As in Parfit's case, memories of the painful shocks will be wiped clean either way.

What would you choose?[1] Of course, (a). It would seem crazy to choose (b)! But in (b), the 10 shocks are in the metaphysical-past.[2] If you think that future-bias is correct in Parfit's *My Future and Past Operations* and that future-bias tracks *metaphysical-time*, you'd have to say that it's correct for you to choose (b).[3] But no one, I believe, wants to say that.

What does future-bias track, if not metaphysical-time? Roughly, it is what David Lewis (1976) calls *personal-time*.[4] The notion of personal-time is familiar to the time-travel literature. It's a useful concept to make sense of "time discrepancies" in time-travel scenarios. We'll better define personal-time shortly, but for the moment we can rely on our intuitive grasp of the concept. For example, if you choose (b), then you *will* experience 10 shocks despite that the ordeal is *past* from a bystander's perspective. This is a sensible description because the ordeal is in *your personal-future*. And we say it's in your personal-future because (insofar as the trip is non-instantaneous) you will have grown a bit older by the time of arrival, and you anticipate the ordeal but have no (possible) memory of it, so on and so forth. The following scenario makes a positive case for the personal-time conception:

> *Time-Travel Shocks (II):* Upon waking up from a coma, you are informed by the philosophical fiend that one of the following scenarios is true about you. Either (a)

[1] I'm assuming that backward causation is possible. And at least on the standard account of time-travel (Lewis 1976), this doesn't absurdly imply that Jamal can *change* the past. Rather, he can *affect* the past just as he can affect the future. But if you are worried that the causal structure of the case, we might well instead just ask what Jamal would prefer. The intuition should be the same.
[2] I've been using "metaphysical past/future" in a A-/B-theoretical-neutral way. If you're a B-theorist, think of the 10 shocks as metaphysically earlier than *the time of the choice*.
[3] If you think future-bias is correct only if it's relatively weak, tweak the payoff in this scenario.
[4] Lewis calls metaphysical-time "external-time" and *defines* time-travel involving mismatch between personal-time and external-time. But we need not endorse this definition of time-travel. (More on this shortly.)

> You had just experienced 10 electric shocks before you were put in an amnesic coma, or, (b) you *will* experience 5 electric shocks five hours ago – the fiend will send you back in time by five hours using his time machine so that you can live through and experience these past shocks.

Insofar as you are future-biased in ordinary scenarios, I predict, you'd prefer (a) rather than (b). That is, you'd prefer more shocks in your personal-past rather than fewer shocks in your personal-future. Moreover, it seems that you have such a preference for *the very same reason* as your future-bias in ordinary scenarios. Personal-time is what matters to you all along. This is true despite that metaphysical-time may have served you well as a proxy of personal-time, given that they normally (roughly) align.

There is no empirical study about time-travel future-bias yet. So, it may be too rushed to proclaim that *our* future-bias actually tracks personal-time rather than metaphysical-time. But I trust you, the philosopher, in your judgement that future-bias *should* track personal-time rather than metaphysical-time. (This is what defenders of future-bias should say. The time-neutralists, on the other hand, should say that we should be *time-neutral* about personal-time.)

In the above backward time-travel scenarios, metaphysical-time and personal-time come apart both in directions and in metrics (i.e., distances). In the following *forward* time-travel scenario, they come apart only in metrics. I take this case to show that near-bias tracks, or at any rate should track distances in personal-time:

> *Time-Travel Shocks (III):* The fiend offers you the following choice. Either (a) You will experience 5 electric shocks *5 hours later*, or, (b) You will experience 5 electric shocks *10 hours later* (in metaphysical-time) – the fiend will send you onto a *sped-up* time-machine trip that in every respect feels like to you that only *5 minutes* have elapsed – i.e., a trip that takes 5 minutes of your *personal-time*.

You may not be near-biased. And you may judge near-bias to be incorrect. Regardless, you should agree that *if* you're near-biased, then you'd choose (a) rather than (b). You should also agree that *if* you ought to be near-biased, then you ought to choose (a). And this is because by choosing (a), you place the electric shocks in your more distant personal-future (but in a more immediate metaphysical-future) than by choosing (b).

I take these cases to have established that time-biases are really about personal-time rather than metaphysical-time. It's thus unsurprising that however the metaphysical PF-asymmetry is embellished, it fails to be reason-giving. (§3.3) Instead, if future-bias (and near-bias) were ever correct, it must be due to features concerning personal-time.

But not everyone would be convinced. Some would protest against the above time-travel scenarios. Maybe given presentism, it's metaphysically impossible to time-travel (backward or forward). And maybe presentism is necessarily true. So, some would say that the above time-travel scenarios are impossible. But it's unclear if intuitions about impossible scenarios should tell us anything.

There are a host of contentious issues surrounding the possibility of time-travel that I cannot address in detail here.[1] I believe that time-travel is possible on all plausible metaphysics of time (presentism included), but I concede that this may be false. However, personal-time is not a fancy device *ad hoc* for making sense of time-travel. It measures familiar changes and processes we

[1] The standard objection to presentist time-travel is the "nowhere objection": The (absolute) past and future are non-existent, so there is no destination to travel to. In response, Keller & Nelson (2001) argue that presentist time-travel is coherent: It doesn't require the destination to exist *as of now*. Suppose you choose (b) in *Time-Travel Shocks (I)*. Your present departure *makes it true* that it *was* the case that you are experiencing 10 electric shocks (when *that* time was present and existent, that is). The obtaining of such a backward causal relation, according to Keller & Nelson, suffices to make you a time-traveller.

Even if presentist time-travel is possible and it works this way, you might think that our normal prudential concern is disrupted. With alternative (b) in *(I)*, for example, it may seem that one simply cannot care about the metaphysically-past (but personally-future) shocks in a normal, egocentric way. Sider (2005) raises such a worry. He thinks that the time-machine is more like an annihilation machine with respect to our prudential concern. You cannot anticipate the past experience as yours since it makes no sense to say that you *will* experience it. If Sider is correct, then our intuitions regarding these cases should be discredited – for the cases just aren't about prudential concern in any proper sense. (Similar problems can be raised about other non-eternalist views.)

There is one more worry to consider. I've been assuming that time-travel works in a roughly Lewisian way. But some (non-eternalists) think of time-travel differently. I cannot consider all these alternative models here. But insofar as the model acknowledges that personal-time and metaphysical-time come apart in the above time-travel cases, the cases remain relevant. For instance, Bernstein (2021) and van Inwagen (2010) think of A-theoretic time-travel as involving the time-traveller resetting the present. On a growing block, for instance, a backward time-traveller would make her destination the "new" present and thereby *wipe out* the portion of the block between the "old" and the "new" present. The time-traveller thus *changes* the past by deleting and regenerating the block. Time-travelling this way involves something that amounts to mass murder. Setting aside the ethical considerations, however, insofar as your own pain is concerned, this model seems to make no relevant difference. Suppose you choose (b) in *(I)*. At the time of your choice, it's nonetheless true that the shocks await in the metaphysical-past (as you haven't reset the present yet) – which is your personal-future. That said, it's unclear if every model of time-travel involves a discrepancy between metaphysical-time and personal-time.

For a more detailed discussion about how such time-travel cases may or may not work for our purposes, see Miller (2021, 2022).

undergo constantly in everyday life: "First come infantile stages. Last come senile ones. Memories accumulate. Food digests. Hair grows." (Lewis 1976: 146)[1] These processes, even without time-travel, can go more quickly or slowly, can pause and reconvene, and can even go backwards (think of Benjamin Button). (That's why I said that metaphysical-time and personal-time *roughly* align in ordinary life.) Time-travel is thus dispensable for my argument here.

Consider this scenario adopted from Todd Karhu (2022: 880):

> *Short and Long Stases*: The philosophical fiend offers you the following choice. Either (a) You will enter a deep freeze for *one day*, during which all your biological and psychological processes will come to a hiatus. Upon your revival, you will immediately undergo 10 hours of excruciating pain. Or (b) Similar to (a) except that the deep freeze lasts for *100 years*.

All else being equal[2], I'd be indifferent between the two options – despite the fact that I *am* near-biased. Regardless of what I choose, pain will impinge on my consciousness immediately as *per* my personal-time. And just to cement the intuition:

> *Stasis and Business-as-Usual:* The philosophical fiend offers you the following choice. Either (a) You will enter a deep freeze for *100 years*. Upon your revival, you will immediately undergo 10 hours of excruciating pain. Or (b) You live a normal life for *50 years*, and then undergo 10 hours of excruciating pain.

Being near-biased, I'd choose (b) rather than (a). That is, I'd prefer to postpone pain in my personal-time rather than in metaphysical-time. And the staunch time-neutralist should say that I'm making a mistake by virtue of being *near-biased* rather than being far-biased.[3] Again, this shows that near-bias is about personal-time rather than metaphysical-time.

---

[1] Lewis speaks of personal-time as what's measured by the time-traveller's wristwatch (ibid.). But this shouldn't be taken as a *definition* of personal-time, just as metaphysical-time isn't defined by your living room clock.

[2] It's a bit difficult to make "all else equal" here. But we may try to imagine that you care about nothing else except for pain. Alternatively, everything and everyone you care about in your local environment freeze along with you. (The latter suggestion raises a question concerning the locality and relativity of personal-time. More on this see 105, fn.1.)

[3] I gather that some opponents of near-bias would even think that it's virtuous to choose to undergo bad experiences as soon as possible, at least when all else is equal. But if far-bias is about metaphysical-time, they would have to find my choice virtuous. This sounds really wrong.

When it comes to future-bias, although I'd love to also describe a time-travel-free case to cement the intuition, I recognize that such a case would be no less vexed than time-travel scenarios. What I have in mind is a Benjamin Button-like figure (residing in a normal world). But he has to be more Button-ish than Button: In addition to a reversal in appearance and physiology, all his psychological processes shall run backwards. This would involve, *inter alia*, precognition of the metaphysical-future that has the same causal and evidential profiles as memory of us normal folks. I consider such a scenario to be obviously metaphysically possible. All that requires is some clever manipulation of local microphysical processes. But I am aware that such a scenario might overly stretch our imagination. So, I'm not sanguine that we can have clear-cut intuitions about that – though I refer to Kristie Miller (2022: 225-28) who describes in detail a Buttonish *world* (rather than merely odd individuals) and argues that such a world vindicates the personal-time conception.

But I think enough has been said. For one, we have mundane and simple cases showing that *near*-bias tracks personal-time rather than metaphysical-time. Absent good countervailing reason, we should expect this pattern to be uniform across near-bias and future-bias. For another, the time-travel cases compel the personal-time conception. Even if you judge these cases to be impossible, it seems that the *apparent sensibility* of your responses can still reveal something significant. The time-travel-denial may query herself with a counterpossible: What would I choose if I, *per impossible*, time-travel? There should be an *intuitive* answer to that question, and this answer should be in favour of the personal-time conception.[1] Such an intuition clearly shows something even if due to some philosophical technicality, all counterpossibles are trivially true. Finally, however one finds perplexing a scenario in which the metaphysical-past is one's personal-future (with or without time-travel), it's undeniable that metaphysical-time and personal-time are conceptually distinct. So, there remains a question: Should we care about time under the metaphysical "guise" or under the personal "guise"? And that's a question about what's genuinely reason-giving.

More on the very concept of personal-time. In the time-travel literature, personal-time tends to appear as an *ad hoc* device. But in fact, it's something we are all familiar with. It's a measurement

[1] To see the point, you don't have to be future-biased or regard future-bias as correct. The cases also show that time-neutralism is about personal-time.

of the direction and speed of our bodily and psychological processes. We judge that the (metaphysically-past) destination of a backward time-traveller is in her *personal-future* because her bodily and psychological processes during the journey (e.g., the accumulation of memory) go in the same direction as how these processes propagate into the *metaphysical-future* in an ordinary non-time-traveller. That is, we assign personal-time to individuals by comparing their changes and functions with how people normally change and function in relation to metaphysical-time. More technically, personal-time can be understood as a special kind of *local time*, i.e., a kind of local time that's relativised to the kind of creatures *we* are. And your personal-time is further relativised to the particular person *you* are. Local time, in general, can be defined as follows:

> The [local] time of an object *o* of kind *K* is an entity *T* such that the regularities in *o*'s career hold with respect to *T* in the same way that the regularities in the career of a typical member of kind *K* hold with respect to [metaphysical-time]. (Gilmore 2016: 3271)[1]

This is rather reminiscent of persistence conditions, i.e., the conditions under which an object *o* of kind *K* remains the very same object over time. This is certainly no accident. The claim that the electric shock is in *your* personal-future presumes that *you* will be around to suffer. This concerns the persistence conditions of you (and more generally the kind *K* you belong to). Moreover, we can say that the shocks are in your personal-*future* because we know how people usually persist over time. We know that normal persons accumulate memories about the metaphysical-past but not the future, that certain kinds of cells die but never regenerate, and so on.

[1] Local time applies to all sorts of stuff – for instance, the time-traveller's wristwatch. To say that the time-traveller's personal-time is measured by her wristwatch (Lewis 1976) is to say that her personal-time aligns with the local time of her wristwatch.

One concern about local time defined this way is that it's relative and therefore unstable (Bernstein 2021). Suppose that everyone is time-travelling to the past. It would seem to follow that your personal-time (and everyone's) aligns with metaphysical-time after all. *Time-Travel Shocks* would then fall apart. That's a difficult question that I don't have a definitive answer. But two brief points. (1) One might resist calling time-travellers "typical" persons in such a world. Perhaps a statistically typical person can fail to be typical in the relevant sense. Analogy: The proper function of dishwashers doesn't change in a world in which all dishwashers malfunction in a weird way. (2) Even if "typical" just means statistically typical, it seems that in order for time-travellers to be statistically typical, it needs to be the case that people are travelling to the past for most of the time throughout the history. I find such a world a bit difficult to imagine. Also, in such a world, it might be that our (appropriate) egoistic concern is just different. At least, I'm sceptical that we can have clear intuitions about such a world.

This is not to say that the determinants of personal-time just are the determinants of personal identity. Personal-time, as I understand it, is a rough and all-encompassing measurement of bodily and psychological processes. But presumably, *not* all these elements *determine* personal identity. You might think that persons persist by virtue of retaining psychological connections but not bodily processes (Parfit 1984). In ordinary life, these two aspects align. But it's only psychological connections that make it the case that you persist over time. That said, personal-time is closely related to personal identity on a superficial but nonetheless important level: Our ordinary *judgements* about personal-time and personal identity track the very same palpable features, even though not all these features determine personal identity.

Here then comes an important qualification of the personal-time conception. I said that time-biases should roughly track personal-time. "Roughly" because presumably not all those features measured by personal-time are reason-giving, though all these features give us *evidence* that *something* reason-giving is present. The fact that your elderly self has lots of gray hair and wrinkles, arguably, is not a *reason* for you to care less about her. Instead, one who takes psychological connections to determine personal identity may very naturally claim that it's exactly and only psychological connections that give reason for near-bias, even though diminished psychological connections happen to coincide with gray hair and wrinkles (so that the latter is evidence that something is reason-giving).

This is the general picture that I am attracted to and will explain in detail in the final chapter (Ch.5): There are certain intra-personal relations (including but not restricted to psychological connections) that give reason for prudential partiality, just as certain inter-personal relationships give reason for moral partiality. Perhaps the resulting picture of prudential partiality is not exactly a future-biased or near-biased one – for these intra-person relations need not perfectly correlate with personal-time (let alone metaphysical-time). Nonetheless, this refined picture doesn't mean that our intuitions about the above cases are inaccurate. Those compelling intuitions tell us two things. The first is that the metaphysical-time conception is false. The second is that there is something *in the vicinity of* personal-time that's potentially reason-giving – something that those palpable features of bodily and psychological processes roughly track and are evidence of. A detailed discussion about these reason-giving intra-personal relations will be offered in the final chapter.

4. Time-Biases and Wellbeing

4.1. Preferring for the Worse

Prudence is the domain of value and normativity that concerns how one treats oneself. It's a familiar thought that time-biases are in tension with prudence. In everyday contexts, near-biased behaviours are chastised as paradigmatically *imprudent*. It's imprudent, we often say, to squander in one's youth at the expense of a secure late life, or to procrastinate on important tasks and get oneself into dire urgencies, and so on and so forth. More philosophically, Henry Sidgwick takes prudence to aim for *an overall good life* – but near-bias frustrates this aim:

> The commonest view of the principle would no doubt be that the present pleasure or happiness is reasonably to be foregone with the view of obtaining greater pleasure or happiness hereafter; but the principle need not be restricted to a hedonistic application, it is equally applicable to any other interpretation of "one's own good", in which good is conceived as a mathematical whole, of which the integrant parts are realised in different parts or moments of a lifetime. (1907: 381)

Sidgwick makes three important points. Firstly, there is the common-sense observation that near-biased agents make intertemporal tradeoffs that are detrimental to *lifetime wellbeing*. Secondly, Sidgwick points out that prudence is first and foremost concerned about lifetime wellbeing. This is a widely shared sentiment, echoed by David Brink who claims that "[p]rudence demands that an agent act so as to promote her own overall good" (2011: 354). Thirdly, the above two points are taken to be *axiologically neutral*. That is, they hold no matter what wellbeing consists in – hedonic pleasure, knowledge, friendship, or what have you.

An argument against near-bias thus emerges. Meghan Sullivan formulates it this way:

> (1) At any given time, a rational agent prefers that her life *going forward* go as well as possible.
> (2) If you are near-biased, then in distant tradeoffs you will prefer and choose the present, lesser good over the greater, future good.

(3) Your life going forward would go better if you preferred and chose the greater, future good in a distant tradeoff.
(4) So, near-biased preferences are not rationally permissible, insofar as you face a distant tradeoff. (2018: 22-23, emphasis mine)

As it's stated, however, the argument is not entirely satisfactory. For a start, it focuses on the consequences of near-biased tradeoffs. But as mentalists of preferences, we want to criticize near-bias by virtue of its content rather than its actual (or potential) consequences (§1.4.1). Insofar as near-bias is imprudent, we want to criticize it as imprudent even when no tradeoff is on offer. Also, premise (1) is a principle concerning life *going forward*. While this principle may suffice for an argument against prospective near-bias, it wouldn't help if one wants to extend the argument to *future-bias* (and perhaps also retrospective near-bias). Future-bias and near-bias alike, as Sidgwick points out, exhibit the same sort of lack of concern for life *as a whole*. Future-bias, however, would be commended rather than condemned by a principle like (1). So, if one regards future- and near-bias as imprudent for the same reason, one needs a principle concerning lifetime wellbeing. (There is also the apparent impossibility of making past-future *tradeoffs*; more on this in §4.5.)

The time-neutralist might argue against near-bias *and* future-bias like this:

The Argument from Wellbeing against Time-Biases

(1) Prudentially[1], one ought *not* prefer that her whole life *does not* fare as well as possible. (*No Tragedian*)
(2) If one is prudentially future-biased *beyond tie-breaking*, then one prefers that her whole life does not fare as well as possible.
(3) If one is prudentially near-biased beyond tie-breaking, then one prefers that her whole life does not fare as well as possible.
(4) Therefore, one prudentially ought not be prudentially future-biased beyond tie-breaking. (from (1) and (2))

[1] I focus on prudential oughts rather than all-things-considered oughts. There might be countervailing non-prudential (moral, etc.) reasons that are weighty enough so that one ought to all-things-considered prefer otherwise. In what follows, I assume that no non-prudential considerations are in play unless indicated otherwise.

(5) Therefore, one prudentially ought not be prudentially near-biased beyond tie-breaking. (from (1) and (3))

Several clarifications. In line with David Brink's remark that "prudence and its demand of temporal neutrality are, at least in the first instance, claims about what we have objective reason to do" (2011: 371), this argument is about the *correctness* of time-biases (just as *the Argument from Arbitrariness* in the previous chapter). This is not to say that the *rationality* of time-biases is completely irrelevant. Presumably, knowing what's correct would inform judgments about what's rational in various epistemic predicaments. But I cannot delve into those vagaries here.

No Tragedian[1] seems a fairly minimal requirement of prudence. At least, it seems that if there are any prudential *oughts* at all, this should be one of them. One who prefers her life to fare worse as a whole would seem to be failing prudence in a fundamental way. (But more reflections on *No Tragedian* in §4.6.) The formulation of *No Tragedian* is slightly different from Sullivan's premise (1). Apart from an extended focus on life as a whole, I speak of what's forbidden whereas she speaks of what's required. Nothing substantive hangs on this except that my formulation renders time-biases in more direct violation of a principle of prudence.

In the previous chapter, I argued that given the metaphysical-time conception, time-biases are arbitrary. But I've also argued for the personal-time conception. As I understand it, the present argument is better interpreted as one concerning *personal-time*. This is because wellbeing is a concept that tracks personal identity and therefore personal-time. When one is concerned about her future wellbeing, she's concerned about the (hopefully large) segment of *her* life she is anticipating and planning about; and what she anticipates may be in the metaphysical past should she time-travel (cf. §3.4.2). So, the argument is not beating a dead horse (if you're convinced that time-biases about metaphysical-time are arbitrary).

The argument as stated is restricted in scope in two respects. First, it targets only time-biases that are *beyond tie-breaking* – that is, those involving *unequal payoffs*, e.g., Parfit's preference for greater past pain over lesser future pain. One might want to extend the argument to cover merely tie-breaking time-biases. If *No Tragedian* is true, then presumably it's also true that one ought to be

[1] I took the term from Sullivan (2018: 24), though by a "tragedian" she means something different: one who believes her life to "go best overall if it will end with a period of very low wellbeing". I shall address such structural features of life shortly in the next section.

indifferent between two courses of life that are equally good. But here I shall focus on time-biases beyond tie-breaking, and from here on, often suppress this qualification. I also confine my discussion to prudential normativity. The argument as stated is restricted to *first-person* time-biases, given that prudence is about how one treats oneself. One might want to extend the argument to third-person time-biases by a moral principle that parallels *No Tragedian* but concerns other-directed attitudes about lifetime wellbeing. But I set that aside.[1]

Finally, it's worth re-emphasising that the argument is supposed to be axiologically neutral. By this I mean that it makes no assumption about what it takes for one's life to "fare well". In other words, it's supposed to be compatible with all *theories of wellbeing*. Like Sidgwick, Sullivan is committed to axiological neutrality: "[W]e do not need to take a stand on the nature of the relevant goods here, since we face [tradeoff] scenarios for any of them" (2018: 23-24). This appears true. Indeed, it's natural to think of time-neutralism as a "formal" requirement of prudence that cuts across substantive issues about wellbeing. Just as we can debate over (say) the just way to distribute utilities among society members while abstracting away what these utilities consist in, there is no need to dig into theories of wellbeing here. Or so it seems.

4.2. Wellbeing, Temporal and Lifetime

It's a mistake, however, to be indifferent to axiological issues. Quite the contrary; how we might think about time-biases in light of wellbeing is intricately connected with our theories of wellbeing. I am in good company here. For instance, Dale Dorsey (2016: 4) suggests that time-neutralism is incompatible with his preferred theory of wellbeing according to which "past benefits are at least fungible against present and future benefits from the perspective of prudence" (2018: 1914). Similarly, Robert Brandt claims that desire-satisfactionism of wellbeing cannot take a "time-neutral" form (1979: 249). At the end of the day (§4.4-4.5), I *disagree* with Dorsey and Brandt regarding these specific claims. But the general sentiment is shared. In order to adequately understand how time-biases relate to wellbeing, we need to get into theories of wellbeing. Among other things, I shall argue that *the Argument from Wellbeing against Time-Biases* can be reasonably rejected by *subjectivists* of wellbeing (§4.6-4.7).

[1] I say more about the normative status of third-person time-biases in §3.4.1 and §5.7.

For now, however, more on how the argument is supposed to work. Begin with (2) and (3). They appear very compelling. Suppose that I'd rather suffer more pain in the more distant future. That amounts to, it seems, preferring a life that contains greater pain in total, and thereby a life that's overall worse off. But we need to be careful here. Our official definition of time-biases (§1.1, §1.3) has no explicit mention of wellbeing. If we focus on (pure) prudential near-bias that's beyond tie-breaking[1], a definition should look like this:

> *Prudential Near-Bias (Beyond Tie-Breaking):* An agent A is (purely) positively prudentially near-biased iff for two exclusive states-of-affairs S_1 and S_2, where S_1 is *prudentially* more valuable than S_2, A prefers S_2 over S_1 (solely) because S_2 is in the nearer future than S_1; An agent A is (purely) negatively prudentially near-biased iff for two exclusive states-of-affairs S_1 and S_2, where S_1 is *prudentially* more disvaluable than S_2, A prefers S_1 over S_2 (solely) because S_1 is in the more distant future than S_2.

But arguably, this is translatable into the language of *temporal wellbeing*, i.e., wellbeing at a time. Suppose that hedonism is true, and that one's level of wellbeing at t is determined by the *net amount* of pleasure experienced at t. An episode of pain at t constitutes a *welfare harm* to one's wellbeing-at-t in the sense that her wellbeing-at-t would be greater had that episode not occurred. (Likewise for pleasure as prudential benefit, in terms of counterfactual increase of wellbeing[2].) But to say that a welfare harm is greater, it seems, just is to say that it's of greater welfare disvalue. Thus, my preference for greater pain at t_2 over lesser pain at t_1 ($t_1 < t_2$) amounts to a preference for a greater welfare harm to my wellbeing-at-t_2 over a lesser welfare harm to my wellbeing-at-t_1. Understood this way, prudential time-biases amount to biased cross-temporal distribution of welfare harm and benefits. Indeed, some might want to *define* time-biases as *wellbeing discounting*.[3] Regardless of if

[1] I set aside retrospective near-bias, but not because it is uninteresting. One interesting question is what the hybridists should say about retrospective near-bias, if hybridism is taken as the position that *prospective* near-bias is incorrect but future-bias is correct. Should they say that it's incorrect because it's *near*-bias or it's correct because it's about the *past* (and thus more like future-bias)?

[2] That is, prudential benefits and harm (or welfare benefits and harm) are understood as inherently comparative. The kind of comparison that matters here is counterfactual and synchronic. This is to be distinguished from cross-temporal comparison as in the claim that my life now is going better than last year.

[3] Broome (1994: 128-29) distinguishes wellbeing discounting from commodity discounting. He states that wellbeing is the "fundamental" good that philosophers (but perhaps not economists) are concerned about. I'd like to add that wellbeing discounting has "purity" built into it, insofar as prudential harm and benefits are defined in terms of counterfactual comparisons with *ceteris paribus* conditions.

time-biases should be thus defined, the following conditional does ring true (NB meaning near-bias):

> *Wellbeing Discounting (NB):* If one is prudentially near-biased beyond tie-breaking, then she prefers a lesser benefit to her wellbeing-at-t_1 over a greater benefit to wellbeing-at-t_2, or a greater harm to wellbeing-at-t_2 over a lesser harm to wellbeing-at-t_1, for some future $t_1<t_2$.

A similar conditional – *Wellbeing Discounting (FB)* – can be formulated for future-bias. But we have yet to connect *temporal wellbeing* with *lifetime wellbeing*. It's the latter that *No Tragedian* is about. The connection may appear obvious:

> *Tragic Preference (NB)*: If one prefers a lesser benefit to her wellbeing-at-t_1 over a greater benefit to wellbeing-at-t_2, or a greater harm to wellbeing-at-t_2 over a lesser harm to wellbeing-at-t_1, for some future $t_1<t_2$, then one prefers her level of lifetime wellbeing to be lower.

The antecedent of *Tragic Preference (NB)* just is the consequent of *Wellbeing Discounting (NB)*. Taken together, we get (2) of the argument. Similarly for (3) of the argument concerning future-bias.

Should we accept *Tragic Preference*? Well, it is obvious that temporal wellbeing and lifetime wellbeing are intimately related.[1] But how? Most straightforwardly, one might think that lifetime wellbeing just is a *simple aggregation* of wellbeing at all times. That is, to determine how one's life fares as a whole, we just need to add up each slice of her temporal wellbeing. I find simple aggregation incorrect. Consider a single-life version of Derek Parfit's repugnant conclusion.[2] An extremely long life that is at each point barely worth living doesn't look overall better than a steadily fantastic life of normal length. Regardless, holding fixed longevity, a weaker principle of *shape-indifference* seems sufficient for *Tragic Preference*. Shape-indifference is a constraint on how temporal wellbeing determines lifetime wellbeing. It says that if two equal-length lives have the

[1] Bramble (2017) argues that there is no such thing as temporal wellbeing. He holds that no welfare harm or benefit harm can be located at any sublifetime. Nonetheless, he might accept (2) and (3) because temporally local states-of-affairs may *directly* feed into lifetime wellbeing, so to speak, without being mediated by temporal wellbeing.
[2] Sullivan (2018: 29-30, 160-61); cf. Parfit (1984: 387-88).

same total amount of temporal wellbeing, then they are overall equally good. This rules out lifetime wellbeing being sensitive to how temporal wellbeing distributes over time.

But shape-indifference is controversial, to say the least, as Sullivan (2018: 30-33) acknowledges. Many have the intuition that "shapes of life" or structural features of life matter for lifetime wellbeing.[1] Consider Life *A* and Life *B* of equal length. They are mirror images with respect to the temporal orderings of temporal wellbeing. Life *A* has an "uphill" trajectory, beginning with a wretched childhood, getting better and better afterwards, and culminating in a content old age. Life *B* is reversed, running constantly "downhill" from blithefulness to disaster. That is, there is a one-to-one mapping between slices of temporal wellbeing of the two lives, so that the total amounts of temporal wellbeing are equal. In other words, the two lives have the same "area under the curve". But to many, the uphill Life *B* appears overall better.

Other sorts of shape variations may also matter. You might think that holding fixed the area under the curve, a life with high peaks of triumphs is overall better than one with even distribution. Or you might think the contrary: Oscillations between peaks and troughs are bad. At any rate, it seems plausible that lifetime wellbeing is something over and beyond aggregating (or averaging) slices of temporal wellbeing. There are different ways to accommodate structural features (e.g., an uphill trajectory). One might think that structural features are of *sui generis* welfare value, so that lifetime wellbeing is a function of both temporal wellbeing and structural values. Alternatively, wellbeing at times might make *weighted* contribution to lifetime wellbeing. For instance, to make the uphill Life *B* overall better, it might be the case that slices of temporal wellbeing at later stages of life (or those that manifest cross-temporal increases) get their weight "boosted" when factoring into lifetime wellbeing.

If shape-indifference is false, then *Tragic Preference* would seem to be false.[2] Consider for instance a future-biased person at the midpoint of her life. She prefers poverty and sadness to be past but

[1] Slote (1982), Velleman (1991); cf. Dorsey (2015), Hersch & Daniel (2023).

[2] Alternatively, it might be thought that such cases are inconsistent with *Wellbeing Discounting* instead of *Tragic Preference*. There are at least two ways in which shapes-of-life might be relevant to lifetime wellbeing. Suppose that hedonism is true. And suppose that it's an overall better life, all else being equal, to have pain congregated in the earlier half and pleasure congregated in the later half. The first way in which such a shape can be lifetime-wellbeing-relevant, as suggested already, is that it directly contributes to lifetime wellbeing without affecting temporal wellbeing. (That is, wellbeing at any time *t* is completely determined by the amount of pleasure and pain at *t*.) Alternatively, it might be the case that temporal wellbeing itself is altered by structural features. This can make a difference to lifetime wellbeing even if lifetime wellbeing is simply temporal wellbeing aggregated. In this case, if I at mid-life prefers pain in the

success and happiness to be future. But if – as the comparison between Life *A* and Life *B* shows – uphill trajectories make overall better lives, then it's certainly possible for this person to end up preferring a life that's overall better. It's certainly possible for structural values to (over)compensate for decreases in temporal wellbeing. Such counterexamples can be multiplied (depending on which structural features matter).

I happen to prefer a *debunking* explanation of our shape-of-life intuitions. I suspect that insofar as we find Life *B* better, we haven't been really imagining two lives with the same aggregated temporal wellbeing. (Hersch & Daniel 2023; cf. Dorsey 2015) By varying the shapes of the curves, we are inadvertently also varying the areas under the curves. Just to illustrate, consider memories of childhood experiences. They fade away easily, not reverberating as much as those of the prime age. So, what we are really imagining might well be a life repleted with stressful memories *versus* a life that's not. Also, one who lives Life *B* can constantly look back upon her trajectory and *see* it as a satisfying and prideworthy story of success and redemption, whereas Life *A* is more likely filled with a sense of loss and despair for the future. Differently put, Life *B* is expected to have a larger area under the curve thanks to better mental health and a brighter outlook on life.

This is of course a contentious claim. And I have no proof that it's *impossible* to alter the shape without altering the area – which is required for preserving shape-indifference. So, let us consider the possibility that shape-indifference is indeed false. How much of a problem would that be for our present argument?

Here is an obvious but unsatisfactory time-neutralist response: It's one thing to be time-biased and quite another to prefer that life has a certain shape. Time-biases are perspectival preferences that privilege certain segments of life in light of their temporal locations relative to where "now" is. But shapes of life are perspective-indifferent descriptions of the temporal and relational structures of life as a whole. So, insofar as an agent is merely being time-biased, she cannot be said to be preferring an overall better life because she is *not* concerned about her lifetime wellbeing in a non-perspectival manner.

---

past or pleasure in the future, what I prefer might turn out to be a *lesser* decrease in temporal wellbeing or a *greater* increase in temporal wellbeing. This proposal is inconsistent with *Wellbeing Discounting* though consistent with *Tragic Preference*. Though slightly different, these two accounts of shapes-of-life are both inconsistent with (2) and (3) of the original argument.

Well, it's true that time-biases and shape-of-life preferences are crucially different. A merely time-biased agent does not have shape-of-life considerations in mind. She could end up preferring an overall better life *only accidentally*, so to speak. But that suffices to cause trouble for the argument. To make this point clearly, let us distinguish two readings of *the Argument from Wellbeing* – the *de facto* reading and the *de dicto* reading. I take it that the *de facto* reading is the desired reading and I have been assuming it. But the above response would work only given the *de dicto* reading.

Lois Lane fancies a date with Clark Kent, not knowing that Clark Kent is Superman. Does she fancy a date with *Superman*? She does *de facto*: The state-of-affairs she fancies is in fact a date with Superman. She does not *de dicto*: She never thinks of Clark Kent "under the guise" of Superman. The above response presumes a parallel *de dicto* reading of the argument. It in effect says that the time-biased agent does *not* prefer bad things in the past (or good things in the future) "under the guise" of shapes-of-life, even if such states-of-affairs would *in fact* entail an overall better life. But the *de dicto* reading is problematic. Consider for example (2): "If one is prudentially future-biased beyond tie-breaking, then one prefers that her whole life does not fare as well as possible." Read *de dicto*, it's evidently untrue. A time-biased agent may be like Lois Lane. A time-biased agent may simply lack the concept of lifetime wellbeing or any wellbeing-concept at all. Even conceptually more sophisticated agents presumably don't always *conceptualize* their time-biases under those guises. Worse still, one might conceptualize her time-biases as preferences for an overall *better* life.

The *de facto* reading evades the problems above and appears more motivated. Insofar as *correctness* – as opposed to rationality which is subjective – is concerned, it seems irrelevant how the agent conceptualizes (or fails to conceptualize) what is entailed by the content of their preferences. To illustrate, suppose that there are some undesirable features of dating Clark Kent/Superman that Lois Lane would only realize by identifying Clark Kent with Superman. (Dating Superman would expose her to evil forces.) Her ignorance might excuse her from the charge of irrationality. But it's true regardless that she *ought not to* fancy a date with Clark Kent – that would be very harmful! The same goes for time-biases. The time-neutralist would *not* let my future-bias off the hook simply because I don't take it to entail a worse life. But given the *de facto* reading, shape-of-life considerations become worrisome. Toggling cross-temporal distribution of goods and bads in a time-biased way does sometimes – however accidentally – result in certain life structures that make life overall better.

So, the argument turns out to be flawed. But the problem doesn't seem too devastating. For a start, one might insert *ceteris paribus* clauses into the argument. Meghan Sullivan makes such a suggestion: "To accommodate structuralism, we need only assume that the relevant tradeoffs do not involve any significant structural changes to a life." (2018: 33) I surmise that for most of our time-biases, this assumption is indeed true. Most of our time-biases concern temporally punctuated and confined states-of-affairs. Such local and trivial tradeoffs, presumably, can hardly result in an overall better life by making a better shape.[1] Of course, our preferences about cross-temporal distribution can in principle be rather global. However, when I try to imagine a realistic instance of a preference about global distribution, I end up imagining a non-perspectival shape-of-life preference (and an arguably appropriate one). This is a speculative claim. But it seems to me that when one says that she'd rather have a wretched past than a tragic future, very likely, what she really cares about is an uphill trajectory or a rags-to-riches narrative rather than pastness and futurity *per se*.

An alternative strategy is to formulate the argument in terms of not lifetime wellbeing but some less comprehensive wellbeing measurement that *is shape-indifferent*. Call that measurement lifetime wellbeing$^{\text{MINUS}}$ – whatever is left off by subtracting the contribution of structural features from lifetime wellbeing. Lifetime wellbeing$^{\text{MINUS}}$ gets different interpretations depending on how exactly structural features affect lifetime wellbeing. The argument can be neutral on this. Suppose that as suggested above, structural features are of *sui generis* value. Or suppose that they make slices of temporal wellbeing more or less weighty when factoring into lifetime wellbeing. Either way, lifetime wellbeing$^{\text{MINUS}}$ can be interpreted as the unweighted sum of temporal wellbeing (or the unweighted average, or some other function weighted by lifespan so as to avoid the repugnant conclusion). *No Tragedian* can then be modified as: There is *pro tanto* reason against preferring a life with lesser lifetime wellbeing$^{\text{MINUS}}$.[2] Lifetime wellbeing$^{\text{MINUS}}$ does not necessarily correlate with lifetime wellbeing. This *pro tanto* reason might be outweighed when structural features kick in. Accordingly, the conclusions shall be modified in terms of *pro tanto* reasons against time-biases

[1] The truth of this predication somewhat depends on what would make a desirable shape-of-life. For example, Dorsey (2015) holds that what underlies a desirable shape-of-life is a good narrative structure – e.g., a rags-to-riches story arc. A good narrative structure takes a village. It is very unlikely generated by punctuated time-biases. On the other hand, Glasgow (2013) holds that any local cross-temporal increase in pleasure or decrease in pain would confer structural value. If so, then the *ceteris paribus* condition might be excluding a lot.

[2] Sullivan seems to be making this suggestion by saying that "we can make 'prefer a life with the best story' another principle of rational planning, alongside Success" (2018: 33). By Success she means that "at any given time, a rational agent prefers that her life going forward go as well as possible" (2018: 22).

instead of straightforward oughts. The modified conclusions, though weaker, seem bad enough for friends of time-biases. Moreover, if I am correct that time-biases usually do not result in structural alterations that matter, then these *pro tanto* reasons usually amount to *all-things-considered* reasons.

It's admittedly a tricky matter how best to modify the argument in light of shapes of life. But as I see it, the problem is relatively minor and more on the side of technicality. And in what follows, for simplicity, I shall stick to the original formulation of the argument (on the *de facto* reading) and set aside shape-of-life considerations unless relevant.

A more fundamental problem arises from how time-biases affect wellbeing given *subjective* theories of wellbeing. More specifically, a case can be made that lifetime wellbeing is *shifty* and *perspectival* for agents who are in fact time-biased. If so, time-biased agents need not violate anything like *No Tragedian*. I turn to this argument in §4.6. However, some think that subjective theories of wellbeing pose a more straightforward problem: Subjective theorists cannot endorse time-neutralism. It's argued that given subjective theories, there are ample cases eliciting very compelling intuitions against time-neutralism. But these intuitions, seems to me, are merely apparently anti-time-neutralism. They can be accommodated by time-neutralist subjective theorists.

4.3. Preferentialism and Now-for-Past Sacrifice

According to some, time-neutralism delivers unacceptable results when combined with subjective theories of wellbeing. In particular, time-neutralism demands *now-for-past* sacrifices that are intuitively objectionable. If so, time-neutralism cannot be axiologically neutral. While I share the intuition against now-for-past sacrifice, I do not see it as an intuition against time-neutralism. Instead, I argue that a subjective theorist can be a time-neutralist and avoid now-for-past sacrifice. By understanding what these intuitions are really about, we can better appreciate what's really at stake between time-neutralism and its denial, and moreover, reach a more sophisticated picture of the diachronic structure of preferences in the context of wellbeing.

A theory of wellbeing is in the business of pinpointing the "value atoms" of wellbeing, i.e., what is intrinsically and fundamentally good or bad for someone (Bradley 2009: 6-7). Consider for instance an "objective list theory" according to which all that's of intrinsic (positive) prudential

value is health, friendship and knowledge. Such a theory is, as its name indicates, *objective*. It says that health, friendship, and knowledge are intrinsically (i.e., non-instrumentally) good for everyone, no matter if one cares about them. A *subjective* theory of wellbeing, on the other hand, makes my value atoms dependent upon my *own attitudes*. Friendship might well be good for me – but only if and *because* I care about friendship. If I just don't care about knowledge, then knowledge isn't good for me. (Though knowledge is often *instrumentally* good for me, since it's conducive to lots of things I really care about. I shall often suppress the "intrinsic" qualification when the context is clear.) On subjective theories, our attitudes "gild and stain the world" with prudential values (Sobel 2011: 67).[1] Subjective theories thus allow for *inter-subjective variation*. What's good for you may fail to be good for me since our attitudes may differ.

One major advantage of wellbeing subjectivism, as it's often advertised, is that it honours *the resonance constraint*. As Peter Railton puts it:

> [It] does seem to me to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware[2]. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him. (1986: 9)

Indeed, subjectivism does more than fulfill the resonance constraint. It also explains why what's good for you always resonates: Your attitudes are *value-conferring*; their objects cannot fail to resonate by virtue of the nature of the attitudes you have.

---

[1] I take bearers of prudential values to be states-of-affairs. To say that knowledge is good for me is thus a shorthand for saying that the states-of-affairs of me gaining knowledge are good for me.

There is some debate over whether *hedonism* falls on the objective or the subjective side. Some think that the *nature* of pleasure and pain consists of our attitudes: What makes some experiential state a state of pleasure is that we hold certain pro-attitudes towards it (Heathwood 2006). But even if that's true, hedonism may still be an objective theory. As I understand subjectivism, it's not sufficient for our attitudes to *compose* what's good for us. Rather, these attitudes should be value-*conferring*. It seems that unless the hedonist is claiming that pain is bad for me *because* I dislike pain (rather than some constituent of pain), hedonism is in effect an objective list theory with a very short list. There might be alternative ways to draw the subjectivism/objectivism distinction. But the distinction I care about here counts a theory as subjective only if it lets one's own attitudes to confer prudential value to states-of-affairs.

[2] More on "rational and aware" (that is, idealizing conditions) in §4.7.

Which attitudes are value-conferring? Some subjective theorists talk about desires, others preferences, and still others "valuing attitudes" (Dorsey 2021). We need to pick one for the sake of a more formal representation. As we have been talking about preferences all along, let's settle on *preferences* being the relevant value-conferring attitudes. Call this version of subjective theory *preferentialism*. My suspicion is that all subjective theories can be translated into some form of preferentialism. Desires and aversions (or a subset of which that are welfare-relevant) seem interdefinable with preferences. Valuing attitudes can be seen as a privileged sort of preferences. So, feel free to read preferentialism as standing for all subjective theories. (Though I shall sometimes speak of desires and aversions for convenience.)

The tension between time-neutralism and preferentialism (or subjective theories generally) arises from two mundane facts about preferences. First, preferences change over time. Preferentialism thus seems to allow for *cross-temporal variation* of prudential value, in addition to inter-subjective variation. Second, preferences can be directed at times other than the times of preferences themselves. Time-biases are themselves such preferences. There are also plenty of ordinary preferences like this. I now dream of throwing a techno-pop party on my 50th party. This preference wouldn't be satisfied or frustrated anytime soon. In addition to such future-directed preferences, there are also past-directed preferences. I told a terrible lie to my parents when I was 17. I wish I hadn't.

But now consider a series of cases involving *merely past* preferences – preferences that were had in the past but do not persist to the present – coming into *direct conflict* with *present* preferences:

> *The Rollercoaster:* Suppose my six-year-old son has decided he would like to celebrate his fiftieth birthday by taking a roller-coaster ride. [And suppose on his fiftieth birthday, he'd rather chill with a bourbon.] This desire now is hardly one we think we need to attend to in planning to maximize his lifetime well-being. Notice that we pay no attention to our own *past desires*. (Brandt 1979: 249, emphasis mine)

> *The Deathbed:* [A] convinced sceptic who has rebelled against a religious background wants, most of his life, no priest to be called when he is about to die. But he weakens on his deathbed, and asks for a priest. Do we *maximize his welfare* by summoning a priest? (Brandt 1979: 250, emphasis mine)

> *The Aspiring Poet*: When I was young what I most wanted was to be a poet. This desire was not conditional on its own persistence. I did not want to be a poet only if this would later still be what I wanted. Now that I am older, I have lost this desire. I have changed my mind in the more restricted sense that I have changed my intentions. But I have not decided that poetry is in any way less important or worthwhile. Does my past desire give me a reason to try to write poems now, though I now have no desire to do so? (Parfit 1984: 157)

All these cases involve potential *now-for-past sacrifice*, where one faces the choice of whether to sacrifice her present preference for the sake of a merely past preference. (More on the structure of now-for-past sacrifice in the next section.) Now-for-past sacrifice seems to showcase a tension between time-neutralism and preferentialism. In a nutshell, if preferentialism were time-neutralist, then past, present, and future preferences all matter, and all matter equally. But that would demand now-for-past sacrifice – which many find unacceptable.

The basic intuition, as Krister Bykvist (2006: 132) observes, is that "when past preferences are replaced by new ones, it seems absurd to say that we ought to respect the preferences that we once had." Bykvist goes on to claim that any plausible theory of wellbeing should avoid now-for-past sacrifice. Robert Brandt takes *the Rollercoaster* to show that preferentialism is false, at least if it takes a "time-neutral" form:

> The idea seems to be that we consider all the desires a person has ... *at some time or other, or many times, over a lifetime*, and what that person ... should aim at is to maximize the satisfaction of these desires. This conception is unintelligible. (1979: 249, emphasis mine)

Similarly, commenting on a case like *the Rollercoaster*, Dale Dorsey finds the implication of time-neutralism unsavory:

> Why should I care about benefiting myself at t_1, when my desires at t_2 have changed? Surely at t_2 it is all-things-considered prudentially rational *not* to [ride the rollercoaster], given my slight aversion to so doing at t_2. (2013: 163)

Dorsey then turns that insight into a case against time-neutralism: Time-neutralism is *not axiologically neutral*. (2016: 4, 2013: 163-64) In particular, it's incompatible with his favoured kind of preferentialism according to which "past benefits are at least fungible against present and future benefits from the perspective of prudence" (2018: 1914). Given his version of preferentialism, prudence should be at least to some extent *future-biased*: Past preferences do not matter as much as present and future ones. More in this regard, Thomas Nagel (1970: 70), a time-neutralist, feels the need to soften his position in light of cases like *the Aspiring Poet*.

A lot needs unpacking here. Indeed, the meaning of "time-neutralism" and "future-bias" becomes somewhat opaque in light of now-for-past sacrifice. In what follows, let us carefully walk through the various temporal dimensions of preferentialism. Ultimately, I propose that the purported "anti-time-neutralist" intuitions are merely apparent. One can be a time-neutralist preferentialist and avoid now-for-past sacrifice.

4.4. Preferentialism and Time

A theory of wellbeing, by virtue of identifying the value atoms of prudence, delivers a calculus of levels of wellbeing. Consider again the objective list consisting of health, friendship, and knowledge. Given this theory, all else being equal, one's life goes better to the extent that she is healthier, has more friends, and is more knowledgeable. (A complete calculus would need to tell us how to weigh these different value atoms against each other.)

Similarly, on preferentialism, my life goes well to the extent that my preferences are *satisfied* rather than *frustrated*. By virtue of being comparative valuing attitudes, preferences confer comparative value relations to (pairs of) states-of-affairs that are objects of preferences. I prefer to go to the movie rather than not. This makes it the case that all else being equal, I'd be better off if I go to the movie rather than not. More generally, a preference for *A* over *B* (where *A* and *B* are mutually exclusive) is satisfied iff *A* rather than *B* obtains, and is frustrated iff *B* rather than *A* obtains.[1] An agent *S*'s life fares well to the extent that it contains a greater balance of preference-satisfaction over preference-frustration. Notice that instances of preference-satisfaction and preference-frustration

[1] I take that in order for the theory to be adequate, preference-satisfaction/-frustration should come in degrees (Skow 2009). But for simplicity, I will henceforth work with satisfaction and frustration to the fullest extent.

are *conjunctive* states-of-affairs: There is *preference*, and there is *satisfaction* or *frustration* thereof. An instance of preference-satisfaction, as a conjunctive state-of-affairs, constitutes a *welfare benefit*: I'd be worse off should this conjunctive state-of-affairs fail to obtain.[1] If I preferred *not* to go to the movie, then I would not be made better off by going to the movie. (I'd instead be made worse off by virtue of having my preference frustrated.) If I have missed the movie despite really wanting to go, then no welfare benefit either. Parallelly, an instance of preference-frustration, as a conjunctive state-of-affairs constitutes a *welfare harm*. Finally: All else equal, the *stronger* the preference, and the *longer* the preference is held, the greater welfare benefit or harm if it's satisfied or frustrated. (I take this to be intuitive. A hedonist would say something similar: The more intense and the longer pain is, the greater the harm.)

Turn now to the temporal structure of *now-for-past sacrifice*. Now-for-past sacrifice is a subgenre of *direct conflict* of preferences. The common structure of direct conflict is this. At t_A, one prefers that, at t_S, A rather than *B*. But at t_B, one prefers, at t_S, *B* rather than *A*. As a characterization of direct conflict in general, it may be the case that t_A *or* t_B coincides with t_S but it need not be. All these temporal stamps can be of extended periods, and they can take any temporal order. *Direct* conflict lies in that the objects of the preferences are about the very same time t_S but with "reversed" content. To satisfy the t_A-preference just is to frustrate the t_B-preference and *vice versa*.

Now-for-past sacrifice is a special kind of direct conflict. Consider *the Deathbed* for example. The protagonist – call him Oscar[2] – has preferred from t_0 to t_{10} that no priest is called at t_{11} (when he is dying). But at t_{11}, he prefers the opposite – that a priest is called at t_{11}. The conflict here is considered a now-*versus*-past conflict because as of t_{11}, Oscar's theist preference is present, but his atheist preference is merely past. Satisfying his merely past preference means frustrating his present preference and *vice versa*.

Note that there are two – or rather *three* – kinds of temporal features involved. The first is the *time-of-attitude*. The time-of-attitude of Oscar's earlier atheist preference is t_0 to t_{10}. The second is the *time-of-object*. The time-of-object of that preference is t_{11}. The time-of-attitude and the time-of-

[1] This is not true on anti-frustrationism (Fehige 1998), according to which preference-satisfaction confers no benefit but preference-frustration confers harm. Nonetheless, in cases of direct conflicts of preferences, anti-frustrationism raises the same problems because *some* preferences have to be frustrated.
[2] Rumour has it that Oscar Wilde had a deathbed conversion. Incidentally, there was a famous line in his 1892 play *Lady Windermere's Fan*: "In this world there are only two tragedies. One is not getting what one wants, and the other is getting it."

object may not overlap, as we've seen. Finally, there is a subtle difference between the time-of-object and the *time-of-obtaining*. I spoke as if the time-of-object of Oscar's past atheist preference carries a *precise* time-stamp. But this is unrealistic (as he's not prognosticating his own death). More realistically, the time-of-object should be something like "when I'm dying". But if this preference is satisfied, the time-of-obtaining – the time at which the preferred state-of-affairs obtains – is precise (namely, at t_{11}). This subtle distinction is worth keeping in mind.

Now, since Oscar's merely past atheist preference lasts sufficiently longer than his present theist preference – enough to outweigh whatever disparity in strengths there may be (let's suppose) – it would seem that satisfying his atheist preference would confer greater welfare benefits than satisfying his theist preference. Time-neutralism would then seem to imply a prudential requirement *not to* call a priest. For past and present (and future) preferences are regarded as on equal footing given time-neutralism. But such a requirement for now-for-past sacrifice seems objectionable.[1] Or suppose you are Brandt's son on your fiftieth birthday. Assume that all else is equal but your merely past preference for rollercoaster was stronger. Time-neutralism would seem to imply that you ought to take the rollercoaster ride rather than chill with a bourbon (which is what you now prefer). Or consider Parfit's lost aspiration for poetry. All else being equal, neither philosophy nor poetry is to be prioritized. Perhaps Parfit should just flip a coin. But that sounds unacceptable.

You may not share these intuitions. And that would be understandable. Truth be told, I find the evidential status of intuitions especially shaky in these cases.[2] One major problem is that presumably, not all preferences are welfare-relevant. It's not the case that satisfying or frustrating *any* preferences would confer welfare benefit or harm, as preferentialists tend to think. Instead, only a privileged subset of preferences is welfare-relevant. The above cases, as stated, might give off the impression that some of the preferences are welfare-irrelevant for reasons that are unrelated to time. And that might mess up our intuitions.

The conditions of welfare-relevance are a messy issue I cannot dig into here (more on which in §4.7). But just to illustrate, it's easy to get the impression that Brandt's son's childhood preference

[1] In case the intuition isn't clear, we may instead stipulate that Oscar's atheist preference lasted *only a minute* longer (and was equally strong) to strengthen the intuition.
[2] Sarch (2013: 241-45), Dorsey (2013: 164-66), and Bykvist (2003: 130-32) also point out ways in which intuitions might go awry with these cases.

is somehow *ill-informed* and therefore welfare-irrelevant.[1] If the childhood preference is based on a baseless projection that the middle-aged man will still enjoy rollercoasters, then presumably it should be disregarded, but not because it's past. It's also natural to read that preference as one that's *conditional* on its own persistence – one has a conditional content in the form of "take a rollercoaster ride if I still enjoy it". That preference would be welfare-irrelevant because it cannot be satisfied or frustrated. It's instead *cancelled* because the persistence condition isn't met.

To "fix" our intuitions, these cases had better be finessed so that the preferences in question satisfy whatever restrictions the preferentialists may want to impose.[2] For instance, in addition to being well-informed and not cancelled, they may have to be non-instrumental, self-regarding, and not "trivial, idiosyncratic, masochistic or immoral" (Sarch 2013: 223). Moreover, they may have to survive suitable *idealizing processes*. I shall have more to say about idealization in §4.7. But here I wish to presume that idealization does not guarantee to eradicate cross-temporal direct conflicts of preferences. This is a common ground shared by many subjectivist theorists[3] and I shall take it for granted. So, the problem of now-for-past cannot be avoided by idealizing.

Modifying these cases is no easy task, and I shall not attempt it here. Also, I do not claim that the reader must share the intuition against now-for-past sacrifice. There may be sensible disagreement in rock-bottom intuitions, and there is little point arguing about them. I can only proceed with the *presumption* that now-for-past sacrifice should be avoided.

But why consider now-for-past as a problem for time-neutralism? As I see it, there are two ways to spell out this idea. The first begins with the observation that time-neutralists are committed to *No Tragedian*. Or alternatively, a positive version which says that one ought to always prefer that one's whole life goes as well as possible. What goes for preferences, presumably, also goes for actions. So, time-neutralism requires now-for-past sacrifice, or at least rules its refusal as impermissible. (The action component, however, does not seem essential to me. What's correct or

[1] It's also easy to get the impression that Oscar's *present* preference that a priest is called is "irrational". He may be demented, overwhelmed by the fear of death, etc.
[2] There are at least two more factors that might render our intuitions unreliable. The first is that one might have *objectivist* prejudices. (cf. Brink 2003) One might, for instance, (tacitly) consider theist values as objectively wrong. But that's misunderstanding what's at stake with these cases. The second is to confuse the question of correctness with the question of rationality. (cf. Parfit 1984: 153-56, Nagel 1970: 74) Perhaps Oscar, on his deathbed, has very good *apparent* reason to disregard his past atheist preference as welfare-irrelevant. But that doesn't entail that he ought to disregard it.
[3] Brandt (1979: 245-60), Bykvist (2003: 118), Heathwood (2011a: 97-98), cf. Sobel (2011).

incorrect to prefer, arguably, aligns with what's correct or incorrect to do in such situations.) These cases also seem to suggest that *No Tragedian* should be replaced with a *forward-looking* principle along the lines of: One ought to prefer that her life *going forward* go as well as possible. (cf. Sullivan 2018: 22, Dorsey 2013: 164)

Alternatively, and more directly, our intuitions tell us that *one ought to be future-biased* in such situations (or so it seems). In *the Rollercoaster*, the correct thing for Brandt's son to do (as of his fiftieth birthday) is to privilege his *present* preference over his *merely past* preference – by chilling with a bourbon rather than riding a rollercoaster. And that's future-biased choice. (Think of present-bias as a limiting case of future-bias.)

To be more precise, notice that given preferentialism, time-biased preferences can be understood as *higher-order* preferences of some sort.[1] They are perspectival preferences over the satisfaction and frustration of lower-order preferences that are held at different times. This is true of cases of direct conflicts of (lower-order) preferences and also true of more mundane time-biases. In *the Deathbed*, Oscar is said to be future-biased if he prefers that his present theist preference rather than his past atheist preference is satisfied. In Parfit's *My Future and Past Operations*, Parfit prefers the satisfaction of his future preference for not being in pain over the satisfaction of his past preference for not being in pain. Interpreted this way, a natural solution is to adopt a *future-biased (rather than time-neutral) axiology*. (Dorsey 2018, Brandt 1979) That is, the time-neutralists are mistaken to think that past, present, and future preferences are all (equally) welfare-relevant. Rather, we should think of past preferences as welfare-irrelevant or at least to a lesser extent welfare-relevant. If satisfying merely past preferences confers no or lesser welfare benefit, then now-for-past sacrifice can be avoided.

Either way, now-for-past sacrifice pushes us away from time-neutralism. The preferentialist should either adopt a future-biased axiology, or (while retaining a time-neutralist axiology) adopt a

[1] I'm stretching the ordinary meaning of "higher-order preferences" a bit. An example of a paradigmatic higher-order preference: I prefer not to have the preference for pizza over pho. Time-biases concern the satisfaction and frustration of lower-order preferences, not the presence or lack thereof. Nonetheless, they can be said "higher-order" because they have lower-order preferences as constituents of their (*de facto*) content. (We are sticking to the *de facto* reading here. A time-biased agent need not believe in preferentialism or conceptualize her time-biases as such in order to be attributed with these higher-order preferences.)

future-biased principle about prudence – one that pays no or lesser attention to how life goes in the past.

All this, however, is a red herring. And I am in good company. Denis McKerlie states that "it is unfortunate that discussions of temporal neutrality often use examples of changing goals" (2007: 56, fn.8). Krister Bykvist points out that the intuition against now-for-past sacrifice "does not seem to have anything to do with *time as such*" (2006: 116). It does not concern "the tensed location of a preference, whether it is past, present, or future" (ibid.). Instead, it supports a sort of bias, so to speak, that's *non-perspectival*.

Let me motivate this alternative understanding with one obvious observation: Insofar as you find now-for-past sacrifice objectionable, you sense that there is something *peculiarly objectionable*. Many of us have mundane pro-future-bias intuitions. We find it appropriate to care less about past pain than future pain. Translating into subjectivist language, we find it appropriate to prefer that one's *future present-directed* aversion to pain is satisfied rather than that one's *past present-directed* aversion to pain is satisfied. (Here by present-directed I mean coincidence between the time-of-attitude and the time-of-object. Likewise, a future-/past-directed preference is one such that the time-of-object is later/earlier than the time-of-attitude.) But it should be clear that something different is going on with now-for-past sacrifice. The intuition cannot be (all) about the pastness and futurity (in terms of times-of-*attitude*) of the lower-order preferences in question. Rather, it should have something to do with the fact that the *past* preferences are *future-directed* and that they come into conflict with the *present present-directed* preferences.

Brink captures the intuition in terms of *authenticity*: "But then an agent who fails to act on ideals she accepts at the time of action would seem to display bad faith or lack of authenticity" (2003: 227). (By ideals he means preferences that are not conditional on their own persistence.) Similarly, Bykvist takes now-for-past sacrifice to violate the *autonomy* of person-stages: "[A] person-stage should not be allowed to poke its nose into another person-stage's business" (2003: 127). That is, each person-stage should have authority over how best to conduct *her* segment of life and not be hijacked by the opinions of other person-stages. [1]

[1] The term "person-stage" is used as a convenient way to pick out persons at different times without committing to the doctrine of temporal parts.

I shall have more to say about the autonomy of person-stages in due course. For now, if you agree that the intuition is latching onto *not* garden-variety future-biases but something concerning (however vaguely) authenticity or autonomy, then you might find attractive *concurrentism*. Concurrentism states a *necessary condition* for preference-satisfaction or preference-frustration to be *welfare-relevant*. It avoids now-for-past sacrifice and is consistent with time-neutralism.

> *Concurrentism:* An instance of preference-satisfaction confers welfare benefit *only if* there is overlap between the time-of-*obtaining* and time-of-*attitude*. (Likewise for frustration and welfare harm.)

As Chris Heathwood, who first proposed concurrentism, explains:

> [I]n order for a state of affairs to count as a genuine instance of desire-satisfaction, the state of affairs desired must obtain at the same time that it is desired to obtain. If I desire fame today but get it tomorrow, when I no longer want it, my desire for fame was not satisfied. A desire of mine is satisfied only if [I] get the thing while I still desire it, and continue to have the desire while I'm getting it. (2005: 490) [1]

Heathwood speaks of concurrence as a necessary condition for preference-satisfaction/-frustration. This sounds off. Unlike conditional preferences that get cancelled (and therefore not satisfied or frustrated), when preferences are satisfied albeit not concurrently, I'd like to say that these are genuine instances of preference-satisfaction, given that the conjunctive states-of-affairs do obtain. But instances of preference-satisfaction may fail to confer any *welfare benefit*. This may happen when the preference is ill-informed, etc. According to concurrentism, preference-satisfaction also fails to confer welfare-benefit when there is no overlap between the time-of-

[1] The label concurrentism has been used to mean many different things. Here I follow what's suggested by the above quote from Heathwood. (cf. Mariqueo-Russell 2023, Sarch 2013, Lin 2017) However, another comment by Heathwood seems to be suggesting something else:

> [T]he concept of concurrence may be less than perfectly clear. For instance, can present desires about the future or past (so-called now-for-then desires) ever be concurrently satisfied? (2006: 542)

If Heathwood has in mind a present desire that ceases to exist before its time-of-obtaining, then there should be no question given our understanding of concurrentism. So, I wonder if what Heathwood means by concurrentism is instead that the time-of-obtaining must be "within" the time-of-object. But that seems trivially true. Also, "concurrent" is often defined as time-of-*object* overlapping with the time-of-attitude. This seems inaccurate to me. I may from Monday to Wednesday prefer that I get a date somewhen between Tuesday to Friday. It would seem against the spirits of Heathwood (2005: 490) to say that I get benefited by getting a date on *Thursday*.

obtaining and the time-of-attitude. (Note that concurrentism as defined here only states a necessary condition for welfare-relevance. It doesn't specify the *temporal location* of welfare-benefit conferred by concurrent preference-satisfaction. I turn to the latter issue shortly.)

Concurrentism avoids now-for-past sacrifice. Consider for example *the Deathbed*. Satisfying Oscar's past atheist preference would be non-concurrent as the time-of-obtaining (t_{11}) does not overlap with the time-of-attitude (from t_0 to t_{10}). On the other hand, satisfying Oscar's present theist preference would confer welfare benefit since he gets what he prefers when he prefers it (insofar as the preference satisfies whatever other restrictions on welfare-relevance).

Importantly, the concurrentist response is perfectly compatible with time-neutralism. For one, *No Tragedian* can be upheld. Since non-concurrent preference-satisfaction/-frustration is welfare-irrelevant, it simply does not factor into the calculus of lifetime wellbeing, just as (say) ill-informed preferences. Merely past preferences do not get to compete with present present-directed preferences to begin with.

Moreover, concurrentism is a non-perspectival restriction. It's past/future-indifferent. Oscar's past atheist preference is welfare-irrelevant *not* because it's past, but because it cannot be concurrently satisfied.[1] To determine if an instance of preference-satisfaction meets the concurrent condition – that is, if the time-of-obtaining and the time-of-attitude ever overlap – all we need to know about is *calendar time*. It can be seen from any temporal perspective – or from God's atemporal perspective – that Oscar's atheist preference cannot be concurrently satisfied. For it's a non-perspectival matter that t_{11} is later than t_{10} and therefore there is no overlap between the time-of-attitude and the time-of-obtaining. An inter-personal analogy may help to illustrate the non-perspectival nature of concurrence: What I prefer about myself determines my welfare value. What you prefer about yourself determines your welfare value. What you prefer about *me*, however, does not determine *my* welfare value. But such a "restriction" on welfare-relevance is in no way biased towards any individual perspective. It's not as if the magical "I" picks out a privileged subject whose preferences have a say about everyone's wellbeing.

[1] See also Bykvist (2006: 129-30) on the time-neutrality of a view similar to concurrentism (to be discussed shortly).

Naturally, concurrentism does not imply that Parfit ought to be future-biased in *My Future and Past Operations*. On preferentialism, the badness of being in pain is analyzed in terms of a present-directed aversion to being in a painful state. Such an aversion, obviously, can be concurrently satisfied. And the satisfaction of such an aversion counts for wellbeing, by the light of time-neutralism, regardless of whether it's past or future.

4.5. Concurrentism and the Harmony View

However, concurrentism is a minority view. Many find it too implausible to be taken seriously. It's also been pointed out that it's unmotivated. Particularly, it's accused of being against the spirit of subjective theories. These objections need to be taken seriously. If concurrentism is such a terrible version of preferentialism, then presumably one would have to avoid now-for-later sacrifice on other grounds – say, giving up on time-neutralism. But concurrentism is not as implausible as it may appear. And in light of certain objections, the view can be improved without sabotaging time-neutralism or reinstating now-for-past sacrifice. Finally, it coheres well with the motivations for subjectivism.

Ben Bradley finds concurrentism implausible: "[T]here will be far less momentary well-being than we might have thought… Surely a great many of our desires are future-directed." (2009: 23) Similarly, Dale Dorsey finds it absurd that "all valuing attitudes I had *prior* to [2025] for things that occur in [2025] are irrelevant to my good" (2021: 194). Moreover, concurrentism seems to rule out *posthumous* harm and benefit (Bradley 2009: 22). Suppose that I, as of now, seriously want this thesis published. Unfortunately, I will drop dead right after sending the draft to the publisher. But it should matter if the publisher eventually publishes my thesis or tosses it into the trash folder; or so some would say. Finally, some find it obvious that non-current satisfaction of *past-directed* preferences can be welfare-relevant. Alexander Sarch considers such a case. During his college tenure, Jeremy "desired to drop out and do something else" but "grudgingly stuck it out" due to parental coercion (2013: 232). But in his thirties, "he starts to see his education as an important source of personal pride and comes to attach a great deal of value to it" (ibid.). It's hard to believe, Sarch claims, that Jeremy gains no welfare benefit whatsoever by coming to have this past-directed preference.

There are also complaints that concurrentism is unmotivated. To see this, let's begin by noting that the "experiential requirement" is rejected by most subjective theorists. The experiential requirement roughly says that a necessary condition for a state-of-affairs to confer welfare benefit is that it involves some positive experience. Subjectivists tend to reject this: Preference-satisfaction and -frustration can be welfare-relevant without being accompanied by any *feeling* of satisfaction/frustration. Indeed, most subjectivists do not even require knowledge of the said satisfaction/frustration in order for it to be welfare-relevant. For example, I can benefit from this thesis being published without ever knowing this – let alone experiencing any associated positive feelings. This may happen if I'm trapped in prison for the rest of my life or if I untimely drop dead. At any rate, by rejecting the experiential requirement, preferentialism is explicitly *anti-hedonist*: Pleasure or good feeling isn't necessary for welfare benefit.

But concurrentism seems to be motivated by hedonic intuitions or the experiential requirement (Bradley 2016: 11, Dorsey 2021: 194). Generally, non-concurrent satisfaction tends to leave one cold or even cause bad feelings, whereas concurrent satisfaction tends to feel great. Indeed, Heathwood himself has much sympathy for hedonism, and he regards it as a nice feature of concurrentism that it comes close to hedonism.[1] The obvious response: Concurrentism doesn't really satisfy the experiential requirement or collapse into hedonism. It allows concurrent satisfaction without good feelings. My thesis may get published without my knowledge when I'm spending my life in solitary confinement – while still longing for that. The concurrentist can say that I'm made better off. (It's not to say that a preferentialist must reject the experiential requirement. A preferentialist sympathetic to hedonism might adopt some hybrid view that satisfies the experiential requirement. But it remains that the experiential requirement is distinct from the concurrent condition.)

But that's exactly what's objectionable about concurrentism, according to Dorsey: Concurrentism comes *close to* satisfying the experiential requirement (extensionally) without really embracing it.[2] Put slightly differently, since concurrentism allows for welfare impact at an *experiential* distance

[1] Heathwood (2006) advocates for a version of preferentialism that embraces the experiential requirement and argues that it's equivalent to his preferred version of hedonism. But I should mention that Heathwood (2005) certainly does not propose concurrentism merely as a solution to now-for-past sacrifice or as a façade for his hedonist propaganda. One potential virtue of concurrentism is that it's able to solve the "Dead Sea apple" problem without giving up on actualist preferentialism (2005: 493). I'll say more about the general motivation for concurrentism-*like* theories later.
[2] Bradley (2016: 9) seems to concur this objection.

but not at a *temporal* distance, it is arbitrary and unmotivated: "If we needn't experience the objects of desire, what difference does it make when these objects occur?" (Dorsey 2013: 158) Concurrentist preferentialism is thus an unstable position trapped in a twilight zone. One had better either just embrace the experiential requirement or give up on concurrentism, but not occupy an unprincipled intermediate position.

So, the problem with concurrentism is two-fold. It leaves us with too little welfare benefit and harm. It's also unmotivated. Let me respond in turn. Consider first Bradley's claim that "there will be far less momentary well-being than we might have thought" given concurrentism.[1] There is some truth in this claim; for concurrentism disregards all preference-satisfaction and frustration that's non-concurrent. But it's doubtful if concurrentism implies a *massive* reduction in temporal wellbeing. For one, it needs to be clarified that *concurrent* preference-satisfaction does not mean satisfaction of *present-directed* preferences (cf. Sarch 2013). There is a subtle distinction that refers back to the distinction between time-of-object and time-of-obtaining. Present-directed preferences have their times-of-object identical to their times-of-attitude. An example is a preference to be not *presently* in pain (when one is in pain). But concurrent preference-satisfaction requires only overlap between time-of-attitude and time-of-obtaining. Say that from 2010 to 2025, I persistently long for publishing this PhD thesis. This preference has a "vague" time-of-object and is in a good sense *future*-directed rather than present-directed: At each time I have the preference, I'm hoping that its object obtains somewhen in the future relative to that time. And this preference can be concurrently satisfied by having the thesis published in 2025. Also, it seems to me that such concurrently satisfiable future-directed preferences should be plenty, since our preferences are expected to be somewhat stable – otherwise we wouldn't find cases of now-for-past sacrifice particularly interesting and troublesome in the first place.

[1] Bradley later changes his mind: "I am no longer sure whether this is an objection or merely a mildly interesting implication of the view." (2016: 9) But I shall proceed as if this is an unwelcome implication.

Also notice that Bradley focuses on temporal wellbeing. Concurrentism as defined here also disallow non-concurrent satisfaction to have *atemporal* wellbeing impact. But one might think that it's not so bad that there is little temporal wellbeing insofar as there is enough lifetime wellbeing. This suggests a modification of concurrentism as a necessary condition about *temporal* wellbeing only, consistent with non-concurrent satisfaction conferring *atemporal* wellbeing. But I agree with Dorsey (2013: 158, fn.19) that this is a bad view. For sure, even non-concurrentists might want to acknowledge atemporal welfare impact (in light of shapes-of-life, say), but we are for now focusing on welfare impact that, if exists, should be temporally local.

Moreover, notice that concurrentism doesn't commit to any particular view about the *temporal location* of welfare benefit and harm.[1] Concurrentism, at least as defined here, merely specifies a necessary condition for welfare-relevance. It leaves open how welfare-relevant preference-satisfaction contributes to *temporal* wellbeing. The latter question concerns time-of-*benefit*, to which there are two salient answers. There is firstly *the time-of-attitude view*, endorsed by Dorsey.[2] This view (in its unrestricted form) says that preference-satisfaction confers welfare benefit at all times the relevant preference is held. There is also *the time-of-obtaining view*.[3] It says that preference-satisfaction enhances the slice(s) of temporal wellbeing coinciding with the obtaining of the preferred state-of-affairs.[4] These two views are manifestly different *without* the concurrentist restriction. Suppose that *no* priest is called for dying Oscar. This satisfies his merely past atheist preference. The non-concurrentist time-of-attitude view says that Oscar is made better off from t_0 to t_{10} (when the atheist preference is being held) whereas the non-concurrentist time-of-obtaining view says that Oscar is made better off at t_{11} (when he is dying without a priest but preferring otherwise). (We are talking about *pro tanto* welfare benefit rather than all-things-considered welfare impact. The latter factors in the frustration of his present preference.) But contrary to what many seem to think (Dorsey 2013, Heathwood 2005), these two views still behave differently when coupled with concurrentism. Like Dorsey, I favor the time-of-attitude view (though I cannot defend it here). Combining concurrentism and the time-of-attitude view, however, it's false that "all valuing attitudes I had *prior* to [2025] for things that occur in [2025] are irrelevant" (Dorsey 2021: 194). Suppose that I get my thesis published in 2025 *right before* I cease to long for that. The concurrent time-of-attitude view says that my temporal wellbeing *throughout 2010 to 2025* gets enhanced. This seems the desirable result. Suppose otherwise that my preference lasted for a shorter period – say, from 2020 to 2025. The total amount of welfare benefit would be lesser. Insofar as the concurrentist condition is satisfied, the amount of welfare benefit turns out *ceteris paribus*

[1] Cf. Heathwood (2005: 490), Paul (2023).
[2] Also endorsed by Bruckner (2013) and Sarch (2013).
[3] Bradley (2016) favors this view over the time-of-attitude view but does not fully endorse it. Also recall the subtle difference between time-of-object and time-of-obtaining. While sometimes the view is branded as "the time-of-object view", this seems to be a mistake. I now prefer to run a marathon at *somewhen* between t_2 to t_8 in the future, or just vaguely "somewhen in the future". And I run a marathon at t_5. It seems that proponents of this view should say that I'm made better off at t_5 – rather than throughout t_2 to t_8 (which would be very bizarre). Also, unlike the time-of-attitude view, this view doesn't sit well with the very compelling idea that the amount of welfare benefit is positively correlated with the *lifespan* of the *preference* and its strength. And it's unclear whether its proponents want to reject the compelling idea.
[4] There are more sophisticated views that I cannot afford to discuss here. Lin (2017) defends asymmetrism: Whichever of the time-of-attitude or the time-of-obtaining that comes later in time. Purves (2017) defends fusionism: At the fusion of these times. But these views do not raise new questions that matter for our purposes.

proportionate to the length of the time-of-attitude. And this seems exactly what a preferentialist would want to say: The longer I want it, the better if I get it.[1] In all, concurrentism does not imply a massive reduction of wellbeing impact. And where it does predict reduction, it's where at least some find it welcome (as in now-for-future sacrifice).

Turn now to the problems concerning posthumous harm and past-directed preferences. Tossing my manuscript into the trash bin after my untimely death, according to concurrentism, doesn't harm me because there is no concurrent preference-frustration. The dead have no preferences. And poor Jeremy gains nothing by retrospectively endorsing his college years. These are clearly wrong verdicts; or at least some think.

Concurrentism can accommodate these cases – but perhaps with some modification. But before anything, a caution about intuitions. I take it as a fair assumption that one's intuitions about these cases are expected to *co-vary* with her intuition about now-for-past sacrifice. The more you are inclined to think that Oscar ought to disregard his merely past atheist preference, for instance, the more likely you'd deny posthumous harm. Not everyone thinks that now-for-past sacrifice should be avoided.[2] And not everyone considers posthumous frustration welfare-relevant.[3] Very plausibly, such variations in judgements are not isolated. After all, now-for-past sacrifice and posthumous harm are similar in various respects. Oscar's atheist preference and my desire to get the thesis published, from the temporal perspectives of their potential satisfaction, are both merely past. And these past selves can no longer feel anything. Jeremy's retrospective endorsement is also structurally similar. From the perspective of Jeremy in college, that retrospective endorsement is *merely future*. Thus, as I am only advancing concurrentism as a time-neutralist solution for those who *want* to avoid now-for-past sacrifice and therefore less inclined to regard these counterexamples as genuinely counterexamples, the objections have reduced dialectical force.

[1] Proponents of the time-of-obtaining view may also say that although I'm only benefited in 2025, the magnitude of benefit gets boosted by the time-of-attitude. But I dislike the time-of-obtaining view and also find this a weird suggestion.
[2] Dorsey (2021), Brink (2011), Sarch (2013).
[3] For example, Mark Overvold states that "it is hard to see how anything which happens after one no longer exists can contribute to one's self-interest"(1980: 108). Also note that while concurrentism implies that no preference held when alive can be satisfied posthumously, it does not imply that death cannot be bad for you. An anti-Epicurean concurrentist may say that death is bad for you because it deprives you of lots of *concurrent* preference-satisfaction that you would enjoy had you lived longer.

But in case you think concurrentism is overkill, here are some suggestions. Consider first Jeremy who comes to view his college years in a positive light. Sarch's case can be seen as a mirror image of now-for-past sacrifice, i.e., past-for-now sacrifice. A past preference is, so to speak, concurrently frustrated for the sake of the non-concurrent satisfaction of a present one. I share Sarch's intuition. Whether Jeremy retrospectively endorses his college years should make some welfare difference. But I think this is a case that can be accommodated by concurrentism *in its current form*. For it seems that what's really making the difference is not the mere presence of retrospective endorsement but the *context* thereof – which turns on the fact that Jeremy *has been through* college. The reason Jeremy comes to have the present preference, presumably, has much to do with what he has gained from his college education, which reverberates into the present and shapes his life as a whole. It then seems that what's making the difference is not so much his past-directed preference but a (weighty) *present-directed* preference. Jeremy is having a content life as of now (and presumably also in the future), we might say, thanks to all these gloomy days having been paid off.

Sarch objects to this "present-directed surrogate" response: It would reinstate now-for-past sacrifice. Sarch thinks that by parity, one must also say that during his atheist years, Oscar has always held a "*present-directed desire* to be such now that he *will* not receive any such visit at the end of his life" (2013: 234, original emphasis). But that present-directed desire is concurrently satisfiable and therefore welfare-relevant. My response is two-fold. I take that "present-directed surrogates" can mean two different things. On one reading, they are shadows of past- and future-directed preferences trivially generated. If you like, you can stipulate that for all past- and future-directed preferences, there is a corresponding present-directed preference in the form suggested by Sarch. But these surrogates are merely (quirky) redescriptions of the original preferences rather than genuinely distinct attitudes. They should therefore inherit the satisfaction conditions of the original preferences. Thus, if all Oscar had is just such a shadowy surrogate preference, then the concurrentist solution to now-for-past sacrifice remains intact. But perhaps Sarch means a present-directed surrogate in a more substantive sense – one that parallels Jeremy's present-directed content with his current life. On this reading, however, there is no reason to posit such a preference on behalf of the younger Oscar. Notice that Jeremy's present-directed preference crucially turns on the fact that he *has been through* college and that this experience has been *formative of* his current life. The past feeds into the present, so to speak. While it may be possible for young Oscar to form a present-directed preference in this way, we know it's unrealistic. This is due to a mundane asymmetry in life that past ordeals may positively affect the present but not, or

at least very rarely, the other way around. We could, if you like, send a message back to the young atheist Oscar informing him that no priest is called at the end. And he could be very thrilled – despite the agony of his dying self – and hold onto that future prospect as a lifetime treasure. But clearly, that's far from a natural way to fill in the case.[1]

The posthumous problem, it seems to me, is more damaging. (Though remember that the intuition is far from universal.) I propose a revision of concurrentism in light of this. The case of publishing my PhD thesis posthumously, just as now-for-past sacrifice, involves a merely past future-directed preference. But there is one important difference. By dropping dead, I do not come to *disprefer* publishing my thesis. It thus seems plausible that whether merely past preferences matter hinges on whether they are *contested* by one's present present-directed preferences. In *the Deathbed*, the dying Oscar has disavowed atheism and become a theist. Hence "direct conflict". But I don't come to *change* my preferences by dropping dead. The corpse simply cannot have preferences.[2] To amplify this point, consider Ostara who is otherwise identical to Oscar but for being utterly *indifferent* on her deathbed: Ostara simply doesn't care whether a priest is called. Here it seems more plausible that her merely past atheist preference can carry some weight. It makes sense for her to reason: "Well, although I don't care right now, I've always been an atheist. So, I should honor my past selves by declining a priest."

This suggests a loosening of concurrentism along the lines of Krister Bykvist's *harmony view* (2003, 2007). The harmony view, in a nutshell, says that we "ought to choose that action that will best satisfy (i.e., maximize the total sum of satisfaction of) all personal wants that are grounded in *sustained priorities*" (2003: 125, emphasis mine). While Bykvist proposes his view as one concerning what one ought to do, it can be easily translated into one concerning which preferences

[1] It makes much sense to talk about redeeming the past but not so much redeeming the future (though one need not understand the former as conferring welfare benefits to the past). This brings us to a related point. Jeremy's case is one where structural features become salient. Jeremy's past ordeal has become meaningful by virtue of how it reverberates into the present. Jeremy's life has a pleasing narrative structure compared with one in which his college years are "wasted". How exactly to pin down the welfare benefits of structural features is a vexed issue (§4.2). But presumably a subjective theorist would like to say that shapes-of-life matter because we value them. (Bruckner 2018; cf. Dorsey 2015: 310)

[2] I submit that things may become more complicated if counterfactual considerations are brought into the picture. What if, you may wonder, I would suddenly come to detest academics were I to hang on a bit longer? Should that counterfactual future preference affect what ought to be done in the actual world? I think a positive response is sensible. (cf. Portmore 2007) But for our purposes, we may stipulate that I wouldn't change my mind. Throughout this chapter, I've been assuming away inter-world conflicts. In particular, all instances of preference change are not choice-dependent. This is certainly unrealistic, but it makes the discussion manageable. But see Bykvist (2006) and Pettigrew (2019) on counterfactual considerations and choice-dependent preference change.

are welfare-relevant.[1] Central to the harmony view is the notion of *consensus* among *person-stages*[2] at different times. As Bykvist explains, "[E]ach person-stage has a *veto* over what they should do with their lifestage so any conflicting preferences of other persons or other temporal stages of the same person are to be completely disregarded" (2007: 74, emphasis mine). On the other hand, a past future-directed preference or a future past-directed preference *is* relevant only if it's in consensus with one's present present-directed preference. So stated, the harmony view appears not so different from concurrentism. But Bykvist allows for posthumous harm and benefit. Regarding my PhD thesis, he would say that the priority of having it published is "trivially sustained" because I have no preference whatsoever when I'm dead (2003: 126-27). There is *silent* consensus, as it were.

Now, I wonder what Bykvist would say about dying Ostara who is utterly indifferent. The situation, it seems to me, is relevantly similar to posthumous cases. At least, the intuition about Ostara should be to some extent in line with posthumous cases if not identical.

No priority is sustained in Ostara's case (as she has no priority in her deathbed); but nor is she *vetoing* her past preference (again, as she has no priority). This situation seems to occupy a middle ground between Oscar's case and cases with sustained priorities. It is thus apt to say that his merely past atheist preference is neither completely irrelevant nor relevant to the fullest extent. If that's correct, then we should find attractive a *generalized* harmony view on which consensus and priority come *in degrees*. The relevance of future- and past-directed preferences, in turn, depends on the extent to which they are in consensus with present-directed preferences and also on the relative priority of these present-directed preferences.

Indeed, this generalization suggests itself if we consider two obvious facts. The first is that one's pairwise preferences at a given time, taken together, constitute a *preference ordering* at that time. We've been focusing on certain pairwise preferences with the unrealistic presumption that they exhaust one's preference ordering at a time. But surely in *the Aspiring Poet*, Parfit might have considered other career choices and had a more comprehensive ranking. Suppose that Parfit used to have the preference ranking {poet > plumber > director > philosopher} but now has the ranking

[1] That is, Bykvist thinks that "all" preferences are relevant but sometimes one ought not to maximize lifetime wellbeing. Concurrentism could have received this alternative formulation as well. But I take that *No Tragedian* is very compelling. It is to be preserved unless there is very good reason against it (which I turn to in §4.6). So, I prefer formulating concurrentism as an axiological thesis.
[2] Again, the term is used as a convenient way to pick out persons at different times without any ontological commitments.

{philosopher > poet > plumber > director}. Neither plumber nor director has ever been his *absolute* priority. But obviously, there should be a difference in welfare impact whether Parfit chose to become a plumber or a director. The second fact is that preferences come in different *strengths*. (This applies to pairwise comparisons as well as preference orderings.) Accordingly, we should think of consensus as a matter of degree. Indeed, we might think of indifference between A and B as in consensus with – to the *lowest* degree – both $A > B$ and $B > A$. In all, the welfare-relevance of future- and past-directed preferences is a matter of degree because both consensus and priority are matters of degree. How exactly to fix such a function is a complicated matter, but hopefully the attraction of such a generalized account is clear (for those who find concurrentism too restrictive.)

Call the view I've motivated above *concordism*. Though underspecified, it has three distinctive features. Firstly, unlike concurrentism, it allows for posthumous harm and benefit. Secondly, unlike Bykvist's harmony view, it doesn't require *full* consensus in order for future/past-directed preferences to be welfare-relevant. Ostara's past atheist preference can be welfare-relevant despite that she's now indifferent. Thirdly, present-directed preference orderings are privileged. Suppose that Parfit's past future-directed and present present-directed orderings are otherwise identical except that the positions of poet and philosopher are reversed (in line with the original story). However the welfare-relevance function is specified, it would have to be the case that being a philosopher (as of now) confers more welfare benefit than being a poet.

And to be clear, I am not wedded to concordism *per se*. I've argued that insofar as one finds now-for-past sacrifice objectionable, one should find attractive *some* restriction on non-concurrent preference-satisfaction. The restriction can take the form of concurrentism, the harmony view – or even less austerely, concordism. The reader is welcome to find concurrentism the most attractive after all because (say) she regards posthumous harm as impossible. I don't see any definitive reason to favour one over another. And much of what I say below about concordism generalizes to this family of views.

Turn now to the complaint that concurrentism is unmotivated as a subjective theory. This objection seems to carry over to concordism. Let me first reiterate that concordism is not a closet hedonist view. Nor does it satisfy the experiential requirement (cf. Heathwood 2006). This is most obviously seen in what concordism says about posthumous harm. But if the experiential requirement is not what's at issue, wonders Dorsey (2013: 157-58), what else could be motivating concordism? Put differently, concordism without the experiential requirement seems arbitrary: It

discounts preferences by *temporal* distance but not by *experiential* distance. There is no justification for that asymmetry.[1]

It should be clear up to this point that concordism is of some good: It avoids now-for-past sacrifice in a time-neutralist way. But more can be said. Let me start with a quibble: Speaking of "temporal distance" in this context makes concordism sound more arbitrary than it in fact is. Better to think of temporal distance as a proxy for *psychological distance*. Here psychological distance is a measurement of change in preferences – or whatever attitudes that are *prudential-value-conferring* on subjectivist theories. (So, psychological distance is very different from experiential distance about which hedonism is concerned.)

And I argue that discounting by psychological distance comports well with subjectivism. Subjective theorists take seriously each person's authority to determine what's a good life for her. Some stranger finds my career as a philosopher stupid and recommends investment banking. I may give her opinion some serious thought and change my mind. But whatever happens, it cannot be the case that her preference *directly* determines that investment banking is better for me. And such inter-personal *non-interference*, it seems, should carry over to the intra-personal. As Bykvist states, "[a] person-stage should not be allowed to poke its nose into another person-stage's business" (2003: 127). That is, each person-stage should have some authority over how best to conduct *her* segment of life and not be hijacked by the opinions of other person-stages. I see this as a natural corollary of the subjectivist commitment that the variation in psychological inclinations is normatively significant. Unlike the stranger, I just happen to find attractive philosophy instead of investment banking. Such inter-personal psychological variation may result from differences in personal traits, happenstances, choices, and so forth. And all these also happen intra-personally across time, albeit usually less dramatically. So, insofar as I have privileged authority (over the stranger) in determining what's good for *me*, it's unclear why each person-stage of mine should not enjoy some authority in determining what's good for *her*. Merely past and future preferences

[1] Dorsey sometimes speaks as if concurrentism is making an arbitrary distinction between *spatial distance* and temporal distance, as in "there can be no principled reason to treat one form of distance (spatial) as more evaluatively significant than another form of distance (temporal)" (2013: 158). But if I understand correctly, "spatial distance" here should be read as a shorthand for the lack of awareness that one's preference is satisfied/frustrated. Surely, no concurrentist (and no subjective theorist generally) cares about whether the object of one's preference obtains *here or there*.

Forrester (2023) defends a form of concurrentism with an epistemic requirement: Preference-satisfaction is beneficial only if one is aware of the satisfaction. The view implies that posthumous benefit is impossible. Forrester argues that concurrentism and the epistemic requirement go hand in hand. But if what I (and Bykvist) say below is correct, one can happily embrace the former without the latter.

are much like other people's preferences about me in this regard. So, some restriction on these preferences seems a well-motivated subjectivist move.[1] Moreover, this emphasis on the authority of subjective perspectives (of person-stages) has little to do with what motivates the experiential requirement or hedonism. Although preferences and good feelings are both in some sense "subjective", it's only the former that "gild and stain the world" with values and thereby demarcate genuine subjective *perspectives*. Hedonic experiences, on the other hand, are merely receptive and regarded by subjectivism as inessential to prudential value.

The inter-/intra-person analogy is not perfect, of course. An individual, however dramatically her preferences fluctuate, has a single life to lead. One might then worry that granting such "autonomy" to person-stages is to disregard "the significance of the fact of the continuing self-identical person, and… that we live our lives partly from a diachronic perspective" (McKerlie 2007: 62). Such a worry raises complicated questions regarding personal identity and the normative significance thereof.[2] I have much more to say about this in §5.3. But a very brief response here. The worry can be partially assuaged by the fact that intra-personal *causal interference* is pervasive. I am much more inclined to take seriously the recommendation for investment banking from my past or future person-stages than from a random stranger. More generally, what I now find attractive is to a great extent shaped by my past choices as well as what I anticipate about my future person-stages. This vulnerability to past and future preferences can give rise to a diachronic perspective rather spontaneously. This doesn't always happen, of course. Parfit now sees nothing good for him in poetry despite the fervent recommendation from his past person-stages. I think the right thing to say here is that there is *no* diachronic perspective to be found. How his philosophically-minded person-stages conduct their life is thus a private matter that should be protected from interference. The inter-personal analogy is again apt. If you and I somehow manage to (autonomously) reach consensus regarding what each of us prefers about ourselves and what we prefer about one another, then there might be a collective perspective that's authoritative for both of us. We can then happily conduct our lives together. But in case we fail (which is often the case), non-interference seems exactly correct.

---

[1] I'd also love to reframe the above argument as saying that concurrentism or concordism best satisfies the resonance constraint (Railton 1986) – suitably specified. But in light of the many inconsistent proposals for specifying the resonance constraint and the apparent dialectical impasse (Dorsey 2013, Bradley 2016, Lin 2017), it may be difficult to pursue this route.

[2] Very briefly, I regard the normative question of diachronic unity as largely independent from the metaphysical theories of persistence. A stage-theorist who denies numerical identity across time, for example, can acknowledge that distinct person-stages, from a normative perspective, are not as "separate" as distinct individuals.

This roughly concludes my response to the problem of now-for-past sacrifice on behalf of the time-neutralists. The crucial point is that what makes now-for-past sacrifice problematic is not that undue weight is given to *past* preferences, but instead that each person-stage should enjoy certain autonomy and authority over her own segment of life. Thus, preferences about a certain person-stage issued from another person-stage that are not in consensus with her own self-directed preferences should have no, or diminished, welfare-relevance.

However, there is one implication of concordism that one might find objectionable. By being a past-/future-neutral view, it demands a form of *now-for-future* sacrifice in which from the future perspective, one's present preference is merely past.

> *The Russian Nobleman:* In several years, a young Russian will inherit vast estates. Because he has socialist ideals, he intends, now, to give the land to the peasants. But he knows that in time his ideals may fade. To guard against this possibility…[he] signs a legal document, which will automatically give away the land… (Parfit 1984: 327)[1]

One might think that the ideals in question are other-regarding and therefore welfare-irrelevant, or that the nobleman's future bourgeois ideal is somehow ill-informed or defective. But suppose all that is not at issue here. All else being equal, concordism is *against* signing the document since that would be poking noses into the business of future person-stages.[2]

If the problem with now-for-past sacrifice lies in authenticity or autonomy of person-stages, then it seems to be the correct verdict that such now-for-future sacrifice is prudentially required. What matters is the veto power of the preferences that are concurrently satisfiable, not whether the preference is past, present, or future. That said, one might still find the verdict intuitively incorrect despite wanting to avoid now-for-*past* sacrifice. If such past/future-asymmetric intuitions survive

[1] Parfit also raises a question about personal identity and moral obligation here. I've truncated the case to avoid distraction. Let's assume that the nobleman remains, through changing ideals, the very same person in every relevant sense.

[2] Concordism doesn't imply now-for-future sacrifice in a stronger sense. It doesn't imply that the young nobleman ought to actively facilitate the flourishing of his future decadent lifestyle he now distastes – say, by distancing himself from the socialist movement and socializing with the bourgeoises. Presumably, the young nobleman also possesses plenty of (very weighty) *present-directed* preferences that can only be satisfied by retaining the current lifestyle. It would be selling short his present person-stages by prematurely preparing for the rainy day, as it were, in an overly costly manner.

reflective equilibrium, then perhaps time-neutralism isn't for that person after all. In that case, perhaps one would have to say that one's present (and future) preferences are somehow privileged. I pursue this route in the next chapter.

To summarize, the problem of now-for-past sacrifice does not show that time-neutralism is inconsistent with subjective theories. Rather, as I see it, it urges a restriction on welfare-relevant attitudes that should be welcomed by the subjectivists. However, this is not to say that subjectivism poses no problem to time-neutralism. As I shall argue in the next section, subjectivists have an easy way out of *the Argument from Wellbeing against Time-Biases*.

But before moving on, I wish to pick up two points about time-biases in relation to the temporal features of preference-satisfaction. The first is that given subjective theories, there could be a good sense in which future-biases are *not practically inert* after all. Recall that future-bias, given preferentialism, is understood as a sort of higher-order preference. Consider the simple case of *My Future and Past Operations*. This is a paradigmatic instance of hedonic future-bias. But the hedonic/non-hedonic distinction is only superficial according to preferentialism. Ultimately, what makes this an instance of future-bias is that Parfit's past (present-directed) aversion weighs more heavily in the wellbeing calculus than his future aversion. (Since the past operation lasts longer, his past aversion lasts longer or is stronger.) It's these first-order preferences (or desires/aversions) that determine the unadulterated (dis)values about which one is time-biased. Thus, for Parfit to be future-biased is for him to have a higher-order preference over the satisfaction of his past aversion over the satisfaction of his future aversion. Now, future-biases can be practically relevant because past preferences can be future-directed. These past future-directed preferences can, as of now, be satisfied or frustrated and thereby affect wellbeing – without causally affecting the past. This is true of both the time-of-obtaining view and the time-of-attitude view (even on their concordist versions). Therefore, the question of whether one ought to be future-biased concerns not merely mental states but also choice behaviors. Does this shift the dialectic in any way? Well, one might think that the requirement of past-/future-neutrality becomes objectionably demanding when it's practically relevant. Perhaps it's particularly implausible that one sometimes ought to *actively* satisfy merely past preferences (when they are welfare-relevant). Or one might think the opposite: Future-bias appears permissible precisely because it's seen as practically inert. The practical relevance of past preferences thus undermines our pro-future-bias intuitions. I cannot decide on this issue here. But it's worth pointing out this interesting feature of preferentialism, with its potential normative implications.

The other observation is that *Wellbeing Discounting* may be inconsistent with the *time-of-obtaining* view. *Wellbeing Discounting* basically says that time-biases are (*de facto*) preferences over *cross-temporal* distribution of wellbeing. Oscar may prefer satisfying his present theist preference rather than his past atheist preference simply *because* he is future-biased. (Forget about autonomy for a minute.) But given the time-of-obtaining view (in its non-concurrentist form), what Oscar prefers is lesser *present* wellbeing. This is because whichever first-order preference is satisfied, the time-of-obtaining is the present. That is, he is making a tradeoff "within" his present wellbeing rather than between his past and future wellbeing. This is a somewhat weird (and arguably bad-making) feature of the time-of-obtaining view. But at any rate, this is not a damning problem for *the Argument from Wellbeing*. After all, all else being equal, lesser present wellbeing means lesser lifetime wellbeing. It's easy to see how the argument can be modified to accommodate the time-of-obtaining view, and I shall not attempt it here.

4.6. Being A Tragedian

Subjective theorists of wellbeing can ward off *the Argument from Wellbeing*. This is because given subjective theories, *No Tragedian* turns out resting on a false presumption that there are non-perspectival facts about lifetime wellbeing. Instead, lifetime wellbeing should be understood as a measurement that's relative to temporal perspectives. Moreover, *perspectival* lifetime wellbeing as such is *shifty* for time-biased agents. Their time-biased preferences *make it the case* that their lifetime wellbeing is not stable across temporal perspectives. Finally, it's often the case that by virtue of being time-biased, one is *not* preferring a life that's overall worse off from that very time-biased perspective.

Before expounding on the argument, there is an immediate worry about flogging a dead horse. Here is the thought: Subjective theorists are *Humeans* about normative reasons. They deny that there are attitude-independent prudential reasons. Otherwise they would have to agree with the objectivists that some things (e.g., pleasure and friendship) are of objective prudential value and are universally desirable regardless of one's subjective attitudes. Thus, they cannot accept conclusions (4) and (5) of *the Argument from Wellbeing* because these conclusions are objective and universal ought-claims about prudence. Likewise, they cannot accept *No Tragedian*, a principle that states attitude-independent normative reasons about prudence. As you cannot

convince Hume that he shouldn't prefer the destruction of the whole world over the scratching of his finger, you cannot convince the Humeans that they ought not to be future-biased.

Indeed, Sullivan concedes that such arguments "will not have any grip on the neo-Humean who insists he is not making any error in valuation" in having apparently incorrect preferences (2018: 25). (Though Sullivan also claims that her argument against time-biases is axiologically neutral.) But not everyone appears to think that subjectivists are immune to such an argument right off the bat. For example, when discussing the problem of now-for-past sacrifice given subjective theories, Dorsey (2013: 163-64) thinks that there is a genuine question of whether subjective theorists should accept something like *No Tragedian*. That would seem to make sense only if subjective theorists can be non-Humeans. Likewise, Bykvist argues that cases of *choice-dependent* preference change reveal it to be inconsistent to accept "both that your desires determine what is good for you and that you must prefer what is better for you" (2010: 1). I have been setting aside choice-dependent preference change, as I wish to focus on cross-temporal conflicts of preferences and cannot afford the complication of inter-world conflicts (135, fn.2). But the problem of choice-dependent preference change is roughly this: By endorsing *No Tragedian*, the subjectivists would mandate wellbeing-maximizing preference change that is apparently objectionable. For instance, it seems to follow that I ought to take a "complacent pill" which will make me, for the rest of my life, prefer whatever states-of-affair that happen to obtain.[1] It's not the place to evaluate that argument; but the point here is that there is a *non-trivial* question of whether subjective theorists should accept *No Tragedian*, according to Bykvist.

What to make of this? My suspicion is that we are *not necessarily* flogging a dead horse. In particular, I side with Bykvist and Dorsey that there doesn't have to be a *flat-out inconsistency* between subjectivism and *No Tragedian*. If that's true, then there is a substantive question of whether the argument succeeds against subjectivism. Although I cannot argue for that in detail, here are several brief observations.

[1] Dorsey (2021: 211-25) responds by rejecting principles like *No Tragedian*. *No Tragedian* is false because it takes prudence to aim for *levels of wellbeing*. According to Dorsey, prudence is in the business of making preferences satisfied instead of making satisfied preference. (cf. The procreative asymmetry: There is good reason to make people happy but no reason to make happy people.) Nonetheless, both Dorsey and Bykvist think that *No Tragedian* is only threatened by choice-dependent preference change, which I have been setting aside throughout the chapter.

It's hard to deny that subjectivism and Humeanism are natural bedfellows. After all, a major selling point of subjectivism is that it provides a naturalistic reduction of prudential value, which comports well with the Humean project of naturalizing normativity. But it seems that subjectivists *don't have to be* Humeans. This position is explicitly taken by Chris Heathwood (2011a). According to Heathwood, it's coherent for a subjectivist to say that what provides (normative) prudential reasons is not desires *per se*, but the *conjunctive* states-of-affairs of desire-satisfaction. Crucially, the conjunctive states-of-affairs are the sort of thing that one may fail to desire despite having value-conferring first-order desires. Being a non-Humean, the subjectivist can say that one has attitude-independent reasons to care about and promote these conjunctive states-of-affairs even if she lacks the corresponding second-order desires. Heathwood also argues that subjective theorists *had better not be* Humeans. This is because the most plausible version of Humeanism is presentist: It grounds reasons in one's present but not future or past desires. If so, then a Humean subjectivist might have no way to criticise *time-biases*. Subjectivists should be non-Humeans, says Heathwood, *precisely because* they need something like *No Tragedian* to criticise time-biases.

Or perhaps even a *Humean* subjectivist can endorse *No Tragedian*. Michael Smith (2009) argues that a Humean might be able to *derive* substantive normative reasons from purely procedural norms of rationality. If there are any substantive norms so derived, *No Tragedian* would seem a plausible candidate. Or one might take a leaf from the neo-Kantians and argue that having certain attitudes towards one's own wellbeing is *constitutive of* being a welfare subject. Again, given this neo-Kantian commitment, it's not a far-fetched claim that being a tragedian would undermine one's status as a welfare subject. I can see an immediate worry that these views are not genuinely Humean. But however one labels things, these views do seem to preserve the core of subjectivism. To illustrate, there is an important difference between saying that one ought to promote one's wellbeing and saying that one ought to desire knowledge and friendship (because they are attitude-independently valuable). Intuitively speaking, the former norm is higher-order and global, but the latter is first-order and local. And it seems that the dividing line between subjective and objective theories of wellbeing lies precisely in whether norms of the *latter* kind exist. The subjectivists reject the latter kind of local norms because they get the explanatory order in reverse. When it comes to global norms like *No Tragedian*, however, it doesn't seem immediately self-undermining for the subjectivists to endorse them. (Heathwood's argument also seems to latch onto this distinction between local and global reasons.)

The reader may or may not find the above arguments convincing. But the point is that considering these controversies, it's warranted to leave it an open question whether subjectivists can, or even should, get on board with *No Tragedian*. Heathwood, for one, thinks that they can and *should* – conditioning upon that they are time-neutralists.

But I suspect that Heathwood is mistaken. It's doubtful whether *No Tragedian* even makes sense given subjectivism (not for the reason that subjectivism entails Humeanism). The heart of the issue is that time-biased preferences seem to be *welfare-relevant* attitudes. More specifically, they are second-order preferences that affect the welfare-relevance of the first-order preferences they're concerned about. And since time-biases are perspectival, they make perspectival and shifty across time the welfare-relevance of those first-order preferences. Thus, for subjects who are *in fact* time-biased, there are no stable and non-perspectival facts about lifetime wellbeing. Furthermore, a time-biased agent need not be preferring a life that is, *from that very perspective*, overall worse off. Often, the opposite is the case. (I take that on this view, lifetime wellbeing should be perspectival for all possible welfare subjects, including those who are always time-neutral. It's just that for those subjects, their lifetime wellbeing remains stable across all perspectives. It would be pretty strange, at least in this context, if the *meta-ethical* question of whether a kind of evaluative fact is perspectival, is itself an attitude-dependent matter.)

Consider again Parfit's *My Future and Past Operations*. When it's Thursday (call it t_2), rather than having a shorter operation awaiting, Parfit prefers a longer operation over and done with. Given preferentialism, Parfit prefers *de facto* that his future (present-directed) aversion to pain rather than his past (present-directed) aversion is satisfied. *Wellbeing Discounting* implies that Parfit prefers (*de facto*) a greater harm to his past temporal wellbeing over a lesser harm to his future temporal wellbeing. (I assume the time-of-attitude view to streamline the discussion.[1]) Then, setting aside shape-of-life considerations (which are presumably absent in this case), Parfit's preference regarding lifetime wellbeing on Thursday can be represented as: $(w_2, t_2) > (w_1, t_2)$. The only difference between w_2 and w_1 is that in w_2, a longer operation takes place on Wednesday; but in w_1, a shorter operation takes place on Friday. This is a perspectival preference that can only be

[1] The time-of-obtaining view might be extensionally equivalent when restricted to present-directed preferences. It would predict different results, however, when it comes to how the satisfaction or frustration of *time-biased* preferences affect *temporal* wellbeing. But the difference between the two views disappears on the level of lifetime wellbeing, and they make lifetime wellbeing shifty all the same. Shiftiness is due to the perspectival nature of time-biased preferences as higher-order preferences, and not due to any particular view about how preference-satisfaction contributes to temporal wellbeing.

adequately represented by centred worlds. It's *de nunc* and *de se* (though I'm suppressing the *de se* for simplicity).[1] As of Monday (t_1), however, Parfit prefers differently. In terms of lifetime wellbeing, he so prefers: $(w_1, t_1) > (w_2, t_1)$. This reflects that Parfit has undergone "preference reversal" (cf. Ch.2).

All this is familiar. I've merely connected Parfit's future-bias with lifetime wellbeing, following *the Argument from Wellbeing*. But on preferentialism, preferences are *value-conferring*. So, this seems to be what a preferentialist should say: In virtue of being future-biased, Parfit makes *shifty* his lifetime wellbeing. That is, as of t_1, his life fares better as a whole in w_1 than in w_2; but as of t_2, his life fares better in w_2 than in w_1. And this is because at t_2, by being future-biased, Parfit makes it the case that the satisfaction (or frustration) of his past aversion to pain makes lesser welfare contribution than that of his future aversion. That is, from this very future-biased perspective, w_1 contains more lifetime wellbeing than w_2 because the welfare-relevance of his past aversion is discounted relative to that of his future aversion. Indeed, if he were absolutely future-biased, then his past aversion would be completely welfare-irrelevant from his t_2 perspective. But his t_1 perspective is different. At t_1, the longer operation makes more welfare contribution than the shorter one due to a different perspectival, second-order preference.

So, because Parfit's second-order preferences over the welfare-relevance of his first-order preferences shift across perspectives, there should be *no* stable, non-perspectival facts about Parfit's lifetime wellbeing. Rather, there are only perspectival facts about lifetime wellbeing that shift over time. Parfit doesn't violate *No Tragedian* after all. As of t_2, he prefers a life that is, *from the perspective of* t_2, better for him. Should Parfit be (always) time-neutral, his lifetime wellbeing would be stable, in which case he need not violate *No Tragedian* either.

We might make this point more precisely in terms of a *dilemma* for the time-neutralists. *No Tragedian* considers lifetime wellbeing as either non-perspectival or perspectival. If non-perspectival, then it can be straightforwardly rejected by the preferentialists. There just is no such thing as non-perspectival lifetime wellbeing. Or *No Tragedian* is perspectival, formulated as: At any given time *t*, one ought not to prefer that her life, *from the perspective of t*, fares worse as a

[1] A more accurate representation would be in terms of sets of centred worlds. I'm taking a shortcut here by making Parfit having preferences over maximally specified states-of-affairs. Also, it's better to have *person-stages* as centres (Lewis 1976) so as to reflect the personal-time conception, but the standard presentation suffices for my purpose.

whole. But then the conclusions wouldn't follow. A time-biased subject, being time-biased, need not prefer a life that's worse off from this time-biased perspective. Often, the opposite is true: By virtue of being future-biased, for example, one makes it the case that it's overall better for her to have pain in the past and pleasure in the future. (Though it's possible for her to "accidentally" prefer *de facto* a worse life from that perspective due to shape-of-life reasons.) Thus, nor can the time-neutralists rely on a perspectival version of *No Tragedian* to argue against time-biases.[1]

Is this a satisfactory response? Well, there is reason to dislike shifty lifetime wellbeing. Alexander Dietz finds it implausible that "even when we hold fixed all the facts about what happens in a person's life, the truth-value of welfare judgments can depend on the time at which we are making them" (2023: 990). Chris Heathwood makes a similar remark: "But just as it makes no sense to ask, say, how long some life was relative to some time, it makes no sense to ask how good overall some life was relative to some time." (2011b: 29) Moreover, shiftiness would seem to render practically useless the notion of lifetime wellbeing. In the context of distributive justice, Matthew Adler comments that "massive incomparability in the ranking of life-histories" is intolerable "for purposes of moral decision-making" (2012: 427).

If shifty wellbeing is so bad that no reputable theory of wellbeing should allow for it, then *the Argument from Wellbeing* can remain intact. Surely, the argument is not supposed to accommodate any crazy theory of wellbeing (e.g., a theory that the more time-biased one is, the better her life is). Indeed, if shifty wellbeing is something to avoid, then the subjectivists are advised to adjust their theory so that no such thing follows. This sounds feasible. After all, most subjectivists do *not* say that *any* preference is welfare-relevant. And time-biases, one might argue, are just among the preferences that are welfare-irrelevant.

[1] Just to clarify, lifetime wellbeing becomes shifty only if one has certain higher-order preferences that change over time. Mere change in first-order preferences has no such implication. Consider for instance *the Deathbed*, where Oscar prefers from t_0 to t_{10} that no priest is called at t_{11} but prefers at t_{11} that a priest is called at t_{11}. Say that he's not time-biased. Regardless of whether you favor concurrentism or non-concurrentism, the time-of-attitude or the time-of-obtaining view, there are stable facts-of-the-matter concerning Oscar's temporal and lifetime wellbeing (conditional upon the satisfaction/frustration of his preferences) that remain the same throughout t_0 to t_{11}. For example, given the non-concurrentist time-of-attitude view, satisfying Oscar's atheist preference would benefit his wellbeing from t_0 to t_{10} (and not at t_{11}). This is true whether from the perspective of t_0 to t_{10} or of t_{11}. His lifetime wellbeing becomes shifty, however, should he have changing higher-order preferences regarding which first-order preference is satisfied.

I do find shifty lifetime wellbeing counterintuitive. On reflection, however, I am genuinely unsure whether this is a problem or just an interesting (and welcome) implication of subjectivism. In what follows, I consider various proposals for making lifetime wellbeing non-perspectival or stable. None of them turns out to be promising. But I don't see this necessarily indicating that subjectivism is somehow deeply flawed. Whether shiftiness is a vice or a virtue may be in the eye of the beholder.

4.7. Why Not Be a Tragedian?

Let us begin with a relatively *conciliatory* strategy. One may concede that time-biases render wellbeing perspectival but also try to recover some sense of non-perspectival wellbeing. The rough idea is that we can pool the diverging perspectives and reconcile them, just as we can pool different opinions of different individuals and reach a compromised consensus. There is the immediate worry that lifetime wellbeing in this derivative sense is a frivolous add-on rather than a genuine measurement. Regardless, the most damaging problem is that derivative lifetime wellbeing as such would almost inevitably fail to be a time-neutral one. It will, so to speak, inherit the time-biased perspectives. However the different perspectives are weighed against one another, the output would very unlikely converge on a time-neutral evaluation in which no temporal wellbeing is discounted. Consider for instance a *consistently absolutely* future-biased person. From each perspective of hers, her lifetime wellbeing is her *future* wellbeing. There is no plausible function that could deliver a non-perspectival lifetime wellbeing that's time-neutral. Put differently, although time-biases of different sorts may *accidentally* cancel out each other, it's certainly not always the case. But no time-neutralist would say that one ought to always maximize derivative lifetime wellbeing as such.

Alternatively, one might furnish a certain perspective with the privilege of determining non-perspectival lifetime wellbeing. But which one? The obvious answer: The time-neutral one(s). This sounds *ad hoc* and question-begging. Also, there may be *none*. That said, there is a better way to pursue this thought. It might be argued that time-biases are ruled out as *welfare-irrelevant* by certain well-motivated restrictions on the welfare-relevance of preferences. If time-biases are welfare-irrelevant – and therefore leave intact how satisfaction/frustration of lower-order preferences at different times contributes to lifetime wellbeing – then lifetime wellbeing would

remain non-perspectival (or at least stable across perspectives) as required for the argument to go through.

Many of our preferences are "misinformed" or "irrational". I may, for example, desire to hike tomorrow under the false impression that the weather will cooperate. It will in fact rain cats and dogs – a terrible condition for hiking. Or I may have read the (accurate) weather forecast but indulged in wishful thinking. Either way, it doesn't seem that I'd be better off going hiking tomorrow. Such considerations lead some subjectivists to *idealize*.[1] My actual preference for hiking is then considered welfare-irrelevant because it wouldn't survive certain idealizing processes – e.g., exposure to accurate information, clear-headed reasoning, etc. Idealist subjectivism is a variety, and I can only afford a brief overview here. Firstly, idealization may be either a *restriction* on *actual* preferences or a function that *generates counterfactual* preferences that are welfare-relevant. Within the latter camp, opinions differ regarding what sort of counterfactual preferences matter. Perhaps it's what I would prefer if I were suitably idealized, or perhaps it's what my idealized self would prefer my actual self to prefer. (This difference might be characterized as first-order *versus* second-order.) Fortunately, these disagreements do not matter for our purposes. I shall use "idealized preferences" to refer to either actual preferences that survive idealization or (first-order or second-order) preferences that idealized agents would have that are welfare-relevant.

There is also disagreement over what exactly idealization consists of. Existing proposals tend to be (understandably) vague and open-ended. For instance, Peter Railton takes idealization to involve contemplating one's "present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality" (1986: 16). Robert Brandt calls the process of idealization "cognitive psychotherapy", which involves "confronting desires with relevant information, by repeatedly representing it, in an ideally vivid way, and at an appropriate time" in an optimal state of mind (1979: 113).[2] Still others have specified idealization more or less differently.

Partly due to the open-endedness and diversity of idealizing accounts, it's not immediately clear if time-biases can survive idealization. But I am inclined to think that they can. For one, as many

[1] Brandt (1979), Railton (1986), Sobel (1994, 2011), Rosati (1995), Dorsey (2021: §6); cf. Murphy (1999), Heathwood (2019).
[2] More precisely, "without influence by prestige of someone, use of evaluative language, extrinsic reward or punishment, or use of artificially induced feeling-states like relaxation" (ibid.).

authors have emphasized, idealization is a *content-neutral* process. John Bronsteen remarks: "Anything that can be a preference… can be a fully informed preference." (2017: 57) The case studies in the literature confirm this. Consider a person who desires to spend her life counting the blades of grass more than anything else. This is an apparent counterexample to non-idealist subjectivism. The desire is apparently "irrational". The idealists would say that the grass counter is probably epistemically impoverished. Had she experienced other more rewarding activities (say, philosophy), she'd come to appreciate those activities instead. But this need not be the case. Suppose instead that however many alternative lifestyles she's been exposed to, the grass counter just finds grass counting most worthwhile and fulfilling. If so, then grass counting is really best for her, says the subjectivist. Consider then Parfit's future-bias. Some time-biases may be due to misinformation, cognitive defects, or lack of experience and therefore can be "cured" by idealization. But Parfit in *My Future and Past Operations*, for one, resembles the second version of the grass counter. It doesn't seem that Parfit is epistemically or experientially impoverished in any sense.[1] Nor does he exhibit any obvious rational defect. It's unclear what kind of further cognitive therapy could "cure" his future-bias. A related point is that idealization is a causal and naturalist process without "import[ing] any substantive value judgments" (Brandt 1979: 13). Otherwise, the subjectivists would be putting the cart before the horse and converting to objectivism. So, when idealizing Parfit, we cannot smuggle in "substantive value judgments" such as that it's *incorrect* to be future-biased.

But you might wonder, doesn't idealization also smooth out *structural* irrationality? And isn't Parfit's "preference reversal" structurally irrational? The first question is tricky. It's unclear to me what the idealists think of structural rationality. Railton (1986: 16) indicates that *instrumental* irrationality shall be smoothed out. But – for one, this seems in tension with his commitment that

[1] Is Parfit missing some information? It's true that Parfit loses and gains perspectival information as time passes. But even the *idealized* Parfit cannot be expected to possess these two pieces of perspectival information at the same time. Idealized agents do not hold inconsistent beliefs like that.

What if Parfit bases his future-bias on some mistaken metaphysics of time? It might then be hoped that correcting this false belief would make him time-neutral. But, for one, time-biases need not be resting upon any metaphysical beliefs. Heathwood (2008: 61) confesses that he is brutally future-biased. For another, at least according to Brandt (1979: 13), idealization doesn't involve God-like omniscience. Rather, "relevant information" doesn't go beyond the experts' consensus in the present days. So, it's not the case that the idealized Parfit should know about the correct metaphysics of time (which is an ongoing debate). Finally, even if Parfit is granted with *all descriptive* information in the world, I doubt that he'd then be time-neutral. For that to be true, idealization would have to guarantee certain inferential leaps from descriptive knowledge to evaluative attitudes. Such idealized agents would seem God-like in their cognitive capacities and thereby no longer idealized versions of any of *us*. And this would be intension with the subjectivist commitment that prudential values resonate with us and vary inter-subjectively.

idealizing processes can be described in purely causal terms. So, perhaps the idea is that idealization tends to, but doesn't guarantee to, cure instrumental irrationality. For another, even if instrumental irrationality is necessarily cured, I doubt that this can be generalized to all (purported) structural irrationalities. At least, I don't see idealists making such commitments. Despite the ambiguities here, I am inclined to think that time-biases can survive idealization. I've argued in Ch.2 that time-biased preference reversal need not be structurally irrational. But I don't hang my argument on that. Thinking more intuitively, I take it that the sort of idealization relevant here should involve confronting one's different perspectives over time – since those who take time-biases to be irrational often point to some sort of diachronic disunity. Admittedly, after confronting one's future and past perspectives, some time-biased subjects may come to see a tension between these perspectives and settle on a time-neutral perspective once and for all. But this is far from guaranteed. Some time-biased subjects would be perfectly satisfied with their pattern of preferences. Looking back upon his earlier preference from his present future-biased perspective, Parfit may clear-headedly think: "I see that it was *then* good for me to take the shorter operation. But things have changed. The longer operation is *now past*." More generally, he may appreciate that the different temporal perspectives warrant different preferences. He sees no more tension to resolve than (say) a clear-headed egoist who thinks that *everyone* should care about only *herself*.[1] Finally, it doesn't necessarily make time-biases welfare-irrelevant even if idealization necessarily removes *instrumental* irrationality. Parfit just cares more about future pain than past pain, and not for any further end.

I'd like to conclude that time-biases can survive idealization, and therefore lifetime wellbeing can be perspectival and shifty even given idealist accounts. For an agent whose time-biases don't survive idealization, I am willing to admit that her lifetime wellbeing is stable across perspectives. If she wouldn't be time-biased but for misinformation, faulty reasoning, etc., then it sounds correct that she wouldn't be better off should her time-biases be satisfied. But for an idealized time-biased agent, her lifetime wellbeing remains shifty. And it remains true that she might well not violate *No Tragedian* perspectivalized.

That said, I acknowledge that I haven't made a conclusive case that time-biases can survive idealization. Perhaps some specification of idealizing processes, though content-neutral in the

[1] This echoes the observation that idealization does not guarantee constant first-order preferences over time (as in cases of now-for-past sacrifice). (Brandt 1979: 245-60, Bykvist 2003: 118, Heathwood 2011a: 97-98; cf. Sobel 2011).

description of the *function*, would guarantee *outputs* with certain contents. And time-neutrality may be one of these outputs. (This is similar to Smith's suggestion that certain substantive norms can be derived from procedural norms.) Such an account, it seems to me, is unmotivated as a subjectivist account. And it's doubtful if such a function is possible without smuggling in substantive value judgments. In all, given the lack of any positive argument that time-biases cannot survive idealization, there is good reason to think the opposite. I hasten to add that not all subjectivists are idealists. These actualists (Murphy 1999, Heathwood 2019) have their own way of handling unwanted preferences – mainly by restricting welfare-relevant preferences to intrinsic and non-whimsical ones. Evidently, time-biases can satisfy these actualist restrictions.

So much for idealization. But perhaps time-biases are welfare-irrelevant for other reasons. For a start, it may be suggested that *higher-order* preferences do not count. That is, preferences that have other preferences as constituents of their content are welfare-irrelevant. This suggestion is against the conventional wisdom that higher-order preferences are of privileged normative authority (Frankfurt 1971). Even if higher-order preferences are not somehow more important, it's implausible that they are utterly irrelevant (simply for being higher-order).

Or perhaps time-biases are irrelevant because they cannot be *concurrently satisfied*.[1] By being future-biased, Parfit prefers that his future aversion to pain rather than his past aversion is satisfied. When would that future-biased preference be satisfied? It's in the future, apparently, when it's a settled matter that no future operation occurs. But this does not preclude concurrent satisfaction. Presumably, when it's Friday, Parfit shall continue to be glad that his ordeal is over and done with (rather than ongoing). Of course, he could instead just stop caring about the operation immediately after being informed by the nurse (on Thursday) that his ordeal is over. In this case, there seems to be no concurrent satisfaction (except for a present-directed preference for certain information from the nurse). Nonetheless, *concordism* (which I consider superior) would make his future-bias relevant. Even though the preference doesn't persist till Friday, it's not contested by any Friday preference either. More generally, insofar as one's time-biased tendencies are relatively stable, we should expect lots of them to be welfare-relevant.

[1] Concurrentism is the worst-case scenario. If time-biases can be concurrently satisfied, then they will satisfy the weaker restrictions offered by the harmony view, concordism, etc.

The final suggestion to consider is that only *temporally local* preferences are welfare-relevant. A preference is temporally local, we might say, if it concerns the happenings in one's life contained at a time or some duration of time. A preference is *temporally global*, by contrast, if it concerns how different temporal segments of life relate to each other.[1] But it seems arbitrary to disregard global preferences just because they're global. Moreover, this would make *shape-of-life* preferences irrelevant as well. As discussed in §4.2, many think that structural features of life are relevant to lifetime wellbeing, though opinions differ regarding how exactly they make contribution. Insofar as subjective theories are concerned, the most obvious strategy is to let the subjects themselves decide the welfare-relevance of structural features (Bruckner 2018, Bykvist 2006: 119). If I prefer my life with an uphill rather than a downhill trajectory, then my life would be overall better if it goes uphill (all else being equal). But if I lack shape-of-life preferences, then my lifetime wellbeing is insensitive to these features. I do not profess that this must be the correct account (cf. Dorsey 2015), but it's at least a serious contender. So, the subjectivists should not preclude global preferences as welfare-irrelevant. Incidentally, this account of structural features would seem to make lifetime wellbeing perspectival and shifty as well, for shape-of-life preferences can change over time (although their contents are non-perspectival). You might dislike this account for this very reason. But then, if this account is well-motivated on other grounds (most importantly it's heart-and-soul subjectivist), then shifty wellbeing may be cast in a more favourable light.

It's worth emphasizing that I am assuming a *de facto* understanding of how time-biases relate to wellbeing. In particular, given preferentialism, time-biases are higher-order preferences over the satisfaction or frustration of preferences at different times *regardless of* whether or not the subjects prefer them under such *guises*. Parfit could be a hedonist but nonetheless prefers *de facto* greater lifetime wellbeing from his future-biased perspective by virtue of being future-biased. But you might think that the *de dicto* understanding is more apt, given which Parfit's future-bias wouldn't be a higher-order preference that discounts the welfare-relevance of his past preferences. Rather, it's just a garden-variety (first-order) preference that concerns past and future pain. And it's hard to see how such a preference could render lifetime wellbeing shifty.[2]

---

[1] The distinction drawn here is different from Parfit's (1984: 496-99). By "global desires" Parfit means desires directed at *extended* periods of life or life as a whole. He would count as global a desire to (say) be successful throughout college. Also note that higher-order preferences don't have to be temporally global. I prefer that I don't prefer (as of now) lots of sugar in my coffee. This is higher-order but temporally local.

[2] The difference here may be illustrated by how paradigmatic higher-order desires may be welfare-relevant. Although the literature is (to my knowledge) virtually silent on this issue, I gather that there are (at least) two possible understandings. Suppose that I desire that I don't have the desire to eat meat (which I in fact have). On one understanding, my second-order desire alters the welfare-relevance of the first-order desire.

Two points in response. As I have suggested in §4.2, the *de facto* reading is more apt for *the Argument from Wellbeing*. Presumably, the time-neutralists wouldn't let Parfit off the hook simply because he doesn't conceive of his future-bias as one that's for an overall worse life. (He might lack the concept of lifetime wellbeing or hold a deviant conception of what makes his life overall better). So, it seems only fair to assume the *de facto* reading here as well. But secondly, adopting the *de dicto* reading doesn't solve the problem completely. It's certainly possible for one to prefer *de dicto* that her temporal wellbeing is distributed in a certain way. And if that's what one in fact prefers and such preferences change over time, there is no reason to deny that her lifetime wellbeing has become shifty (insofar as these preferences satisfy other restrictions such as idealization). By all accounts, shifty lifetime wellbeing seems unescapable for subjective theorists.

Where does this leave us? For one, shifty lifetime wellbeing might still be thought of as a problem for subjective theories. If shifty lifetime wellbeing is objectionable, then subjective theorists had better find a way to make time-biases welfare-irrelevant. But we have seen that time-biases can survive idealization. Nor is it advisable to exclude them on the grounds that they are non-concurrent, higher-order, or temporally global. Now, it seems that the only solution would be a targeted welfare-relevance restriction on time-biases.[1] Perhaps some subjective theorists would be willing to make such a move. But this must be acknowledged to be *ad hoc*.

But is shifty lifetime wellbeing so obnoxious after all? Might subjective theorists embrace it? Consider again the quote from Chris Heathwood: "But just as it makes no sense to ask, say, how long some life was relative to some time, it makes no sense to ask how good overall some life was relative to some time." (2011b: 29) He's certainly correct about life span. But the analogy is inapt.

Satisfying my first-order desire confers lesser (or even no) welfare-benefit compared with what would be the case without the second-order desire. Alternatively, my second-order desire is welfare-relevant in the same way as a garden-variety first-order desire. It leaves intact the welfare-relevance of the first-order desire but confers positive value to the state-of-affairs that I don't have the first-order desire. Given the *de facto* reading of time-biases and subjectivism, the welfare-relevance of time-biases has to parallel the first account of higher-order desires above. This is because *de facto*, time-biases just are preferences over the *welfare-relevance* of first-order preferences. But without the *de facto* assumption, time-biases might be treated like garden-variety first-order preferences. And change in time-biases wouldn't make wellbeing shifty any more than change in garden-variety first-order preferences.

[1] It wouldn't do to ban all perspectival preferences or all *de nunc* preferences. *De se* preferences are welfare-relevant if any preferences are. First-order *de nunc* preferences should also be allowed. In the above discussion, I have formulated the first-order preferences in eternalist terms for convenience's sake. But obviously lots of our first-order preferences have perspectival content: I may prefer that I get this thesis published right now (without specifying some calendar time).

Lifetime wellbeing is value-laden but life-span is not. And according to subjective theorists, lifetime wellbeing should be determined by the subjects' own attitudes. A more appropriate analogy would be (say) how good a life is relative to some *subject*, or how good a *segment* of life is relative to some *temporal perspective*. Such phrases are far from insensible. I may look at your life and find it meaningless. And I may, as Jeremy who stuck out of college, look back on my past and have an opinion different from what I had back then. Of course, such other-directed and cross-temporal attitudes may lack *prudential-value-conferring authorities* on pain of objectionable nose-poking. Nonetheless, there is a good sense in which (some segments of) a life may be perceived as good from one perspective but bad from another. Shifty lifetime wellbeing takes seriously this insight and makes the further claim that each perspective is value-conferring regarding lifetime wellbeing. If at each moment, all one cares about is her future wellbeing, then her *future* wellbeing is her *lifetime* wellbeing at each moment. For such a consistently future-biased person, it makes sense to grant each future-biased perspective with value-conferring authority. You may still find such a life troubling: As the best course of life for her changes constantly, it's unclear how she could have a stable deliberative perspective. But then, subjective theorists tend to be tolerant of lifestyles that may appear troubling (as they should). If one *really* wants to starve herself to death or count grass all day, then there seems no reason to paternalistically deny that such a life is good for her (cf. Street 2009).

If all this is plausible, then, it seems that *the Argument from Wellbeing* cannot be axiologically neutral. This is because we have well-motivated subjectivist theories which imply that lifetime wellbeing is perspectival and, for time-biased agents, shifty. Neither the non-perspectival nor the perspectival version of *No Tragedian* works for the argument.

To briefly summarize, the central theme of this chapter is to examine time-biases against the background of wellbeing – in particular subjective theories. I've first argued that now-for-past sacrifice does not pose a problem for time-neutralism. It can be avoided by embracing some non-perspectival restrictions on the welfare-relevance of preferences. I then turned to how subjective theorists might respond to *the Argument from Wellbeing* in defence of time-biases. The interesting result is that *the Argument from Wellbeing* fails even if we grant that subjective theories do not preclude all (attitude-independent) normative reasons. Although the subjective theorists might well acknowledge that there is some global norm of prudence, they have good reason to embrace *No Tragedian* in its perspectival rather than its non-perspectival form. But the perspectival version does not pose a problem to (all) time-biases.

Now, it might be wondered whether a similar strategy is available to the *objective* theorists. That is, can an objective theorist make lifetime wellbeing perspectival and shifty as well? Can a hedonist say that pain is objectively worse for you when it's future rather than past? I think they can. But as objective theorists, they cannot appeal to the welfare subjects' own attitudes. They must instead make a case that as an *attitude-independent* matter, prudential value is perspectival and shifty. While the previous chapter reveals this to be unpromising on the metaphysical-time conception, it remains a possibility on the personal-time conception. I explore the promises and difficulties in the next (and final) chapter.

5. From Moral to Prudential Partiality

5.1. Neutrality and Partiality

In *Reasons and Persons*, Parfit argues that Self-Interest Theory (S-Theory) is untenable because it's "incompletely relative". By S-Theory Parfit means a combination of *egoism* and *time-neutralism*. According to S-theory, each of us ought to pursue *one's own* best interest, *considering one's life as a whole*. S-Theory is neutralist with respect to time but (extremely) relativist with respect to personal identity. But a theory should assign equal significance to time and personal identity, says Parfit. So, the internal tension makes S-theory untenable:

> [There is an] analogy between oneself and the present, or what is referred to by the words "I" and "now". This analogy holds only at a formal level. Particular times do not resemble particular people. But the word "I" refers to a particular person *in the same way in which* the word "now" refers to a particular time. And when each of us is deciding what to do, he is asking, 'What should I do *now*?' Given the analogy between "I" and "now", a theory ought to give to both the same treatment. (1984: 140, emphasis in original)

I agree with Parfit that S-Theory is difficult to defend. I agree that incomplete relativity as such is unstable. But not for the reason stated above. (Not that Parfit considers this the final word.) Parfit is certainly correct that there is a good (but trivial) sense in which every decision is indexed at an individual at a particular time. That's just how decisions are made, given our natural constitution. But that tells nothing about what kinds of considerations should weigh in the time- and individual-indexed decisions. The genuinely normatively significant parallel between times and individuals, I believe, lies in the *bonds-of-concern*[1] that obtain in various degrees both inter-personally and intra-personally. To the extent that morality is partial (or relative) due to the presence of these bonds-of-concern – as many of us agree – there is good reason to reject time-neutralism. That is, because these very bonds-of-concern that make morality partial also obtain intra-personally, a moral partialist should also endorse prudential partialism. This is not to endorse egoism or its

[1] I borrow this term from Dorsey (2018, 2021).

prudential counterpart (presumably a form of absolute present-bias). Rather, I'd like to begin with a form of moral relativity that's roughly in line with commonsense morality – which is moderate and far from egoist – and argue that it generalizes to the intra-personal. More specifically, I take it to be part of commonsense morality that we are permitted to be (to some extent) partial towards our "nearest and dearest" (family members, friends, compatriots, etc.). This modest form of partiality is widely accepted (though not entirely uncontroversial), and I will not argue for that. But from there, I argue that a parallel form of prudential partiality is also true.

The idea that time-biases can be justified on a similar ground as moral partiality is not new. It has been echoed, in various forms, by many philosophers[1] – especially those who share the Parfitian sentiment that the unity of persons is not as robust as it may seem (more on which in §5.3). I shall try to be ecumenical in motivating this analogy (§5.2 - §5.5).

However, there are three lacunas in the existing literature I wish to draw attention to.

The first concerns prudential latitude. Some defenders of time-biases take time-biases to be *merely permissible* – i.e., permissible but not obligatory. They claim that it's fine to be time-biased but also fine to be time-neutral, as it were. But not much has been said about how we come to enjoy such latitude. Indeed, in certain normative frameworks – e.g., a prudential maximizing consequentialism – prudential latitude would seem impossible. I sketch an account of latitude in §5.6.

Secondly and perhaps surprisingly, it's not immediately clear how to accommodate *future-bias* (§5.7). While it's easy to see that bonds-of-concern tend to diminish over time (so that near-bias is vindicated), it's hard to see how bonds-of-concern could be past-/future-asymmetric, and few have commented on that. This is upsetting. We are much more inclined to consider as permissible future-bias than near-bias. Anyone who considers near-bias permissible, I surmise, would consider future-bias permissible (or even obligatory). So, it's a pressing question whether and how this justification can be extended to future-bias.

[1] Parfit (1984), Whiting (1986), Shoemaker (1999, 2002), Miller (2014), Sebo (2015), Braddon-Mitchell and Miller (2020), Dietz (2020), Dorsey (2018, 2021), Goodrich (2021), Karhu (2023). For criticisms, see for example: Korsgaard (1989), Brink (1997, 2011), Black (2001), Sullivan (2018: Ch.4), Ahmed (2018).

Finally, we should wonder about *third-person* time-biases (§5.7). To illustrate, suppose that intra-personal bonds-of-concern are past-/future-asymmetric so that we ought to be prudentially future-biased. Does the same hold true for our attitudes towards other individuals? How should we think about the bonds-of-concern that obtain from my present person-stages to your past and future person-stages – or to the past and future person-stages of my loved ones?

These are all important and difficult questions. My aim in this chapter is modest: to articulate the questions and motivate some possible approaches, rather than to have the final word.

5.2. Reasons, Agent-Relative and Stage-Relative

The existence of *agent-relative reasons* is necessary for justified (either *obligatory* or *merely permissible*) *relationship partiality* – i.e., the sort of special and prioritized concern for one's friends, spouses, parents, etc. Let's say that a moral theory is *partial* if it admits justified relationship partiality. Partialists are thus committed to a special sort of agent-relative reasons – i.e., agent-relative reasons concerning one's nearest and dearest. Impartialists deny justified relationship partiality. *Contra* commonsense morality, they say, you really shouldn't be biased towards your friends, spouses, or parents. They thus reject the sort of agent-relative reasons upheld by partialists.[1] (As I am concerned only about relationship partiality, an impartialist as defined here can admit other sorts of agent-relative reasons, e.g., those giving rise to agent-centered constraints against killing, etc. Also, I'll frequently use "agent-relative reasons" to refer to the special kind that gives rise to relationship partiality, since that's the only kind relevant here.)

Very roughly, a reason is agent-relative if it cannot be stated without "includ[ing] an essential reference to the person who has it"; otherwise, it's *agent-neutral* (Nagel 1986: 152-53). How exactly to draw this distinction is a vexed matter (Ridge 2023), but we can have an intuitive grasp by considering an example.[2] Suppose that *someone* is in pain and that a dose of analgesic shall relieve

[1] I count as *impartialists* those who say that relationship partiality is usually justified because we are usually in a better position to take care of our intimates (because we know them better, we are spatially closer to them, etc.). This form of "partiality" is merely apparent because it's ultimately grounded in features that are only contingently associated with special relationships, in a way that parallels merely apparent time-biases in consideration of epistemic uncertainty, external utilities, etc.

[2] The agent-neutral/-relative distinction is often considered as fundamental to demarcating the landscape of moral theories. Importantly, it's supposed to cut across the consequentialism/deontology divide. A

it. This gives me a reason to offer a dose to that someone – or, if I am not able to do so, at least a reason to wish that a dose is offered. Indeed, it gives *everyone* such a reason. It's an agent-neutral reason. But suppose: *That someone* happens to be my spouse. The partialists would then say: The fact that *my spouse* is in pain gives *me* a reason to offer a dose of analgesic to *him* – an agent-relative reason *over and above* the agent-neutral reason I already have. This agent-relative reason doesn't apply to you. But were you and your spouse in a similar situation, you'd have a corresponding agent-relative reason to attend to *him or her*.

I have reservations about conceiving agent-relative reasons as something *extra to* agent-neutral reasons, which I turn to shortly. But let's grant this standard conception for a moment. The purported agent-relative reasons explain a variety of reasonable partiality. To continue with the example, suppose that my spouse is in pain and a stranger is in a bit more pain, and I am in a position to offer only one dose of analgesic so as to make either of them pain-free. If there are only agent-neutral reasons, then I'd be required to unburden the stranger rather than my spouse.[1] But if I have an extra agent-relative reason, the overall balance of reasons might then be tipped. It might then be permissible, if not required, for me to unburden my spouse.

While more needs to be said to make sense of *mere permissibility* (being permissible but not required) in relation to the "overall balance of reasons" (§5.6), the point here is simply that the existence of such agent-relative reasons makes room for morality partial. In what follows, I take for granted that morality is partial and that there are such agent-relative reasons. I proceed by arguing that the *grounds* for agent-relativity in special relationships – call these grounds *bonds-of-concern*, whatever they are – are also present in the intra-personal case. These grounds then give rise to (person-)*stage-relative* reasons. So, prudence is partial.

So, the argument is a conditional one. It says that if one is a moral partialist, then she should be a prudential partialist. It's thus rather modest. There may be more ambitious arguments for prudential partialism that do not rely on a moral antecedent. One might even claim that morality is impartial but prudence is partial (which is not obviously incoherent). But I am a moral partialist.

---

consequentialist may acknowledge agent-relative reasons – which are presumably grounded in agent-relative values given consequentialism. More on how reasons relate to values in §5.6 and §5.7.

[1] Let's suppose that the one dose cannot be shared, that it makes one pain-free regardless of the intensity of pain, and that the strength of the reason is proportionate to the intensity of pain.

And I consider this "companions in innocence" strategy, despite having a circumscribed audience, the most promising one.

By grounds for agent-relative reasons (or grounds for partiality) I mean facts that are fundamentally explanatory of the presence of agent-relative reasons. You ask me about how I'm justified to be partial to my friend. In an everyday context, the mere claim that "she's my friend" might be a satisfactory answer. But here we need a deeper answer. Presumably, the grounds have much to do with what our friendship consists in – our shared history, shared interest, mutual affection, etc. I call these grounds for agent-relative reasons, whatever they turn out to be, *bonds-of-concern*.[1] They are descriptive facts with normative upshots.

This brings us to the second aspect of modesty in my argument. There is no consensus about the scope of partiality. That is, there is no consensus about what relationships call for partiality and what sorts of agent-relative reasons there are. Friendship, familial relationships, and romantic love are considered paradigmatically partial. But it's debatable whether partiality binds these relationships universally or only those non-defective and non-degenerate instances. What if my father has been negligent? What if my romantic partner is abusive? Also, it's very unclear how far partiality extends from this inner circle. Do I owe more to my compatriots than non-compatriots? Should a member of the Ku Klux Klan be partial to their league?

Such disagreement, in part, is due to disagreement over the nature and extent of bonds-of-concern. There are lots of features about my best friend that are somewhat special. For instance, we went to the same college and rank *Six Feet Under* the best TV drama of all time. But obviously, not any old person who happens to stand in such relations to me deserves my partiality. The question of bonds-of-concern is about what's distinctive and fundamental about partial relationships that are reason-giving. And even if we know how friendship typically looks or even what friendship *is*, we may still disagree over the bonds-of-concern. Suppose that shared history, shared interest, and mutual affection are individually necessary and jointly sufficient for a relationship to count as friendship (which is a contested definition[2]). Given this, there is still the further question of which of these features are necessary or sufficient for agent-relative reasons. Perhaps not all of them are necessary.

[1] It's apt to think of bonds-of-concern as *partial* grounds. Consider my agent-relative reason to offer my spouse the analgesic. That she is *in pain* is obviously a partial ground (and presumably a complete ground for the corresponding agent-neutral reason). But here I am interested in the distinctive (partial) ground that gives rise to *agent-relativity*.
[2] See Helm (2023) for an overview.

Perhaps they are not jointly sufficient – perhaps there is some extra feature that distinguishes friendships that are ripe for partiality from those that are not. Moreover, partiality in different kinds of relationships may be grounded in different kinds of bonds-of-concern. If compatriotism is partial, then presumably the grounds there would be very different from those in friendship. It's also plausible that different kinds of bonds-of-concern generate reasons of different flavors. It's plausible that I have a special duty to see to it that my friends are not morally corrupted, but implausible that I have such a duty towards all my compatriots.

Given all these indeterminacies, I do not aim to offer a definitive answer regarding the inter-personal bonds-of-concern that generate agent-relative reasons. The "companions-in-innocence" argument for stage-relativity shall inherit all these indeterminacies. But this need not weaken the argument. My speculation is that all the reason-giving inter-personal bonds-of-concern also obtain, to differing extents, among person-stages. If so, then the case for stage-relative reasons is *overdetermined*. Of course, here I cannot enumerate all the candidate inter-personal bonds-of-concern and prove that they are genuine. I can only point to the most *salient candidates* and show that they also obtain intra-personally. If that much is true, then there is at least *very good reason* to think that prudence is partial – for the very same reason that morality is partial (though short of a *proof*).

So, I argue that there are stage-relative reasons, grounded in the same sort of bonds-of-concern that ground agent-relative reasons. A stage-relative reason is such that a full statement of the reason essentially involves a reference to the stage(s) who has it, and thereby is not necessarily shared by all stages of the person. Talking about (person-)stages here allows us to conveniently pick out persons at times; I am not thereby committing to the doctrine of temporal parts (Sider 2001). An endurantist can interpret stages as identifying the temporal locations of enduring persons. Given endurantism, it's perhaps more apt to speak of "dated" or "time-relative" reasons[1]: Reasons that apply to a person at some times but not necessarily at all times. A friend of temporal parts, on the other hand, may take it as literally true that person-stages are possessors of reasons.

Think of a stage-relative reason, R_{SIX}, which is a reason *for* person-stage P_A that *concerns* another person-stage P_S: That watching *Six Feet Under* will make happy P_A's "nearest and dearest" fellow

[1] Nagel (1970: 58-80), Parfit (1984: 142-44).

stage, namely P_S, is a reason for P_A to prefer and promote[1] that P_S watches *Six Feet Under*. Insofar as this is a prudential reason, P_A and P_S must be stages of the same person. Again, endurantists, perdurantists, and stage-theorists will specify the same-person-relation differently. I also remain neutral on the persistence condition of persons. That may be psychological continuity or something else.

To make life easier, let's define *prudential closeness* (*PC*) as a measurement of the extent to which bonds-of-concern obtain between two stages.[2] Let $PC(P_A, P_S)>PC(P_B, P_S)$ mean P_S is prudentially closer relative to P_A than relative to P_B – in other words, greater bonds-of-concern obtain from P_A to P_S than from P_B to P_S. In turn, *PC* determines the presence or absence of stage-relative reasons and also their strengths. All else being equal, the greater *PC* is, the stronger the stage-relative reason. (For the moment I leave open if prudential closeness is a symmetric relation. That is, I leave open if for all *P*s and *P**s, $PC(P, P^*)=PC(P^*, P)$; cf. §5.6)

Now, I wish to outline a relatively non-standard framework for drawing the stage-neutral/-relative distinction that will be helpful for my purpose. I take it that among any two stages of a single person (living an ordinary life), there tend to be *some* bonds-of-concern.[3] (It will become obvious in §5.3 if not obvious now.) So, stage-relative reasons are very plentiful, with *varying strengths*. Suppose that $PC(P_A, P_S)>PC(P_B, P_S)$; and that $PC(P_B, P_S)$ reflects that *some* bonds-of-concern obtain from P_B to P_S. Then in some sense, P_A and P_B share the very same stage-relative reason R_{SIX} concerning P_S. But speaking more fine-grainedly, the reasons P_A and P_B have respectively are distinct, for they are of different strengths.

How should reasons be individuated, then? In my preferred framework, we should think of P_A and P_B as sharing the very same *stage-neutral* reason but also having distinct *stage-relative* reasons

[1] Two points about promotion. (1) A reason to promote a certain state-of-affairs is a teleological reason. Nagel (1970) thinks that all reasons are teleological. But you may think that there are non-teleological reasons such as a reason to respect one's dignity. It may be that prudential partiality also has non-teleological elements. This is an open possibility that I won't pursue further. (2) Promotion is action. So, to say that *P* has a reason to promote *S* entails that *P* is in a position to or is capable of executing (some of) the relevant actions (not necessarily performing the very action that brings about *S*, according to Nagel's capacious understanding of promotion). Thus, if *S* is past and out of the causal reach of *P*, then *P* may have a reason to prefer *S* but not a reason to promote *S*. More generally, the condition of capability is considered as an enabler of reasons (Dancy 2004) rather than a ground for reasons.

[2] Presumably different sorts of bonds-of-concern may weigh differently in determining *PC*. I leave open the exact function.

[3] Depending on the persistence condition of persons, it may be a conceptual truth that within a single person, some bonds-of-concern obtain between any two stages.

derived from that very same stage-neutral reason. I take stage-relative reasons to be *modified* stage-neutral reasons. And I take bonds-of-concern (in other words, *PC*) to be modifiers: They operate on stage-neutral reasons by modifying their strengths. To see what I mean and why this framework is ideal, recall that when illustrating the agent-neutral/-relative distinction with the analgesic case, I said that I have an *extra* agent-relative reason to unburden my spouse in addition to an agent-neutral reason to unburden her (which is shared by everyone). But I find this way of carving out reasons clumsy and unparsimonious. It gives us *too many reasons*. And it makes weighing reasons too complicated. Better to understand agent-relative reasons as *modified* agent-neutral reasons. (Lord 2016; Bader 2016: 42-44; Löschke 2017) Agent-relative reasons are thus considered not as something extra, but something *derivative upon* agent-neutral reasons. These two kinds of reasons end up playing distinctive theoretical roles and are not to be weighed against each other.

To illustrate this alternative framework by focusing on the stage-neutral/-relative distinction, consider again R_{SIX}. Presumably, just as in the inter-personal case, insofar as watching *Six Feet Under* would make P_S happy, then *every* fellow stage has *some* prudential reason to prefer and promote that state-of-affairs – even if *no* bond-of-concern obtains.[1] In this sense, all stages share the very same stage-neutral reason. Stage-neutral reasons can be thought of as "generic reasons" that we can sensibly specify under a veil of ignorance (of bonds-of-concern). The *basis* for R_{SIX} is the fact that watching *Six Feet Under* would make P_S happy. Bonds-of-concern are *not bases* of any sort. The mere fact some stage is intimate to another does not generate reasons for anything. Instead, bonds-of-concern are *modifiers*.[2] Modifiers operate on "preexisting" reasons by intensifying or diminishing their strengths. In particular, bonds-of-concern are *intensifiers*; they operate on "preexisting" stage-neutral reasons by intensifying their strengths. Think of the stage-neutral reason R_{SIX} as carrying some *default strength* (proportionate to the amount of happiness involved). A modifying function then takes as input R_{SIX} (with its default strength) and generates stage-relative "specifications" of that stage-neutral reason. Those stage-relative reasons are derivative upon the stage-neutral reason: They inherit their bases. Since $PC(P_A, P_S) > PC(P_B, P_S)$, P_A

[1] This may depend on one's theory of personal identity. One some accounts of personal identity – e.g., some versions of psychological reductionism – it may be that personal identity cannot outlive bonds-of-concern. (cf. §5.3) But if you think personal identity consists in (say) the persistence of the same soul, then presumably you can have personal identity without any bonds-of-concern. On this view, it's plausible there can be prudential reasons without bonds-of-concern, just as there can be moral reasons to take care of perfect strangers.

[2] Dancy (1993) introduces the basis/modifier distinction. He calls basis "ground". But since I've used "ground" to refer to the totality of facts that fix the presence *and weight* of reasons, I introduce the term "basis".

and P_B have *distinct stage-relative* reasons concerning P_S. Their respective stage-relative reasons possess different strengths by virtue of having been modified (intensified, more specifically) by different *PC*s (P_A's stage-relative reason being stronger).

This is a more parsimonious and sensible picture. Instead of having a profusion of stage-relative reasons *alongside* stage-neutral reasons, stage-relative reasons turn out to be derivative upon stage-neutral reasons. They serve different roles. Stage-relative reasons are to be *directedly weighed against each other* in determining all-things-considered normative statuses but *not against stage-neutral reasons*.[1] Stage-neutral reasons are just not the sort of things that can directly feed into one's deliberation. Note that even if no bond-of-concern obtains between some fellow stages, there are still stage-relative reasons. A modifier that has a neutral impact on the default weight is still a modifier. (I do not have a specific account about how to assign numbers to *PCs*. Though it's non-negotiable that the extent of bonds-of-concern turns out proportionate to the strength of stage-relative reasons.)

So, if stage-relative reasons derived from the very same stage-neutral reason can *vary in strengths*, then prudence is partial. I now argue for the antecedent – or better, I argue that they actually do vary in strength. I argue that the salient candidate inter-personal bonds-of-concern – which determine *moral closeness* and therefore generate agent-relative reasons with varying strengths – are also found among person-stages to various extents.

5.3. My Nearest and Dearest (1)

Friendship is a paradigmatically partial relationship. All else being equal, my agent-relative reason to benefit my best friend tends to be stronger than the corresponding agent-relative reason I have to benefit a stranger. Therefore, one is often permitted or even required to prioritize the interests of her friends even when this leads to lesser overall good.

[1] This also coheres nicely with the model of utility discounting we have touched upon in the previous chapters. Stage-neutral reasons may be seen as responding to unadulterated utilities (utilities prior to factoring into perspectival temporal information), whereas stage-relative reasons map onto discounted utilities.

Why is friendship partial? Presumably, this has much to do with the kind of *relationship* friendship is. At least, it cannot be all about the intrinsic features of the *individuals*.[1] My best friend may not be particularly virtuous, but this does not invalidate my being partial to her. So, what's in a friendship that grounds reasonable partiality? One salient and common feature of friendship is *like-mindedness*. As David Shoemaker remarks, friends tend to "share highly similar goals, values, beliefs, and experience-memories" (2002: 73). Mere like-mindedness, though, doesn't seem very adequate. Consider a random stranger whose values and beliefs, etc., by sheer fluke, extensively overlap with yours. We may even imagine memory-sharing due to you two always crossing paths at the same events by happenstance. We don't want to say that you should thereby be partial to this stranger. What's inadequate about like-mindedness as such seems to be its overemphasis on *qualitative similarity*. Similarity may be relevant, but only if it comes about in the right way – namely, through a shared history of causal interactions. So, like-mindedness seems more of a symptom of what really matters (though like-mindedness may also matter in itself, to some extent).

As Jennifer Whiting points out, it's typical of friendships (and perhaps intimate relationships in general) that "our interaction with one another causally affects the formation of these desires, interests and values" (1986: 558). This is the point that the genesis of similarity matters. Moreover, it seems that not any old causal interactions will do. What's characteristic of psychological influences between friends is that they are rather immediate and spontaneous. We are *psychologically vulnerable* to our friends, so to speak. To take a mundane example, when my best friend said that I would enjoy *Six Feet Under*, it was automatically on the top of my watch list. I wouldn't be that susceptible to the recommendation from any old person. Or consider planning what to do with your friend over the weekend. Knowing that she has recently got into hiking, you'd be proposing hiking routes even though you don't particularly fancy hiking (yet). From your deliberative perspective, it appears that her desires are already yours.[2] It doesn't feel like self-sacrifice despite conflicting desires. So, it's not only the case that there is mutual influence of

[1] This is even true according to what's called the "individual account" of partiality. Simon Keller defends the individual account:

> While reasons of partiality are reasons to give special treatment within special relationships, the ground of such reasons is in the value of individuals as they stand in their own rights, not in the ethical significance of relationships themselves. (2013: 111)

This may appear to be saying that agent-relative reasons are grounded in the intrinsic features of the individuals. But this is certainly not the complete picture. Otherwise, everyone would end up having the same reasons. Keller goes on to suggest that relationships can be thought of modifiers or enablers (rather than bases) of special reasons.

[2] Here I am merely making an observation about human psychology. But it might be argued that your friend's interest is literally your interest. I turn to this metaphysical point in §5.5.

psychology, but more importantly, that such influence is automatic and unforced. One might put this point slightly differently in terms of *trust*. When debating with myself whether to break up with my partner, I'm willing to consult my best friend and then just follow her suggestion. I trust in her good will even though she's no expert on this matter. (I wouldn't defer to a random stranger on this matter.) All this is *not* to say that one must go along with her friend. I can and sometimes should, after deliberation, repudiate my friend's ideas. As I see it, psychological vulnerability concerns more of the *starting point* of my deliberation than its *end product*.

All the above-mentioned elements form a self-perpetuating cycle. Psychological vulnerability (or trust), insofar as it hasn't been betrayed, leads to more shared history and like-mindedness, which in turn reinforces vulnerability. The totality of these features constitutes what we may call *psychological closeness* or *psychological intimacy*. In addition to friendship, psychological closeness is also typical of romantic and familial relationships, or at least of those non-degenerate instances that are undoubtedly suitable for partiality.

Psychological closeness among person-stages varies. Some fellow stages of my present stage P_0 are psychological intimates relative to P_0, while others are less so. Say that a future (in personal-time) stage of P_0, namely P_1, relates to P_0 in a way that manifests all the hallmarks of psychological intimacy. P_1 and P_0 have much shared history and shared memory. Partly thanks to memory, immediate and spontaneous casual influences of beliefs, desires, values, and characters are rampant. P_1 just inherits whatever psychological features P_0 possesses unless there is a call for change or reconsideration. Notice that psychological intimacy need not coincide with *similarity*. This is true in friendship and also true in the intra-personal case. What matters more is not actual convergence but the susceptibility and willingness to be influenced and impinged upon. Importantly, spontaneous causal influence may manifest as *change and dissimilarity*. Suppose that P_0, being timid and introverted, intends a change. To the extent that P_1 carries out this intention and thereby becomes confident and extroverted, I'd say that psychological closeness is *ceteris paribus enhanced* rather than diminished.

On the flip side, it's equally clear that person-stages can be psychologically rather distant from each other. This may be due to memory naturally fading (which is pivotal to much of causal influence of other sorts), change of environments and circumstances (moving to a new city and meeting new people, etc.), disruptive and life-changing events (losing one's spouse, entering a new career, etc.), and so on. Consider again *the Russian Nobleman*. This is an extreme case in which the old

bourgeoisie is so psychologically distant to the young socialist that the latter considers the former as more of an *enemy* than a friend, as it were.[1] The antagonism in fundamental values is so drastic that the young socialist conscientiously forfeits his vulnerability and trust towards his future stages. This is much like how and why people sometimes come to put an end to a friendship or foreclose entrance to a friendship in the first place. In all, we find psychological intimacy obtaining to lesser and greater extents among stages of a single individual. If psychological intimacy is what grounds moral partiality (as many think), it would seem that prudence is also partial.

All this talk about intra-personal psychological closeness should sound familiar. It's a popular account of personal identity, known as psychological reductionism, that the determinants of psychological closeness – i.e., memory, causal influences of desires and beliefs, and the vulnerability thereof, etc. – are also what make a person persist over time. Indeed, some proponents of psychological reductionism have directly argued from psychological reductionism to prudential partiality (Parfit 1984, Shoemaker 1999, 2002). Roughly, the idea is that if these psychological *connections* (i.e., components of psychological intimacy) are what weave different person-stages into a single person, then there is an intuitive sense in which fellow stages can be "the same person" to different degrees. For example, it makes sense for the young socialist to say that there is "little of him left" in the old bourgeoisie (if they are the same person at all). And it's rather intuitive that prudential concern should be proportionate to the extent to which person-stages are "the same person".

This does not require the metaphysical relation of *personal identity* to be gradable. One may, as Parfit himself does (1984: 302-06), maintain that persons endure and that numerical identity is all-or-nothing, but also admit that there is a good sense in which the old bourgeoisie is only the same person as the young socialist to a very diminished extent. In his words, they are different "selves". They belong to distinct lumps of stages that are internally interwoven by "strong psychological connectedness" but not so related with each other externally. Alternatively, one might think that personal identity – or better, the relation of *being-the-same-person-as* – is literally gradable. Though perhaps in tension with endurantism, such a gradable notion can be easily accommodated by perdurantists and stage theorists. (Lewis 1983, Williams 2014, Sider 2018, Braddon-Mitchell & Miller 2020) For instance, since counterparthood comes in degrees (as

[1] There is an interesting moral question whether one has reason to be *hostile* to one's *adversaries*. (Brandt 2020) This sort of negative partiality is the mirror image of positive partiality towards intimates. A parallel question can be raised about prudence, though I cannot pursue here.

counterparts can be closer or more distant), a stage theorist can say that literally, the relation of being-the-same-person-as obtains to various degrees among person-stages.

Regardless of whether personal identity is literally gradable, we can speak of *personal closeness* (*or distance*) among person-stages being gradable. Personal closeness measures the amount or extent of the basic determinants of personal identity (whatever they may be) that obtain between person-stages. Given psychological reductionism, personal closeness just is psychological closeness. Say that P_1 is psychologically close to P_0 to a high degree D. P_1 is thereby personally close to P_0 to exactly the same degree D. And if psychological closeness is what grounds prudential partiality, then prudential closeness – the metric that proportionately determines the strength of agent-relative reasons – also matches psychological closeness and therefore personal closeness.

So, the above justification for prudential partiality in terms of psychological closeness coheres perfectly with psychological reductionism. There is a pleasing unity between the metaphysical and the normative. The fundamental grounds for personal identity and the fundamental grounds for prudential partiality coincide. The psychological reductionists can say that psychological connections are normatively significant *because* they are metaphysically significant. Slightly differently put, they can say that one person-stage is permitted to be partial to certain fellow stages *because* there is more of her in those stages.

But one does not have to be a psychological reductionist to endorse the view that psychological closeness grounds prudential partiality. That is, psychological closeness can ground partiality without factoring into an account of personal identity. You might think that persons are human animals and that human animals persist by perpetuating the proper biological processes (Olson 1997). For such an animalist, personal closeness and psychological closeness are separate things. Nonetheless, an animalist might well endorse the view that psychological closeness determines prudential closeness. Consider moral partiality. Your reasonable partiality towards your friends and spouse presumably has very little to do with the biological make-up of human organisms.[1] So, it should at least be consistent (and perhaps well-advised) for an animalist to hold that psychological connections are *prudentially* significant despite being metaphysically insignificant

[1] Though it might be the case that familial partiality with blood ties is partially grounded in biological relations. Nonetheless, it's hard to see how *mere* biological relations can ground partiality. It doesn't seem that one is equally morally close to a random sperm donor and her mother who has been a good parent in all respects. Moreover, biological relations are clearly irrelevant to other partial relationships.

(regarding personal identity). In fact, some animalists explicitly deny that prudence – and other "identity-involving" ethical practices such as attributing moral responsibility – should track personal identity (Olson 1997: 54-55). More generally, one may be a "minimalist" about the normative implication of personal identity (Johnston 1992). That is, one might think that the justification of prudence and other apparently "identity-involving" ethical practices just has nothing to do with how persons persist metaphysically. In all, the proposed justification for prudential partiality is not committed to any particular account of personal identity.

Now, it might be wondered whether prudence is thus *time-biased* – that is, future-biased or near-biased. It's one thing to say that stage-relative reasons vary in strength but another thing to say that they give rise to prudence partiality with a certain shape that's sought for. Compare: The mere fact that there are *some* agent-relative reasons does not suffice to show that paradigmatically partial relationships such as friendship are reasonably partial.

I have touched on this question in §2.4.2. There I used a series of imagined cases to pump our intuition that time-biases actually track, or at least should track, personal-time rather than metaphysical-time. However, I also suggested that the intuition is nonetheless vague and should not be taken as determining the *precise shape* of prudential partiality. Personal-time is an inclusive and somewhat superficial measurement of the kinds of changes and processes persons undergo over time. Presumably, not everything measured by personal-time matters prudentially (e.g., the growth of hair). But given the limited information presented in the imagined cases and our (reasonable) presumption that what really matters *roughly coincides* with personal-time, our intuition tells us that prudential partiality should *roughly track* personal-time.

This is not a trivial conclusion, but it clearly needs refinement. The above proposal that prudential partiality should track psychological closeness (rather than other elements of personal-time) is just such a refinement. And I am willing to say that given this proposal, prudential partiality is *not* literally time-biased, for neither metaphysical-time nor personal-time *per se* is what *grounds* prudential partiality. Indeed, the resulting shape of prudential partiality fails to even (perfectly) *coincide with* personal-time (for there are determinates of personal-time that are prudentially irrelevant). In this respect, then, the proposal is revisionary. But this is not changing the subject in any objectionable sense. I take it that the fundamental disagreement between time-neutralists and their opponents is whether prudence "demands equal concern for all parts of that person's life" (Brink 2011: 353). The picture of prudential partiality we have arrived at denies exactly that. Given

that stage-relative reasons vary in strength, it's not the case that from the perspective of any person-stage, all fellow stages must deserve equal concern.

But there is a more substantive problem concerning future-bias. With near-bias, we can say that the resulting picture is not *too much* of a revision because it's *roughly* near-biased. Since psychological closeness tends to diminish over time, it's usually the case that what's prudentially closer is also closer in time. So, the intuition had by some that near-bias is often permissible is vindicated (despite that what makes those instances of "near-bias" permissible is not temporal distance *per se*). However, it's unclear how psychological closeness (or any other salient candidate grounds of partiality) can be past/future-asymmetric. If that cannot be, then it's unclear if we can get anything that remotely resembles future-bias. In other words, we would end up with not a *refinement* but a *debunk* of our intuition. I turn to this problem in §5.7.

5.4. My Nearest and Dearest (2)

I find it compelling that psychological closeness is an important ground for moral partiality – and therefore also for prudential partiality. But I'm less confident that it's all that matters. A common complaint against the above picture of moral partiality is that it slights the *agential* aspect of special relationships. Friendships, you might think, are not mere collections of encounters and mutual causal influences. Friendships are, importantly, long-term projects that we *value*, are *committed* to, and *actively pursue*. Bernard Williams (1981) famously argues that special relationships are normatively significant because they constitute our personal "ground projects"[1] and shape our practical identity. Without ground projects, one would lack an agential perspective to engage with practical reasons – and even literally cease to exist. Alternatively, this agential aspect can be understood less self-centeredly (Helm 2008, Stroud 2010). Friendship, familial relationships, romantic relationships, and political associations, etc. can be seen as *joint* projects that call for and sustain some sort of collective agency. Participants in such collective agency are united by shared values and commitments, mutual recognition of such shared values, and mutual expectation that these mutual commitments perpetuate.

[1] Projects are to be distinguished from mere desires and fleeting preferences. Short of a definition of projects, projects are marked by stable, long-term commitments, are deeply valued by the subjects, and engender considerable constraints on what the subjects are willing to consider as reasons.

However the details are fleshed out, the important point is that the fact a project is *your* project (or a joint project in which you participate) has *per se* normative significance. You don't have to take on any specific projects. But insofar as you are committed to a project, it becomes a source of agent-relative reason for you. I don't have to befriend any particular individual. But by choosing to enter this very friendship and committing to its flourishing, I'm voluntarily taking on its concomitant normative burden.[1] And the agent-relative reasons generated need not be *merely instrumental*. That is, they can be concerning the individuals themselves rather than the projects (cf. Stroud 2010: 147-48). It would be an overly barren conception of partiality, even on the view that relationships are *personal* projects, that it's only reasonable to the extent that it's conducive to promoting the projects. In other words, justified partiality can extend beyond the purpose of perpetuating and flourishing the relationship itself. Indeed, I suspect that there are even project-grounded agent-relative reasons for things detrimental to the flourishing of the projects themselves. Suppose that my best friend has an overly limited friend circle (consisting of only me). Being a good friend, I should have special reason to encourage her to socialize with new people despite the potential cost that she may grow into new relationships and out of me. I have reason to see to it that *she* is well rather than just that the friendship is in good shape. (Though presumably after the relationship has ceased to be, it's no longer a source of partiality.) More generally, *project-grounded* partiality can go beyond being *project-centered* and be intrinsically about the individual's wellbeing.

It's easy to see a similar dynamic playing out on the intra-personal level. Distinct person-stages are often woven together by projects. They can also be committed to different projects that may come into conflict – in the sense that they compete for resources (time, energy, and money) such that the pursuit of one thwarts the other. We may think of the Russian nobleman's youthful stages as being united by a joint commitment to the socialist project and his elderly stages as being united by a bourgeois practical identity.[2] From his youthful perspective, there is an important normative

[1] Scheffler draws a distinction between seeing a relationship *as valuable* and *valuing* a relationship of one's own. The latter valuing attitude is considered as the source of agent-relative reasons:

> The value of the relationships may be—and may be seen to be—the same. But they will affect one's perceived reasons for action in different ways. If I value a relationship in which I am a participant, then I will treat that relationship as presenting me with reasons that differ from the reasons generated by other relationships of the same type in which I am not a participant. (2004: 249)

[2] The nobleman has a Williamsian line: "I regard my ideals as essential to me. If I lose these ideals, I want you to think that I cease to exist." (Parfit 1984: 327) Dietz (2020: 367-38) suggests that the young nobleman and the old nobleman are distinct agents, even if they are metaphysically the same person. This also echoes Williams's thought that ground projects determine one's practical identity. Although we need not equate

difference between his near-future socialist stages and the far-future bourgeois stages. It's not so much because the bourgeois project is worthless, but rather that it's not *his* project. More specifically, regarding his fellow socialist stages, he not only commits to cooperation on his part, but also (quite reasonably) expects such a commitment to be mutual. The same attitude is not and cannot be sensibly taken towards the bourgeois stages. The present stage may take some aggressive measures – say, sign an irrevocable contract to donate the future inheritance to the peasants (as Parfit suggests) – so as to forcefully perpetuate *his* project. But that wouldn't make socialism a *joint* project that unites his present and far-future stages. (See also Dorsey 2018: 283-84)

If projects are a source of moral partiality, they should also ground prudential partiality. Some of these reasons may be merely instrumental. For instance, since a debilitating health condition would be counterproductive to advancing the socialist course, the nobleman's present stage has an instrumental reason to take care of himself health-wise. But if it's correct that project-*grounded* partiality does not have to be project-*centered*, the young nobleman's reasonable partiality towards those comrade person-stages can take the general form of advancing their wellbeing, which includes advancing their interests that are tangential to the socialist project.

The claim that projects ground partiality does not mean that psychological closeness becomes irrelevant. Far from being mutually exclusive, projects and psychological closeness can both be partial grounds that are connected and complementary. As a descriptive matter, psychological closeness and joint commitments to projects tend to be mutually reinforcing. Joint projects require a decent amount of shared values to take off, and moreover, considerable trust and psychological vulnerability in order to sustain and flourish. The convergence in agential perspectives, in turn, tends to strengthen psychological intimacy. On the normative level, there is no principled reason why prudential partiality can only have a single ground. Just as moral partiality in different kinds of special relationships may have distinct (partial) grounds, intra-personal "special relationships" can also take different forms and have multiple potential grounds for partiality. Here, it's natural to think that the strengths of stage-relative reasons depend on both the degrees of psychological closeness and the extent to which the agential perspectives converge. (Though be careful not to double-count in case they interact and overlap on a descriptive level.)

---

practical identity with personal identity (such that a drastic reorientation in ground projects literally kills a person), we can appreciate that the old bourgeois becomes a "different person" in an ethically significant way.

Still some other elements may complement this picture. The final salient candidate ground I wish to draw attention to is *emotional*. It would be a caricature of romantic love and intimate friendship to leave out the complex syndrome of emotions involved. If your romantic partner doesn't rejoice in your delight or feel frustrated by your exasperation, you'll question if she or he really loves you. Likewise, emotional aloofness and asynchronization tend to mark the end of a friendship. But the emotions characteristic of intimate relationships are not merely *sympathetic* in the sense that one is "infected" by the emotions of another party (Helm 2010: 88). Seeing a stranger in distress, sympathy is a natural and appropriate response. What marks the emotional dynamics among intimates is a sort of *empathy*. What underwrites empathy is not so much a susceptibility to the (expressed) emotions of the other party, but a disposition to adopt her perspective when responding to certain situations, given one's understanding of her interests and desires. You are distressed by some bad news about your partner, and aptly so, even if they themself are still in the dark and yet to respond emotionally. Such perspective-taking also engenders a distinctive kind of *vicarious* emotions. You may be proud of some virtuous feat by your partner or feel ashamed about something foolish they've done, *as if* they are your own deeds. Vicarious pride and shame, on a phenomenological level, are verisimilar to being *self-directed*. They are usually sensible among intimates but not among strangers. Finally, it's worth emphasizing that distinctively second-personal emotions may also manifest differently between intimates. For example, one may sensibly feel deeper indignation or frustration when being wronged by one's best friend than by a random stranger. There is also a quirky side to emotional intimacy. "Mischievous delight" at the befuddlement of your loved one or amusement at her embarrassment can be tender expressions of love, but these emotions would be out of place if directed at non-intimates (Baier 1991: 443–44).

Not all the emotions characteristic of intimate relationships are apt in the intra-personal context. Only in rare circumstances one "loves" oneself in the sense of finding oneself romantically or sexually attractive, for instance. Nonetheless, the general point holds that one's fellow person-stages can be more or less emotionally intimate to each other. Anticipating impending pain feels different from anticipating pain ten years later. The stress and fear incurred by anticipating impending pain are just (helplessly) much more intense. More cognitively-loaded emotions also have intimate and estranged variants. Reflecting on some past wrongdoing, I may feel intense regret, remorse, or even self-hatred. These emotions are markedly first-personal and "intimate" in that they manifest an acknowledgment of self-ownership of the past wrongdoing on the part of my present stage. Or I may look back and *resent* my past stage. Although resentment is often no

less affectively intense, it is in this context a more self-estranging emotion. Resentment is paradigmatically second-personal – it involves an addresser and an addressee distanced by differing perspectives. It makes sense only if one is dissociating her present perspective from her past perspective and thereby denounces self-ownership. Or I may simply be emotionally indifferent. I may recognize that some wrong is done but think to myself: "What has this to do with *me*?" Indifference likewise falls on the side of estrangement, and perhaps of a more extreme sort.

Emotional intimacy tends to piggyback on psychological intimacy. Identifying with one's past perspective and acknowledging self-ownership (however spontaneous or volitional) requires a decent amount of shared understanding and values. Emotional intimacy also seems to be closely tied to psychological vulnerability – that is, the susceptibility to be impinged upon by the interests and desires of one's fellow stages. When I resent – rather than regret – my past wrongdoing, it's often because I fail to, or refuse to, understand the underlying motivation and acknowledge it as mine. In case of drastic changes of fundamental values or personal traits, perhaps I just cannot sensibly imagine that this is something that *I* could have done. Nonetheless, emotional intimacy seems to be conceptually distinct from psychological intimacy. We can imagine relatively deep psychological intimacy with relatively shallow emotional intimacy, as between romantic partners who have just begun to lose affection for each other. And it seems reasonable that all else being equal, such "degenerating" romantic relationships tend to demand diminished partiality. It likewise seems reasonable that when it comes to prudence, emotional intimacy may confer extra strength to stage-relative reasons in addition to psychological intimacy (and unification by projects).

One might sense some air of circularity here. It seems that I have been bootstrapping justified partiality from *de facto* partiality. The proposal seems to be saying that (say) since I feel more distressed when anticipating near-future pain than far-future pain, I am justified to prefer pain to be far-future rather than near-future. But that amounts to a viciously circular claim that because I *do care more about* near-future pain, I'm *justified to care more about* near-future pain. A similar circularity problem seems to infect the appeal to psychological vulnerability. It seems to be saying that my present stage *should* weigh heavily the desires of certain fellow person-stages because I *in fact* tend to care more about those desires. In a nutshell, all that we are saying seems to be that prudential partiality is justified because we are in fact partial – or in other words, we have reasons to *care more about* certain stages because we do *care more about* them.

There is some truth in this charge of circularity. Consider friendship again. It makes some sense to say that friendship just (in part) *consists in* partiality. A friend who is indifferent between promoting your interest and a stranger's interest just doesn't seem a genuine friend. It's also not an unreasonable claim that emotional intimacy is a special duty that relationships *call for*. That is, emotional intimacy is considered as what partiality demands rather than what grounds partiality. Quite generally, there doesn't seem a clear-cut line between the descriptive features of friendship and the normative implications thereof. Some philosophers even seem to embrace a sort of circularity. Samuel Scheffler (2004), for instance, claims that the hallmark of intimate relationships is the attitude of *valuing* (as opposed to merely considering something to be valuable). Valuing a relationship, according to Scheffler, just consists in seeing the relationship as a distinctive source of agent-relative reasons. This seems to imply that agent-relative reasons somehow bootstrap themselves: We come to have agent-relative reasons because we take there to be. Partiality is thereby self-vindicating.[1] (Scheffler thinks the same about personal projects.)

Perhaps an account like Scheffler's would be viciously circular. Or perhaps such circularity is innocent. (cf. Keller 2013: 120-23) But I think we can admit a sort of *self-reinforcing* mechanism between what special relationships *consist in* and what they *call for*, but without sliding into outright circularity. To start, it's worth emphasizing that not all the candidate grounds for prudential partiality involve, in any intuitive sense, "caring more" (an ambiguous term to be clarified shortly). Psychological closeness in part consists of memory connections and shared beliefs. These relations may facilitate "caring" but are not themselves "caring" in any sense. Things get murkier with the more conative side of psychological closeness. Shared desires, interests, and values – as well as the underlying vulnerability – are indeed sometimes caused by "caring". Insofar as I care more about certain future fellow stages and are thereby more susceptible to incorporating their desires as mine, our desires tend to converge. But there are other equally important bases of conative psychological intimacy that have little to do with our caring. Memories fade, beliefs change, circumstances alter, and transformative events befall. All these "external" factors may erode or even foreclose psychological intimacy no matter how much caring is on offer. (That's probably why long-distance relationships are difficult to maintain.) So, not all candidate *grounds*

[1] In a similar vein, it's been proposed that loving someone consists in bestowing agent-relative value to the loved one (rather than responding to some antecedent value). (cf. Helm 2023: §4.2)

for partiality can be sensibly folded into what partiality *demands*. The line between the justifier and what's being justified, though perhaps muddled, is not non-existent.

The line can be further un-muddled by disambiguating different senses of "caring". The charge of circularity, recall, is that one cannot bootstrap *reasons for* caring from caring *de facto*. The charge is already mitigated by the observation that some partial grounds for partiality (e.g., the non-conative aspect of psychological intimacy) is not a matter of "caring" in any sense. But we can further distinguish a *primitive* kind of caring from a *sophisticated* kind of caring. The former is folded into the grounds for partiality, I claim, whereas the latter is what partiality calls for. Jennifer Whiting has made a distinction between "primitive concern" and "general concern". Primitive concern, according to her, is "a concern on the part of our present selves that our future selves be doing or experiencing certain things" (1986: 564). It is "in principle no different from that involved in desires for immediate satisfaction" and does *not* depend on "any general conception of, or concern for, the welfare of its object" (ibid.). General concern, on the other hand, is directed at welfare-subjects *qua* welfare-subjects and expressed in more cognitively-loaded attitudes. Though the distinction I intend here may not be identical to Whiting's, I follow her important insight that there is a sort of caring that's just a helpless "matter of psychological fact" (Whiting 1986: 565), to be distinguished from another sort of caring that is more under deliberative control and is therefore subject to a robust form of normative evaluation. Unless I am completely confused about my own experiences, emotional intimacy and psychological vulnerability (or at least a significant part of them) simply "assail", and are not subject to deliberative control. When anticipating immediate (rather than far-future) pain, it's just a helpless matter of human psychology that anxiety and fear impinge on me intensely. Or compare the phenomenology between anticipating your spouse being in agony and your freshly-moved-in neighbor being in agony. Given your shared history with your spouse and other sorts of intimacy, you cannot help but be very agitated when anticipating his or her agony. And you cannot make yourself to feel the same about the neighbor however hard you try. There is a further question, however, if you have better reason to attend to your spouse's agony (or prefer that his or her agony is attended to). This question concerns the allocation of the more sophisticated form of caring that is susceptible to rational reflection and therefore subject to robust normative criticism. It's this sophisticated caring that's at the center of the partialism/impartialism debate. The impartialists are of course *not* making the unreasonable (and rather cruel) claim that your more intense emotional response concerning your spouse is inappropriate or should be eradicated. What they are against is partial sophisticated caring. They urge that despite your partial primitive caring, you should discount your emotion

and decide that your spouse and the neighbor equally deserve the analgesic after all (just as you can come to regard that it's best to get the root canal done asap despite enormous fear). The partialists, on the other hand, deny that sophisticated caring should be impartial. And they might locate the justification for partiality (partly) in primitive caring – which is what special relationships partially consist of. By distinguishing two kinds of caring, the charge of circularity is disarmed. The same is true of prudential partiality. Just to illustrate, I have explained earlier that psychological vulnerability concerns more about the starting point of one's deliberation rather than its product. Psychological vulnerability is a largely involuntary tendency to incorporate the (anticipated or remembered) psychologies of one's fellow stages into one's present deliberative perspective. Their interests and desires just "pop into mind" and automatically come into sight. But, of course, this leaves open the further question of how all these interests and desires ought to be weighed in order to arrive at all-things-considered decisions.

So, there is no circularity. There is a legitimate conceptual line to draw between the grounds for partiality and what partiality calls for. But it should also be acknowledged that these two aspects tend to be (causally) mutually reinforcing. That is, partial sophisticated caring tends to bring about intensified grounds for partiality, which in turn strengthens the agent-/stage-relative reasons for partiality. This is how friendships evolve from being casual and shallow to being intimate and deep. Favoring acts create environments and opportunities for shared history and emotional bonds, which in turn pave the way for deeper commitments and weightier agent-relativity. But such causal cycles do not seem morally vicious. Nor does mutual reinforcement entail that friendship must self-perpetuate indefinitely or that one cannot justifiably discharge the normative burden by exiting the relationship when it gets sour. The same goes for prudential partiality. For instance, if your past stages have persistently taken an aggressive and uncooperative stand towards your present stages – as the young nobleman might have done to his older stages by signing the contract of donation so as to thwart their projects – then you may reasonably cease to treat them as "friends" and discharge yourself from the demands that would otherwise be called for by the relationship. (cf. Dorsey 2018: 283-84)

5.5. Partiality, Prudential and Moral

The above argument for prudential partiality exploits the parallel between morality and prudence. It crucially depends on the observation that the salient candidate grounds for moral partiality –

viz. psychological intimacy, joint projects, and emotional intimacy (and perhaps more) – can also be found to obtain, to various extents, among different person-stages of individual persons. Differently put, relative to a person-stage, certain fellow stages can be more friend-like while others, more stranger-like. Prudential partiality (in proportion with prudential closeness) is thereby justified.

The argument is admittedly schematic but hopefully compelling. At least, it should be compelling to those who are antecedently sympathetic to moral partiality. Differently put, the argument shows that moral partiality is in great tension with prudential neutrality because these two domains share the same potential grounds for partiality. However, it might be argued that the two domains are importantly disanalogous, so that prudential partialism does not naturally follow from moral partialism.

The objection from disanalogy may take a variety of forms, given the multitude of potential differences between morality and prudence. Here I can only briefly set forth several conspicuous versions of the objection and then suggest that they are far from decisive.

First, it may be argued that it's crucial to justified partiality that the relationships are *mutual (or reciprocal)*, but distinct person-stages of an individual do not seem to stand in mutual relationships. Plausibly, not all inter-personal relationships (in a capacious sense of the term) are apt for partiality. Think of parasocial relationships or unrequited infatuation. These relationships do not seem to justify partiality (on either party). And this seems to be because unlike paradigmatic friendships or romantic relationships, they are sadly one-sided and not mutual. If that much is true, then it would be bad news for the prudential partialists if intra-personal relationships are like parasocial relationships or unrequited infatuation. However, it seems clear to me that intra-personal bonds-of-concern can and often do obtain mutually between person-stages. Consider first psychological vulnerability. Normally, one's present mental states are susceptible to being influenced by both her past and future mental states. In the past-to-present direction, it's often a psychological default that one inherits her past desires, beliefs, and intentions – absent anything calling for reconsideration. In the present-to-future direction, it's also common for one to incorporate her anticipated future desires, beliefs, and interests into her present deliberative perspective. That just means that psychological vulnerability is often mutual. Similarly for emotional intimacy. I've mentioned only a few examples of past-directed and future-directed intimate emotions (e.g., remorseful regret and distressed anticipation). But such examples, in both directions, can obviously be multiplied.

Mutuality is even more pronounced when person-stages pursue joint projects. Carol Rovane (1997) has detailed how intra-personal cooperation resembles inter-personal cooperation in all the normatively significant respects. Here I only wish to briefly mention one feature, namely that normally, when one (reasonably) commits to a long-term project, she expects her later stages to carry on with the project with the *recognition* that this is what she's been committed to (instead of using brute causal force to perpetuate the project, as Ulysses who ties himself up to prevent being seduced by the siren). In this sense, intra-personal cooperation is often genuine *co*-operation with mutual commitment – not unlike inter-personal cooperation.

There are still other differences between inter-personal and intra-personal relationships. For one, interactions among fellow stages, however mutual, are hopelessly *temporally detached*. Inter-personal relationships, by contrast, involve interactions that are *concurrent*. Although it's a sad truth that distinct stages of mine cannot literally spend time together (saving time-travelers), I fail to see why this difference is normatively significant. Consider pen-pals who have never met each other but have formed very deep bonds. Despite having *lives* that are temporally overlapping, their *interactions* are desynchronized. They do not literally *spend* time together; but that shouldn't undermine their relationship. However, it might be further suggested that there is a sense in which pen-pals *genuinely interact* but my person-stages don't. Crudely put, there are causal arrows reaching from me to my pen-pal and then back to me, and so forth. But there is no causal arrow reaching from a past stage to a present stage *and then back to* the past stage. Yet again, it's unclear to me why the lack of genuine interactions necessarily undermines partiality. Some inter-personal relationships involve non-overlapping lives, so that genuine causal interaction is impossible. Parental relationships can sometimes, unfortunately, take this form. But it seems that such parental relationships can nonetheless generate agent-relative reasons (even if mere blood ties do not suffice for partiality).[1] And this would seem to have much to do with the fact that the parent "lives on" in the form of memories retold and reconstructed by other family members and people around – and thereby "passes on" his or her expectations, personalities, and ideals to the child. But that seems to be exactly how temporally detached person-stages manage to "interact" with each other – through memories and anticipations, despite lacking causal interactions in the narrow sense. Indeed, "interactions" through memories and anticipations are more close-knit and reliable in the intra-

[1] Though it's true that as of now, there is not much that the child *can do for* her parent, and nothing that the parent can do for the child. I'd like to say that here, the bases of agent-relative reasons obtain; it's just that the reasons lack enablers. (Dancy 2004) And it remains true that there are (enabled) reasons for partial *preferences*.

personal case than in the unfortunate parental relationship just considered. (Unlike how my future stages relate to my present stage, the child cannot literally remember her parent). If the unfortunate parental relationship, despite lacking genuine causal interactions, is apt for partiality, it's hard to see how intra-personal relationships could fail to be partial.

The disanalogy objection may take a quite different form. The above objections are in effect saying that the grounds for moral partiality do not really obtain in the intra-personal case. But one might think that even if these grounds, as a descriptive matter, traverse the two domains, they somehow lose their normative import when it comes to prudence. This objection may be spelled out in different ways, for there are a bunch of (alleged) differences between morality and prudence one may point to. Here I can only address a familiar one.

One often-evoked difference between morality and prudence concerns compensation. (Brink 1997; cf. Rawls 1971) It's suggested that one deciding reason why morality cannot be completely agent-neutral is that compensation is not automatic in inter-personal tradeoffs (though the beneficiary may kindly repay in due course). That is, due to compensation being non-automatic, it matters *who* suffers and *who* benefits. But a parallel reason for stage-relativity does not obtain. In intra-personal tradeoffs, any damage or sacrifice is automatically compensated because the benefactor and the beneficiary are the very same person. So, there is no need for stage-relativity. This appeal to compensation is shown to be problematic on multiple grounds.[1] What's most relevant here, however, is that there is good reason to think that the normative significance of compensation tracks not personal identity *per se* but bonds-of-concern. So, it's untrue that compensation drives a normative wedge between morality and prudence.

Suppose that you can do a huge favor for your best friend. Then consider an otherwise identical scenario in which the beneficiary is a random stranger. Both cases involve sacrifice that is not automatically compensated – insofar as automatic compensation entails personal identity, as it's presumed by the objector.[2] But evidently, you have more reason to sacrifice for the sake of your best friend than for the stranger. This shows that the extent to which compensation is a hurdle to inter-personal trade-offs is a function of bonds-of-concern rather than personal identity *per se*.

[1] See especially Dorsey (2021: 256-60), Pettigrew (2019: 179-83), Jeske (1993); cf. Parfit (1984: 329).
[2] Alternatively, one might say that in benefiting your friend, your sacrifice is *indirectly* compensated because *your* friend is directly benefited. One might even think that this is not much of a sacrifice because her interest is part of *your* interest. No matter how sacrifice and compensation is described in this case, it generalizes to the intra-personal.

Intra-personal trade-offs are no different. Thinking this way, stage-relativity is *supported by* rather than undermined by considerations of compensation, precisely because bonds-of-concern obtain to various extents intra-personally and the normative significance of compensation tracks these bonds-of-concern. My present stage has greater reason to sacrifice for an intimate fellow stage than for an estranged fellow stage precisely because of the difference in bonds-of-concern. And that just is prudential partiality as I have argued for.

Finally, as a general note, it seems that any objection appealing to a distinction between prudence and morality as different domains would face the problem that the borderline between these two domains may not be clear-cut. I do not know how exactly to demarcate these two domains, but very likely such a demarcation shall involve some reference to personal identity. [1] However, many take personal identity to be a vague matter. Given psychological reductionism, for instance, there are borderline cases in which it's indeterminate if personal identity obtains (Parfit 1984, Shoemaker 1999). And as mentioned in §5.3, some even claim that personal identity is *always* a matter of degree. (Lewis 1983; Braddon-Mitchell and Miller 2020). If either of these claims is true, then any personal-identity-involving distinction between morality and prudence cannot be clear-cut. Indeed, there may be a stronger upshot that follows from the above discussions about bonds-of-concern – namely, that there is no normatively meaningful distinction to draw at all. We have seen that the normative significance of bonds-of-concern traverses the boundary of personal identity (and therefore the boundary between morality and prudence). We have also seen that compensation – erroneously thought to mark a distinction between morality and prudence – turns out to track bonds-of-concern. The emerging picture seems to be that bonds-of-concern – which obtain both inter-personally and intra-personally – have done a lot of normative heavy-lifting in both what we intuitively call morality and prudence respectively. It's thus unclear how to draw a non-trivial distinction that can warrant some normative asymmetry between morality and prudence. This is a big claim that I cannot defend here, but David Brink's (1997) metaphysical egoism is one theory of practical reason that in effect removes the line between morality and prudence. Although I disagree with Brink on many issues, we agree that because the grounds for prudence – psychological connections, according to Brink – also obtain among distinct individuals, the line between prudence and morality is thereby muddled. What Brink fails to

[1] There is the truism that prudence is self-regarding but morality is other-regarding. But this seems to fail to accommodate the fact that questions of morality are often questions involving conflicts between *oneself* and other people. Also, it's a plausible claim that there can be self-regarding moral duties.

realize, however, is that if bonds-of-concern vindicate agent-relativity (which he endorses), then they should also vindicate stage-relativity.

5.6. Lifetime Wellbeing and Prudential Latitude

I have so far argued for a rough picture of prudential partiality and defended it against some objections. This section explains what prudence partiality as such may imply about *the Argument from Wellbeing* – and more generally, how a prudential partialist might think of prudential values and reasons. The discussion will inevitably implicate some fundamental debates concerning values and reasons. Where full neutrality is impossible, I can only take a stand without detailed justification.

Recall *the Argument from Wellbeing against Time-Biases*:

> (1) Prudentially[1], one ought *not* prefer that her whole life *does not* fare as well as possible. (*No Tragedian*)
> (2) If one is prudentially future-biased *beyond tie-breaking*, then one prefers that her whole life does not fare as well as possible.
> (3) If one is prudentially near-biased beyond tie-breaking, then one prefers that her whole life does not fare as well as possible.
> (4) Therefore, one prudentially ought not be prudentially future-biased beyond tie-breaking. (from (1) and (2))
> (5) Therefore, one prudentially ought not be prudentially near-biased beyond tie-breaking. (from (1) and (3))

I have suggested how a wellbeing subjectivist might respond to this argument. (§4.6) To summarize briefly: Given subjectivism, time-biased preferences are perspectival and higher-order preferences about the welfare-relevance of first-order preferences at different times. Insofar as they survive idealizing, etc., these time-biases, as higher-order preferences, are themselves welfare-

[1] I focus on prudential oughts rather than all-things-considered oughts. There might be countervailing non-prudential (say, moral) reasons that are weighty enough so that all-things-considered, one ought to prefer otherwise. In what follows, I assume that no non-prudential considerations are in play unless indicated otherwise.

relevant by way of altering the welfare-impact of the first-order preferences (were these first-order preferences satisfied or frustrated). It then follows that lifetime wellbeing is perspectival and, for (idealized) time-biased agents, shifty. The argument is disarmed because it's far from true that a time-biased agent must be preferring a life that is, from her *time-biased perspective*, overall worse.

As I see it, wellbeing subjectivists have pretty much opted out of the debate between prudential partialism and impartialism, for they hold that there are no attitude-independent substantive normative reasons of prudence.[1] The prudential partialism/impartialism divide, at least as I am using the terms, is internal to wellbeing objectivism. The partialists claim that as an attitude-independent matter-of-fact, there is permissible or even obligatory prudential partiality. (This is not to say that psychological intimacy, etc., are not constituted by subjective attitudes. The claim is that once the bonds-of-concern are in place, it's not up to us to decide their normative significance.) The impartialists instead claim that as an attitude-independent matter-of-fact, there is no permissible partiality.

So, which premise(s) in the above argument should the partialists reject? (Though as I have suggested, talking about near-biases and future-biases is sort of a red herring. The shape of prudential partiality is determined by bonds-of-concern instead of time *per se* given the view defended here.)

The partialists may mimic the subjectivist response. That is, they may say that facts about lifetime wellbeing are perspectival and usually shifty. But unlike the subjectivists, the partialists take these perspectival facts to be fixed by facts about bonds-of-concern. More specifically, relative to a particular person-stage P, the welfare (dis)values of states-of-affairs befalling any other fellow stages $P^{\#}$ depend on the bonds-of-concern between P and $P^{\#}$. That is, the prudential value of a state-of-affairs S consisting of $P^{\#}$, relative to P, is an increasing function of both the unadulterated value of S and $PC(P, P^{\#})$. (PC means prudential closeness, which is proportionate to the amount of bonds-of-concern.) Insofar as the partial preferences of a person-stage accurately track facts about prudential closeness, she does not violate a perspectival-ized version of *No Tragedian*.

[1] Though some (idealist) subjectivists may end up with a picture of prudential partiality that's extensionally similar to certain versions of partialism which allow for roomy prudential latitude. See Gert (2016) for a similar claim about moral subjectivism and moral partialism. More on prudential latitude shortly.

This is reminiscent of *agent-relative consequentialism*. (Smith 2003, Portmore 2008, Hammerton 2020) The agent-relative consequentialists account for moral partiality by grounding agent-relative reasons in *agent-relative values*.[1] Just to illustrate, consider again the analgesic case. My being partial to my spouse is justified, say the agent-relative consequentialists, because the deontic statuses of *my* actions and preferences do not depend on an impartial value ranking of states-of-affairs, but instead depend on an agent-relative ranking that is relative to *me*. When determining what some particular individual should do, we need to look at how much the alternative states-of-affair are valued relative to that particular individual. The state-of-affairs that *my* spouse becomes pain-free, *ceteris paribus*, is of greater value *relative-to-me* than the state-of-affairs that a random stranger becomes pain-free. This is because agent-relative value increases when moral closeness increases. The strengths of agent-relative reasons are in turn directly determined by agent-relative values. Importantly, agent-relative values differ from agent to agent. Suppose that the other individual in pain, who is a stranger to me, happens to be your spouse. According to the value ranking *relative-to-you*, it would be better that your spouse becomes pain-free. You thereby have greater agent-relative reason to attend to your spouse.

The first brand of prudential partialism introduced above, by making facts about wellbeing perspectival and dependent on bonds-of-concern, postulates *stage-relative (prudential) values* and accounts for stage-relative reasons in the same way as agent-relative consequentialism does for agent-relative reasons. I'll refer to the view *stage-relative consequentialism*.

There is, however, a problem with agent-relative consequentialism as it stands. It's the problem of failing to accommodate *substantive latitude*. The problem carries over to stage-relative consequentialism.

By *latitude* I mean preference or choice situations in which more than one alternative is *permissible but not obligatory* – more simply, *merely permissible* or *optional*. (I'll use these terms interchangeably depending on the context.) It is completely uncontroversial that there are some situations of optionality in morality as well as in prudence. Agent-/stage-relative consequentialism can accommodate this. But I also take it to be a very compelling thought that morality and prudence allow for *substantive latitude* – and in particular, substantive latitude in the context of

[1] Some think that by postulating agent-relative values (or reasons), a theory ceases to be consequentialist (McNaughton and Rawling 1995). Perhaps so. But I shall not be concerned about labelling issues.

partiality.[1] Very roughly, substantive latitude in moral partiality manifests itself as a *wide range* of scenarios of optionality. It's the idea that partiality to intimates is *very often* permissible but not obligatory – and also that the deontic status of mere permissibility tends to be *"insensitive to mild sweetening" or souring* (Hare 2010: 237). When it comes to prudence, there is the familiar idea that time-biases (future-bias and near-bias), when they are permissible, are often not obligatory.[2]

To illustrate, consider again the analgesic case in which all else is equal but the stranger is suffering a bit more pain than my spouse. It should seem permissible for me to use my only dose of analgesic to unburden my spouse, but also permissible to unburden the stranger.[3] That is, both alternatives are optional. More importantly, if this is a clear case of optionality, then there should also be a substantive range of cases of optionality in its vicinity with slightly variant agent-relative values.[4] Make the stranger's pain a bit worse. Or make my spouse suffer a bit more (but not more than the stranger does). In either case, it seems unlikely that I'd suddenly be *obligated* to unburden the stranger or my spouse. In other words, the deontic status of being merely permissible is at least to some extent *insensitive* to mild sweetening or souring.

Agent-relative consequentialism as stated can allow for *some* latitude. Optionality happens when the agent-relative values of both alternatives are exactly the same, giving rise to equally weighty agent-relative reasons. But it does not allow for *substantive* latitude, for mild sweetening or souring in agent-relative values inevitably topples the balance of reasons and results in a single obligation.

And let me claim upfront that opting for *satisficing* (as opposed to maximizing) consequentialism is a bad solution. Roughly, satisficing is the idea that one is not always required to do what's best – instead, one is permitted to opt for what's "good enough". There are many general problems with satisficing (either in morality or prudence). (Bradley 2006, Dorsey 2021: 229-35) It's also

[1] There are other kinds of latitude that one may wish to accommodate. I wish to briefly mention one kind. Some think that morally, one is sometimes merely permitted (and not obligated) to self-sacrifice for what's impersonally *suboptimal*. A parallel prudential case would be something like: Preferring *greater present* pain over lesser future pain. I don't have a clear intuition if this is permissible. If it's permissible, then the model for latitude I shall propose shortly is severely incomplete. But this does not mean that the dual-weight account, specified differently, cannot accommodate this.
[2] See especially Hedden (2015) and Dorsey (2021).
[3] The reader doesn't have to share the intuition with any particular case. The reader may revise the relative weights of the (unmodified) agent-neutral values to reach a clear case of mere permissibility. What's more important is that given that a case is a clear instance of mere permissibility, it should be at least to some extent insensitive to mild sweetening.
[4] See also Portmore (2001, 2003), Tucker (2021, 2024)

particularly ill-motivated in the context of partiality. This is because substantive latitude in partiality seems to have much to do with a conflict between two perspectives, so to speak. In the moral case, for instance, substantive latitude seems to arise from a conflict between an agent-neutral perspective to optimize the general good and an agent-relative perspective to prioritize the interests of one's intimates – and a sense that these two perspectives are *not easily* "commensurated", so that the conflict doesn't always receive a definite resolution. I take this to be a basic intuition that any decent solution should reflect. Satisficing, however, is completely oblivious to this conflict in perspectives. Given satisficing, what counts as "good enough" is utterly indifferent to the significance of bonds-of-concern on the one hand, and the significance of agent-/stage-neutral values on the other hand. And there's good reason to suspect that by virtue of being ill-motivated, satisficing is destined to be extensionally inadequate as a solution. So, I'll henceforth assume a maximizing framework (not saying that one cannot go satisficing for other reasons). That is, I assume that whenever there is a definitive winner in the game of weighing reasons, there is no optionality. (To anticipate, my preferred account of latitude challenges the very notion of "definitive winner".)

When it comes to prudence, I find the presence of substantive latitude equally compelling, if not more. Comparing greater benefit to an intimate fellow stage with lesser benefit to an estranged fellow stage, I find it very compelling that there should be wide spectrums of cases with slightly variant amounts of benefits and/or extents of prudential closeness that are all cases of optionality. That said, I cannot really argue for it here.[1] Suffice it to say that substantive latitude in prudential partiality is a *very live* possibility, so it would be very disappointing if we cannot accommodate it. Nor do I have a clear idea about *how much* latitude there is. In fact, I do not know whether prudential partiality, when permissible, is *sometimes obligatory* or *always merely permissible*. You might think that one is obligated to be partial when the benefits are equal, or when the discrepancy in benefits is "small enough" in comparison to the discrepancy in prudential closeness.[2] But I myself am inclined to think that prudential partiality is *never obligatory*. I also find it compelling that partiality is *sometimes impermissible*. In the moral case, trivial benefit to one's intimate at the tremendous expense of some stranger does not appear permissible. The same seems true of prudence. For illustrative purposes, the "toy model" offered below will reflect my personal belief

[1] You might object that substantive latitude is merely apparent and it's due to an epistemic failure to pin down where exactly the perfect balance lies.
[2] Or you might think that there is a kind of prudential partiality that's, when permissible, always obligatory because it patterns with *associative duty*. I've been so far setting aside duties (if understood as something different from merely very weighty reasons). I'll briefly discuss duties in §5.7.

that partiality is *sometimes impermissible* but *never obligatory*. This doesn't mean that prudential partiality *must* take this exact shape.

Various strategies have been proposed to accommodate substantive moral latitude, though they usually focus on moral latitude of other sorts (e.g., supererogation). Here I shall focus on substantive latitude in prudential partiality directly. Taking stage-relative consequentialism as a starting point, one might want to revise its axiology, or revise how axiology relates to reasons, or revise how reasons determine deontic statuses (or some combination of these). My preferred account relies on a *dual-scale* or *dual-weight* account of reasons.[1] The account, in its final statement, departs from stage-relative consequentialism in all three aspects. Some of these points of departure are optional; but it's essential to this account that there is a non-trivial distinction between the *permitting weight* and the *requiring weight* of reasons[2]. Given this distinction, deontic status is not in any straightforward sense determined by "the overall balance of reasons".

A quick outline of the dual-scale account of reasons before explaining its attractions. Start from the permitting/requiring weight distinction. A reason carries permitting weight iff it can render *permissible* alternatives that are, without it, impermissible. A reason carries requiring weight iff it can render *obligatory* alternatives that are, without it, not obligatory. It's completely *uncontroversial* that reasons can carry permitting weight and can carry requiring weight. It's just that the dual-weighters take the distinction to be *non-trivial*: They claim that the permitting weight of a reason may come apart from its requiring weight. For instance, there may be reasons with permitting weight *only*; those reasons can push alternatives towards permissibility but never towards obligation. The standard single-weighters, by contrast, assume that the permitting weight of a reason just is its requiring weight. For this reason, latitude occurs only when reasons are tied.

In addition to a non-trivial permitting/requiring distinction, a *dual-scale* is needed to account for substantive latitude. Given the duality of weights, a single-scale can no longer adequately reflect "the overall balance of reasons". Instead, two separate scales are required to weigh reasons against each other (to be detailed below).

[1] Gert (2007, 2016), Greenspan (2005), Little and Macnamara (2021), Tucker (2021, 2024); cf. Portmore (2008).
[2] Some writers prefer to call permitting weight "justifying weight" instead.

Thus stated, the dual-weight account may appear rather alien and overly technical. And you may wonder why we are getting into this. While it's not the place to defend the account, I should say a few words in its favor. First of all, although not many have overtly made a non-trivial requiring/permitting distinction, the underlying intuition is not unfamiliar.[1] Just to illustrate, say that you may donate to charity to relieve a child from famine, but you also need a new car. If you could relieve famine just by a finger snap, then you would be obligated to do so. But given the self-sacrifice involved, you now have an "excuse", as it were, not to fulfill what would otherwise be an obligation.

Indeed, protection of self-interest seems to be a *very effective* "excuse". The need for a new car may even excuse you, for instance, from not relieving *several* children from famine. On the other hand, your self-interest carries *very weak* (or no) moral *"demand"*. If you choose to donate rather than benefit yourself, you are *not being immoral* – on the contrary, you are being morally admirable. Indeed, you won't be immoral (though perhaps irrational) even when the benefit to others is *trivial* relative to your self-sacrifice. The dual-weight account well captures the intuition that there is a significant asymmetry between self-interest as a moral "excuse" and as moral "demand": This is because from a moral perspective, reasons of self-interest carry great permitting weight but little (or no) requiring weight. (Gert 2004: 22-28) And when supplemented with the dual-scale for determining deontic statutes, the dual-weighters can explain why self-sacrifice for the overall good is often supererogatory – that is, often "beyond the call of duty". More generally, the dual-scalers can offer a unified account for a wide range of ethical phenomena that would otherwise be perplexing, including supererogation and moral dilemmas. And as Joshua Gert (2016) and Chris Tucker (2021, 2024) have suggested, the dual-scale account is an ecumenical framework in the sense that it's not so at odds with the purported "alternative" solutions to the various conundrums. Instead, it seems that the "alternatives" will either be supplemented with the dual-scale or can be translated into the dual-scale. (More on comparing different solutions to latitude in partiality by the end of the section.)

Despite the virtues of the general framework, the dual-scalers haven't said much about substantive latitude in moral partiality let alone prudential partiality.[2] Here I shall take some initial steps in offering an account of substantive latitude in prudential partiality (which is hopefully

[1] Gert (2016) offers an overview of how one might arrive at the dual-weight account from various starting points.
[2] Tucker (2024) has an account of latitude in supererogation. I draw much inspiration from that.

generalizable to the moral case). I follow the lead of Dale Dorsey who suggests that "the fact that a particular good ϕ benefits [an intimate fellow stage] in comparison to another good ψ is a reason of permitting strength to promote ϕ rather than ψ" (2021: 251). Dorsey seems to take stage-relative reasons as something *extra to* rather than derivative upon stage-neutral reasons. But I have argued in favor of viewing bonds-of-concern as modifiers of stage-neutral reasons. I shall develop his suggestion by taking bonds-of-concern to be modifiers of *permitting weight* but not requiring weight.

Let me clarify that I do not intend this to be the *only* viable dual-scale account of latitude in prudential partiality. Joshua Gert (2007: 548) makes the "opposite" suggestion that temporal distances modify *requiring* strength but leave permitting strength intact. I'm genuinely unsure if this is more of a difference in technicality or something more substantive, but here is my *very personal* motivation for taking bonds-of-concern to modify permitting weight: I am deep down an old-school consequentialist. And I take it to be an elementary consequentialist intuition that morality and prudence alike make a *prima facie demand* to maximize the overall good. But I also acknowledge that morality and prudence cannot be thoroughly impartial. Bonds-of-concern matter. (There may be some other constraints on impartial maximizing.) Given my old-school consequentialist starting point, at least in the case of prudence, I am inclined to view partiality as something to be *condoned* rather than *commended*.[1] Facing the *prima facie* demand for impartial maximization, bonds-of-concern function much like *excuses*. Bluntly put, regarding reasons and apparent moral/prudential worthiness, prudential *impartiality* is akin to supererogation and prudential *partiality* is akin to permissible self-benefiting options that are impartially suboptimal.[2] (Recall the famine relief case above.)

All this is very impressionist. And the reader may reasonably disagree. If you find Gert's alternative suggestion more attractive (because you consider prudential *partiality* to be intuitively *better* than impartiality when they're both permissible), you are welcome to pursue this alternative suggestion. In my own model, however, I take bonds-of-concern to be modifiers of permitting but not requiring weight.

[1] This sounds less true in the case of morality. Moral partiality may appear more admirable. So, it may be reasonable to apply Gert's suggestion to moral partiality but Dorsey's suggestion to prudential partiality.
[2] On the standard account of supererogation, supererogation is in some sense morally better or morally worthier. Tucker's (2021) dual-weight account explains that by assigning greater *requiring* weight to supererogation; but one might well locate the sense of moral betterness somewhere else.

Suppose that relative to P_0, P_I is an intimate fellow stage and P_S is a "perfect stranger" (for the sake of simplicity). To make things more concrete, let us assign numbers to prudential closeness. Let $PC(P_0, P_S)$=1 so that the complete lack of prudential bonds-of-concern, as a modifier, leaves *neutral* impact on the weight of stage-neutral reasons. (This is partly because it's intuitive to consider perfect strangers as occupying a default middle point in the space of morality, and partly for convenience.[1]) And let $PC(P_0, P_I)$=2, reflecting that P_I is an intimate fellow stage relative to P_0.

Here are the exhaustive and mutually exclusive alternatives for P_0. *A*: Make P_S enjoy great pleasure; and *B*: Make P_I enjoy minor pleasure. Let's assume that the *unmodified* weights of reasons provided by pleasure – that is, the default weights of the *stage-neutral* reasons in play here – are proportionate to the amount of pleasure. So, let's make *NR(A)*=400w and *NR(B)*=200w. Read: The unmodified stage-*N*eutral *R*eason for *A* has 400 units of weight and the unmodified stage-neutral reason for *B* has 200 units of weight. So far this is not a dual-weight account. I've made no distinction between requiring weight and permitting weight. And note that this is a case where the single-weighters *do* say that both alternatives are permissible. Modifying *NR(A)*=400w with $PC(P_0, P_S)$=1 and *NR(B)*=200w with $PC(P_0, P_I)$=2, we get *RR(A)*=400w and *RR(B)*=400w (read *RR* as stage-*R*elative *R*eason). But we do not have *substantive* latitude, as should be clear. This is just an odd case in which the *RRs* happen to be equally weighty. Any mild sweetening or souring in *NRs* or *PCs* would render either *A* or *B* obligatory.

Now, let us distinguish permitting weight from requiring weight. For each of the stage-neutral reasons, I shall stipulate for simplicity that its unmodified permitting weight equals its unmodified requiring weight. [2] That is, *permitting-NR(A)=requiring-NR(A)*=400w and *permitting-NR(B)=requiring-NR(B)*=200w. (This is not to say that there is no substantive permitting/requiring distinction with these stage-neutral reasons. It's just the weights happen to be the same by stipulation.) As explained above, I follow Dorsey's suggestion and take *PCs* to modify *permitting* weights but *not requiring* weights. This results in: *Permitting-RR(A)=requiring-RR(A)*=400w; *permitting-RR(B)*=400w and *requiring-RR(B)*=200w. In modifying, the requiring and permitting

[1] In general, the specific numbers assigned in this toy model are only for illustrative purposes. Also, I am again setting aside the issue of asymmetric prudential closeness. (cf. §5.7)

[2] There may be a substantive issue here. Some (unmodified) reasons, by their nature, carry more requiring weight than permitting weight or *vice versa*. For instance, I've mentioned in the context of supererogation that from the perspective of morality, reasons of self-interest carry great permitting weight but little or no requiring weight. But as far as I can see, there is no obvious reason to think that prudentially, the permitting and requiring weights of stage-neutral reasons concerning pleasure should come apart.

weights of *NR(A)* are left intact, the requiring weight of *NR(B)* is left intact, but the permitting weight of *NR(B)* is intensified. (*RR(A)*, though carrying its default weights, is nonetheless an stage-relative reason. It's just that the modifier has left neutral impact on its weights.)

Suppose that *RR(A)* and *RR(B)* are the only reasons relevant. How are these reasons to be weighed to determine the deontic statuses of *A* and *B*, on the *dual-scale*? A full illustration can be found in Chris Tucker (2021); here I shall be brief. The business of determining deontic statuses consists of two steps: (i) Two scales to *separately* determine the respective *permissibility* of *A* and *B*. We weigh *permitting-RR(A)* against *requiring-RR(B)* to determine if *A* is *permissible* for P_0. Parallelly, we weigh *permitting-RR(B)* against *requiring-RR(A)* to determine if *B* is *permissible* for P_0. This is rather intuitive, and something the single-scalers also admit. Think of the first scale which determines the permissibility of *A*. The permitting weight for *A* just is what may potentially make permissible *A* which would otherwise be impermissible. The requiring weight for *B* just is what may potentially make obligatory *B* which would otherwise not be obligatory – and by the same token, may potentially make *impermissible* the other alternative *A* which would otherwise be *permissible*.[1] Then, let's say that an alternative is permissible iff its permitting reasons are *at least as weighty as* the requiring reasons for the other alternative. It follows that *A* and *B* are both permissible, for *permitting-RR(A)*=400w>*requiring-RR(B)*=200w and *permitting-RR(B)=requiring-RR(A)*=400w. Step (ii) is the easy step of tallying the results of the two scales. Since both of the (exhaustive and mutually exclusive) alternatives are permissible, they are both *merely permissible* – that is, permissible but not obligatory. (If, say, *A* were permissible and *B* were *im*permissible, then *A* would be obligatory.)

How does the dual-scale account for substantive latitude? Note first that I've made the example an "edge case" of permissible partiality. I mentioned that I personally view partiality as sometimes impermissible. I find it impermissible to let an estranged stage suffer tremendously just to spare an intimate from trivial pain. (In case you disagree, feel free to propose alternative modifying functions or alternative assignments of default weights.) This case shows the limit of partiality: If we instead make the pleasure of P_1 (the intimate) very *trivial*, then it would become that *permitting-RR(B)<requiring-RR(A)*. *A* would then be obligatory and *B* would be impermissible because the two scales respectively tell us that *A* is *permissible* and *B* is *impermissible*. On the other hand, there

[1] I am speaking very loosely here without much attention to the possibility of dilemmas and scenarios with more than two alternatives. (cf. Tucker 2021, 2024)

is substantive latitude. For example, we may sweeten the pleasure befalling P_1 up all the way until the alternatives contain equal amounts of pleasure, during which process both alternatives remain merely permissible. (Such sweetening strengthens both *requiring-RR(B)* and *permitting-RR(B)*. When the alternatives are stage-neutrally equal, we get an opposite "edge case" where *permitting-RR(A)=requiring-RR(B)*=400w and *permitting-RR(B)*=800w>*requiring-RR(A)*=400w.[1]) But above this point, when P_1's pleasure exceeds P_S's pleasure (so that *permitting-RR(A)<requiring-RR(B)*), it becomes obligatory for P_0 to make P_1 happy. This seems to be the desired result: Partiality at the expense of what's stage-neutrally optimal is to be directed towards intimates but not towards strangers.

Finally, notice that in this toy model, any sweetening of $PC(P_0, P_1)$ will not make partiality obligatory, for holding everything else fixed, it's always that *permitting-RR(A)>requiring-RR(B)*. Although I find this a welcome result, you may instead think that partiality towards intimates is sometimes obligatory (when the discrepancy in *PC*s is great enough relative to the discrepancy in stage-neutral values, say). If so, then you might have to, contrary to my proposal, make bonds-of-concern as modifiers of both requiring and permitting weights (or of requiring weights only). As I said earlier, despite my personal inclination, I don't have a real argument for favoring Dorsey's rather than Gert's proposal. In general, there are various points of reasonable disagreement regarding the exact shape of justified partiality and its latitude. What I have offered is just a toy model with features that are satisfying to me – and hopefully many others. The reader is welcome to alter the model in case of disagreement.

(A quick sidenote: Here I've been taking reasons to be "monadic" and non-comparative facts which count in favor of alternatives which are also "monadic". For example, the fact that *A* is a pleasurable state-of-affairs is taken as a reason for *A*. This may be best seen as a model for weighing

[1] There is room for minor disagreement. In this toy model, on the assumption that *A* turns out permissible when its permitting reasons are *at least as weighty as* the requiring reasons for *B*, it follows that in cases of equal payoff, it's permissible to favour the stranger (and also permissible to favour the intimate). But you might think that partiality is obligatory here. I do not have clear opinion on this. But in case you disagree, you may instead make it that the permitting reasons for *A* have to be *weightier than* the requiring reasons for *B* in order for *A* to be permissible. This is a technical detail concerning how weights relate to deontic statutes. But if you make this change, the original case turns into one where partiality is impermissible. This is not unwelcome; I intend it to be an edge case.

This may make you worry about the limit of insensitivity to mild sweetening. Repeated mild sweetening, as it seems, topples deontic statuses at some arbitrary point. This is unappealingly sorites-like scenario. I don't have much to say here except for that borderline cases, if a problem, are a problem for virtually everyone.

reasons for *actions*. And it is structurally different from how I regiment reasons for preferences in §2.2. There I took reasons to consist in *differences* between alternatives which count in favor or against *preferences* (as comparative attitudes ranging over pairs of alternatives). For example, the fact that *A* contains more pleasure than *B* is a reason for preferring *A* over *B*. But as I see it, these different regimentations are not making inconsistent commitments on the nature of reasons – e.g., whether they are fundamentally monadic or contrastive. I take them to be inter-translatable schemas that only disagree on a surface level. Here I adopt the monadic formulation because this is how the dual-scale is standardly presented and it makes illustration convenient. And insofar as there are only two exhaustive and mutually exclusive alternatives, it can be easily read as a model for weighing reasons for preferences.)

The dual-scale account reflects the intuition that latitude in prudential partiality has much to do with a deep conflict between a stage-neutral perspective and a stage-relative perspective (which satisficing cannot reflect). On my model, there is latitude because although the stage-neutral perspective issues a *prima facie* demand for maximizing, it can be countered by the permitting (or "excusing") force of the stage-relative perspective. There are other alternative proposals that also reflect this intuition of conflicting perspectives. But before comparing these proposals, I need to clarify that the dual-weight account itself doesn't say anything about values. If I must take a stand, I'd say that all values are *stage-neutral* (*contra* stage-relative consequentialism). Although there is no stage-relative value, it's a *stage-relative* matter how one should *respond to* stage-neutral values; hence stage-relative reasons with modified strengths. I find this account attractive because some view agent-relative values as weird creatures (Schroeder 2007). On the other hand, it's a familiar idea that the strengths of reasons need not directly map onto the values promoted. That's what modifiers are made for: The fact that I am the only person around doesn't make more valuable the state-of-affairs that the drowning child is saved, but it strengthens my reason to save the child.

This is not to say that one cannot combine the dual-weight account with stage-relative values. One similar account of *moral* latitude has been made by Douglas Portmore (2011). The very rough idea is to posit a dual-ranking of values which directly determines the weights of reasons on the dual-scale. In our case, this would mean positing stage-relative values *alongside* stage-neutral values. Although I find such a surplus of values unappealing, one with a different palate may pursue this route. One may even want to kick off the dual-scale of reasons and let an isomorphic dual-ranking of values *directly* fix deontic statuses (Portmore 2008), though my suspicion is that such an account would be covertly appealing to the dual-scale after all. There are other revisionary accounts of

values that appear to accommodate substantive latitude. Perhaps stage-neutral and stage-relative values are partially *incomparable*.[1] (Though Norcross (2020) argues that partial incomparability is incomprehensible.) Or perhaps stage-neutral and stage-relative values, though comparable, sometimes stand in a non-standard comparison relation – namely, being *on a par* (Chang 2004).[2] These two accounts likewise reflect the intuition that latitude resides in a conflict between the two perspectives that's subject to no easy resolution. But both accounts make controversial claims about values, and it appears to me that both accounts still need something like the dual-weight account to serve as a link between values and deontic statuses.

Finally, back to *the Argument from Wellbeing*. Stage-relative consequentialism responds to the argument by making facts about wellbeing perspectival and shifty. In contrast, the dual-weight account – when coupled with a stage-neutral axiology that I favor – denies *No Tragedian*. As the permitting weights for partiality are modified by bonds-of-concern and can sometimes outweigh the countervailing requiring weights for the greater overall good, it's sometimes permissible to be partial. Given considerations of bonds-of-concern, rejecting *No Tragedian* isn't odd at all. It's a familiar claim in ethics that one does not always have most reason to promote what's impartially best – for reasons including (but not restricted to) bonds-of-concern.

5.7. Lingering Issues: Future-Bias and Other-Directed Partiality

Given the above defense of prudential partiality in terms of intra-personal bonds-of-concern, it's easy to see that prudential partiality has a roughly near-biased shape. This is because the components of bonds-of-concern – viz. psychological intimacy, emotional intimacy, and unification by projects – all tend to diminish over time. Of course, on this account, what really matters is not temporal distance but bonds-of-concern. But given the rough correlation, it's not *too* revisionary an account. If you have the intuition that near-bias is often permissible, the intuition is by and large vindicated.

But when it comes to future-bias, it doesn't seem that our account of partiality has a remotely desired shape. Not only is our intuition about future-bias entrenched, but it also indicates that

[1] But not completely incomparable on pain of *thorough* latitude.
[2] By definition, parity is different from being equally good, and it does not contain indeterminacy of any sort. I personally find parity difficult to fathom.

justified future-bias is *robust* and *pervasive*. It's robust in the sense that future-bias can outweigh *extremely unequal* payoffs and remain permissible (as in *My Future and Past Operations*). Indeed, our intuition may even appear to favor *absolute* future-bias: Past pain is over and done with, so it doesn't matter *at all*. (Heathwood 2008; cf. Sullivan 2018: 78-82) Justified future-bias also seems pervasive. We do not think that future-bias is justified only in some unusual scenarios. Instead, it should be *always or at least nearly always* justified. Finally, though short of a universal intuition, future-bias strikes some as not merely permissible, but instead *obligatory*. (Parfit 1984; cf. Prior 1959) All these three features are *prima facie* in tension with our account of prudential partiality. I'll focus on the first two – robustness and pervasiveness – for the time being.

In order to vindicate the intuition that future-bias is robust and pervasive, there has to be a *past/future-asymmetry* in prudential closeness that's pervasive and robust. That is, we are expecting *forward-looking* bonds-of-concern to, routinely, greatly outweigh *backward-looking* bonds-of-concern. Otherwise – if such past/future-asymmetry is only mild and rare – then it doesn't vindicate future-bias as we know it. But as we'll see shortly, pervasive and robust past/future-asymmetry by no means naturally falls out of how we naturally think of bonds-of-concern. This is a particularly weird and unfortunate situation: Lots of writers who appeal to bonds-of-concern seem to talk as if once we've listed the bonds-of-concern, we immediately see a unified vindication of near-bias and future-bias. But much more needs to be said about future-bias.

To better appreciate pervasive and robust past/future-asymmetry in prudential closeness, let us sharpen our terminologies by explicitly making *PC* a non-symmetric relation. Let $PC(P_1, P_2)$ represent how prudentially close P_2 is *relative to* P_1 and $PC(P_2, P_1)$ represent how prudentially close P_1 is *relative to* P_2. *PC* being non-symmetric means that for any such pair of fellow stages, $PC(P_1, P_2)$ does not have to equal $PC(P_2, P_1)$. Whereas $PC(P_1, P_2)$ determines the stage-relative reasons P_1 has concerning P_2, $PC(P_2, P_1)$ determines the stage-relative reasons P_2 has concerning P_1. And to say that prudential closeness is pervasively and robustly past/future-asymmetric (in the direction that coheres with future-bias) is to say that for all pairs of fellow person-stages P_1 and P_2 (in normal life), if P_1 is past relative to P_2 (in personal-time), then it's *at least very often* the case that $PC(P_1, P_2) > PC(P_2, P_1)$ *to a significant degree*.[1]

[1] While there is no formal obstacle to construing *PC* itself as non-symmetric, things may get trickier if one takes *PC* to be (extensionally) equivalent with personal identity. (Sider 2018, Braddon-Mitchell & Miller 2020, Lewis 1983)

It should be conceded it's not difficult to imagine, or even find in real life, some instances of robust past/future-asymmetries. Psychological intimacy, emotional intimacy, and unification by projects – or the components thereof – can obtain only in one direction or obtain more extensively in one direction than in the other. Take emotional intimacy for instance. Given that P_1 is emotionally very empathic and vulnerable towards P_2, it's a still further matter if such empathy and vulnerability is requited. And as an empirical matter, anticipation tends to be emotionally more intense than retrospection.[1] Or consider memory and intention. It's not an unreasonable suggestion that memory establishes only backward-looking bonds-of-concern but intention establishes only forward-looking bonds-of-concern. That is, the fact that P_2 remembers the experiences of some past stage P_1 only adds to $PC(P_2, P_1)$ but not $PC(P_1, P_2)$. On the other hand, the fact that P_1 (reasonably) intends some future stage P_2 to do certain things adds to $PC(P_1, P_2)$ but not $PC(P_2, P_1)$ – although, presumably, $PC(P_2, P_1)$ would be strengthened if P_2 picks up and carries out the intention. Also, when memory loss is severe, other components of psychological intimacy will also be badly compromised. Without memory serving as a background, one cannot "inherit" her past mental states and form backward-looking psychological connections with ease.

But all this doesn't suffice for pervasive and robust past/future asymmetry.[2] For one, anticipation/retrospection is just one component of emotional intimacy among many, and emotional intimacy presumably doesn't exhaust bonds-of-concern. (cf. Karhu 2023) While it might be common for one to be intensely distressed when anticipating future pain but fairly nonchalant when retrospecting on past pain, it's definitely *not* common for one to completely lack backward-looking emotions. Indeed, it seems that some emotions very important to bonds-of-concern are *predominantly (or exclusively) backward-looking* rather than forward-looking – e.g., pride, self-endorsement, regret, and shame. Moving beyond emotional intimacy, it's far from obvious, to say the least, that psychological intimacy and unification by projects tend to obtain predominantly forward-lookingly. Severe memory loss is rare. In general, in ordinary scenarios, it seems that psychological intimacy and unification by projects tend to be roughly symmetric – or worse, asymmetric in the other direction. Speaking of myself, I do not put more trust in my future mental states more than my past mental states (I don't even anticipate my future desires and beliefs much), and I frequently honor sunk costs without willingly cooperating with my future plans as much.

[1] Van Boven & Ashworth (2007), Harris (2010), Hardisty & Weber (2020).
[2] Sider (2018) discusses an interesting (and far-fetched) case of extreme asymmetry. And that is a case of *backward-looking* bonds-of-concern dominating forward-looking ones.

Although I don't have a precise calculus of *PC*, I am inclined to think that in ordinary life, backward-looking bonds-of-concern tend to (slightly) *outweigh* forward-looking ones. This is indicated by the temporal structures of our *self-narratives*. (Sider 2018: 131-32, Schechtman 2001) Put crudely, in telling our life stories, we identify with our past selves more easily than with our future selves. This point is mostly obviously seen when beliefs, desires, and values change drastically. Consider again the Russian nobleman. In his later years, although he no longer has socialist ideals and is content with the present decadent lifestyle, it's still fairly sensible for him to reminisce about his youthful passions with joyful nostalgia. On the other hand, it strains psychological reality for the young nobleman to *not* anticipate his later "degeneration" with horror and antagonism. In general, from a narrative perspective, it's easy and common for later stages to view earlier experiences (even regrettable ones) as "formative" and an essential part of the narrative "leading me here". But it's in comparison much more difficult for me, as of now, to incorporate my (anticipated) future selves into a coherent self-narrative; very often, they feel like strangers to me. Although I do not have a detailed story about the psychological underpinnings of self-narratives, it's intuitive to think of self-narratives as a "master" bond-of-concern that incorporates, depends upon, and is indicative of, all the component bonds-of-concern we've canvassed. In all, pervasive and robust past/future asymmetry *in the correct direction* does not seem something we can reasonably expect.[1]

This is a surprising result. Our pro-future-bias intuition is certainly more entrenched and robust than our pro-near-bias intuition. If *even* near-bias can be largely vindicated, we should certainly expect future-bias to be vindicated. However, the picture of prudential partiality we've so far outlined, though cohering with near-bias, is in tension with future-bias.

I'm genuinely unsure what to make of this. On the one hand, one might just accept the counterintuitive upshot that future-bias is only occasionally permitted – and that permissible future-bias tends to be weak rather than robust. Following the argument where it leads, one might endorse this revisionary picture of prudential partiality. On the other hand, I can also understand the impulse to resist. But to resist the revisionary upshot, it has to be argued that the past/future-asymmetry is robust and pervasive after all. And this means that the above description of bonds-

[1] For sure, one might suggest that (say) anticipatory/retrospective emotions of pain and pleasure are all that matters, or that they weigh extremely heavily in the overall calculus of prudential closeness. But I hope that the reader will agree, that's a desperate claim.

of-concern is incomplete, to say the least. Is there a feasible way, as it were, to insert a past/future-asymmetry into bonds-of-concern?

Here I can only make two tentative suggestions, but neither of them is manifestly promising. One natural thought is that the direction of *causation* matters. Past mental states cause present mental states, but future mental states do not cause present mental states (speaking in personal-time). This asymmetry in causal direction is indeed pervasive and robust (better, omnipresent and absolute). It may then be suggested that psychological intimacy is by its nature *one-directional* and *only forward-looking*, in line with the fact that causation propagates only forward-lookingly. But I don't find this a promising solution. First, I don't find it intuitive at all that the direction of bonds-of-concern should align with that of causation in this way. Say that my best friend has enormous causal influence on my mental states. As I like her and trust her a lot, I come to desire and believe what she desires and believes. It sounds implausible that by my becoming like-minded as her, I am *not at all* making her more of an intimate of mine, despite that she *is* thereby making me more of an intimate of hers. Second, although later mental states cannot directly cause earlier mental states, my *anticipation* of later states is causally effective. This is just the previous point that although distinct person-stages do not temporally overlap, they can nonetheless stand in mutual relationships. But if "backward causation" mediated by anticipation should count as backward-looking bonds-of-concern, the alleged asymmetry would be to a great extent neutralized. Finally, even if certain components of bonds-of-concern can be understood as exclusively forward-looking due to the arrow of causation, this might well leave us plenty of other components that tend to be mutual or predominantly backward-looking. Retrospective identification weaved by self-narratives is a salient example, where causation doesn't seem to be what matters.

Then there is an even more speculative suggestion which exploits causation in a different manner. It points to a causal asymmetry in the very *existence* of person-stages. My present stage is *causally responsible* for the existence of my future stages but not for the existence of past stages. Might this asymmetry be normatively significant? Well, causal responsibility seems intimately connected with moral responsibility (or normative responsibility in general). So, one might explore the idea that by virtue of "creating" the future stages, the present stage somehow incurs some very weighty stage-relative reasons or even *prima facie duties* to take care of them. Such a suggestion has precedents in the moral domain. Many think that parents have *special duties* towards their children. And parental duties, at least for biological parents, seem to be (in part) grounded in the fact that parents are causally responsible for the very existence of their children. As Lionel McPherson

remarks, "[P]arents have basic substantive reasons to give priority to their children's interests, reasons tied almost inextricably to… [that] parents have brought into the world children who need care." (2002: 33) This is not to say that there aren't filial duties. But arguably children are not "born with" filial duties; they don't owe their parents much if they have been maltreated. And this asymmetry in parental/filial duties seems to have much to do with an asymmetry in causing-to-exist.

There is something nice about accounting for future-bias in terms of an asymmetry in causing-to-exist. It allows one to go beyond mere reasons and appeal to special duties. When generalizing moral partiality to prudential partiality, I focused on *permissible* (and occasionally obligatory) partiality. But some think that special relationships also give rise to *special duties* – e.g., parental and filial duties – and that duties possess very *stringent deontic force* that is not reducible to, and cannot be (easily) outweighed by, mere aggregations of reasons. You may feel uneasy about positing duties alongside reasons (cf. Kagan 2019). But if this deontic mixology can be made to work, and if the asymmetry in causing-to-exist gives rise to *an asymmetry in duty*, then it would be unsurprising that the asymmetry in partiality is *robust and pervasive*. Indeed, it would vindicate the intuition that future-bias is *obligatory* – though "obligation" is hereby explicated in terms of duties instead of mere very weighty (requiring) reasons.

However, the analogy with parental duty is contentious, to say the least. It is not at all obvious that causing a *person-stage* to exist should carry the same normative significance as causing a *person* to exist. The term "causing-to-exist" is seductively indifferent to what intuitively matters for duty. Consider this asymmetry: If I push you under the bus and cause you impairment, I may have a special duty to take care of your impairment– but not if I am just a bystander not having come to your rescue. Not getting into to the vexed details of how exactly to spell out the relevant distinction here (though presumably one along the lines of doing *versus* allowing), I find it natural to say that causing-to-exist a *person* is more like pushing you under a bus, but causing-to-exist a future *person-stage* is more like standing by idly. Crudely put, causing-to-exist person-stages just is *persistence*: It's not really *doing* anything; it's just inaction so as to maintain the *status quo* of continuing to exist. Persistence, I'd say, is the *status quo* both *causally* and *normatively* (the alternative being suicide), at least for a life that's worth living. It thus strains credibility to say that

one takes on a special duty by persisting in the same way as parents who take on a special duty by bringing persons into existence.[1]

The above objection is as tentative as the proposal itself. Although I do not want to proclaim that the proposal is hopeless, it had better do more than just a dubious analogy. More generally, I don't take myself to have proved that we cannot find pervasive and robust past/future-asymmetry in prudential closeness. I've merely taken some initial steps to illuminate a gaping lacuna in the literature: Philosophers who defend prudential partiality in terms of bonds-of-concern, for some reason, have not bothered to indicate what would be left with future-bias – which happens to be the most conspicuous form of prudential partiality which has largely motivated the whole enterprise in the first place. Maybe those philosophers find pervasive and robust asymmetry to be obvious. The proceeding discussions have shown that it's *far from obvious*.

Finally, I (re)turn to third-person time-biases: If self-directed prudential partiality is at least sometimes permissible, what should be said about other-directed partiality which discriminates among the person-stages of other people? I have suggested in §2.4.1 that it's unclear if there is a *de facto* first-/third-person asymmetry in future-bias; nor is it clear if there is a corresponding asymmetry in normative status. The present account of prudential partiality casts a new light on third-person time-biases. Let us suppose that Parfit's *My Past and Future Operations* is a case in which future-bias is permissible (setting aside the above worry about robust and pervasive asymmetry). Then focus on this question: What should you prefer about Parfit?

Todd Karhu (2023: §4) suggests that you ought to be *time-neutral* (or impartial). You ought to prefer the upcoming shorter operation for Parfit, according to Karhu, because Parfit's past and future person-stages are *equally strangers* to your *present person-stage*. This sounds plausible. But what if Parfit is your best friend? A natural extension of Karhu's suggestion would be saying that third-person time-biases should depend on *stage-relativized moral closeness*. To make things concrete, say that the past and future person-stages of Parfit's in question are respectively P_P and P_F and that your present person-stage is $P^\star$. What you shall permissibly prefer would then hinge on how morally close P_P and P_F are relative to $P^\star$ respectively – that is, $MC(P^\star, P_P)$ and $MC(P^\star, P_F)$.

[1] One disanalogy may be telling: Many think that there is very weighty moral reason *against* creating miserable people but little or no reason *for* creating happy people. But hopefully no one thinks that although there is *very weighty* prudential reason against continuing a miserable life but *little or no* reason for continuing a happy life.

The underlying idea just is that *moral partiality* hinges on moral closeness. But here we are concerned about a more fine-grained sort of moral partiality that discriminates among different person-stages of a single individual. Thus, moral closeness is accordingly relativized to person-stages.

But it remains obscure how stage-relativized moral closeness is to be measured. One suggestion is that between intimates, stage-relativized moral closeness *mirrors* prudential closeness of the *target*. Say that Parfit's present person-stage is P and that $PC(P, P_P) < PC(P, P_F)$ (otherwise his own future-bias wouldn't be permissible). Given that you and Parfit are intimates, the present suggestion delivers that $MC(P^*, P_P) < MC(P^*, P_F)$. Is there any good motivation for mirroring? I have a vague idea in mind: Prudential closeness and moral closeness are commensurable and inter-relatable. In particular, one might think that how P^* relates to P_P and P_F is *mediated by* the immediate intimate relationship between P^* and P (which are co-present). *Via* the co-present intimate person-stages, the intra-personal prudential closeness between Parfit's person-stages is transferred and transformed, as it were, into the stage-relativized moral closeness between different individuals. In a nutshell, relative to P^*, P_F is morally closer than P_P because P^* is a close intimate of *P, and* relative to P, P_F is prudentially closer than P_P. Notice that thinking this way, mirroring *wouldn't* hold if Parfit is a complete stranger to you. A close intimate of a stranger (relative to you) is not an intimate of yours of any sort. This is consistent with Karhu's original suggestion.

Perhaps there are other plausible ways to determine stage-relativized moral closeness, but I wish to highlight a potential attraction of the present proposal. It coheres nicely with the hypothesis that *empathetic engagement* facilitates future-bias. (§2.4.1) The present proposal in effect says that whether and to which extent it's reasonable for a third party to defer to the target's own (reasonable) time-biases depend on moral closeness. But empathetic engagement is *associated with* moral closeness. As a matter of fact, you're more prone to empathize with your intimates than with strangers. It's also a plausible claim that you *ought to* empathize with your intimates but take a more detached stance towards strangers. The present proposal thereby pieces together empathetic engagement, moral closeness, and the variability of third-person time-biases.

That said, I also see some plausible thoughts *inconsistent with* the above proposal. For one, one might think our preferences should *universally* defer to those of the targets' (insofar as their own preferences are permissible). Suppose that Parfit is a perfect stranger in a situation in which it would be permissible but not obligatory for him to be near-biased. And he is in fact near-biased.

Unfortunately, for some reason he cannot act to make inter-temporal trade-offs for himself; you need to choose on behalf of him. Karhu's proposal implies time-neutrality on your part – for Parfit's near- and far-future stages are equally strangers to you. But this sounds weird. Surely you ought not to contradict what Parfit himself permissibility prefers.[1] For another, perhaps determining stage-relativized moral closeness is a more complicated matter. Suppose again that Parfit has been your best friend and that $PC(P, P_P) < PC(P, P_F)$ (P being Parfit's present stage). Parfit is permitted to be future-biased for himself. But consider this sad situation: *A few seconds from now*, Parfit will do something diabolical and friendship-undermining to you. Here it seems apt to say that although P (his present stage) is an intimate to P^* (your present stage) – and likewise for the past co-present stages of you two – the future stages of you two are *no longer* friends. In this case, it doesn't seem reasonable for you to care more about P_F (his future stage you'll no longer befriend) than P_P (his past stage you befriended). If anything, it should be the other way around. Mirroring thus seems too simplistic. It anchors stage-relativized moral closeness solely on the present, as it were, but fails to consider the cross-temporal dynamics of special relationships.

A recurring theme of the proceeding discussions is how *prudential closeness* relates to and compares with *moral closeness*. This is an issue with far-reaching practical implications. As Leora Urim Sung (2024) observes, conflicts between self-interest and benevolence are frequently complicated by temporal factors. For instance, charity giving sometimes involves little cost to your present self-interest but may have considerable impact on your future wellbeing. (You could have invested the money and made a huge fortune.) So, there is the need to weigh the prudential closeness to one's future stages against the moral closeness towards the beneficiaries. Or consider a conflict between a significant sacrifice by your distant-future self and a mediocre amount of benefit to your best friend in the near future. What ought you to do? How much moral latitude do you have? The answer would depend on, among other things, the "exchange rate" between moral closeness and prudential closeness as well as how much permitting weight they can confer respectively. Here I can only raise these questions without answering.

5.8. Epilogue

[1] But this counterexample might be resisted by distancing what one ought *to do* from what one ought *to prefer*. You ought to defer to his own preference in *your choice* on pain of violating his *autonomy*, you may think, despite that you ought to be time-neutral in *your preference*.

We have come a long way to reveal the nature and rational status of time-biases. Just to briefly recap: In Chapter 1, I propose for a conception of time-biases as a kind of preferences in the sense of evaluative mental states. This mentalist rather than behaviourist conception enables us to go about applying more robust forms of rational evaluation to time-biases. I also disentangle future-bias from two similar past/future-asymmetric attitudes, namely, temporal relief and the temporal valuation asymmetry. These phenomena, though undoubtedly closely related, should not be taken for granted to share the very same underlying mechanisms or the same rational status. Finally, I rebut a recent argument that time-biases are not rationally evaluable as mental states because the temporal features of the alternatives cannot be adequately represented.

Chapter 2 focuses on criticisms of time-biases from the perspective of structural rationality. Despite the almost truism that non-exponential near-bias and future-bias (typically) lead to preference reversal and are, for this reason, irrational, the literature has neither made clear what preference reversal means, nor provided good arguments that preference reversal is irrational. In light of the lack of a well-motivated general constraint on preference change that doesn't prove too little or too much, I argue, it's ill-advised to criticize time-biases on structural grounds. Instead, it would be more fruitful to look at individual time-biases, taken in isolation, and see if they are well-grounded by reasons.

Chapter 3 investigates time-biases in relation to the metaphysics of time. Focusing on future-bias, conceived as preferences that are sensitive to the past/future-asymmetry in metaphysical-time, I argue that we lack good reason to regard future-bias as correct (or objectively rational). Although there might be some metaphysically deep asymmetries between the past and the future, it remains mysterious how these asymmetries could provide normative reasons for future-bias. But this is somewhat expected, for our intuition is clear that time-biases should track not metaphysical-time, but instead (roughly) what we may call personal-time. This means that we shouldn't seek reasons for time-biases in the metaphysics of time to begin with.

Chapter 4 takes on the personal-time conception and investigates how time-biases relate to wellbeing. I've advanced two important arguments. The first is that the tension between time-neutralism and subjective theories of wellbeing in the context of changing preferences is merely apparent. The apparent tension can be expelled by adopting a well-motivated, non-perspectival constraint on the welfare-relevance of preferences (along the lines of concurrentism), and is therefore not a reason for subjective theorists to reject time-neutralism. However, subjective

theorists need not endorse time-neutralism either. They have a distinctive response to the Argument from Wellbeing, an argument against time-biases that rests on a non-perspectival conception of lifetime wellbeing. Subjectivist theorists can reasonably reject this non-perspectival conception.

Chapter 5 turns to how objectivists of wellbeing may respond to the Argument from Wellbeing. The account I favor parallels a popular approach to accommodate partiality in special relationships in the moral domain. I argue that the very same bonds-of-concern which ground agent-relative reasons for moral partiality also obtain intra-personally, and therefore ground stage-relative reasons for prudential partiality. By *not* positing corresponding stage-relative *values*, I deny that to be prudent is to maximize lifetime wellbeing, just as one may deny that to be moral is always to do what's impersonally optimal. I also provide a dual-weight account for accommodating substantive latitude in prudential partiality. Finally, in relation to the issue of structural rationality in Chapter 2, I should add that prudential partiality justified this way does not take an "exponential" form. There is no reason to believe that the dynamics of bonds-of-concern among person-stages in real life should map onto the discounting functions dictated by the structural constraint against preference reversal. This is expected and should not be taken to indicate something wrong with the account of prudential partiality defended here. An inter-personal analogy is telling: Since different individuals may participate in distinct special relationships, there is no reason to expect their justified partiality to converge without conflicts. And there shouldn't be a "structural constraint" dictating that there is no "cross-personal preference reversal".
Indeed, thinking in terms of bonds-of-concern, there is more reason to reject the idea that preference reversal is a "cardinal sin".

The whole thesis, though perhaps a bit of a patchwork, hopefully will advance the debate over time-biases in important respects. As I said in the beginning, I do not aspire to offer a simple and conclusive answer to the question "Are time-biases rational?". Lots of conclusions I draw are very tentative. Given the multiplicity of conceptions of what time-biases are and ways in which one may approach their rationality, I take the primary goal of the thesis to be more on the side of preliminary groundwork.

Picking up some loose ends in the previous chapters, I wish to sketch some avenues for future exploration. A somewhat striking upshot of Chapter 5 is that prudential partiality grounded in intra-personal bonds-of-concern doesn't appear to have a future-biased shape because there

doesn't seem to be robust and pervasive past/future-asymmetry in bonds-of-concern. I'm unsure if this is indeed true. One might argue that, despite appearances, the past/future-asymmetry is indeed pervasively robust. Although the two proposals I have considered don't turn out promising, there may be alternative routes to explore. But if it's instead acknowledged that future-bias is rarely justified, a mystery arises regarding our robust intuition that future-bias is rational. We should be curious about why future-bias appears to be rational when it's not.

One obvious response is to appeal to the affective asymmetry as an evolved heuristic, and suggest that it explains not only our future-bias, but also our pro-future-bias intuition. If it's generally advantageous for us to anticipate future pain with more intense emotions (and thereby be prompted to form future-biases), then it should be no wonder that we cannot shake the intuition that future-bias is rational. That is, our pro-time-bias intuition should be part and parcel of the evolutionary advantage. But I suspect that the affective asymmetry is not the whole story. The main reason is that the affective asymmetry, if explanatory of our future-bias and pro-future-bias intuition, should presumably be equally explanatory of our near-bias and pro-near-bias intuition. After all, it's also true that events in the near future tend to be more certain and controllable, so it's evolutionarily advantageous for us to attend to them with greater affective response and thereby have the tendency to be near-biased. But our pro-near-bias intuition is far less robust than our pro-future-bias intuition, if existent at all. At least, while we are easily convinced that near-bias is irrational after all, our pro-future-bias intuition tends to remain entrenched under rational reflection. (Latham, Li, Miller and Yu manuscript find that ordinary folks generally agree with quasi-philosophical arguments for future-bias.) So, the affective asymmetry by itself cannot explain why it's our pro-*future-bias* intuition that's *particularly* entrenched. In other words, a full "debunking" explanation of future-bias must go beyond merely pointing to this evolved heuristic. This also indicates that more (empirical) work needs to be done to uncover the underlying psychological mechanisms of future-bias. For all we know, a full answer might be very complicated.

The second venue for future exploration concerns how time-biases relate to *happiness*. By happiness – sometimes also called "subjective wellbeing" – I mean a kind of mental state or a constellation of different but related kinds of mental states. (Proposed candidates include hedonic pleasure, satisfaction with one's life[1], certain positive emotions, and some combinations thereof).

[1] Satisfaction with one's life is to be distinguished from satisfaction of preferences or desires (as figured in subjective theories of wellbeing). The former is a kind of mental state – presumably a kind of cognitively loaded feeling.

Presumably, happiness is something we deeply care about and consider valuable.[1] So, it would be nice to see if there is some correlation between time-biases and happiness, which may in turn indicate the existence of some causal relation. In particular, you might think that if time-biases promote (or frustrate) happiness, then this is some *state-given* reason – rather than object-given reason – for (or against) being time-biased. It's true that a state-based reason for being future-biased would not make future-bias the *correct* attitude. But if being future-biased makes one happy, then pragmatics considerations may count in favor of future-bias. (cf. §1.4) That is, it might be that future-bias, though incorrect, is not something we should seek to eradicate for pragmatic reasons. Caruso, Latham, Miller and Yu (manuscript) have taken an initial step in investigating the correlation between time-biases and happiness, using four popular scales for measuring happiness. The result is a bit disappointing. We found no robust correlation between time-biases and the four measurements, either positive or negative. But this shouldn't be taken as the final word. Perhaps the correlation didn't manifest itself because happiness is to a great extent influenced by other factors. Or perhaps whether time-biases promote or frustrate happiness is a very contingent matter – contingent upon age, socioeconomic environments, and so on. Future researches can disentangle these factors.

Finally, given the account of prudential partiality according to which prudential concern should track intra-personal bonds-of-concern (psychological intimacy, emotional intimacy, unification by projects, and perhaps more), it might be wondered whether other sorts of "identity-involving" attitudes and practices should also track these bonds-of-concern. I have in mind attitudes and practices such as attribution of moral responsibility, blame and praise, feelings of guilt, legal punishment, etc.[2] They are "identity-involving" in the sense that, like prudential concern, they are often presumed to track precisely personal identity. Take moral responsibility for example. It's commonly assumed that once one is morally responsible for some act at the time of the act, then she continues to be responsible for the rest of her life, and the level of responsibility doesn't wax or wane over time. However, if you believe prudential concern shouldn't track personal identity, then you probably should think the same about moral responsibility. This doesn't mean that all the candidate bonds-of-concern are relevant to moral responsibility. For instance, it's plausible that when one's (morally relevant) values have changed drastically, moral responsibility may diminish;

[1] How exactly happiness relates to wellbeing (in the sense used in Chapter 5) is a vexed matter. I'm inclined to think that happiness is somewhat correlated and indicative of wellbeing. But at any rate, happiness is not presumed to be of fundamental prudential value, though some might argue that it is (say, if one is a hedonist of wellbeing and also a hedonist of happiness).

[2] See also Khoury & Matheson (2018) and Carlsson (2022).

but it's far less plausible that mere emotional detachment (not underwritten by value change) should lead to diminished moral responsibility. In general, once we get into the mindset that those apparently identity-involving practices and attitudes shouldn't track identity after all, lots of work needs to be done to uncover what exactly they should track.

References

Adler, Matthew (2011). *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford: Oxford University Press.

Ahmed, Arif (2018). Rationality and Future Discounting. *Topoi* 39 (2):245-256.

Ainslie, George. (2001). *Breakdown of Will*. Cambridge: Cambridge University Press.

Andreou, Chrisoula (2007). There Are Preferences and Then There Are Preferences. In Barbara Montero and Mark D. White (ed.), *Economics and the Mind*.

Anscombe, G. E. M. (1957). *Intention.* Harvard University Press.

Bacharach, Julian (2022). Relief, Time-bias, and the Metaphysics of Tense. *Synthese* 200 (3):1-22.

Bader, Ralf. (2016). Conditions, Modifiers, and Holism. In *Weighing Reasons*. Oxford University Press.

Baier, Annette. (1991). Unsafe Loves. In Solomon, R. C. & Higgins, K. M. (eds.), *The Philosophy of (Erotic) Love*. Kansas University Press: 433–50.

Baker, Derek Clayton (2010). Ambivalent Desires and the Problem with Reduction. *Philosophical Studies 150* (1):37-47.

Barnes, Elizabeth & Cameron, Ross. (2008). The Open Future: Bivalence, Determinism and Ontology. *Philosophical Studies* 146 (2):291-309.

Bernstein, Sara (2022). Paradoxes of Time Travel to the Future. In Helen Beebee & A. R. J. Fisher (eds.), *Perspectives on the Philosophy of David K. Lewis*. Oxford, UK: Oxford University Press.

Binmore, Ken (1994). *Playing Fair.* Cambridge, MA: MIT Press.

Black, Sam (2001). Altruism and the Separateness of Persons. *Social Theory and Practice* 27 (3):361-385.

Bnefsi, Sayid (2023). Future Bias and Regret. In David Jakobsen, Peter Øhrstrøm & Per Hasle (eds.), *Logic and Philosophy of Time: The History and Philosophy of Tense-Logic*. Aalborg: Aalborg University Press: 1-13.

Boonin, David (2019). *Dead Wrong: The Ethics of Posthumous Harm*. New York, NY: Oxford University Press.

Braddon-Mitchell, David & Miller, Kristie (2020). Surviving, to Some Degree. *Philosophical Studies* 177 (12):3805-3831.

Braddon-Mitchell, David; Latham, Andrew J. & Miller, Kristie (2023). Can We Turn People Into Pain Pumps?: On the Rationality of Future Bias and Strong Risk Aversion. *Journal of Moral Philosophy* 1:1-32.

Bradley, Ben (2006). Against Satisficing Consequentialism. *Utilitas* 18 (2):97-108.

Bradley, Ben (2009). *Well-being and Death*. New York: Oxford University Press.

Bradley, Ben (2016) Well-Being at a Time. *Philosophic Exchange* 45 (1).

Bramble, Ben. (2016). A New Defense of Hedonism about Well-Being. *Ergo, an Open Access Journal of Philosophy* 3. 10.3998/ergo.12405314.0003.004.

Bramble, Ben (2017). *The Passing of Temporal Well-Being*. New York, NY: Routledge.

Brandt, Josh. (2020). Negative Partiality. *Journal of Moral Philosophy* 17 (1):33-55.

Brandt, Richard B. (1979). *A Theory of the Good and the Right*. Amherst, N.Y.: Prometheus Books.

Brandt, Richard B. (1992). *Morality, Utilitarianism, and Rights*. New York: Cambridge University Press.

Bricker, Phillip (1980). Prudence. *Journal of Philosophy* 77 (7):381-401.

Brink, David O. (2011). Prospects for Temporal Neutrality. In C. Callender (ed.), *The Oxford Handbook of Philosophy of Time*. Oxford University Press: 353–81.

Brink, David O. (1997). Rational Egoism and the Separateness of Persons. In J. Dancy (ed.), *Reading Parfit*. Oxford: Blackwell.

Brink, David O. (2003). Prudence and Authenticity. *Philosophical Review* 112 (2):215-245.

Brink, David O. (2011). Prospects for Temporal Neutrality. In C. Callender (ed.), *The Oxford Handbook of Philosophy of Time*. Oxford University Press: 353–81.

Bronsteen, John (2017). Preferring to Decrease One's Own Well-Being. *Utilitas* 29 (1):52-64.

Broome, John (1991). *Weighing Goods*. Hoboken, NJ: Wiley-Blackwell.

Broome, John (1994). Discounting the Future. *Philosophy and Public Affairs* 23 (2):128-156.
Broome, John (2006). Reasoning with preferences? *Royal Institute of Philosophy Supplement 59*:183-208.
Broome, John (ed.) (2013). *Rationality Through Reasoning*. Malden, MA: Wiley-Blackwell.
Bruckner, Donald W. (2013). Present Desire Satisfaction and Past Well-Being. *Australasian Journal of Philosophy* 91 (1):15 -29.
Bruckner, Donald W. (2018). The Shape of a Life and Desire Satisfaction. *Pacific Philosophical Quarterly* 100 (2):661-680.
Brueckner, Anthony & Fischer, Martin (1986). Why is Death Bad? *Philosophical Studies* 50 (2): 213–221.
Burns, P., McCormack, T., Jaroslawska, A., Fitzpatrick, Á., McGourty, J., & Caruso, E. M. (2019). The Development of Asymmetries in Past and Future Thinking. *Journal of Experimental Psychology: General*, 148(2): 272–288.
Bykvist, Krister (2003). The Moral Relevance of Past Preferences. In Heather Dyke (ed.), *Time and Ethics: Essays at the Intersection*. Kluwer Academic Publishers: 115-136.
Bykvist, Krister (2006). Prudence for Changing Selves. *Utilitas* 18 (3):264-283.
Bykvist, Krister (2007). Comments on Dennis McKerlie's 'Rational choice, Changes in Values Over Time, and Well-being'. *Utilitas* 19 (1):73-77.
Bykvist, Krister (2010). Can Unstable Preferences Provide a Stable Standard of Well-being? *Economics and Philosophy* 26 (1):1-26.
Callender, Craig (2021a). The Normative Standard for Future Discounting. *Australasian Philosophical Review* 5 (3):227-253.
Callender, Craig (2021b). Response to Critics. *Australasian Philosophical Review, 5*(3): 309–321.
Callender, Craig (2022). Is Discounting for Tense Rational? In Hoerl, Christoph; McCormack, Teresa & Fernandes, Alison (eds.) *Temporal Asymmetries in Philosophy and Psychology.* Oxford: Oxford University Press.
Cantwell, John (2003). On The Foundations of Pragmatic Arguments. Journal of Philosophy 100 (8):383-402.
Carlsson, Andreas Brekke (2022). Deserved Guilt and Blameworthiness over Time. In Andreas Carlsson (ed.), *Self-Blame and Moral Responsibility*. New York, USA: Cambridge University Press.
Caruso, Eugene M.; Latham, Andrew J. & Miller, Kristie (2024). Is Future Bias Just a Manifestation of the Temporal Value Asymmetry? *Philosophical Psychology*, DOI: 10.1080/09515089.2024.2360673
Caruso, Eugene M.; Gilbert, Daniel T. & Wilson, Timothy D. (2008). A Wrinkle in Time: Asymmetric Valuation of Past and Future Events. *Psychological Science*, 19(8): 796–801.
Caruso, Eugene M; Latham, Andrew J.; Miller, Kristie & Yu, Wen (manuscript). Do Time-Biases Promote or Frustrate Wellbeing?
Casati, Roberto. & Torrengo, Giuliano (2011). The not So Incredible Shrinking Future. *Analysis* 71 (2):240-244.
Chang, Ruth. (2002). The Possibility of Parity. *Ethics*, *112*(4): 659–688.
Cockburn, D. (1998). Tense and Emotion. In R. Le Poidevin (Ed.), *Questions of Time and Tense*. Oxford: Clarendon Press: 77-91.
Craig, William Lane (2000). *The Tensed Theory of Time: A Critical Examination*. Kluwer Academic.
Dancy, Jonathan (2004). *Ethics Without Principles*. Oxford: Oxford University Press.
Davidson, Donald; McKinsey, J. C. C. & Suppes, Patrick (1955). Outlines of a Formal Theory of Value, I. *Philosophy of Science* 22 (2):140-160.
DeGrazia, David (2005). *Human Identity and Bioethics*. Cambridge: Cambridge University Press.
Dietrich, Franz & List, Christian (2016). Mentalism *versus* Behaviourism in Economics: A Philosophy-of-Science Perspective. *Economics and Philosophy 32* (2):249-281.
Dietz, Alexander (2020). Are My Temporal Parts Agents? *Philosophy and Phenomenological Research* 100 (2):362-379.
Dietz, Alexander (2023). Making Desires Satisfied, Making Satisfied Desires. *Philosophical Studies* 180 (3):979-999.

Dorato, Mauro (2006). The Irrelevance of the Presentist/Eternalist Debate for the Ontology of Minkowski Spacetime. In Dennis Geert Bernardus Johan Dieks (ed.), *The Ontology of Spacetime*. Boston: Elsevier: 93-109.
Dorsey, Dale (2013). Desire-satisfaction and Welfare as Temporal. *Ethical Theory and Moral Practice* 16 (1):151-171.
Dorsey, Dale (2015). The Significance of a Life's Shape. *Ethics* 125 (2):303-330.
Dorsey, Dale (2016). Future-Bias: A (Qualified) Defense. *Pacific Philosophical Quarterly 98* (S1):351-373.
Dorsey, Dale (2018). A Near-Term Bias Reconsidered. *Philosophy and Phenomenological Research* 99 (2):461-477.
Dorsey, Dale (2018). Prudence and Past Selves. *Philosophical Studies* 175 (8):1901-1925.
Dorsey, Dale (2021). *A Theory of Prudence*. Oxford: Oxford University Press.
Dorsey, Dale (2020) Prudence and Self-Concern. In Sauchelli, Andrea (ed.), *Derek Parfit's Reasons and Persons: An Introduction and Critical Inquiry*. Routledge.
Dougherty, Tom (2011). On Whether To Prefer Pain to Pass. *Ethics 121* (3):521-537.
Dougherty, Tom (2015). Future-Bias and Practical Reason. *Philosophers' Imprint* 15.
Dreier, James (1996). Rational Preference: Decision theory as a Theory of Practical Rationality. *Theory and Decision 40* (3):249-276.
Dyke, Heather & Maclaurin, James (2002). 'Thank Goodness That's Over': The Evolutionary Story. *Ratio* 15 (3):276–292.
Fehige, Christoph (1998). A Pareto Principle for Possible People. In Christoph Fehige & Ulla Wessels (eds.), *Preferences*. New York: De Gruyter: 508–543.
Fernandes, Alison (2021). Does the Temporal Asymmetry of Value Support a Tensed Metaphysics? *Synthese 198* (5):3999-4016.
Fernandes, Alison. (2022). Caring for Our Future Selves. In Hoerl, Christoph; McCormack, Teresa & Fernandes, Alison (eds.) *Temporal Asymmetries in Philosophy and Psychology.* Oxford: Oxford University Press.
Forrest, Peter (2004). The Real but Dead Past: A Reply to Braddon-Mitchell. *Analysis* 64 (4):358–362.
Forrester, Paul (2023). Concurrent Awareness Desire Satisfactionism. *Utilitas* 35 (3):198-217.
Frankfurt, Harry G. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy* 68 (1):5-20.
Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, *40*(2): 351–401
Friedrich, Daniel (2017). Desire, Mental Force and Desirous Experience. In Julien Deonna & Federico Lauria (eds.), *The Nature of Desire*. Oxford University Press.
Gallois, André (1994). Asymmetry in Attitudes and the Nature of Time. *Philosophical Studies* 76 (1):51-69.
Gert, Joshua (2007). Normative Strength and the Balance of Reasons. *Philosophical Review* 116 (4):533-562.
Gert, Joshua (2016). The Distinction between Justifying and Requiring: Nothing to Fear. In *Weighing Reasons*. Oxford University Press.
Gilmore, Cody (2016). The Metaphysics of Mortals: Death, Immortality, and Personal Time. *Philosophical Studies 173* (12):3271-3299.
Glasgow, Joshua (2013). The Shape of a Life and the Value of Loss and Gain. *Philosophical Studies* 162 (3):665-682.
Goldman, Alan H. (2006). Desire Based Reasons and Reasons for Desires. *Southern Journal of Philosophy* 44 (3):469-488.
Goodrich, James (2021). Separating Persons. In Jeff McMahan et al. (eds), *Principles and Persons: The Legacy of Derek Parfit*. Oxford: Oxford University Press.
Greene, P., A. Holcombe, A. J. Latham, K. Miller, and J. Norton (2021). The Rationality of Near Bias towards both Future and Past Events. *Review of Philosophy and Psychology* 12: 902–22.

Greene, P., Latham, A. J., Miller, K., and Norton, J. (2021a). Hedonic and Non-hedonic Bias Towards the Future. *Australasian Journal of Philosophy, 99*:1, 148-163. DOI: 10.1080/00048402.2019.1703017
Greene, P., Latham, A. J., Miller, K., and Norton, J. (2021b). On Preferring that Overall, Things are Worse: Future-Bias and Unequal Payoffs. *Philosophy and Phenomenological Research* 105 (1):181-194.
Greene, P., Latham, A. J., Miller, K., and Norton, J. (2022a). How Much Do We Discount Past Pleasures? *American Philosophical Quarterly* 59 (4):367-376.
Greene, P., Latham, A. J., Miller, K., and Norton, J. (2022b). Capacity for Simulation and Mitigation Drives Hedonic and Non-hedonic Time Biases. *Philosophical Psychology* 35 (2):226-252.
Greene, P., Latham, A. J., Miller, K., and Norton, J. (2022c). Why are People so Darn Past Biased? In Christoph Hoerl, Teresa McCormack & Alison Fernandes (eds.), *Temporal Asymmetries in Philosophy and Psychology.* Oxford: Oxford University Press: 139-154.
Greene, Preston (2021). 'Pure' Time Preferences Are Irrelevant to the Debate over Time Bias: A Plea for Zero Time Discounting as the Normative Standard. *Australasian Philosophical Review 5* (3):254-265.
Greene, Preston & Sullivan, Meghan (2015). Against Time Bias. *Ethics 125* (4):947-970.
Greenspan, P. (2005). Asymmetrical Reasons. In M. E. Reicher and J. C. Marek (eds.), *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium*. Vienna: Austrian Wittgenstein Society: 387–94.
Gregory, Alex (2017). Might Desires Be Beliefs About Normative Reasons? In Julien Deonna & Federico Lauria (eds.), *The Nature of Desire*. Oxford University Press: 201-217.
Grüne-Yanoff, T. (2015). Models of Temporal Discounting 1937–2000: An Interdisciplinary Exchange between Economics and Psychology. *Science in Context, 28*(4): 675–713.
Grüne-Yanoff, Till (2021). In Defence of Intertemporal Consistency. A Discussion of Craig Callender's 'The Normative Standard for Future Discounting'. *Australasian Philosophical Review* 5 (3):266-276.
Guala, Francesco (2019). Preferences: Neither Behavioural nor Mental. *Economics and Philosophy, 35*(3): 383–401.
Gustafsson, Johan E. (2010). A Money-Pump for Acyclic Intransitive Preferences. *Dialectica* 64 (2):251-257.
Gustafsson, Johan E. (2013). The Irrelevance of the Diachronic Money-Pump Argument for Acyclicity. *Journal of Philosophy* 110 (8):460–464.
Halabi, M. E., Chan, Y., Tunca, B., Ziano, I., & Feldman (2022). Replication: Unsuccessful replications and extensions of Temporal Value Asymmetry in monetary valuation and moral judgment. *Journal of Economic Psychology*. DOI: 10.17605/OSF.IO/XCY9F
Hammerton, Matthew (2020). Relativized Rankings. In Douglas W. Portmore (ed.), *The Oxford Handbook of Consequentialism*. New York: Oxford University Press: 46-66.
Hansson, Sven Ove & Grüne-Yanoff, Till (2022). Preferences. In *the Stanford Encyclopedia of Philosophy (Spring 2022 Edition)*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/preferences/>.
Hardisty, D. J., & Weber, E. U. (2020). Impatience and Savoring vs. Dread: Asymmetries in Anticipation Explain Consumer Time Preferences for Positive vs. Negative Events. *Journal of Consumer Psychology, 30*(4): 598-613.
Hare, Caspar (2007). Self-Bias, Time-Bias, and the Metaphysics of Self and Time. *Journal of Philosophy* 104 (7):350-373.
Hare, Caspar (2008). A Puzzle about Other-directed Time-bias. *Australasian Journal of Philosophy* 86 (2):269 – 277.
Hare, Caspar (2013). Time – The Emotional Asymmetry. In Heather Dyke & Adrian Bardon (eds.), *A Companion to the Philosophy of Time*. Chichester, UK: Wiley: 507–520.
Hare, Caspar (2010). Take the Sugar. *Analysis*, 70(2): 237–247.
Harman, Elizabeth (2009). "I'll Be Glad I Did It" Reasoning and the Significance of Future Desires. *Philosophical Perspectives* 23 (1):177-199.

Harris, Christine R. (2010). Feelings of Dread and Intertemporal Choice. *Journal of Behavioral Decision Making 25*: 13–28.
Hausman, Daniel M. (2012). *Preference, Value, Choice, and Welfare.* Cambridge: Cambridge University Press.
Heathwood, Chris (2005). The Problem of Defective Desires. *Australasian Journal of Philosophy* 83 (4):487 – 504.
Heathwood, Chris (2006). Desire Satisfactionism and Hedonism. *Philosophical Studies* 128 (3):539-563.
Heathwood, Chris (2008). Fitting Attitudes and Welfare. *Oxford Studies in Metaethics* 3:47-73.
Heathwood, Chris (2011a). Desire-Based Theories of Reasons, Pleasure and Welfare. *Oxford Studies in Metaethics* 6:79-106.
Heathwood, Chris (2011b). Preferentism and Self-Sacrifice. *Pacific Philosophical Quarterly* 92 (1):18-38.
Heathwood, Chris (2019). Which Desires Are Relevant to Well-Being? *Noûs* 53 (3):664-688.
Hedden, Brian (2015). *Reasons Without Persons: Rationality, Identity, and Time*. Oxford, United Kingdom: Oxford University Press UK.
Helm, Bennett W. (2010). *Love, Friendship, and the Self: Intimacy, Identification, and the Social Nature of Persons.* Oxford, GB: Oxford University Press UK.
Helm, Bennett (2023) Friendship. *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/friendship/>.
Hersch, Gil & Weltman, Daniel (2023). A New Well-being Atomism. *Philosophy and Phenomenological Research* 107 (1):3-23.
Hoerl, Christoph (2015). Tense and the Psychology of relief. *Topoi* 34 (1):217-231.
Hoerl, Christoph (2022). Past/Future Attitude Asymmetries: Values, Preferences and the Phenomenon of Relief. In Christoph Hoerl, Teresa McCormack & Alison S. Fernandes (eds.), *Temporal asymmetries in philosophy and psychology.* Oxford: Oxford University Press: 204-222.
Jeske, Diane (1993). Persons, Compensation, and Utilitarianism. *Philosophical Review* 102 (4):541-575.
Johansson, Jens & Rosenqvist, Simon (2016). 'Pure Time Preference': Reply to Lowry and Peterson. *Pacific Philosophical Quarterly* 97 (3):435-441.
Johnston, Mark (1992). Reasons and Reductionism. *Philosophical Review* 101 (3):589.
Kagan, Shelly (2019). *How to Count Animals*. Oxford: Oxford University Press.
Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy*, *98*(6): 1325–1348.
Karhu, Todd (2023). What Justifies Our Bias Toward the Future? *Australasian Journal of Philosophy* 101 (4):876-889.
Kauppinen, Antti (2018). Agency, Experience, and Future Bias. *Thought: A Journal of Philosophy* 7 (4):237-245.
Keller, Simon & Nelson, Michael. (2001). Presentists Should Believe in Time-travel. *Australasian Journal of Philosophy* 79 (3):333 – 345.
Keller, Simon (2013). *Partiality*. Princeton: Princeton University Press.
Khoury, Andrew C. & Matheson, Benjamin (2018). Is Blameworthiness Forever? *Journal of the American Philosophical Association* 4 (2):204-224.
Kiesewetter, Benjamin & Worsnip, Alex (2023). Structural Rationality. In *the Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/rationality-structural/>.
Kolodny, Niko (2007). How Does Coherence Matter? *Proceedings of the Aristotelian Society* 107 (1pt3):229 - 263.
Korsgaard, Christine (1989). Personal Identity and the Unity of Agency: A Kantian Response to Parfit. *Philosophy and Public Affairs* 18 (2):103-31.
Lange, Benjamin & Brandt, Joshua (2023). Partiality, Asymmetries, and Morality's Harmonious Propensity. *Philosophy and Phenomenological Research* 109 (1):1-42.
Latham, Andrew J.; Miller, Kristie; Norton, James & Tarsney, Christian (2020). Future Bias in Action: Does the Past Matter More When You Can Affect It? *Synthese 198* (12):11327-11349.

Latham, Andrew J.; Miller, Kristie; Oh, Jordan; Shpall, Sam & Yu, Wen (2024). Exploring Arbitrariness Objections to Time Biases. *Journal of the American Philosophical Association*, *10*(3): 588–614.
Latham, Andrew J.; Miller, Kristie & Norton, James (2021). Pure and Impure Time Preferences. *Australasian Philosophical Review*, *5*(3): 277–283.
Latham, Andrew J.; Miller, Kristie; Tarsney, Christian & Tierney, Hannah (2022). Belief in Robust Temporal Passage (Probably) Does Not Explain Future-bias. *Philosophical Studies 179* (6):2053-2075.
Latham, Andrew J.; Li, Rongting; Miller, Kristie & Yu, Wen (manuscript). Exploring People's Normative Judgements About Future Bias and The Temporal Value Asymmetry.
Lee, R., Shardlow, J., O'Connor, P. A., Hotson, L., Hotson, R., Hoerl, C., & McCormack, T. (2022). Past-future Preferences for Hedonic Goods and the Utility of Experiential Memories. *Philosophical Psychology 35*(8): 1181–1211. https://doi.org/10.1080/09515089.2022.2038784
Lee, Ruth; Hoerl, Christoph; Burns, Patrick; Fernandes, Alison Sutton; O'Connor, Patrick A. & McCormack, Teresa (2020). Pain in the Past and Pleasure in the Future: The Development of Past-future Preferences for Hedonic Goods. *Cognitive Science* 44 (9): e12887.
Lewis, David (1976). The Paradoxes of Time Travel. *American Philosophical Quarterly, 13*(2), 145–52.
Lewis, David (1979). Attitudes De Dicto and De Se. *Philosophical Review* 88 (4):513-543.
Lewis, David (1988). Desire as Belief. *Mind* 97 (418):323-32.
Lin, Eden. (2017). Asymmetrism about Desire Satisfactionism and Time. In Mark Timmons (ed.), *Oxford Studies in Normative Ethics, vol. 7*. Oxford, UK: Oxford University Press: 161-183.
Little, Margaret Olivia, & Macnamara, Coleen (2021). Non-requiring Reasons. In *The Routledge Handbook of Practical Reason*. Routledge: 393–404.
Loar, Brian (1990). Phenomenal states. *Philosophical Perspectives* 4:81-108.
Loewenstein, George (1987). Anticipation and the Valuation of Delayed Consumption, *Economic Journal, Royal Economic Society*, 97(387): 666-684.
Lord, Errol (2016). Justifying Partiality. *Ethical Theory and Moral Practice* 19 (3):569-590.
Lord, Errol (2018). *The Importance of Being Rational*. Oxford, UK: Oxford University Press.
Löschke, Jörg (2017). Relationships as Indirect Intensifiers: Solving the Puzzle of Partiality. *European Journal of Philosophy* 26 (1):390-410.
Lowry, Rosemary & Peterson, Martin (2011). Pure Time preference. *Pacific Philosophical Quarterly* 92 (4):490-508.
MacBeath, Murray. (1983). Mellor's Emeritus Headache. *Ratio*, 25: 81–88.
MacFarlane, John (2003). Future Contingents and Relative Truth. *Philosophical Quarterly* 53 (212):321–336.
MacIntosh, Duncan (2010). Intransitive Preferences, Vagueness, and the Structure of Procrastination. In Chrisoula Andreou & Mark D. White (eds.), *The Thief of Time: Philosophical Essays on Procrastination*. New York, US: Oxford University Press.
Mariqueo-Russell, Atus (2023). Well-being and the Problem of Unstable Desires. *Utilitas* 35 (4):260-276.
McClennen, Edward Francis (1990). *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge, England: Cambridge University Press.
McGahhey, Marcus & Van Leeuwen, Neil (2018). Interpreting Intuitions. In Julie Kirsch Patrizia Pedrini (ed.), *Third-Person Self-Knowledge, Self-Interpretation, and Narrative*. Springer Verlag.
Mckerlie, Dennis (2007). Rational Choice, Changes in Values Over Time, and Well-being. *Utilitas* 19 (1):51-72.
McNaughton, David & Rawling, Piers (1995). Value and Agent-Relative Reasons. *Utilitas* 7 (1):31.
McPherson, Lionel K. (2002). The Moral Insignificance of 'Bare' Personal Reasons. *Philosophical Studies* 110 (1):29 - 47.
Mellor, David Hugh (1981). Real Time. New York: Cambridge University Press.
Mellor, David Hugh (1998). Real time II. New York: Routledge.
Merricks, Trenton (2006). Good-Bye Growing Block. *Oxford Studies in Metaphysics* 2: 103-110.

Miller, Kristie (2013). Presentism, Eternalism, and the Growing Block. In Heather Dyke & Adrian Bardon (eds.), *A Companion to the Philosophy of Time*. Chichester, UK: Wiley-Blackwell: 345-364.
Miller, Kristie (2021). What Time-Travel Teaches Us about Future-Bias. *Philosophies* (Basel), 6(2), 38–. https://doi.org/10.3390/philosophies6020038
Miller, Kristie (2022). Tensed Facts and the Fittingness of our Attitudes. *Philosophical Perspectives* 36 (1):216-232.
Miller, Kristie; Greene, Preston; Latham, Andrew J.; Norton, James; Tarsney, Christian & Tierney, Hannah (2022). Bias Towards the Future. *Philosophy Compass* 17 (8):e12859.
Moller, Dan (2002). Parfit on Pains, Pleasures, and the Time of Their Occurrence. *Canadian Journal of Philosophy* 32 (1):67 - 82.
Molouki, S., Hardisty, D. J., & Caruso, E. M. (2019). The Sign Effect in Past and Future Discounting. *Psychological Science*, 30(12):1674-1695.
Mulligan, Kevin (2015). Is Preference Primitive? In Johannes Persson, Göran Hermerén & Eva Sjöstrand (eds.), *Against Boredom.* Tanke Förlag.
Murphy, Mark C. (1999). The Simple Desire-Fulfillment Theory. *Noûs* 33 (2):247-272.
Nagel, Thomas (1970). *The Possibility of Altruism*. Princeton: Princeton University Press.
Nguyen, Anh-Quân (2022). Future-bias and Intuition Shifts Between Moments and Lifetimes. *Inquiry (Oslo)*: 1–28. https://doi.org/10.1080/0020174X.2022.2126394
Ninan, Dilip (2008). *Imagination, Content and the Self.* PhD Dissertation, MIT.
Norcross, Alastair (2020). Value Comparability. In *The Oxford Handbook of Consequentialism*. Oxford University Press.
Norton, John D. (2015). The Burning Fuse Model of Unbecoming in Time. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 52 (Part A):103-105.
Oaklander, L. Nathan. (1984). *Temporal Relations and Temporal Becoming: A Defense of a Russellian Theory of Time*. University Press of America.
Olson, Eric T. (1997). *The Human Animal: Personal Identity Without Psychology*. New York, US: Oxford University Press.
Olson, Eric T. (2007). *What are We?: A Study in Personal Ontology*. New York: Oxford University Press.
Ostapiuk, Aleksander (2022). Weakness of Will. The Limitations of Revealed Preference Theory. *Acta Oeconomica, 72*(1): 1–23.
Overvold, Mark Carl (1980). Self-interest and the Concept of Self-sacrifice. *Canadian Journal of Philosophy* 10 (1):105-118.
Parfit, Derek (1984). *Reasons and Persons*. Oxford: Oxford University Press.
Parfit, Derek (2001). Rationality and Reasons. In Dan Egonson, Jonas Josefson, Bjȍrn Petterson, and Toni Rønnow-Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values.* Aldershot: Ashgate, 17-39.
Parfit, Derek (2011). *On What Matters: Volume Three*. Oxford University Press UK.
Paul, Laurie Ann. (2014). *Transformative Experience*. Oxford: Oxford University Press.
Pearson, Olley (2018). Appropriate Emotions and the Metaphysics of Time. *Philosophical Studies* 175 (8):1945-1961.
Perry, John (1979). The Problem of the Essential Indexical. *Noûs* 13: 3–21.
Perry, John (2010). Critical Study Velleman: Self to Self. *Noûs* 44 (4):740-758.
Pettigrew, Richard (2020). *Choosing for Changing Selves*. Oxford: Oxford University Press.
Pettit, Philip. (2006). Preference, Deliberation and Satisfaction. *Royal Institute of Philosophy Supplement, 59*: 131–154.
Phillips, Callie K. (2021). Why Future-bias Isn't Rationally Evaluable. *Res Philosophica* 98 (4):573-596.
Portmore, Douglas W. (2007). Desire Fulfillment and Posthumous Harm. *American Philosophical Quarterly* 44 (1):27 - 38.
Portmore, Douglas W. (2008). Dual-ranking Act-consequentialism. *Philosophical Studies* 138 (3):409 - 427.

Portmore, Douglas W. (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York, USA: Oxford University Press USA.
Prior, Arthur N. (1959). Thank Goodness That's over. *Philosophy* 34 (128):12 - 17.
Purves, Duncan (2017). Desire Satisfaction, Death, and Time. *Canadian Journal of Philosophy* 47 (6):799-819.
Rabinowicz, Wlodek (2008). Pragmatic Arguments for Rationality Constraints. In Maria-Carla Galavotti (ed.), *Reasoning, Rationality and Probability*. CSLI Publications: 139-163.
Rachels, Stuart (1998). Counterexamples to the Transitivity of Better Than. *Australasian Journal of Philosophy* 76 (1):71–83.
Railton, Peter (1986). Facts and Values. *Philosophical Topics* 14 (2): 5–31.
Rawls, John (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Read, Daniel (2004). Intertemporal Choice. In D. Koehler and N. Harvey (eds.), *Blackwell Handbook of Judgment and Decision Making*. Oxford: Blackwell.
Ridge, Michael (2023). Reasons for Action: Agent-Neutral vs. Agent-Relative. *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2023/entries/reasons-agent/>.
Roelofsma, P. H. M. P., & Read, D. (2000). Intransitive Intertemporal choice. *Journal of Behavioral Decision Making*, *13*(2): 161–177.
Rosati, Connie S. (1995). Persons, Perspectives, and Full Information Accounts of the Good. *Ethics* 105 (2):296-325.
Rovane, Carol Anne (1997). *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton University Press.
Sarch, Alexander (2013). Desire Satisfactionism and Time. *Utilitas* 25 (2):221-245.
Schafer, Karl (2013). Perception and the Rational Force of Desire. *Journal of Philosophy 110* (5):258-281.
Schechtman, Marya (2001). Empathic Access: The Missing Ingredient in Personal Identity. *Philosophical Explorations* 4 (2):95-111.
Scheffler, Samuel (2004). Projects, Relationships, and Reasons. In R. Jay Wallace (ed.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. New York: Oxford University Press: 247-69.
Scheffler, Samuel (2021). Temporal Neutrality and the Bias Toward the Future. In McMahan, J et al. (eds), *Principles and Persons: The Legacy of Derek Parfit*. Oxford University Press.
Schick, Frederic (1986). Dutch Bookies and Money Pumps. *Journal of Philosophy 83* (2):112-119.
Schroeder, Mark (2007). Teleology, Agent-relative Value, and 'Good'. *Ethics* 117 (2):265-000.
Schroeder, Mark (2018). Getting Perspective on Objective Reasons. *Ethics* 128 (2):289-319.
Sebo, Jeff (2015). Multiplicity, Self-narrative, and Akrasia. *Philosophical Psychology* 28 (4):589-605.
Sen, Amartya (1977). Rational fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs* 6 (4):317-344.
Shoemaker, David W. (1999). Selves and Moral Units. *Pacific Philosophical Quarterly* 80 (4):391-419.
Shoemaker, David W. (2002). Disintegrated Persons and Distributive Principles. *Ratio* 15 (1):58–79.
Sider, Theodore (2005). Travelling in A- and B- Time. *The Monist 88* (3):329-335.
Sider, Theodore (2006). Quantifiers and Temporal Ontology. *Mind* 115 (457):75-97.
Sider, Theodore (2018). Asymmetric Personal Identity. *Journal of the American Philosophical Association* 4 (2):127-146.
Sidgwick, Henry (1907). The Methods of Ethics. *Hackett Publishing Company.*
Skow, Bradford (2009). Preferentism and the Paradox of Desire. *Journal of Ethics and Social Philosophy* 2009 (3):1-17.
Slote, Michael (1982). Goods and Lives. *Pacific Philosophical Quarterly* 63 (4):311-326.
Smartt, Tim (2021). Arbitrariness Arguments against Temporal Discounting. *Australasian Philosophical Review* 5 (3):302-308.
Smith, Michael (2003). Neutral and Relative Value after Moore. *Ethics* 113 (3):576-598.
Smith, Michael (2009). Desires, Values, Reasons, and the Dualism of Practical Reason. *Ratio* 22 (1):98-125.

Smith, Quentin (2002). Time and Degrees of Existence: A Theory of 'Degree Presentism'. *Royal Institute of Philosophy Supplement* 50: 119-136.
Snedegar, Justin (2017). *Contrastive Reasons*. New York, NY: Oxford University Press.
Snedegar, Justin (2018). Reasons For and Reasons Against. *Philosophical Studies* 175 (3):725-743.
Snedegar, Justin (2021). Reasons, Competition, and Latitude. In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics Volume 16*. Oxford University Press.
Sobel, David (1994). Full Information Accounts of Well-being. *Ethics* 104 (4):784-810.
Sobel, David (2011). Parfit's Case Against Subjectivism. In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, volume 6. Oxford University Press.
Soman, D., Ainslie, G., Frederick, S., Li, X., Lynch, J., Moreau, P., Mitchell, A., Read, D., Sawyer, A., Trope, Y., Wertenbroch, K., & Zauberman, G. (2005). The Psychology of Intertemporal Discounting: Why Are Distant Events Valued Differently from Proximal Ones? *Marketing Letters, 16*(3/4): 347-360.
Stokes, Patrick (2017). Temporal Asymmetry and the Self/Person Split. *Journal of Value Inquiry* 51 (2):203-219.
Street, Sharon (2009). In Defense of Future Tuesday Indifference: Ideally coherent Eccentrics and the Contingency of What Matters. *Philosophical Issues* 19 (1):273-298.
Strotz, Robert H. (1955) Myopia and Inconsistency in Dynamic Utility Maximization. *Review of Economic Studies*, 23, 165–80.
Stroud, Sarah (2010). Permissible Partiality, Projects, and Plural Agency. In Brian Feltham & John Cottingham (eds.), *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*. Oxford: Oxford University Press.
Suhler, Christopher & Callender, Craig (2012). Thank Goodness That Argument Is Over: Explaining the Temporal Value Asymmetry. *Philosophers' Imprint* 12:1-16.
Sullivan, Meghan (2018). *Time Biases: A Theory of Rational Planning and Personal Persistence.* Oxford, UK: Oxford University Press.
Sullivan, Meghan (2022a). Temporal Discounting in Philosophy and Psychology: Four Proposals for Mutual Research Aid. In Christoph Hoerl, Teresa McCormack, and Alison Fernandes (eds.), *Temporal Asymmetries in Philosophy and Psychology*. Oxford: Oxford University Press.
Sullivan, Meghan. (2022b). Scheduling Deliberation. *Philosophical Perspectives*, 36: 329–344.
Sung, Leora Urim (2024). Time Bias and Altruism. *Mind*, fzae045, https://doi.org/10.1093/mind/fzae045
Tarsney, C. (2017). Thank Goodness That's Newcomb: The Practical Relevance of the Temporal Value Asymmetry. *Analysis*, 77(4): 750–9.
Tenenbaum, Sergio (2007). *Appearances of the Good: An Essay on the Nature of Practical Reason.* Cambridge University Press.
Thaler, Richard H. (1980). Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior & Organization*, 1 (1): 39-60.
Thoma, Johanna. (2018). Temptation and Prefence-Based Instrumental Rationality. In J. Bermudez (ed.) *Self-Control, Decision Theory, and Rationality*. Cambridge: Cambridge University Press.
Torre, Stephan (2011). The Open Future. *Philosophy Compass* 6 (5):360-373.
Tucker, Chris (2021). The Dual Scale Model of Weighing Reasons. *Noûs* 56 (2):366-392.
Tucker, Chris (2024). Parity, Pluralism, and Permissible Partiality. In Eric Siverman (ed.), *Virtuous and Vicious Expressions of Partiality.* Routledge.
van Boven, Leaf, & Laurence, Ashworth (2007). Looking Forward, Looking Back: Anticipation Is More Evocative Than Retrospection. *Journal of Experimental Psychology.* General 136(2): 289–300.
van Inwagen, Peter. (2010). Changing the Past. In Dean W. Zimmerman (ed.), *Oxford Studies in Metaphysics 5*.
van Weelden, Joseph (2019). On Two Interpretations of the Desire-Satisfaction Theory of Prudential Value. *Utilitas* 31 (2):137-156.
Velleman, J. David (1991). Well-Being And Time. *Pacific Philosophical Quarterly* 72 (1):48-77.
Velleman, J. David (1996). Self to self. *The Philosophical Review 105*(1): 39–76.
Vogelstein, Eric (2012). Subjective Reasons. *Ethical Theory and Moral Practice 15* (2):239-257.

Whiting, Jennifer (1986). Friends and Future Selves. *Philosophical Review* 95 (4):547-80.
Williams, Bernard (1973). Imagination and the Self. In *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge: Cambridge University Press: 26–45.
Williams, Bernard (1979). Internal and External Reasons. In Ross Harrison (ed.), *Rational Action: Studies in Philosophy and Social Science.* Cambridge: Cambridge University Press: 17–28.
Williams, Bernard (1981). Persons, Character, and Morality. In *Moral Luck: Philosophical Papers 1973–1980*. New York: Cambridge University Press.
Williams, J. Robert G. (2014). Nonclassical Minds and Indeterminate Survival. *Philosophical Review* 123 (4):379-428.
Worsnip, Alex (2022). Making Space for the Normativity of Coherence. *Noûs* 56 (2):393-415.
Yehezkel, Gal (2013). Theories of Time and the Asymmetry in Human Attitudes. *Ratio* 27 (1):68-83.
Yi, Richard; Gatchalian, Kirstin M. & Bickel, Warren K. (2006). Discounting of Past Outcomes. *Experimental and Clinical Psychopharmacology 14* (3): 311–317.
Yu, Wen (2024). On the Rational Evaluability of Future-bias. *Inquiry*: 1–22. https://doi.org/10.1080/0020174X.2024.2446246
Zimmerman, Dean W. (2005). The A-Theory of Time, The B-Theory of Time, and 'Taking Tense Seriously'. *Dialectica* 59 (4):401-457.