Parfit on Personal Identity and Ethical Theories

JUSSI SUIKKANEN

Final author copy; To be published in *Oxford Studies in Normative Ethics*

Abstract: In his early works, Derek Parfit famously defended revisionary reductionism about personhood. According to this view, facts about personal identity consist in the holding of more particular psychological facts, which can be described wholly impersonally. He also argued that, in some cases, the truth of this view makes questions about diachronic personal identity empty questions to which no meaningful answers can be given. Yet, in his later works, Parfit defends several ethical theories such as contractualism and rule-consequentialism, which seem to rely on exactly the kind of determinate notion of personal identity to which he objected earlier. Parfit, furthermore, never explored reductionism's consequences for such theories in his later works. In order to solve this interpretative puzzle, this chapter tries to argue that, even if they are seemingly conflicting, Parfit's views on personal identity and rule-consequentialism, Scanlonian contractualism, and Kantian contractualism do form a coherent and unified whole.

Keywords: Contractualism; Ethical Theories; Derek Parfit; Personal Identity; Revisionary Reductionism; Rule-Consequentialism.

1. An Interpretative Puzzle

Derek Parfit published two hugely influential books: *Reasons and Persons* (hereafter '*R&P*', in 1984) and *On What Matters* (hereafter '*OWM*', Vols. 1 and 2 published in 2011, Vol. 3 in 2017). If we want to understand Parfit's philosophy as a whole, we need to consider the connection between these two books. Do their theories fit nicely together, is there perhaps evidence of changes in Parfit's views, or can we find inconsistencies between the views and arguments put forward in them?

Addressing these questions comprehensively would require a whole book, and so this chapter focuses, instead, only on just one connection. Part 3 of *R&P* is celebrated for its investigation of personal identity.[1] Here Parfit defends revisionary reductionism about personhood according to which (i) facts about personal identity consist in the holding of more particular psychological facts, which (ii) can be described in an impersonal way (*R&P*, 210). Parfit then argues that this view has radical, revisionary consequences. It implies both that, sometimes, questions about personal identity have no determinate answers and that identity should not matter when we consider our own survival. Parfit furthermore suggests that reductionism has important implications for first-

[1] This part itself is a further development of Parfit (1971).

order ethical theorizing as it, for example, helps the utilitarians to address the classic objections based on the intuitive idea of separateness of persons.

Yet, in *OWM*, Parfit never explores reductionism's implications for ethical theories further – personal identity is not even mentioned. Furthermore, in *OWM*, instead of act-utilitarianism Parfit defends his versions of rule-consequentialism, Scanlonian contractualism, and Kantian contractualism. Especially the last two theories have, however, always been assumed to rely on exactly the kind of separateness of persons that *R&P* rejects. This suggests that, at least *prima facie*, Parfit's views in *R&P* and *OWM* seem to conflict with one another (Hédoin 2021, sec. 1). This, I believe, is an interesting interpretative puzzle concerning Parfit's philosophy. One simple way around this problem would be to think that Parfit changed his views on personal identity between *R&P* and *OWM*, but this interpretation is questionable because Parfit (2012) continued to defend those views in print (see Hédoin (2021, sec. 1)).[2]

This chapter argues that Parfit did not need to change his views on personal identity because the *prima facie* tension between *R&P*'s discussion of personal identity and *OWM*'s ethical theories is not genuine. If we look more closely, we can see that Parfit argues, on first-order ethical grounds, against the versions of Kantian and Scanlonian contractualism that assume separateness of persons, whereas the versions he defends are wholly compatible with rejecting that thesis. This is why *OWM*'s ethical theories cohere nicely with *R&P*'s radical claims about personal identity, and in fact are supported by them (though Parfit never makes such an argument). It is therefore not accidental that Parfit defends his own, original versions of the three ethical theories of *OWM*: he had to or his later views in normative ethics would have conflicted with his earlier work on personal identity. In addition to helping us to understand how Parfit's views on different topics fit together, this conclusion also enables us to appreciate the constraints which reductionism about personhood sets for those ethical theories that have traditionally relied on separateness of persons.

This is how this chapter will proceed. §2 briefly outlines Parfit's revisionary reductionism. §3 then considers how that view and Parfit's rule-consequentialism fit together. It argues that, even if Parfit's rule-consequentialism is compatible with his views on personal identity, those views still impose a significant constraint on how rule-consequentialism is to be formulated. §4 argues that, even if T.M. Scanlon's contractualism conflicts with Parfit's revisionary reductionism, Parfit's modifications to Scanlon's view mean that the resulting version of Scanlonian contractualism is compatible with Parfit's reductionism about personhood. §5 similarly suggests that, even if the traditional formulations of Kantian ethical principles may have assumed separateness of persons, Parfit's own version of Kantian contractualism does not do so. §6 concludes by considering why Parfit chose not to rely in *OWM* on arguments based on personal identity.

2. Parfit on Personal Identity

This section summarises *R&P*'s revisionary reductionism.[3] Parfit (*R&P*, 210) understands the view as the following theses:

[2] Parfit briefly adopted animalism about personhood in the 1990s before reverting to a more sophisticated version of his previous view (Edmonds 2023, 236).

[3] Before Parfit, similar positions had been defended by Shoemaker (1959) and Wiggins (1967, part 4).

> (i) the fact of a person's identity over time consists in the holding of certain more particular facts [and]
>
> (ii) these facts can be described without either presupposing the identity of this person […] or even explicitly claiming that this person exists. These facts can be described in an *impersonal* way.

Reductionism is thus opposed to non-reductionism according to which facts about identity involve some further facts, for example, about Cartesian Egos or souls, which cannot be described impersonally.

What are the underlying facts of which the facts about diachronic personal identity consist on this view? According to physical reductionism, identity over time depends on having 'enough' of the same body or brain or on the continuation of a single biological life (Olson 1999). Parfit, however, prefers psychological reductionism (*R&P*, secs. 78–79). On his view, facts about a person's identity over time consist of the holding of the psychological Relation R, which itself consists of psychological connectedness and continuity. For two person-stages to be psychologically connected, there must be direct psychological relations between them, direct connections of quasi-memories, intentions, character, beliefs, values, and so on. Psychological connectedness thus comes in degrees, and for example to be 'strongly connected' requires at least half the number of connections that hold 'over every day, in the lives of nearly every actual person' (*R&P*, 206). Psychological continuity then holds between person-stages in virtue of the overlapping chains of strong connectedness between them. Given that (ii) is a central part of reductionism, these psychological connections that constitute the Relation R (and hence also the facts about diachronic personal identity) must be describable without any presuppositions of the relata, the underlying mental states, being states of the same person.

Parfit argues that reductionism has radical, revisionary consequences for everyday thinking. Firstly, it makes certain questions about personal identity empty questions to which there are no determinate answers (*R&P*, 241). Reductionists should thus not try to settle these questions by comparing different physical or psychological criteria of personal identity. Secondly, Parfit also argues that, in practical deliberation, what should matter to us with respect to survival is not identity-relations but rather psychological connectedness and continuity with any cause (*R&P*, chs. 12–13).

Parfit uses two cases to motivate these claims. The Combined Spectrum case concerns a range of conditions we can imagine obtaining in succession (*R&P*, §86).[4] In the first state, there is a future person, who is both physically and psychologically continuous with me now. We can, however, imagine that, in the next state, a few of my body's cells (including some brain cells) have been replaced by cells similar to Greta Garbo's cells. The resulting person will be almost like me, but they will also have some physical similarities to Garbo and some quasi-memories of living Garbo's life (plus some other mental states similar to hers). Thereafter, at each stage, few more of my cells are replaced by Garbo-like cells, and so the following person-stages are gradually less like

[4] Sorensen (1988, 250–252) objects to this argument based on epistemicism about vagueness, but, for a defence of Parfit's reasoning, see Alter and Rachels (2004).

my now and more like Garbo, until the last person-stage is wholly like Garbo both physically and psychologically.

Parfit uses this case to argue against non-reductionism and for reductionism. If you are a non-reductionist and think that facts about personal identity are grounded in Cartesian egos or souls, then for you there must be a sharp borderline in the previous spectrum where the next person-stage is no longer you. At this point, minute physical and psychological differences between the two stages would constitute the difference between life and death (even if we could never locate those differences).

Parfit's argument for reductionism is that it can avoid the previous absurd consequence. Reductionists can claim that the first stages will be the future me, the last stages will definitely not be me, and in between there are various person-stages where it is indeterminate whether they are future stages of me or not. Here, according to Parfit, we can know all the impersonal facts about physical and psychological continuity, but any further questions about identity-relatedness are empty questions to which no meaningful answers can be given.

The second case is the famous fission case (*R&P*, 254–5):

> *My Division*. My body is fatally injured, as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes that he is me, seems to remember living my life, has my character, and is in every other way psychologically continuous with me. And he has a body that is very like mine.

Here the non-reductionists seem to have three options (*R&P*, sec. 89). They could claim that (i) I don't survive, (ii) that I am identical to one of the resulting people (either 'Lefty' or 'Righty'), or (iii) that I survive as both Lefty and Right. The argument against non-reductionism then is that all these options are problematic. The problem with (i) is that intuitively you can survive if your brain is successfully transplanted to another body, and people have survived when half of their brain has been destroyed. But, if this is right, it would be odd if you could not survive when both halves of your brain are successfully transplanted to different bodies. (i) would, furthermore, imply that fission would be as bad as death.

The problem with (ii) is that the choice between whether Lefty or Righty is me would be arbitrary given that they are qualitatively identical. Finally, (iii) is problematic because, if I am identical with both Lefty and Righty, then by transitivity of identity Lefty and Righty would be identical with each other. Yet, this cannot be true if they remain distinct persons. Thus, the only coherent way to accept (iii) would be to think that after the division there is just one person with two bodies and 'two separate spheres of consciousness.' According to Parfit (*R&P*, 257), this would, however, distort the ordinary concept of personhood beyond recognition.[5]

Reductionism, in contrast, avoids these problematic options. The question of what happens to me in My Division becomes an empty question to which no determinate answer can be given as we know all the facts there are to be known about the case before that question is even raised.

[5] According to Lewis's (1976) response to Parfit, in My Division the Relation R-related person-stages form two four-dimensional persons that share some of their pre-fission person-stage parts. For Parfit's response, see Parfit (1976).

Parfit also argues that division is as good as ordinary survival. From this he concludes that identity does not matter in survival but rather the Relation R, which, unlike identity, can branch from one to many in the previous kind of cases (*R&P*, sec. 90).

Let me conclude this outline of reductionism with two observations. Firstly, in *R&P* (secs. 111–117) Parfit considers reductionism's consequences only to one ethical theory, utilitarianism – the view that right actions maximise the total amount of wellbeing.[6] One classic objection to utilitarianism has always been that it ignores the separateness of persons by aggregating benefits and burdens across different persons whilst being blind to how they are distributed to different persons.[7] In *R&P*, Parfit argues that reductionism undermines the force of this objection for three reasons.[8]

Firstly, those who object to utilitarianism based on separateness of persons tend to allow persons to maximise the total sum of their own well-being over their lifetimes. Parfit (*R&P* 334–5) then claims that '[i]f the unity of a life is less deep, it is more plausible to claim that this unity is not what justifies [such] maximisation.'[9] Rather, because reductionism leads to a partial disintegration of persons, we should focus in practical deliberation on 'selves', that is, person-stages without much temporal extension. This suggests that questions of distributive justice should focus on distribution between all person-stages that are R related rather than on distribution between different persons. As an impersonal view, this gets us, according to Parfit, closer to utilitarianism.

Secondly, Parfit also thinks that reductionism undermines the separateness of persons on which the traditional objections to utilitarianism are grounded. As he puts it (*R&P*, 281):

> There is still a difference between my life and the lives of other people. But the difference is less. Other people are closer. I am less concerned about the rest of my life, and more concerned about the lives of others.

Parfit's thought here is that, due to reductionism, we have the same significant psychological relations to other people around us as we have to our own future selves. Just as we are in the Relation R to our own future person-stages, we are in this same relation to other people, which reduces the gap between us and them. As a consequence, we can no longer rely so much on the idea of separateness of persons to block the utilitarian inter-personal aggregation.

Finally, Parfit's third argument uses reductionism itself to challenge the moral significance of the remaining unity of individuals and separateness of persons. Reductionism is, after all, the view that a person's existence can be described impersonally by stating the underlying psychological and physical facts. Beyond those facts, there are no further facts about persons based on Cartesian Egos or souls. Yet, because a person's life consists of merely psychological and physical events that can be captured impersonally, there is no reason to care especially about the facts that have to do with personal identity. There just is less to being a person that could make a moral difference.[10]

---

[6] He also considered the views implications for more specific ethical issues such as paternalism and autonomy, abortion, promises and commitments, and retribution and desert (*R&P*, secs. 107–110).

[7] See, e.g., Gauthier (1963, 125–126), Rawls (1999, 23–26), and Nozick (1974, 32–33).

[8] For a critical discussion of these arguments, see Steuwer (2020)

[9] For objections according to which something else than Relation R, such as agency, can unify persons, see Korsgaard (1989), Blackburn (1997) and Brink (1997).

[10] Wolf (1986, 705–708), Adams (1989, 454–460), and Johnston (1997, 159) argue that, even if reductionism were true, this would not diminish the importance of persons. For a response, see Parfit (1995, 29–31).

Based on these considerations, Parfit concludes (*R&P*, 346):

> This [i.e., reductionism] gives some support to the Utilitarian View, making it more plausible than it would have been if the Non-Reductionist View had been true… The impersonality of Utilitarianism is therefore less implausible than most of us believe.

R&P thus argues that reductionism about personhood can undermine the objections to utilitarianism that draw from the separateness of persons, which makes the view more compelling. Yet, in *R&P*, Parfit never considers reductionism's consequences for other ethical theories: there are no discussion of what the view would entail for Kantian ethics, contractualism, or rule-consequentialism.

The second observation I want to make is that there are, of course, many objections to Parfit's reductionism and his arguments for it.[11] Yet, this chapter is only interested in the internal coherence of Parfit's views. Therefore, it is not relevant here whether Parfit's reductionism really is true, but rather all that matters in this context is how the view fits Parfit's discussions of the three ethical theories in *OWM*. I will therefore not attempt to defend reductionism here.

3. Parfit's Rule-Consequentialism and Personal Identity

This section investigates the relation between *R&P*'s reductionism and *OWM*'s formulation of rule-consequentialism. Parfit (*OWM*, Vol. 1, 375) formulates rule-consequentialism in the following way:

> (F) everyone ought to follow the principles whose universal acceptance would make things go best.

On this view, actions are right when they are authorised by the previous 'optimific' principles and wrong otherwise. I will assume both standard rule-consequentialist understandings of the relevant principles and universal acceptance and that these elements do not raise any questions concerning personal identity.[12] We can, for example, take universal acceptance to mean the circumstances in which (almost?) all person-stages accept the relevant principles, whilst remaining neutral on which of those person-stages are identity-related. The part about making things go best, however, raises questions about personal identity and so needs to be investigated further.

In (F), the principles whose universal acceptance would make things go best are the principles the universal acceptance of which would bring about an outcome that would be impersonally better than the outcomes brought about by the universal acceptance of any other principles. Parfit defines the impersonally best outcome as the outcome 'that, from an impartial point of view, everyone would have most reason to want, or hope will come about' (*OWM*, Vol. 1, 372). The impartial point of view is a hypothetical perspective from which we consider different outcomes as if they would only involve strangers. This enables us to exclude reasons grounded in our own self-interests, personal relationships, and agential involvement.

---

[11] In addition to footnotes 4–5 and 9–10 above, see, e.g., Cassam (1993) and Merricks (1997).

[12] See, e.g., Hooker (2000, §3.3 and §3.5).

This might suggest that Parfit's rule-consequentialism cannot conflict with his reductionism about personhood, but this is not necessarily so. The problem is that the previous sense of 'best' 'leaves it entirely open which are the ways in which we would have most reason to want things go' (*OWM*, Vol. 1, 374). Because of this, (F) enables us to formulate both versions of rule-consequentialism that cohere with reductionism about personhood and ones that conflict with it. I will next introduce a simple version of the former kind of views and then an example of the latter kind of views. After this, we can finally consider whether *OWM*'s rule-consequentialism conflicts with *R&P*'s reductionism about personhood.

Let us assume a specific first-order theory of the reasons we would have from the impartial point of view. According to this view, only facts about wellbeing provide us with impartial reasons to prefer an outcome. There are, however, two ways in which we can aggregate every individual's momentary levels of wellbeing together to form the total amount of wellbeing an outcome contains: the 'people route' and the 'snapshot route' (Broome 2004, 104). In the snapshot route, the total amount of wellbeing a given outcome contains at each distinct time is determined first by aggregating together each person-stage's momentary level of wellbeing at that time. How much total wellbeing the outcome contains is then the sum of these snapshots of aggregate wellbeing. This route, therefore, assumes both that individuals have momentary levels of wellbeing and that the amount of wellbeing a time-slice contains is independent of the other times.[13] If our impartial reasons for preferring the universal acceptance of certain principles are based on how much total wellbeing, determined in this way, the acceptance of those principles brings about, the resulting version of rule-consequentialism is wholly compatible with Parfit's reductionism. No facts about which person-stages are identity-related are assumed, but rather the relevant reasons to prefer some outcomes over others are wholly provided by the underlying physical and psychological facts described impersonally.

It is, however, equally easy to formulate versions of rule-consequentialism that conflict with reductionism.[14] In the people route, each person's total lifetime wellbeing is first determined by aggregating together the levels of wellbeing of their person-stages. The total amount of wellbeing an outcome contains is then the sum of the total amounts of wellbeing which each person's life contains. This view too might seem to be compatible with Parfit's reductionism. There are, as we have seen, different reductionist views of which person-stages are identity-related. However, as long as every person-stage counts as a part of some person's life and no person-stage counts as a part of many different persons' lives, the views that take the 'people route' to aggregation would seem to come to the same conclusions about the total value of outcomes as the previous snapshot route view. After all, the well-being of each person-stage would be aggregated to the total sum, and no person-stage would be double-counted. Yet, by changing our assumptions, we can also formulate versions of the person route views that will conflict with reductionism about personhood.

Let us first assume that how impersonally good a person's life is in total is not determined merely by the simple sum of the wellbeing contained in the different stages of the life but also by both the

[13] It can be argued that this separability of times thesis follows from Parfit's reductionism. See Broome (2004, §15.2).

[14] See also Shoemaker (1999, §5).

shape and length of the life. On this view, longer lives are disproportionally better than shorter ones and an improving life is better than a deteriorating one even if otherwise the total amount of momentary wellbeing is the same in both lives.[15] Let us also imagine that we are comparing, from the impersonal point of view, outcomes that contain My Division-like cases.

Under these assumptions, questions about personal identity would have consequences for how good outcomes would be from the impersonal point of view. Consider the following four distributions of lifetime wellbeing in My Division (where 'Ω' stands for non-existence):

A) I do not survive:

Me = (3, 3, Ω, Ω)

Lefty = (Ω, Ω, 2, 1)

Righty = (Ω, Ω, 4, 5)

B) I survive as Lefty:

Me = (3, 3, 2, 1)

Righty = (Ω, Ω, 4, 5)

C) I survive as Righty:

Me = (3, 3, 4, 5)

Lefty = (Ω, Ω, 2, 1)

D) I survive as both:

Me = (3, 3, 6, 6)

In all these scenarios A)–D), the fundamental physical and psychological facts are the same. It's just that, in these evaluations, we are assuming different views of which person-stages are identity-related. We are also assuming a holistic axiology according to which the length and the shape of a life matters to how good the life is impartially speaking. The numbers in these tables then represent how much wellbeing a given person-stage contains.

Under these assumptions, which person-stages we take to be identity-related affects how good an outcome is from the impartial perspective. If I do not survive, there are three short lives: one improving, one deteriorating, and one that does neither; if I survive as Lefty, there is one long deteriorating life and one short improving life; if I survive as Righty, there is one long improving life and one short deteriorating; and if I survive as both, there is one long life that has a sudden jump in wellbeing in the middle. These alternatives contain the same amount of total wellbeing (18 units), but because we are assuming that the lengths and shapes of lives matter these outcomes will not be equally good from the impartial perspective. C) could, for example, be a better outcome

[15] See, e.g., Slote (1982), Velleman (1991), and Dorsey (2015).

than the others because it contains one long gradually improving life and just one short deteriorating one.

Yet, because Parfit's reductionism entails that it is an empty question whether Lefty, Righty, both, or neither is identity-related to me before the division, we cannot answer the question of whether A), B), C), or D) would be the correct evaluation of the case assuming the previous holistic axiology. Because of this, we could not evaluatively compare the moral code that is responsible for this distribution of wellbeing to person-stages to other codes. If reductionism is true, there just is no fact of the matter of how the codes would rank. This example illustrates more generally that any axiology (i.e., a theory of the impartial reasons to prefer outcomes) according to which facts about identity-relations can make a difference to the value of outcomes conflicts with *R&P*'s reductionism about personhood. Axiologies according to which the shape and length of a life makes an evaluative difference are, as we have seen, examples of such axiologies.

Does Parfit, in *OWM*, take identity-relations to make a difference to how good outcomes are in the impartial sense and thus does his own rule-consequentialism thereby conflict with his earlier reductionism about personhood? Parfit never addresses this question explicitly, but there are some passages that suggest a conflict. For example, in his discussion of rule-consequentialism, Parfit (*OWM*, Vol. 1, 373–4) claims that:

> the goodness of some outcomes might depend in part on facts about the past. It might be better, for example, if benefits went to people who had earlier been worse off, or if we kept our promises to those who are dead, or if people are punished only if they earlier committed some crime.

If rule-consequentialists adopted such an axiology, their view would conflict with Parfit's reductionism about personhood. Imagine I committed a crime before my division and Lefty is punished for this crime after the division. According to B) and D) above, the resulting outcome would be good impartially because a guilty person has been punished, whereas according to A) and C) this would be a bad outcome because an innocent person has been punished. However, according to reductionism it would again be an empty question which of these evaluations would be correct. This suggests that the previous axiology would conflict with Parfit's reductionism about personhood. It would lead to us having to ask the kind of allegedly empty questions that Parfit believed cannot be answered.[16]

Parfit's own reductionism in *R&P* thus sets a constraint on what kind of an account of the impartial value of outcomes he could accept for the purposes of his rule-consequentialism in *OWM*. To avoid any inconsistencies, his rule-consequentialism cannot allow any alleged identity-relations between person-stages affect how good different outcomes are impartially. It is not clear whether Parfit always respected this constraint in *OWM*'s discussions of rule-consequentialism, but there would have been a way for him to avoid this problem. He merely would have needed to restrict

[16] In *R&P* (ch. 109), Parfit considers the view according to which desert should not track identity-relations but rather psychological continuity. This view would be compatible with reductionism, but it would entail that all the future person-stages who are psychologically continuous with the person-stage that committed the crime would deserve at least some degree of punishment. More generally, in *OWM*, Parfit's rule-consequentialism could also accept axiologies according to which the degree to which person-stages are R related makes an evaluative difference (for example, in the contexts of promise-keeping, benefitting the worst off, and so on). In this way, the view could be argued to be able to accommodate our intuitions about at least real-world cases.

the kind of things that can affect the value of outcomes to facts that can be described impersonally.[17]

4.Parfit's Scanlonian Contractualism and Personal Identity

Let us next contrast *OWM*'s Scanlonian contractualism to *R&P*'s reductionism. This section first introduces Scanlon's (1998) own contractualism and explains why it conflicts with reductionism about personhood. It then considers Parfit's amendments to the view and how Parfit's amended Scanlonian contractualism better coheres with reductionism.[18]

According to Scanlon's (1998, chs. 4–5) contractualism, an action is right if and only if it is authorised by the principles no one could reasonably reject (and wrong otherwise). Reasonable rejection is a function of how strong personal objections individual persons can make to different principles from their personal standpoints. Individuals are allowed to aggregate *intra-personally* burdens that they have to bear at different times in their lives, but they cannot aggregate other people's burdens to their own to form stronger objections to a given principle (Scanlon 1998, 237). The non-rejectable principles are then such that there are stronger intra-personally aggregated personal objections to all alternatives to those principles.

This view conflicts with Parfit's reductionism about personhood. Consider the following division case in which we are comparing Principle P to Principle Q to see which of these principles could not be reasonably rejected, whilst using numeric values to represent momentary personal burdensomeness of living under a given principle at a specific time:

E) I do not survive:

Under Principle P:	Under Principle Q:
Me = (5, 5, Ω, Ω)	Me = (1, 1, Ω, Ω)
Lefty = (Ω, Ω, 5, 5)	Lefty = (Ω, Ω , 1, 1)
Righty = (Ω, Ω, 5, 5)	Righty = (Ω, Ω, 1, 1)
Ann = (1, 1, 1, 1)	Ann = (5, 5, 5, 5)

F) I survive as Lefty:

Under Principle P:	Under Principle Q:
Me = (5, 5, 5, 5)	Me = (1, 1, 1, 1)

[17] This presumably has first-order consequences for how well Parfit's rule-consequentialism can fit our commonsense moral intuitions.

[18] David Shoemaker (2000) has argued that Parfit's reductionism about personhood can be used to support Scanlon's contractualist account of moral motivation if we take the social contract to be between selves rather than persons (which fits Parfit's reformulation of Scanlonian contractualism below). For objections to Shoemaker's proposal and for an alternative, see Hédoin (2021).

Righty = (Ω, Ω, 5, 5)	Righty = (Ω, Ω, 1, 1)
Ann = (1, 1, 1, 1)	Ann = (5, 5, 5, 5)

G) I survive as Righty:

Under Principle P:	Under Principle Q:
Me = (5, 5, 5, 5)	Me = (1, 1, 1, 1)
Lefty = (Ω, Ω, 5, 5)	Lefty = (Ω, Ω, 1, 1)
Ann = (1, 1, 1, 1)	Ann = (5, 5, 5, 5)

H) I survive as both:

Under Principle P:	Under Principle Q:
Me = (5, 5, 10, 10)	Me = (1, 1, 2, 2)
Ann = (1, 1, 1, 1)	Ann = (5, 5, 5, 5)

In these scenarios E)–H), the fundamental underlying physical and psychological facts are again the same. The only difference is which person-stages are taken to be identity-related. Let us then assume that persons are allowed to intra-personally aggregate their burdens to more serious personal objections, and that the non-rejectable principle is the one to which there is the weakest most serious such personal objection.

In scenario E) in which I do not survive the division, the non-rejectable principle is P because Ann's intra-personally aggregated objection (20 units) to Q is stronger than anyone's objection to P as I, Lefty, and Righty all have merely 10 unit personal objections to P. In F) and G), it is initially not clear which principle is not reasonably rejectable because my objection to P (20) is just as strong as Ann's objection to Q (20). If we take the second strongest objections to be decisive, in both cases the non-rejectable principle is Q as Righty's objection to P in F) and Lefty's objection to P in G) (10 units) are both stronger than my objection to Q (4). Finally, in H, the non-rejectable principle is Q, as my objection to P (30 as here I am identical to both Lefty and Righty and so I can intra-personally aggregate their momentary burdens to my own) outweighs Ann's objection to Q (20).

The problem this scenario illustrates is that Scanlon's view entails that which person-stages are identity-related in division cases determines which principles cannot be reasonably rejected. Yet, according to Parfit's reductionism questions about identity-relations in these cases are empty questions to which no determinate answers can be given. Therefore, because Scanlon's view relies on the kind of unity and separateness of persons assumptions that must be given up according to Parfit's reductionism, it leads to unanswerable questions about which principles cannot be reasonably rejected and hence also to unanswerable questions about which actions are right and

wrong.[19] This why Scanlon's view conflicts with *R&P*'s reductionism, and hence why, in *OWM*, Parfit could not have adopted Scanlon's version of contractualism.

Yet, instead of Scanlon's contractualism, in *OWM* Parfit defends a version of *Scanlonian* contractualism, but this is because he modifies Scanlon's view based on first-order ethical considerations. Parfit's key case here is the *Case Six* (*OWM*, Vol. 3, 200):

	Blue will live to the age of	Each of some number of other people live to
We do nothing	30	30
We do A	70	30
We do B	35	35

According to Parfit, on Scanlon's view we would have to do A because Blue's personal objection to B (loss of 35 years of life) is stronger than anyone else's objection to A (loss of 5 years). Yet, intuitively we should choose B both (i) because it provides more years of life (as long as the group contains more than 7 people), and (ii) because it leads to a fairer distribution of goods between different people.

From this, Parfit concludes that a more plausible version of Scanlonian contractualism would drop the 'Individualist Restriction' and thus allow individuals to aggregate their personal burdens together to stronger group objections. Such a view would thus allow them to combine their burden-based objections not only intra-personally but also *inter-personally*. With this amendment, in Case Six the other people can aggregate their personal losses of 5 years of life to a stronger objection they can make as a group, which then outweighs Blue's objection to doing B. With this and other amendments such as giving additional weight to the objections of the worst off, Parfit (OWM, Vol. 2, chs. 21–22) believes that Scanlonian contractualism can be made more extensionally adequate than Scanlon's original version.

The previous amendment to Scanlon's contractualism, however, also makes the resulting view wholly compatible with *R&P*'s reductionism about personhood. It does this by making identity-relations between person-stages irrelevant to reasonable rejectability. We can now stipulate that the objections to the compared principles are not made by persons but rather by momentary person-stages. We can then allow these objecting person-stages to aggregate other person-stages' objections to their own to form stronger aggregate objections which they can then make on behalf of the whole group of person-stages. Here, however, it makes no difference whether those person-stages, whose objections get aggregated together, are identity-related to the other person-stages making the objection.[20] And, because this makes no difference, we do not need to answer the empty questions concerning which future person-stages, if any, are identical to me in My Division. For example, in the previous case, we now know without answering that question that the principle

[19] One problem with Lewis's (1986) four-dimensionalism here would be that the pre-fission burdens to me would be double-counted in the two four-dimensional person's intra-personal aggregations of their burdens.

[20] There would also be room for a middle-ground view according to which only groups of R related person-stage groups can aggregate their burdens together to objections.

Q cannot be reasonably rejected, because there is a group of person-stages that has a stronger aggregate objection to P than any group of person-stages has to Q.

We can thus conclude that, even if Scanlon's contractualism conflicts with *R&P*s reductionism about personhood, Parfit's Scanlonian contractualism does not do so. The latter view defended in *OWM* is wholly compatible and even supported by reductionism. This does, however, lead to another interpretative question. Why did Parfit object to Scanlon's contractualism merely based on first-order moral intuitions given that he could have equally well objected to it directly based on his reductionism about personhood? I will return to question below, but let us first consider *OWM*'s third ethical theory, Parfit's Kantian Contractualism.

5. Parfit's Kantian Contractualism and Personal Identity

To arrive at his Kantian Contractualism, Parfit (*OWM*, Vol. 1, ch. 12) first considers several formulae that are closer to Kant's original statement of the Universal Law formulation of the Categorical Imperative.[21] He criticises these formulae on first-order ethical grounds and through the process of amending them gradually arrives at his own formula of Kantian Contractualism.

Let me illustrate this process with one example. Consider (*OWM*, Vol. 1, 281):

> (H) It is wrong to act on any maxim whose being universally accepted and acted upon would make it impossible for anyone to act upon it.

Take then the maxim 'Have no children, so as to have more time and energy to work for the future of humanity' (*OWM*, Vol. 1, 282–3). (H) would make acting on this maxim wrong in the actual world because, if everyone acted on it, there would be no one left to act upon it and so acting on this maxim would be impossible and hence wrong. Yet, acting on that maxim must be permissible at least as long as not everyone acts on it.

For reasons of space, I will set aside whether the Kantian formulae Parfit critically examines conflict with reductionism about personhood.[22] If they did, this would again provide Parfit with another, more direct reason to object to these Kantian formulae. Instead, I want to focus on the version of Kantian Contractualism which Parfit endorses in *OWM* (vol. 1, 342):

> The Kantian Contractualist Formula: Everyone ought to follow the principles whose universal acceptance everyone could rationally will.

We are to imagine that each person is granted a magical power such that, when they will the universal acceptance of some principle, they thereby *see to it* that this principle is universally accepted. Whether a person can rationally will the universal acceptance of some principle then depends on whether the outcome that results from its universal acceptance provides sufficient object-given reasons for the agent to see to it that the principle is universally accepted. This is the

[21] For a discussion of how well these formulae capture Kant's (1998/1785) view, see, e.g., Morgan (2009).

[22] The fact that Christine Korsgaard (1989) needed to provide an alternative view of personal identity in response to Parfit's reductionism suggests that there is a conflict between reductionism and the Kantian formulae. On her view, it is practically necessary to see yourself as a unified agent over time.

case when those reasons to will the universal acceptance of the principle are not outweighed by any reasons for willing that some other principle is universally accepted instead.

Thus far, nothing in the previous formula conflicts with *R&P*'s reductionism as we can, again, take 'everyone' in the Kantian Contractualist Formula to refer to every momentary person-stage and remain neutral about which person-stages are identity-related. This, however, is not the whole story because Parfit has a substantial theory about which principles everyone could rationally will to be accepted universally.

One of *OWM*'s key arguments attempts to show that the principles whose universal acceptance everyone could rationally will are the very principles whose universal acceptance would make things go best (*OWM*, Vol. 1, sec. 57). These are principles that are optimific according to Parfit's rule-consequentialism, the principles whose acceptance brings about the outcome that, from an impartial point of view, everyone would have the strongest reasons to want to come about. Parfit's Kantian argument to this conclusion proceeds in two stages.

Firstly, according to Parfit (*OWM*, Vol. 1, 379) it is trivially true that the optimific principles are such that everyone could rationally will their universal acceptance. This is because the universal acceptance of these principles brings about the impartially best outcome: the outcome that contains most of the impartially good things such general wellbeing, friendships, promise-keepings, reduction of suffering, kind actions, and so on. That the outcome of the universal adoption of these principles contains the maximum amount of these goods then provides everyone with very strong reasons, from the impartial point of view, to will that these principles are universally accepted.

As explained in §3, Parfit's reductionism imposes a constraint on the considerations that can make outcomes of the universal adoption of principles at the same time both (i) impartially best and (ii) rationally willable by everyone. To avoid any inconsistencies, Parfit cannot recognise here any considerations that would allow identity-relations between person-stages to affect how good an outcome is impartially. But, insofar as the relevant good-makers are provided by different aspects of person-stages and the relevant underlying physical and psychological facts described impersonally, Parfit's Kantian contractualism too can here fit his earlier reductionism about personhood.

The second stage of Parfit's Kantian argument for rule-consequentialism is more interesting in this context. In it, Parfit argues that no one has stronger self-interested, altruistic, or non-deontic reasons (based on the wrong-making features of actions) to will the adoption of any other principles than the optimific ones.

Firstly, with respect to the self-interested reasons, Parfit (*OWM*, Vol. 1, 380) considers a case in which I am stranded on one rock and five others on another, where you could save either me or the group. Here, the optimific principle is the 'numbers principle' which requires everyone to always save the group that contains more people in these cases rather than, say, the 'nearness principle' that would require everyone to save those nearer to them (and let us assume that I would be closer to you). The question then is, would I still have sufficient reasons to will that the numbers principle, rather than the nearness principle, is accepted universally to govern this type of cases? Would my self-interested reason to want the latter principle to be universally accepted outweigh my impersonal reasons for willing that the former principle be accepted? According to Parfit

(OWM, vol. 1, 381), the answer to this question is 'No!', because my self-interested reason for preferring my own survival cannot compete with how many more people would survive if the numbers principle were universally adopted instead of the nearness principle.

Parfit (*OWM*, vol. 1, 385–8) adopts a different approach with regards to the strongest possible partial reasons. This time assume that you can either save your own child or five strangers. Here, Parfit argues that the optimific principles themselves require saving your own child. This is because, even if the universal adoption of such a partial principle led to less people being saved over time, this cost would be outweighed by the benefits that follow from the existence of the strong bonds of love, which the principle requiring everyone to save their own children supports and protects. World is a much better place with these bonds than it would be without them.

The final type of cases is based on the non-deontic reasons provided by the wrong-making features of actions, and these cases are the most difficult for Parfit. Here he (*OWM*, Vol. 1, 390) considers:

> Bridge, in which you cannot save the five except by causing me to fall in front of the runaway train, thereby killing me.

Here the optimific principles seem to require you to push me in front of the train, but intuitively that principle is not one we could rationally will everyone to adopt. The fact that you would be killing me as a means of saving other people seems to give us a decisive (non-deontic, i.e., not based on the wrongness of the action itself) reason not to choose the universal adoption of the previous optimific principle. Thus, we seem to have a counterexample to the claim that everyone could rationally will the universal acceptance of the optimific principles.

Yet, Parfit (*OWM*, Vol. 1, 392) claims that this objection would require the following claims to be true simultaneously:

> (S) You would have a decisive non-deontic reason not to save the five by killing me.
>
> (T) You would also have most reason to want or hope that some stranger would arrive and act instead of you, saving five by killing me.

The objection thus requires that, even if we had decisive reasons not to kill someone as a means of saving many others, we would also have, from the impartial point of view, most reason to want that everyone who can save five by killing one does so. Parfit (ibid.), however, claims that this view and hence also the objection it grounds are incoherent. Therefore, either it has to be the case (i) that whatever consideration gives me sufficient reason not to save the five by killing one also gives everyone a sufficient impartial reason to prefer the outcome in which no one saves five by killing one, or (ii) that whatever provides everyone with sufficient impartial reason to prefer that the group is always saved also gives you sufficient reason to kill me to save the five.

From these arguments, Parfit (*OWM*, Vol. 1, 400) concludes that everyone would have sufficient reason to will the universal acceptance of the optimific principles, and that there are no other conflicting principles that everyone could rationally will to be universally accepted. Here, however, the most important thing is how Parfit arrives at this conclusion. This happens through comparing the strengths of the impersonal reasons to accept the optimific principles to the strengths of the self-interested, partial, and non-deontic reasons that have to do with prudence, relationships, and

agential involvement (such as you committing acts of killing to save five others). The argument thus proceeds wholly at the level of first-order normative considerations.

The compared self-interested, partial, and non-deontic reasons are, however, intimately tied to facts about personal identity. They are reasons based on (i) what would maximize the total wellbeing of all the person-stages that are identity-related to you, (ii) what would benefit the person-stages that are both identity-related with each other and with whom your own past person-stages have been intimately connected, and (iii) what kind of actions the person-stages identity-related to you would have to carry out.

There is a reason why considering these identity-relation-based reasons in this context does not make Parfit's Kantian Contractualism conflict with his reductionism about personhood. This is because these reasons are only considered as grounding potential objections to the claim that everyone would have sufficient reasons to will the universal acceptance of the optimific principles. Parfit, thus, can be understood to be making the argument that, even if we recognised these reasons at face value, they still would never outweigh anyone's impartial reasons to will the universal acceptance of the optimific principles.

Yet, here, we are also again led to ask a different question. In §2, we saw that Parfit argued in *R&P* that reductionism about personhood entails (i) that unity of persons is less deep and so cannot justify maximising the total sum of wellbeing over your lifetime, (ii) that the difference between us and others is less deep and so we should be less concerned with our own lives and more concerned about the lives of others, and (iii) that, because person's existence can be described impersonally by stating the underlying mental and physical facts, there is no reason to care especially about facts about personal identity. The mystery is just why Parfit did not rely on these implications of reductionism directly to argue that the reasons based on self-interest, partiality, and non-deontic considerations (such as your own agential involvement) can never outweigh the impartial reasons we have for willing the universal acceptance of the optimific principles.[23] Insofar as Parfit still accepted reductionism about personhood when writing *OWM*, just why did he not use this more direct route from *R&P* to vindicating his Kantian argument for rule-consequentialism? The next section concludes by considering this question.

6. Conclusion

This chapter has argued that the *prima facie* contradiction between the three ethical theories of *OWM* and *R&P*'s reductionism about personhood is not a genuine one. At a closer look, those ethical theories turn out to be fully compatible with reductionism about personhood. This is insofar as (i) Parfit's rule-consequentialism is able to rule out identity-relations from affecting how impersonally valuable different outcomes are, (ii) Scanlon's contractualism is amended so that momentary burdens are aggregated inter-personally (rather than merely intra-personally) to stronger objections, and (iii) we understand Parfit to be considering the self-interested, partial, and non-deontic reasons to will the universal acceptance of non-optimific principles merely as a potential objection to his

[23] This also suggests that, if Kantian contractualists want to defend a version of their view that is not extensionally identical with rule-consequentialism, they must challenge Parfit's reductionism about personhood.

Kantian Contractualist argument for rule-consequentialism. This means that Parfit is at no point in *OWM* committed to anything that would deeply conflict with *R&P*'s reductionism.

One interpretative puzzle, however, remains. If Parfit's ethical theories in *OWM* do not conflict with *R&P*'s reductionism about personhood, why does he never bring up reductionism in *OWM*? More importantly, why does he never argue for his ethical theories with arguments based on reductionism? Why does he not use reductionism to attack Scanlon's contractualism or the self-interested, partial, or non-deontic reasons to will the universal acceptance of the non-optimific principles? Moreover, given his dismissiveness of separateness of persons in *R&P* (346):

> These principles [distributive principles] are often held to be founded on the separateness … of different persons. This fact is less deep on the Reductionist View, since identity is less deep.

why does Parfit (*OWM*, Vol. 3, 236) then seemingly endorse that claim in his argument against Scanlon in Case Three?

> And if we fail to distinguish between Dick and Harry, regarding them as merely parts of a general person, we are ignoring the separateness of persons, which has been called 'the basic fact for ethics'.

According to Mark Schroeder's (2011) interpretation, both *R&P* and *OWM* are fundamentally about the possibility of moral progress. Yet, whereas *R&P* (x) wants to show that moral progress is possible by making genuine progress itself, *OWM* argues that moral progress is possible at least in principle because there is less disagreement between Kantians, contractualists, and consequentialists than usually assumed. In the context of *R&P*'s project, it thus makes sense to defend a controversial and radical view about personhood and attempt to make moral progress by deriving first-order ethical conclusions on that basis. However, for *OWM*'s ecumenical project, that approach would simply not work. The defenders of the traditional versions of the three traditional ethical theories would simply reject Parfit's reductionism about personhood based on the separateness of persons thesis to which they are committed. This is why, I believe, Parfit chose to defend the ethical consequences of reductionism about personhood, his own versions of the three theories, merely by relying on first-order intuitions about cases.[24] In this way, Parfit did not need to challenge the separateness of persons explicitly even if in *OWM* we are led to ethical theories that do not presuppose it either. In his later works, Parfit thus deliberately set aside the question of personhood to avoid getting entangled again in the disagreements about personal identity.

Bibliography

Adams, R.M. (1989): 'Should Ethics be More Impersonal?' *Philosophical Review* 98: 439–484.

Alter, Torin and Stuart Rachels (2004): 'Epistemicism and the Combined Spectrum'. *Ratio* 17: 241–255.

[24] That the three ethical theories defended in *OWM* are compatible with *R&P*'s reductionism also shows that the latter view is less revisionary than Parfit initially assumed.

Blackburn, Simon (1997): 'Has Kant Refuted Parfit?' In J. Dancy (ed.): *Reading Parfit*. Oxford: Blackwell, 180–201.

Brink, David (1997): 'Rational Egoism and the Separateness of Persons'. In J. Dancy (ed.): *Reading Parfit*. Oxford: Blackwell, 96–134.

Broome, John (2004): *Weighing Lives*. Oxford: Oxford University Press.

Cassam, Quassim (1993): 'Parfit on Persons'. *Proceedings of the Aristotelian Society* 93: 17–37.

Dorsey, Dale (2015): 'The Significance of Life's Shape'. *Ethics* 125: 303–330.

Edmonds, David (2023): *Parfit*. Princeton: Princeton University Press.

Gauthier, David (1963): *Practical Reasoning*. Oxford: Clarendon Press.

Hédoin, Cyril (2021): 'A Social Contract without Persons? Revisionary Psychological Reductionism and Contractualism'. An unpublished manuscript.

Hooker, Brad (2000): *Ideal Code, Real World*. Oxford: Oxford University Press.

Johnston, Mark (1997): 'Human Concerns without Superlative Selves'. In J. Dancy (ed.): *Reading Parfit*. Oxford: Blackwell, 149–179.

Kant, Immanuel (1998/1785): *Groundwork of the Metaphysics of Morals,* ed. and trans. By M. Gregor. Cambridge: Cambridge University Press.

Korsgaard, Christine (1989): 'Personal Identity and the Unity of Agency: A Kantian Response to Parfit'. *Philosophy & Public Affairs*18: 101–132.

Lewis, David (1976): 'Survival and Identity'. In A. Oksenberg Rorty (ed.): *The Identities of Persons*. Berkeley: University of California Press, 17–40.

Merricks, Trenton (1997): 'Fission and Personal Identity over Time'. *Philosophical Studies* 88: 163–186.

Morgan, Seiriol (2009): 'Can There Be Kantian Consequentialism?' *Ratio* 22: 19–40.

Nozick, Robert (1974): *Anarchy, State, and Utopia*. New York: Basic Books.

Olson, Eric (1999): *The Human Animal*. Oxford: Oxford University Press.

Parfit, Derek (1971): 'Personal Identity'. *Philosophical Review* 80: 3–27.

Parfit, Derek (1976): 'Lewis, Perry, and What Matters'. In A. Oksenberg Rorty (ed.): *The Identities of Persons*. Berkeley: University of California Press, 91–108.

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

Parfit, Derek (1995): 'The Unimportance of Identity'. In H. Harris (ed.): *Identity*. Oxford: Oxford University Press, 13–45.

Parfit, Derek (2011): *On What Matters*, Vols. 1 and 2. Oxford: Oxford University Press.

Parfit, Derek (2012): 'We Are not Human Beings.' *Philosophy* 87: 5–28.

Parfit, Derek (2017): *On What Matters*, Vol. 3. Oxford: Oxford University Press.

Rawls, John (1999): *A Theory of Justice*, rev. ed. Oxford: Oxford University Press.

Scanlon, T.M. (1998): *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schroeder, Mark (2011): 'Review of *On What Matters*, vols. 1 & 2.' *Notre Dame Philosophical Reviews*, available online at https://ndpr.nd.edu/reviews/on-what-matters-volumes-1-and-2/.

Shoemaker, Sydney (1959): 'Personal Identity and Memory'. *The Journal of Philosophy* 56: 868–882.

Shoemaker, David (1999): 'Utilitarianism and Personal Identity'. *The Journal of Value Inquiry* 33: 183–199.

Shoemaker, David (2000): 'Reductionist Contractualism: Moral Motivation and the Expanding Self'. *Canadian Journal of Philosophy* 30: 343–370.

Slote, Michael (1982): 'Goods and Lives'. *Pacific Philosophical Quarterly* 63: 311–326.

Sorensen, Roy (1988): *Blindspots*. New York: Oxford University Press.

Steuwer, Bastian (2020): 'Why it Does not Matter What Matters: Relation R, Personal Identity, and Moral Theory'. *Philosophical Quarterly* 70: 178–198.

Velleman, David (1991): 'Wellbeing and Time'. *Pacific Philosophical Quarterly* 72: 48–77.

Wiggins, David (1967): *Identity and Spatio-Temporal Continuity*. Oxford: Blackwell.

Wolf, Susan (1986): 'Self-Interest and Interest in Selves'. *Ethics* 96: 704–720.