Introduction to Part Two: Rationality and Time

Brian Hedden

1 Introduction

Parfit describes his task in Part Two of *Reasons and Persons* as a sustained attack on S, the Self-interest Theory. S makes its first appearance on the first page of the book, where Parfit says that 'S gives to each person this aim: the outcomes that would be best for himself, and that would make his life go, for him, as well as possible' (3). Restricting ourselves to evaluation of acts, S says that a person ought to perform act ϕ only if there is no alternative act ψ such that his life would go better, for him, if he were to ψ than if he were to ϕ.[1]

It bears emphasizing that S does not say that people ought to be selfish or egotistical, at least as we ordinarily understand these characteristics. One's life might go best by having deep friendships, loving one's children, and giving to charity. This can be true, albeit for different reasons, on any of the three kinds of theory of well-being that Parfit considers, namely hedonistic theories, desire-fulfilment theories, and objective list theories (Parfit discusses these theories in depth in Appendix I).

Parfit advances three main objections against S. First, it can be rational to care most about things other than one's overall lifetime well-being, such as doing one's duty or achieving some goal, and so, *contra* S, it can be rational to perform some act that will not maximize one's overall lifetime well-being. Second, S is an objectionably hybrid view in that it is agent-relative but time-neutral. It is agent-relative in that it gives to each agent a distinct aim, namely

[1]Parfit uses masculine pronouns in stating such principles. Rather than employ generic or feminine pronouns, I will follow Parfit's practice for the sake of avoiding confusion.

that *his* life goes as well as possible. But it is time-neutral in that it prohibits privileging any temporal part of his life over any other. Insofar as well-being can accrue to temporal parts of a life, the well-being of each of those temporal parts is to count equally. In his 'Appeal to Full Relativity,' Parfit argues that, unlike S, a theory should be either fully neutral or fully relative. Third, time-neutrality may not be rationally required. It may be rational to care more about some parts of one's life than about others.

I will examine each of these arguments in what follows. But first, we must look at the theory that constitutes the main foil for S, namely P, the Present Aim Theory.

2 The Critical Present Aim Theory and Intrinsically Irrational Desires

The Present Aim Theory, or P, 'tells each to do what will best achieve his present aims' (92). (Parfit also distinguishes various versions of P at the outset of Part Two; we will soon consider the version he favours.) P differs from S by being both agent-relative and time-relative, whereas S is agent-relative but time-neutral. This difference plays a key role in Parfit's arguments against S, especially his Appeal to Full Relativity.

But there is another important difference between P and S. S is about maximizing well-being, whereas P is about achieving aims. Thus P and S are not a 'minimal pair,' differing only with respect to time-neutrality. A theory that differed from S only by being time-relative where S is time-neutral would say that an agent A at time t ought to perform the act that will maximize the well-being of A-at-t. This view may be implausible for two reasons: first, perhaps time-slices cannot themselves have levels of well-being (this is an issue we will revisit in Section 4), and second, perhaps how A acts at t cannot affect the well-being of A-at-t, but only that of A-at-t+ϵ and successive time-slices (this is an issue that will come up in Section 5 in a slightly different form).

A theory that differed from P only by being time-neutral where P is time-relative would

say that agent A at time t ought to perform the act that will best achieve all of the aims that A has throughout his life, not just those that A has at time t (Kagan 1986). This theory runs into trouble in cases where what A does at t affects what aims A will later have. Suppose that if I now enrol in business school, I will acquire the aim of making lots of money, while if I now enrol in a philosophy PhD program, I will acquire the aim of writing a philosophy book. Setting aside all other aims I might have, our time-neutral analogue of P says that what I ought to do depends on what I will in fact do. If I will in fact enrol in business school, then I will in fact later have the aim of making lots of money, and so I ought to enrol in business school, since this will best achieve the aims I have over the course of my life. Similarly, if I will in fact enrol in a philosophy PhD program, then I will in fact later have the aim of writing a philosophy, an aim which I will now best help achieve by enrolling in a philosophy PhD program. So, our time-neutral analogue of P says that I ought to enrol in the philosophy PhD program. Each of my possible decisions is thus self-reinforcing in such a case. Worse, suppose that if I enrol in business school, I will acquire the aim of writing a philosophy book, while if I enrol in a philosophy PhD program, I will acquire the aim of getting rich. Then, our time-neutral analogue of P says that whichever of these actions I will in fact perform, I ought to perform the other. Each of my possible decisions is self-frustrating in this case. Thus, our time-neutral analogue of P has the odd implication that what one ought to do at time t can depend on what one will in fact do at time t. This implication is certainly odd. It is a further question whether it constitutes an objection to the theory; see Hare and Hedden (2016) for discussion.

I will follow Parfit in continuing to focus on S and P, but it is important to keep in mind that they are not a minimal pair, but rather differ along two dimensions, namely time-neutrality vs. time-relativity and well-being maximization vs. aim-achievement.

Parfit's preferred version of P is the Critical Present Aim Theory, or CP. Parfit argues that some desires are intrinsically irrational, while others are (intrinsically) rationally required. Intrinsically irrational desires do not provide reasons for action. And while he is not explicit

on this point, he presumably thinks that if one is rationally required to have some desire, then one has a reason to try to make true the content of that desire regardless of whether or not one in fact has that desire. Then, CP says that agent A at time t ought to perform the act that will best satisfy the set of desires whose members include the rationally permissible desires that A has at t as well as any rationally required desires that A lacks at t.

In claiming that some desires are intrinsically irrational, while others are rationally required, Parfit is opposing Hume, who is commonly interpreted as holding that desires cannot be irrational unless they are based on a false (or better: irrational) belief. One's ultimate ends cannot be criticized as irrational.

What about Parfit's argument for his anti-Humean position? The first thing to note is that he does not actually argue for the second part of his claim, that some desires are (intrinsically) rationally required. He only argues for the claim that some desires are intrinsically irrational. Still, if one is convinced that some desires are intrinsically irrational, this should at least lessen one's resistance to the thought that some desires are rationally required.

Parfit's argument that some desires are intrinsically irrational is based on examples, the most famous of which is the man with Future Tuesday Indifference:

> A certain hedonist cares greatly about the quality of his future experiences. With one exception, he cares equally about all the parts of his future. The exception is that he has Future-Tuesday-Indifference. Throughout every Tuesday he 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. (123-4)

Considered as a mere intuition pump, Future Tuesday Indifference is compelling but may not add much to the debate. After all, Hume himself gave an example of an intuitively irrational preference—that of the man who preferred the destruction of the world to the scratching of his finger—but insisted that the preference was not in fact irrational. And Rawls (1971) gives the example of someone whose main goal in life is to count the blades of grass in his yard.

But Parfit's key point is that Future Tuesday Indifference is arbitrary. It involves drawing a sharp line between cases that are similar in all respects worth caring about. The same is true of his other examples, Bias Towards the Next Year and Within-A-Mile Altruism. In arguing that such arbitrariness is irrational, Parfit thus provides more of an argument than we get from just considering intuition pumps like those of Hume and Rawls.[2]

3 The 'Best Objection' to S

In criticizing S, Parfit leads with what he calls his 'best objection' to the theory. He argues that according to S, each person should be 'governed by the desire that his life goes, for him, as well as possible,' regardless of the costs to others (131). He refers to this desire as the *bias in one's own favour*. And he argues that the bias in one's own favour is not supremely rational; it can be rational to have other desires that are as strong as, or stronger than, the bias in one's own favour.

[2]See Street (2009) for critical discussion of all these 'ideally coherent eccentrics.' See also Broome (1991) for another prominent argument againt Humeanism about preference. Broome argues that some *substantive* constraints on preferences are required if the *formal* constraints of transitivity and the like are to have any bite. For criticism of Broome's argument, see Dreier (1996).

He considers two sorts of desires that might lead one to act in a way that violates S and that, he claims, are no less rational than the bias in one's own favour. First are moral desires. One might desire to sacrifice one's life in order to save several other people, even though this self-sacrifice would make one's own life go worse overall.[3] Second are desires for achievement. One might desire to write a great novel, even knowing that one will thereby make one's life go worse, given the stress, uncertainty, and self-doubts that one will suffer while writing.

Parfit claims, plausibly, that these desires are no less rational than the bias in one's own favour, and that it would not be irrational for one to act on these desires, even in cases where doing so would make one's own life go worse overall.

Kagan (1986) objects to Parfit's argument. On his way of reading Parfit, Parfit is objecting to S on the grounds that it 'elevates one particular pattern [of concern]—the bias in one's own favor—and gives it a unique theoretical status,' and S is therefore 'intolerant in its attitude toward the rationality of different patterns of concern' (750). But, he argues, P also elevates a particular pattern of concern—we might call it the *bias in favour of one's present desires* —and gives *it* a unique theoretical status.

The situation, as Kagan sees it, is this: Both S and P (and indeed any theory of what one ought to do) elevate some particular pattern of concern (a 'metadesire,' as Kagan terms it) to unique theoretical status. For S, this is the bias in one's own favour. For P, this is the bias in favour of one's present desires. But neither needs to deem other particular desires, such as moral desires, or desires for achievement, to be irrational. Thus, for Kagan, S and P are on a par, and Parfit's 'best objection' gives no grounds for preferring P to S.

As I see it, though, Kagan misinterprets Parfit's objection. Parfit is not objecting to the mere fact that S elevates some pattern of concern or other to unique theoretical status. He is objecting to the fact that it elevates *this particular pattern of concern*—the bias in one's own

[3]This self-sacrifice might make one's life go better in some respects. Perhaps if one did not sacrifice oneself, one would suffer unpleasant feelings of guilt or shame. Moreover, if one desires to sacrifice oneself, then satisfying this desire itself constitutes a boost in one's well-being. And some objective list theories say that acting morally itself makes one's life go better. Thus, we must be careful in spelling out the case to ensure that the painfulness of one's death, and the shortening of one's life, sufficiently outweigh these benefits of self-sacrifice.

favour—to this unique theoretical status. He thinks that it can be rational to act on certain desires, like moral desires and desires for achievement, even when doing so makes one's life go worse overall. If this is so, then S is false.[4]

For S and P to be on a par in this respect, we would need to make an analogous case against P. We would need a case where performing some action is rational even though this action will not best satisfy one's present desires, and even though one's present desires are all rationally permissible and include all rationally required desires (since we are considering the Critical version of P).[5] It is not clear what such a case would look like. And absent such a case, Parfit can claim that S and P are not on a par.

4 The Appeal to Full Relativity

We noted earlier that S combines agent-relativity with time-neutrality. Parfit objects to this structure in his Appeal to Full Relativity.[6] Theories that combine agent-neutrality with time-neutrality, or agent-relativity with time-relativity, are pure. They are either fully neutral, or fully relative. Examples of fully neutral theories include Consequentialism (C) and the agent-neutral modification of Common Sense Morality (N) that Parfit considers toward the end of Part One. The Present Aim Theory is a fully relative theory. By contrast, S is a hybrid theory, being neither fully neutral nor fully relative. It is *incompletely relative*.

Parfit suggests that by virtue of this hybrid, incompletely relative structure, 'S can be

[4]As Parfit (1986, 845) says in his 'Comments,' P's metadesire or 'master function' is not restrictive, whereas S's metadesire or master function is restrictive. This is because P tells us one to do whatever will best achieve one's present aims, whatever those aims may be. By contrast, S tells one to do whatever will best achieve the aim of maximizing one's overall lifetime well-being, even when that conflicts with one's other aims. Parfit then claims that in such conflict cases, S is committed to regarding those other aims as in some sense irrational or inferior to aim of maximizing lifetime well-being, for otherwise it should be rational to act on those aims rather than the aim of maximizing lifetime well-being.

[5]Compare Parfit (1986, 845-6). Note also that it might be possible to devise a case against IP, the instrumental version of P, which says that one ought to act so as to best satisfy one's present desires, whatever those desires happen to be. We might think it would be rational for the man with Future Tuesday Indifference to schedule a less painful operation for next Wednesday rather than a more painful one for next Tuesday, even though doing so would not best satisfy his bizarre present desires.

[6]The Appeal to Full Relativity appears briefly in Sections 34 and 35 of Part One. It has antecedents in Sidgwick's *The Methods of Ethics* (1907) and also in Nagel's *The Possibility of Altruism* (1970, 60-71).

charged with a kind of inconsistency' (140). But it is not clear why exactly incomplete relativity constitutes a kind of inconsistency. As Kagan (1986) notes, the most Parfit seems to do to argue that theories should be either fully neutral or fully relative is to note a formal analogy between personhood and time. The words 'I' and 'now' are both indexicals. The way in which the word 'I' picks out a person is similar to the way in which the word 'now' picks out a time. An utterance of 'I' refers to the speaker of the utterance, and the word 'now' refers to the time of the utterance. This is just a linguistic point, however. Parfit adds that 'When each of us is deciding what to do, he is asking, 'What am *I* to do *now*?'' (140), and concludes that a theory of rationality should treat 'I' and 'now' in the same way.

This formal analogy is not entirely convincing. But let me first point out that Parfit has overlooked a way in which even C and N are not fully neutral. Just as there is a formal analogy between 'I' and 'now,' so there is a formal analogy between these two and 'actual.' 'Actual' is also an indexical, utterances of which refer to the possible world of the utterance. And when each of us is deciding what to do, he is asking not just 'What am *I* to do?', and not just 'What am *I* to do *now*?', but 'What am *I* *actually* to do *now*?' So one could argue that even C and N are incompletely relative. A fully neutral theory must be not only agent-neutral and time-neutral, but also world-neutral. Such a theory would be implausible in the extreme. A genuinely fully neutral version of Utilitarianism, for instance, would say that we ought to act so as to maximize the sum-total of happiness across all people, times, and possible worlds. But this sum-total of happiness is fixed regardless of what we do (if we increase happiness in our world, we thereby decrease it in another, simply by making the former rather than the latter actual), and so it would say that all acts are on a par. Of course, if genuine full neutrality is absurd, this might just provide support specifically for P (and other fully relative theories). But it might also make us doubt Parfit's contention that incomplete relativity is objectionable.

Let us set aside modality and world-neutrality. Why should the mere existance of a formal analogy between 'I' and 'now' mean that they should be treated alike by a theory of

rationality? Perhaps simplicity and elegance are theoretical virtues, in normative theorizing as in scientific theorizing. And pure theories are simpler and more elegant than hybrid ones. This yields a *pro tanto* reason for preferring pure theories over hybrid ones. Thus, other things being equal, we should opt for a pure theory over a hybrid one. This, I think, is the most plausible interpretation of Parfit's Appeal to Full Relativity.[7]

However, in Part Three Parfit provides further support for favouring pure theories over hybrid ones by arguing for reductionism about personal identity over time, according to which facts about personal identity over time are neither metaphysically deep nor normatively significant.

But it also makes the Appeal weak, in that an opponent could respond that other things are not equal. Perhaps personhood and time are different in sufficiently important respects as to outweigh the *pro tanto* reason for preferring pure theories over hybrid ones.

Brink (2011) responds along these lines, invoking the notion of *compensation* to break the analogy between 'I' and 'now.' If one person is harmed for the sake of providing a greater benefit to another person, the first person is not thereby compensated for the harm. This is the upshot of the so-called *separateness of persons* emphasized by Rawls (1971) and Nozick (1974). Nor is there any larger entity, like the mereological sum of the two persons, that is both benefactor and beneficiary, and thus automatically compensated for the harm. As Nozick (1974, 32-3) writes, 'there is no *social entity* with a good that undergoes some sacrifice

[7]Toward the end of Part Two, Parfit goes beyond the formal analogy between 'I' and 'now' and notes that with respect to some requirements of rationality, the relation between a person now and himself at other times *is* relevantly similar to the relation between different people' (191). For instance, while consistency of beliefs at a time is a requirement of rationality, it is no more irrational for one person to believe P and another to believe $\neg P$ than it is for a person to believe P at one time but believe $\neg P$ at another time. Similarly, it is a requirement of rationality that one's preferences be transitive at each time. It is thus irrational for a person at a time to prefer A over B, B over C, and C over A. But just as there need be no irrationality in one person preferring A over B and B over C while another prefers C over A, so there need be no irrationality in a person at one time preferring A over B and B over C while at another time preferring C over A. Does this observation bolster the Appeal to Full Relativity? I am doubtful. Parfit is correct that with respect to *some* requirements of rationality, the relation between a person now and himself at other times is relevantly similar to the relation between different people. But the Appeal claims that this is so with respect to *all* requirements of rationality. Thus far, then, we are left with just a *pro tanto* preference for pure theories over hybrid ones. In my view, the strongest arguments for favouring pure theories over hybrid ones come in Part Three, where Parfit argues for reductionism about personal identity over time. See also Hedden (2015).

for its own good.'

By contrast, when one person is harmed at an earlier time for the sake of a providing a greater benefit to the same person at a later time, the person is thereby automatically compensated for the earlier harm.[8] According to Brink, this fact—that compensation is automatic when benefactor and beneficiary are the same person, but not when they are different persons—justifies the hybrid structure of S.

Brink addresses an important objection to his suggestion. Might we go further than appealing to the separateness of persons and insist on the 'separateness of different periods within a person's life' (Brink 2011, 364)? We could then claim that 'me-now is [no] more compensated for its sacrifices on behalf of me-later than I am compensated by my sacrifices for you' (ibid). Compensation would then be no more automatic in cases of intrapersonal tradeoffs than in interpersonal ones.

Brink's own response is unsatisfactory in my view. Here is what he says:

> [T]his challenge to temporal neutrality requires thinking that we can and should adopt a sub-personal perspective when reckoning compensation. But there are problems with this idea. Once we go sub-personal and appeal to full relativity, there seems no reason to stop until we reach the sub-personal limit—a momentary time slice of the person. But notions of compensation have no application to momentary time slices, which do not persist long enough to act or receive the benefits of earlier actions. Moreover, many of the goods in life, especially the pursuit and achievement of worthwhile projects, seem to be realized only by temporally extended beings. (364)

I agree with Brink that if we go sub-personal in thinking about compensation, it would be arbitrary to stop short of the limiting case—momentary time-slices. But he is mistaken in thinking notions of compensation cannot be applied to such time-slices. First, he says that

[8]This is not to say that such compensation always justifies the harm. It may be wrong to impose the earlier harm if the person does not consent to it, even if the later and greater benefit constitutes compensation.

these momentary time-slices do not persist long enough to act. But while time-slices are not sufficiently long-lived to perform physical actions, it is less clear that they cannot perform mental actions like making decisions and forming intentions. More to the point, it seems that what is crucial is whether a given entity can *receive* compensation, and it is unclear why this should require the ability to act.

Second, he claims that many goods can be 'realized only by temporally extended beings.' But there is an ambiguity here. We must distinguish between which entities suffice to cause some good to exist and to which entities that good (understood as a unit of well-being) accrues. There are many goods that cannot be realized by a single person, such as friendship. But this does not prevent the corresponding good—the well-being associated with engaging in friendship—from accruing to each individual friend. Similarly, it may be that there are goods that require for their creation the existence and cooperation of many successive time-slices. But that does not prevent the corresponding well-being from accruing to each of those time-slices. And it is worth noting that none of the major theories of well-being seem to entail that well-being cannot accrue to time-slices. Time-slices can presumably have both phenomenal states and desires (meaning that they can have levels of well-being according to hedonistic and desire-fulfilment theories), and they can have at least some of the items mentioned by objective list theorists, such as knowledge and health.

So I think that Brink is incorrect in thinking that notions of compensation make no sense when applied to momentary time-slices. But a better response to the objection is available. The compensation theorist should insist that what is at issue, when one entity is harmed so as to give a greater benefit to another, is whether they comprise some larger entity that is both benefactor and beneficiary, and so is automatically compensated for the harm. Recall that, in arguing that it is impermissible to harm one person for the sake of benefiting another, Nozick emphasizes that 'there is no *social entity* with a good that undergoes some sacrifice for its own good' (see also Gauthier 1962, 126). The implication is that if there were such a social entity with the two persons as parts, it *would* be permissible to harm the one person

to benefit the other, for the social entity would be both benefactor and beneficiary; this is so even though the one person is not compensated by the benefit gained by the other. If that is correct, then it is irrelevant, for Brink's purposes, that me-now is not compensated when it is harmed so as to provide a greater benefit for me-later. What matters is rather that me-now and me-later are parts of a larger entity—me—which *is* compensated for the early harm by the later benefit. (Of course, we might well deny that the existence of a larger entity which is compensated when one of its parts is harmed for the sake of benefiting another makes this harm permissible. This would make Nozick's rejection of a social entity a *non sequitur* and would resuscitate the 'separateness of time-slices' objection to Brink's compensation argument.)

Where does that leave us? Suppose we reject this separateness of time-slices objection but take seriously the separateness of persons and the importance of compensation. In my view, this may still not suffice to justify S. First, there may be agent-neutral theories which prohibit harming one person just for the sake of providing a greater benefit to another (or, more generally, for the sake of maximizing the good). A consequentialism of rights (of which Parfit's N, an agent-neutral modification of Common Sense Morality, may be an example) might have this implication. It would allow one agent to be harmed, in violation of his rights, only for the sake of reducing the overall number (and magnitude) of rights violations that occur.

Second, the appeal to compensation may not justify time-neutrality as a rational *requirement*. Lack of compensation may make it impermissible to impose a harm on someone for the sake of benefiting another (without the former's consent), but it is far from clear that the presence of compensation can make it rationally required for one to undergo a harm for the sake of a greater benefit. A person might grant that it is him who will be compensated if he undergoes some harm now for some greater benefit later, but insist that it would nonetheless not be irrational for him to decline to do so. He might say that while the fact that he will be compensated for the harm makes it *permissible* to undergo the harm, it does not make

it rationally *required*. And time-relative theories can easily grant that doing so would be permissible. So while Brink has identified a potentially important disanalogy between 'I' and 'now,' facts about when compensation is automatic and when it isn't do not clearly justify either agent-relativity or time-neutrality.

5 Time-Bias

Amid the myriad fascinating observations about our attitudes to time, Chapter 8 includes perhaps the most powerful of Parfit's arguments against S. S gains some of its plausibility from its condemnation of what Parfit calls the *bias toward the near*, whereby we care more about our near futures than our far futures and are willing to trade a larger benefit in the farther future for a smaller one in the nearer future, and to trade a smaller harm in the near future for a larger one later on. This form of time-bias is apt to strike us as irrational, and it has been widely condemned by philosophers. If we accept that the bias toward the near is irrational, this lends some support to S's claim that time-neutrality is a requirement of rationality.

But Parfit notes that there is another form of time-bias as well, namely bias toward the future. We prefer that our pleasures be in the future and our pains in the past, even if this means a worse overall balance of pleasure and pain throughout our lives. This time-bias is illustrated in the case *My Past or Future 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. (165-6)

Parfit thinks not only that most (all?) of us have the bias toward the future, but that most of us think that it is rational. Indeed, many of us probably think that it is not just rationally permissible, but rationally required. (Parfit later argues that it would be better for us if we lacked the bias toward the future, but he notes, correctly, that an attitude can be rational even if it is bad for us.)

Clearly, if the bias toward the future is rational, this is a problem for S. But Parfit's discussion of exactly whether it is a problem is confusing. Here is how Parfit seems to view the matter (see p. 153, 157, and 163): The rationality of bias toward the future does not entail the falsity of S. Rather, it undercuts one motivation for S. The S-theorist must claim that we ought to give equal weight to the well-being of all our present and future time-slices. But he cannot support this claim by appeal to time-neutrality. He cannot claim that a mere difference in timing has no rational significance. After all, time-neutrality entails the irrationality of bias toward the future. And so the S-theorist must find some other motivation for the claim that it is irrational not to give equal weight to the well-being of all our present and future time-slices. In a nutshell, Parfit sees S as committed to the irrationality of the bias toward the near, but not to the irrationality of the bias toward the future. But it is *prima*

facie difficult to motivate the claim that the bias toward the near, but not the bias toward the future, is irrational (we will shortly consider a few attempts to do so).

In my view, Parfit is understating his case. If the bias toward the future is rational, then this simply entails that S is false. Recall that 'S gives to each person this aim: the outcomes that would be best for himself, and that would make his life go, for him, as well as possible' (3). S does *not* say that each person should aim for the outcomes that would be best for his present and future selves, and that would make his life go, for him, as well as possible in the future. It says that each person should aim for the outcomes that would be best for himself *simpliciter*, and that would make his life *simpliciter* go, for him, as well as possible. There is nothing in the statement of S that allows it to acknowledge the rational permissibility of bias toward the future.

Thus, Parfit should instead claim that the rationality of bias toward the future decisively refutes S, but that nonetheless one previously sympathetic to S could switch to defend a future-directed modification of S, S*. But S* will be difficult to defend, since the S*-theorist cannot condemn the bias toward the near by invoking full-blown time-neutrality, which would likewise implausibly condemn the bias toward the future.

What might the S*-theorist say in response? Is it possible to narrowly target the bias toward the near without overshooting and condemning the bias toward the future? Not obviously. Consider some ways to defend the bias toward the future. First consider *control*. The past is outside our control (we will shortly question this claim), while the future is at least somewhat within our control. But this is a bad justification for bias toward the future. It's not clear why we should prefer that our pains be in the past and hence outside our control, and moreover, as Parfit notes (p. 168), we are not unconcerned about future pains that are outside our control in the way that we are unconcerned about past pains. Worse, even if considerations of control could justify the bias toward the future, they might also justify the bias toward the near, since we often have more control over the near future than the far future.

Second consider *epistemology*. We have different, and usually better, epistemic access to facts about the past than to facts about the future. But again, it is not clear why this would justify the bias toward the future. Why should we prefer that our pains be such that we are in a better position to know about them, and our pleasures to be mired in uncertainty (Hare 2013)? Moreover, this consideration (if it were a good one) would implausibly justify the mirror image of the bias toward the near. Since we typically have better epistemic access to facts about the near future than about the far future, we would be led to prefer that our pains be in the near future and our pleasures in the far future, rather than *vice versa*.

Third consider the *metaphysics of time*. Parfit suggests that belief in the genuine passage of time, or 'objective becoming,' might help justify the bias toward the future. Without going too far into the details, he seems to be referring to the A-theory, on which properties like *being past*, *being present*, and *being future* are not reducible to relations like *being earlier than*, *being simultaneous with*, and *being later than*. On one version of the A-theory, *now* is like a 'moving spotlight' that illuminates different times with a special property—now-ness—as time passes.

But as Parfit notes (p. 180), it is not clear what the *argument* would be from the truth of the A-theory to the rationality of bias toward the future. Without such an argument, it is impossible to say whether the A-theory, if it supports the bias toward the future, would also support the bias toward the near.[9] And it is worth noting that popular ways of fleshing out the A-theory do not seem to support bias toward the future. Presentism, the view that only present things exist, would seem to support an absolute bias toward the present if it supports any form of time-bias at all. The growing block view, on which past and present things exist, with the block universe growing and adding more and more time-slices as time passes, would seem to support a bias toward the past if it supports any form of time-bias at all. As Hare (2013) notes, the only sort of A-theory that seems to have any chance of

[9]For starters, if *being past* and *being future* are special properties not reducible to relations of earlier than and later than, it would seem that *being in the near future* and *being in the far future* are likewise special properties not reducible to such relations.

supporting bias toward the future is the shrinking block view, on which present and future things exist, so that the block universe sheds time-slices and thereby shrinks as time passes. But the shrinking block view is believed by almost no one.

Parfit concedes that the S*-theorist could insist that 'in appealing to time's passage, we do not need arguments' and that it could be a 'fundamental truth that, since time passes, past suffering simply cannot matter—cannot be an object of rational concern' (180-1). This is of course true. One can always insist that some facts are brute. But such an appeal to brute facts would constitute a somewhat unattractive feature of S*.

Perhaps it was a mistake to abandon S in favour of S*. Instead, the S-theorist should stick to his guns and say that the bias toward the future, like the bias toward the near, is irrational. He could soften the blow by saying that the bias toward the future is harmless, since it can't be acted upon. And that helps explain why we mistakenly think that it is rationally permissible. (It would hardly explain why we think it is rationally *required*, however.) Alternatively, but equivalently for present purposes, he could restrict S to a claim about rational *action* rather than about rational *aims* and *preferences*; then, since the bias toward the future can't be acted upon, the S-theorist need neither condemn nor approve of it.

But it is false that the bias toward the future can't be acted upon. Parfit observes that on a desire-fulfilment theory of well-being, we *can* affect the well-being of our past selves (and of past people generally). If in the past I had a desire to skydiving at some point in my life, I can fulfil this desire, and boost the well-being of my past self, by going skydiving. And he argues, persuasively, that most of us are biased toward the future with respect to desire-fulfilment. We think that we may have *no reason* to try to fulfil such mere past desires.

Parfit (p. 163) suggests that the S-theorist could maintain that the bias toward the future cannot guide action by rejecting desire-fulfilment in favour of a hedonistic or objective list theory of well-being. But even here matters are complicated. Some of the items typically mentioned by objective list theorists are such that we can affect whether they were had by past selves. For starters, desire-fulfilment itself often appears on such lists. Consider also

accomplishment. If my earlier self laboured in obscurity to try to write a great novel, I can make it the case that my earlier self's labours were in service of a genuine accomplishment, namely by finishing the novel and getting it published. Next consider *knowledge*. Suppose my past self had a justified belief in P. I may be able to make it the case that this past self had knowledge, and not mere justified false belief, by making P true. Only if the S-theorist strikes items like desire-fulfilment, accomplishment, and knowledge from his objective list can he maintain, on an objective list view, that the bias toward the future cannot guide action.

It may seem obvious that on a hedonistic view, the bias toward the future is practically inert, since we cannot causally affect the happiness of past selves. This is true at least on phenomenal conceptions of happiness. But if we consider decision-making under uncertainty, the bias toward the future may nonetheless be action-guiding.

Dougherty (2011) argues that if we are both biased toward the future and risk averse, then this can affect how we act. For instance, consider a modification of Parfit's case from above:

> A coin was flipped to determine which of two surgery regimes you will undergo. If it landed heads, you will undergo the Early Course—4 hours of painful surgery on Tuesday and 1 hour of painful surgery on Thursday. If it landed tails, you will undergo the Late Course—no surgery on Tuesday and 3 hours of painful surgery on Thursday. It is now Wednesday (and you know this), but you have been given selective amnesia whereby you don't remember whether you had surgery on Tuesday. You are offered a gamble: If you accept the gamble, then if you are in the Early Course, your Thursday surgery will be extended by 30 minutes, while if you are in the Late Course, your Thursday surgery will be shortened by 30 minutes.

If you are risk averse, you will prefer options that better your worst-case scenario while worsening your best-case scenario, at least if doing so makes no difference to your *expected* utility. Now, if you are risk averse in this way and also future-biased, then you will accept the

gamble, while if you are risk averse but time-neutral, you will decline it. This is because what you deem your worst- and best-case scenarios differs depending on whether or not you are future-biased. If you are future-biased, your worst-case scenario is being in the Late Course, since that involves the most future pain. And accepting the gamble betters that worst-case scenario while worsening your best-case scenario, without affecting your expected total or future pain at all. So, being risk averse as well, you will accept the gamble. But if you are time-neutral, then your worst-case scenario is being in the Early Course, since that involves more total pain (albeit less future pain). And accepting the gamble will worsen that worst-case scenario while bettering your best-case scenario, without affecting your expected total or future pain at all. So, being risk averse, you will decline the gamble.

Thus, we have a case where, assuming you are risk averse, how it is rational for you to act depends on whether or not you are future-biased. Thus, it seems future-bias can be action-guiding even on a hedonistic view of well-being.

Greene and Sullivan (2015) object, arguing that we aren't risk averse with respect to pleasures and pains in the standard sense of having diminishing marginal utility for pleasures and pains. It is not the case that each additional unit of pleasure or pain matters less to us than the one that came before, as is the case with dollars, say. Note, however, that if we adopt Buchak's (2013) Risk-Weighted Expected Utility Theory, then Dougherty's argument goes through.

Similarly, if we accept evidential decision theory, then the bias toward the future can be action-guiding even given hedonism and even without any risk aversion. Evidential decision theory says that you should evaluate actions by looking at what evidence the performance of those actions would constitute for how good or bad things will be with respect to whatever you think matters (represented by a utility function). Formally, it says that you should perform the action with highest evidential expected utility, defined thus:

$$EEU(A) = \sum_i P(O_i \mid A) \times U(O_i)$$

where A is an action, the O_i are all the possible outcomes, $P(O_i \mid A)$ is your credence that

O_i is true, given that you perform A, and $U(O_i)$ is your utility for O_i.

Now suppose that there are two buttons in front of you. You're told that people who had lots of pain in their childhood (since forgotten) tend to push the left button, while people who had lots of pleasure in their childhood tend to push the right button. Pushing the right button thus gives one strong evidence that one had a pleasure-filled childhood, while pushing the left button gives one strong evidence that one had a pain-filled childhood. (Neither button is such that pushing it gives one evidence about how things will be in the future.) Therefore, whether evidential decision theory tells you to push the left button or the right one depends on whether or not you are biased toward the future. If you are, then it says that you can push either button. But if you are not biased toward the future, then it says to push the right one, since pushing the right button gives you strong evidence that things were good for you in the past (which, being non-future-biased, you care about) and gives no evidence either way about how things will be for you in the future. Thus, given evidential decision theory and hedonism, whether or not you are biased toward the future can make a difference to how you ought to act.

Thus, in order for the S-theorist to defend the claim that the bias toward the future cannot be acted upon, he must endorse both (i) hedonism or an objective list theory that doesn't include things like desire-fulfilment, accomplishment, or knowledge, and (ii) non-risk-weighted causal decision theory. This is not necessarily a bad consequence; the latter is perhaps the orthodox decision theory, and the former are not implausible theories of well-being. Nevertheless, it is important to be clear on what the S-theorist must commit to in order to claim that the bias toward the future cannot be acted upon.

We have been considering the option of the S-theorist maintaining, against our intuitions, that the bias toward the future is irrational, but softening the blow by claiming that it is practically inert and hence harmless. But we have seen that it is difficult to defend the claim that bias toward the future is in fact practically inert. This defeats the S-theorist's attempt to softening the blow that comes with claiming that bias toward the future is irrational.

But actually, the fact that the bias toward the future *can* be acted upon may actually help the S-theorist's case. This is because a policy of acting on one's bias toward the future can be predictably disadvantageous. In the article mentioned above, Dougherty (2011) shows that if one is biased toward the future and also risk averse, then one will sometimes perform a sequence of actions that is predictably disadvantageous in the following sense: it is foreseeable that at the time each member of that sequence is available, one will prefer to perform it rather than not, but at all times, one will prefer performing no member of that sequence over performing all of them. This is the same sort of diachronic inconsistency that is involved in violations of other purported requirements of rationality, such as transitivity of preferences and Bayesian Conditionalization. The details of the case are somewhat intricate, so I relegate an overview to a footnote.[10] For other arguments that bias toward the future is irrational because of the actions it recommends, see Greene and Sullivan (2015) and Dougherty (2015).

Let us step back. Common sense suggests that bias toward the near is irrational while

[10]Consider a slight modification of the surgery case above. The setup is the same: A fair coin it tossed to determine whether you will undergo the Early Course (4 hours of pain on Tuesday and 1 hour on Thursday) or the Late Course (none on Tuesday and 3 hours on Thursday). And you know throughout that on Wednesday you will be given amnesia which will make you forget whether or not you had surgery on Tuesday. In this new version of the case, you know that on Monday you will be offered Help Early and Help Late, respectively. If you take Help Early, then if you're in the Early Course it will reduce your Thursday surgery by 29 min, while if you're in the Late Course it will increase your Thursday surgery by 31 min. If you take Help Late, then if you're in the Early Course it will increase your Thursday surgery by 30 min, while if you're in the Late Course it will decrease your Thursday surgery by 30 min. If you take both pills, the result is simply that you will suffer one more minute of pain on Thursday than if you had refused both pills. Hence it seems irrational to take both. But Dougherty argues that if you are biased toward the future and risk averse, then on Monday you'll prefer to take Help Early and on Wednesday you'll prefer to take Help Late, regardless of what you do on the other day. Why? Taking Help Early reduces the difference between the highest and lowest amounts of possible future pain while increasing your expected future pain by only one minute. (This preference for reducing the difference between the worst-case and best-case scenarios at a small cost to expected value is characteristic of risk aversion.) And taking Help Late reduces the difference between the highest and lowest amounts of possible future (i.e. future relative to Wednesday) pain without any change in expected future pain.

Thus, Dougherty argues that the combination of bias toward the future and risk aversion is irrational since it yields the threat of diachronic inconsistency—on Monday you prefer to take Help Early, on Wednesday you prefer on to take Help Late, but on both days you prefer rejecting both pills over taking both of them. He regards risk aversion as clearly rational and therefore lays the blame at the feet of bias toward the future.

As noted, Greene and Sullivan (2015) reject Dougherty's argument on the grounds that we are not risk averse with respect to pleasures and pains in the sense of having diminishing marginal utility for pleasures and pains. It is true that bias toward the future can still yield diachronic inconsistency if one is risk averse in the sense of following Buchak's Risk-Weighted Expected Utility Theory. But Risk-Weighted Expected Utility Theory yields diachronic inconsistency on its own (see Briggs 2015 and Buchak 2015 for discussion). So this diachronic inconsistency may not constitute a good argument against bias toward the future in particular.

bias toward the future is rational. This constitutes a *prima facie* problem for the S-theorist, who must condemn both. It seems to me that the best approach for the S-theorist is to argue that bias toward the near and the bias toward the future are in fact both irrational, and moreover irrational for the same reason, namely that they sometimes recommend courses of action that predictably makes one's life as a whole go worse.

This is a powerful argument. But I do not think it is fully convincing. First, attitudes can be predictably disadvantageous despite being rational, or predictably advantageous despite being irrational; exaggerated self-confidence may be greatly beneficial, but no less irrational for it. Indeed, Parfit himself argues that we would be better off if we were not biased toward the future, since it would yield more equanimity in the face of death (see Section 74). But he stops short of condemning the bias toward the future as thereby irrational.

Second, in arguing that bias toward the future (and bias toward the near) are irrational because they can predictably make one's life as a whole go worse, the S-theorist is begging the question against his opponent. An opponent who does not think that one must be concerned principally for how one's life goes on the whole, and who cares more about some parts of his life than others, will not be convinced. And a defender of full neutrality will agree with the S-theorist's condemnation of bias toward the future on the grounds that it can predictably make one's life go worse overall. But he will go further and condemn the bias in one's own favour on analogous grounds. For example, if two people are both biased in their own favour, then they will sometimes be led to perform a *combination* of actions that is predictably bad for both, in the sense that each would be worse off if each performs his part of that combination of actions than if neither does. The Prisoner's Dilemma is one such example. See Parfit (1984, 187-191) and Hedden (2015, Ch. 7) for further discussion.

6 Conclusion

We have looked at Parfit's main arguments in Part Two against the Self-interest Theory, S. In my view, his arguments are, on the whole, successful, despite various troubles. But Part

Two contains a multitude of observations about our attitudes to time that are fascinating and important independently of the debate over S, for instance about how our attitudes about the importance of past desires depends on whether they were based on deeply held values or mere tastes, about how our other-directed time-biases depend on our proximity to, and ability to communicate with, our loved ones. And his discussion of the relationship between rationality and time continues in Part Three, where he discusses the implications that the metaphysics of personal identity over time has for the rational permissibility of time-bias.

References

Briggs, R.A. 2015. 'The Costs of Abandoning the Sure-Thing Principle.' *Canadian Journal of Philosophy* 45 (5):827-840.

Brink, D. 2011. 'Prospects for Temporal Neutrality.' In Callender, C. (ed). *Oxford Handbook of the Philosophy of Time*. New York: Oxford University Press, pp. 353—381.

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty, and Time*. Wiley-Blackwell.

Buchak, L. 2013. *Risk and Rationality*. New York: Oxford University Press.

Buchak, L. 2015. 'Revisiting Risk and Rationality: A Reply to Pettigrew and Briggs.' *Canadian Journal of Philosophy* 45 (5):841-862.

Dougherty, T. 2011. 'On Whether to Prefer Pain to Pass.' *Ethics* 121 (3):521—537.

Dougherty, T. 2015. 'Future-Bias and Practical Reason.' *Philosophers' Imprint* 15 (30): 1-16.

Dreier, J. 1996. 'Rational Preference: Decision Theory as a Theory of Practical Rationality.' *Theory and Decision* 40 (3): 249-276.

Gauthier, D. 1962. *Practical Reasoning*. Oxford: Clarendon Press.

Greene, P. and Sullivan, M. 2015. 'Against Time-Bias.' *Ethics* 125 (4):947—970.

Hare, C. 2013. 'Time: The Emotional Asymmetry.' In Bardon, A. and Dyke, H. (eds) *A Companion to the Philosophy of Time*. Wiley-Blackwell. pp. 507—20.

Hare, C. and Hedden, B. 2016. 'Self-Reinforcing and Self-Frustrating Decisions.' *Noûs* 50 (3):604—628.

Hedden, B. 2015. *Reasons without Persons: Rationality, Identity, and Time*. Oxford: Oxford University Press.

Kagan, S. 1986. 'The present-aim theory of rationality.' *Ethics* 96 (4): 746—759.

Nagel, T. 1970. *The Possibility of Altruism*. Princeton: Princeton University Press.

Nozick, R. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Parfit, D. 1986. 'Comments.' *Ethics* 96 (4): 832–72.

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

Sidgwick, H. 1907. *The Methods of Ethics*, 7th Edition. London: Macmillan.

Street, S. 2009. 'In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters' *Philosophical Issues* 19 (1): 273-298.