Agency, Experience, and Future Bias

Antti Kauppinen

University of Helsinki

To appear in *Thought*

In *Reasons and Persons*, Derek Parfit (1984) observed that most people are biased towards the future at least when it comes to pain and pleasure. That is, they regard a given amount of pain as less bad when it is in the past than when it is in the future, and a given amount of pleasure as less good. While Parfit (implicitly) held that this bias is rational, it has recently come under effective attack by temporal neutralists, who have offered cases that with plausible auxiliary assumptions appear to be counterexamples to the rationality claim. I'm going to argue that these cases and the rationale behind them only suffice to motivate a more limited rejection of future bias, and that constrained future bias is indeed rationally permissible. My argument turns on the distinct rational implications of *action-guiding* and *pure* temporal preferences. I'll argue that future bias is rational when it comes to the latter, even if not the former. As I'll say, Only Action Fixes Utility: it is only when you act on the basis of assigning a utility to an outcome that you rationally commit to giving it the same value when it is past as when it is in the future.

1. Future Bias

Let's start with Parfit's most famous case. I will borrow Tom Dougherty's formulation of the setup:

> *Parfit's Operations*
>
> On Monday, you are admitted into a hospital. You are told you will have one of two operations, but you are not told which. If you have the early operation, then you will

> have a painful, four-hour operation on Tuesday. If you have the late operation, then you will have a painful, two-hour operation on Thursday. After either operation, you will have amnesia for several days, and so you will not be able to remember if you have just had the operation. There is a calendar next to your bed, and so you always know what day it is. (Dougherty 2011, 522; cf. Parfit 1984, 165)

Now, suppose you wake up in a fog, and see that it is Wednesday. You don't know whether you've had the operation or not. How would you feel if a nurse came by and told you that you had the more painful operation on Tuesday? Most people, Parfit's critics included, admit they would be relieved. But if so, they prefer a greater pain in the past to a smaller pain in the future. This means, on standard decision-theoretical assumptions, that they assign a lower disutility to the same amount of pain in the past than in the future. They are *future-biased*.

Notice that future bias entails *preference reversals* as one moves forward in time. On Monday, when both operations are in the future, you'd prefer to have the operation on Thursday rather than Tuesday, while on Wednesday, you have the opposite preference. (However, and this will be crucial to my argument, in Parfit's original case you don't get to choose.)

While most people are future biased, the question is whether such bias is *rational*. Recently, this assumption has been challenged by several critics who hold that other things being equal, we should be neutral between the past and the present, just as we should be neutral between near future and far future.

2. Against Future Bias

I'm going to start presenting the case against the rationality of future bias by introducing a few counterexamples from the literature. Some objections draw on the notion of regret, understood thinly as preference that one would have chosen differently.

As Preston Greene and Meghan Sullivan emphasize, anticipating regret can influence rational choice. According to a plausible principle they call Weak No Regrets, it is rationally permissible for agents to avoid options they know they will regret, if they have full and accurate information about their options (Greene and Sullivan 2015, 958). This creates a problem for a future-biased agent in a case like the following:

> *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. (Greene and Sullivan 2015, 959)

Insofar as Jack is future-biased, he knows he will regret choosing the meal for any but the last night, so in accordance with Weak No Regrets, it's rational for him to postpone it as far as he can. If there is no fixed endpoint (if he can have the fancy meal any day he wants), it is rational for him to just keep waiting. But this is a terrible idea; as Greene and Sullivan put it, "future-biased agents who avoid regret will postpone positive experiences for no good reason" (2015, 959–960). So to be rational, Jack needs to give up his future bias.

Future bias will also lead to irrational-seeming regret when agents make trade-offs between goods that depreciate with the passage of time and goods that don't, which results in a diachronically inconsistent exchange rate. To quickly sum up a case from Dougherty (2015), suppose Victoria thinks on Saturday morning that the value of having a mown lawn is 5 utils and the value of a pleasant afternoon spent sunbathing with no garden work is 6 utils. However, if future-biased, she will in the evening regard her past pleasure as being worth less, say 4 utils, and consequently regret she didn't mow the lawn. Victoria's regret seems irrational. After all, she always knew how much she values an orderly lawn, but still chose the sunbathing, which was just as nice as expected. As Dougherty puts it, "If the pleasure of sunbathing is insufficient to justify foregoing gardening, then she should not choose sunbathing in the first place." (2015, 8)

As Dougherty observes, sometimes our choice of what to do depends more directly on how we evaluate what has happened. Here is one case based on his 2015 paper:

Volunteer Tom

> Tom thinks we should spend some time doing unpleasant things that help others. This week, he figures he'll have done his bit if he gives up five utils for others. Looking ahead, he thinks each hour spent volunteering at the soup kitchen costs him one util, so he decides to spend five hours there on Saturday. It's just as unpleasant as he thinks. But Tom is future-biased, and regards five hours of past pain as having the same disutility as one hour of future pain. Consequently, when he's about to finish his shift, he realizes that he's after all only suffered one util for others, so his duty calls him to stay another four hours. The next four hours are as unpleasant as he thought, but once they're in the past, he finds that his additional sacrifice amounts to only 4/5 utils, so he'll have to stay yet longer… This sad story keeps repeating itself.

Here Tom's future bias catches him in a Zeno-like situation. Prospectively, he takes the disvalue of five unpleasant hours of volunteering to equal the sacrifice he should make, but retrospectively it will not suffice, so he ends up doing more and more. This pattern seems irrational.

What, then, is the alternative to future bias that avoids the counterintuitive consequences? The critics hold that it is the following:

> *Temporal Neutrality*
>
> The location of goods and harms within a life has no rational significance except insofar as it contributes to the value of that life. (Cf. Brink 2011) In particular, rationality requires you to regard the same amount of pain as just as bad for you whether it occurs in the past, present, or future, and pleasure equally good.

An agent who has the pattern of valuing required by Temporal Neutrality will not discount past pains or pleasures, so they won't regret their choices in scenarios like the above or end up like Tom.

Temporal Neutralists offer various general rationales for their view. The core idea is that we're temporally extended agents, and must rationally value what happens to us at different times in a consistent way (Brink 2011). Otherwise we will forego pleasure or suffer pain without being in any way compensated for it, in spite of full information and rationality, and may even be brought to make tradeoffs that make us worse off without benefiting us in any way (Dougherty 2011). To avoid such an undesirable situation, we need an overall life-plan that we can endorse at all times (Dougherty 2015).

Finally, Temporal Neutralists offer *debunking explanations* of why we're future-biased. Greene and Sullivan, in particular, argue that future bias is an evolved heuristic: "future-biased emotions and preferences evolved to track asymmetries in control" (2015,

968). The idea is simple: we care more about future pleasures and pains, because we can often do something about them, while we can't do anything about past pleasures and pains. It is clearly advantageous to focus on what we can control. Nevertheless, according to critics, this bias can systematically lead us astray in circumstances in which our attitudes or actions hang on our evaluations of what has already happened.

3. Only Action Fixes Utility

I'm fairly convinced by the counterexamples that the critics present (though as Dorsey 2017 shows, the intuitions are less firm than they may seem). I'll grant that unrestricted future bias isn't rational. Nevertheless, I'm not convinced by Temporal Neutralism. In the rest of this paper, I'm going to argue that the cases only motivate a more narrow rejection of future bias. In particular, I'm going to defend the following principle of diachronic rationality:

> *Only Action Fixes Utility*
>
> If you act on the basis of assigning utility *u* to state of affairs S, rationality requires you to assign *u* to S whenever it is relevant to action or attitude, unless you gain new information about S.[1] However, if you do not act on the basis of assigning *u* to S (nor have acted or ever will), it is rationally permissible to assign a different utility *u'* to S at different times without gaining new information about S, at least when the underlying preferences are hedonic.

As I'm using the terms, the *utility* of a state for an agent is a value that is derived from the agent's preferences between possibilities, provided that they meet constraints like completeness and transitivity. I'll say an agent *assigns* a utility to S when she has preferences with the right structure between possibilities that include S. A change in S-regarding

[1] We might also add the condition that one doesn't discover that one's earlier preferences were mistaken or corrupted. I'll ignore this complication in what follows.

preferences thus entails a change in the utility assigned to S. An agent *acts on the basis of* assigning a certain utility to S if and only if her S-regarding preferences explain, at least in part, her choice between acts whose outcomes include S.[2] Finally, by *preference* I mean an attitude of favouring one state over another that may be manifest both in dispositions to choose and in emotions and attitudes like regretting, hoping, and wishing.[3]

It follows from the principle that if at t_1 you prefer A to B at *t*, and act on that basis, then at any later t_2, you rationally must still prefer A to B at *t* in the context of action, unless you've learned something new about A or B at *t*. However, if you *don't* act on a preference, you're rationally permitted to change it as time goes by. I'll call the latter kind of bias *pure* future bias.

Only Action Fixes Utility has the right implications for the above scenarios. Greene and Sullivan's Jack's problem is that he'll regret eating the fancy meal on any night other than the last possible night, since while a meal on any night is prospectively equal, a meal on any but the last is retrospectively downgraded. But if Only Action Fixes Utility is true and Jack chooses to dine on Monday on the basis of assigning 10*u* to the pleasure of doing so, he's rationally committed to assigning the same value to it subsequently. Insofar as he does so, he won't regret it, since on Tuesday he remains indifferent between having had the meal and having it that night or later. And this means that even if Weak No Regrets is true, it is not rational for Jack to postpone positive experiences just because of temporal location. Similarly,

---

[2] A radical holist might challenge the present view by claiming that we always act on the basis of *all* of our preferences, so that it is not possible to isolate non-action-guiding utility-assignments. But as a very helpful reviewer for this journal pointed out, focusing on preferences that *explain* a choice offers a response to this challenge: not even the radical holist can credibly claim that all preferences play an explanatory role of this sort in each choice.

[3] As a reviewer for this journal observed, some might want to distinguish between practical and emotional preferences, where the latter do not involve a disposition to act. These different kinds of preferences might then be subject to different rational requirements. I believe that my thesis and arguments could be formulated in these terms as well, but given that the critics of future bias treat preferences as forming a unified kind (e.g. Dougherty 2015, 2fn1), I will continue to do so myself.

Dougherty's Victoria, having chosen to sunbathe on the basis of assigning it a higher utility than lawn-mowing, will never come to regret her choice, since she won't vary the exchange rate. Finally, Tom, too, decides to do five hours of volunteering on the basis of prospectively considering it to be the right amount of sacrifice. His problem is that while it is just as hard as he expected, he retrospectively revises down its disvalue, which sets him on an Eleatic path. But again, if Only Action Fixes Utility and Tom is rational, he will continue to regard five hours of volunteering as having a sufficient amount of disutility, when it turns out as anticipated, and he'll be satisfied that he's fulfilled his duty after a good afternoon's work.

Unsurprisingly, Only Action Fixes Utility yields the same verdict regarding the above cases as Temporal Neutrality. Unlike Temporal Neutrality, however, Only Action Fixes Utility is compatible with pure future bias. In particular, it permits preference change in Parfit's Operations, because in it, the subject doesn't have a choice, and thus isn't rationally compelled to assign the same utility to past and future pain. So even if on Monday you prefer the Thursday operation, it's rationally okay to prefer the more painful Tuesday one, come Wednesday, and thus be relieved if you discover the doctors did indeed perform the operation on Tuesday.

The significance of Only Action Fixes Utility comes out clearly if we consider a variant of Parfit's original case in which the subject *does* get to choose which operation to have:

> *Chosen Operations*
>
> On Monday, you are admitted into a hospital. You are told you will have one of two operations, and *you get to choose which one*. If you have the early operation, then you will have a painful, four-hour operation on Tuesday. If you have the late operation, then you will have a painful, two-hour operation on Thursday. After either operation, you will have amnesia for several days, and so you will not be able to

remember if you have just had the operation. There is a calendar next to your bed, and so you always know what day it is.

If you're rational, you'll choose the Thursday operation on Monday. Now suppose you again wake up in a fog in a hospital bed, not knowing what has happened, and notice that it is Wednesday. First question: do you regret having chosen the operation on Thursday rather than Tuesday? Second question: should you? Speaking for myself, I think I might feel a twinge of disappointment, and wish it was Friday already. But on the whole, I think I'd be fine with my choice, and certainly wouldn't reproach my past self for having made it. I would not wish I had chosen otherwise, and thus wouldn't regret my choice in the technical sense used in this debate. I'd say to myself "Well, it's too bad it's not over and done with, but there's just one more day to the less painful option I chose." If that's how I think, I *own* my choice, as we might say. And I think that's the rational attitude to take here.

To further confirm the principle's fit with considered judgments about cases, imagine the following scenario:

Mistaken Operations

Your situation is the same as in Chosen Operations, and you choose the Thursday operation. As you wake up in a fog, you see that it is Wednesday. A nurse brings you a tub of ice cream, asking you if you feel okay. Baffled, you learn that due to a clerical error, the operation has already been performed after all.

In Mistaken Operations, you made a choice, but it turned out to be irrelevant to what actually happened – you might as well have not made a choice. If the outcome is out of your hands, I think it is both predictable and rationally permissible for you to feel *relieved*, though you might insist on a double-check the next time. Since switching from passive experience to

active agency (and back) switches intuitions, there's good *prima facie* reason to think it is agency that plays an explanatory role when it comes to the impermissibility of future bias.

4. Defending Pure Future Bias

The obvious criticism of Only Action Fixes Utility is that it is *ad hoc*. Indeed, as long as there are diachronic norms of rationality, it is hard to see any other possible fault in it, since it captures both Parfit's original intuition (which even his critics nearly always acknowledge to be strong) and the intuitions of his critics, and is in this respect clearly superior to Temporal Neutrality.[4] For it not to be *ad hoc*, we need some further justification for it. Fortunately, there are several rationales available.

First, let me emphasize again that the general rejection of future bias is not motivated by the counterexamples the critics have presented. They motivate precisely Only Action Fixes Utility, and no more. It is perhaps less obvious that the same is true of the *rationales* the critics give for temporal neutrality. But that turns out to be the case. Consider, first, the general claim made by Dougherty in his earlier paper: "The reason why an inconsistency in preferences is a rational defect is that this inconsistency will lead to problems when acting" (2011). If the preferences aren't or can't be acted upon, temporal inconsistency isn't a problem on these grounds. In later work, too, Dougherty emphasizes temporally extended agency. The idea is that since our future selves are as much us as our present ones, we're under rational pressure to make experience-affecting choices from a viewpoint that is equally satisfactory for our past, present, and future selves (Dougherty 2015, 7–8). And the only assignment of utility that is equally satisfactory from different temporal perspectives is one that is based only on intrinsic features, not relative temporal location. So Dougherty holds that

[4] There are those who deny the existence of diachronic norms of rationality, to be sure (see Hedden 2015). I'm skeptical of time-slice rationality for independent reasons, but cannot discuss the issue here.

"By conceiving of ourselves as temporally extended agents, we form temporally neutral preferences for goods that are based on how these goods contribute to how well our temporally extended lives go." (2015, 14)

While there is much that I agree with in this, the last formulation is revealing in that it involves a subtle but illegitimate generalization from the need for diachronic coordination and endorsement of actions to temporally neutral preferences for *any kind* of goods, not just those that our choices and actions affect. Here is a better motivated variant: by conceiving of ourselves as temporally extended agents, we form temporally neutral preferences for *goods that are at stake in our actions* that are based on how these goods contribute to how well our temporally extended lives go. And this is a rationale for Only Action Fixes Utility, not for Temporal Neutrality across the board.

So as far as I can see, there is a good rationale or two for Only Action Fixes Utility that is independent of the intuitions about cases, and it is given by Temporal Neutralists themselves. But that's not all. There is good reason to think that there is something special about how agency contributes to the value of our lives. Consider here again the fact that most people's intuition with respect to shameful actions is temporally neutral, as Parfit already observed. You probably won't prefer a world in which you did something shameful yesterday to one in which you'll do something less shameful tomorrow, other things being equal. The same goes for actions that merit pride. I'm just as happy to think I have done something worthy of pride as to think I will do something worthy of pride, other things being equal. So it is not only the case that our intuitions are temporally neutral when it comes to active choices regarding pleasure and pain, but also when it comes to *non-experiential* goods or bads that result from exercising agency and can serve as grounds for choice. This again points to the importance of agency to temporal neutrality – which is captured in Only Action Fixes Utility.

The emphasis on agency also has the more controversial (and non-Parfitian) implication that it matters *whose* choices brought about an outcome. Someone objected to my view by pointing out that if I wake up on Wednesday and I'm told that my mother chose on my behalf to have the operation on Thursday, I can rationally wish she had chosen the Tuesday slot. But if it is then revealed that it was in fact myself who chose the Thursday operation, I can no longer rationally wish I'd already had the operation. The objection is that this is an odd pattern of emotional reaction. But I don't think it is. It already makes a great deal of difference to us whether, say, we earn some money through our own work, or get the same amount from our mother. It's not odd at all for us to own up to our choices – after all, if someone signs up for the army, fully knowing what to expect, a sergeant can rightly and effectively respond to complaints by saying "You didn't have to sign up if you didn't want to haul these missiles around". If anything is odd about the case where my mother chooses the Thursday operation, it's my wish that she had chosen otherwise. It is, of course, *permissible* according to Only Action Fixes Utility, but it's not *required*. And when I recognize that someone made, on my behalf and in my best interests, the same choice I would myself have made, had I had the chance, it's natural for me to retrospectively endorse it – to treat it *as if* it was my own choice. So it would be less surprising if I said to myself "It's too bad it's still ahead of me, but mama knows best, and I'm glad she chose the less painful operation for me".

Finally, I have argued elsewhere (Kauppinen 2015) that it is of fundamental significance to prudence that we have a dual nature as both temporally extended agents and as subjects of experience. While these two dimensions of prudential value can interact, as when we choose to act to gain or avoid an experience, they remain distinct. In particular, experience as such has an element of passivity: fundamentally, pleasures and pains *happen* to us. They are not under our voluntary control, as our actions may be. That is why our preferences regarding them are not subject to the same intertemporal consistency requirements, insofar as

they are untethered from actions. In such cases, nothing forces us to prefer a particular pattern of hedonic states. For example, rationality permits preferring a larger pleasure in the past to a smaller pleasure in the future. But most of us don't have this preference, perhaps for the kind of evolutionary reasons that Greene and Sullivan point to. If I were to hazard a little psychological speculation, I'd say that perhaps anticipated future pains, say, *de facto* impact on our present preferences more than past ones, because our *attention* is ordinarily directed more towards the future rather than the past, and in some measure makes anticipated pains present right now, while past pains quickly vanish from the arena of presence.

But isn't preferring pain to be in the past still *arbitrary*, and therefore irrational? I see no reason to think so. It is rationally permissible to prefer chocolate ice cream to peppermint ice cream (or vice versa) because one simply happens to please you more. In such cases, there is no more fundamental reason for the preference, so it is in a sense 'arbitrary', but that doesn't make it irrational. Pure future bias, or its absence, may be similar. You just happen to be pleased that a pleasure is to come or a pain is no longer. Since in the case of pure future bias there is no reason not to have such preference and no rationale for a rational prohibition, it is rationally permissible without further grounds.

5. Conclusion

Critics of future bias make a convincing case that if my preferences regarding pleasure and pain change with the mere passage of time, making choices on their basis can make me predictably worse off. As temporally extended agents, we should own up to our choices, and hold fixed our preferences between hedonic outcomes once we've acted on them. But not everything that happens to us is due to our own choice. That's why there's nothing wrong from the perspective of rationality if I'm relieved to discover that an unpleasant experience beyond my control is already in the past, even if I would have earlier preferred a less

unpleasant one in the future, or if my non-action-guiding wishes manifest a preference for a lesser pleasure in the future over a greater pleasure in the past. Given that Only Action Fixes Utility, pure future bias remains rationally permissible.[5]

References

Brink, David 2011. Prospects for Temporal Neutrality. In C. Callender (ed.), *The Oxford Handbook of Philosophy of Time*. Oxford: Oxford University Press, 353–381.

Dorsey, Dale 2017. Future Bias: A Qualified Defense. *Pacific Philosophical Quarterly* 98 (1), 351–373.

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 (30): 1–16.

Greene, Preston and Sullivan, Meghan 2015. Against Time Bias. *Ethics* 125 (4): 947–970.

Hedden, Brian 2015. Time-Slice Rationality. *Mind* 124 (494): 449–491.

Kauppinen, Antti 2015. The Narrative Calculus. *Oxford Studies in Normative Ethics* 5: 196–220.

Parfit, Derek 1984. *Reasons and Persons*. Oxford: Oxford University Press.

[5] I'd like to thank Anh-Quan Nguyen for helpful discussion on the topic, as well as three anonymous reviewers for this journal.