DEREK PARFIT'S *REASONS AND PERSONS*

An Introduction and Critical Inquiry

Edited by Andrea Sauchelli

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2020
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
52 Vanderbilt Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Sauchelli, Andrea, editor.
Title: Derek Parfit's Reasons and persons : an introduction and
critical inquiry /edited by Andrea Sauchelli.
Description: Abingdon, Oxon ; New York, NY : Routledge, 2020. |
Includes bibliographical references and index. |
Identifiers: LCCN 2019045380 (print) | LCCN 2019045381 (ebook) |
ISBN 9781138595118 (hardback) | ISBN 9781138595132 (paperback) |
ISBN 9780429488450 (ebook)
Subjects: LCSH: Parfit, Derek. Reasons and persons. | Ethics. | Rationalism. | Self.
Classification: LCC BJ1012.P393 D47 2020 (print) |
LCC BJ1012.P393 (ebook) | DDC 170–dc23
LC record available at https://lccn.loc.gov/2019045380
LC ebook record available at https://lccn.loc.gov/2019045381

ISBN: 978-1-138-59511-8 (hbk)
ISBN: 978-1-138-59513-2 (pbk)
ISBN: 978-0-429-48845-0 (ebk)

Typeset in Bembo
by Newgen Publishing UK

1

TOWARD A UNIFIED THEORY OF MORALITY

An introduction to Part One of *Reasons and Persons*

Ben Eggleston[1]

Introduction

Part One of *Reasons and Persons* discusses a wide variety of topics, and a first-time reader could be pardoned for not seeing the topics as unified in any clear way. This part of the book can seem, instead, like a tour de force of disconnected insights: brilliant, but haphazard. The closing pages of Part One reveal, however, that nearly all of it can be seen as articulating and arguing for a particular theory of morality. In retrospect, this gives Part One a coherent agenda that, for most readers, is not initially evident.

The moral theory advocated blends consequentialism and common-sense morality, and Parfit's support for it has three main components: rebutting a particular set of objections to a standard form of consequentialism, arguing that this standard form of consequentialism should be modified in a particular way, and claiming that common-sense morality should also be modified in a way that makes it similar to the modified form of consequentialism.

It would have been fitting for Parfit to call Part One something like "Toward a Unified Theory of Morality" (which is why I have chosen that phrase for the title of this chapter). Why, then, does he call it "Self-Defeating Theories"? That title is apt for two interconnected reasons. First, both the standard form of consequentialism that Parfit discusses and common-sense morality are self-defeating in ways that Parfit describes at length. Indeed, practically all of Part One is concerned with self-defeating theories. The second reason is that many of the arguments of Part One focus on the self-defeating character of the theories being discussed. In the case of the standard form of consequentialism that Parfit discusses, Parfit's main purpose is to argue that it is not self-defeating in a way that warrants its rejection. In the case of common-sense morality, his purpose is the opposite.

In fact, according to Parfit it is not just those two theories that are self-defeating in certain ways: he writes that "all of the best-known theories are in certain ways self-defeating" (p. 3).[2] He goes on to say, however, that some theories are self-defeating in harmless ways while the self-defeating character of others necessitates their revision or rejection. To appreciate these subtleties in Parfit's views, some terminological points should be kept in mind. When Parfit first mentions the idea of a theory being self-defeating, he implies that such a theory thereby "fails in its own terms, and thus condemns itself" (p. 3). In later passages, however, Parfit diagnoses some theories as self-defeating, but also holds that they do *not* fail in their own terms, and do not condemn themselves. In other words, contrary to initial appearances, self-defeat does not imply self-condemnation – the latter verdict depends on what kind of self-defeat the theory suffers from.

I mentioned above that Parfit's argument for a unified theory builds on discussions of both a standard form of consequentialism and common-sense morality – emphasizing, in each case, concerns about self-defeat. He begins, however, by exploring several aspects (especially the self-defeat) of yet another theory: the Self-interest Theory.

The Self-interest Theory

What the Self-interest Theory is

Parfit characterizes the Self-interest Theory in terms of several claims, the "central" one being that "For each person, there is one supremely rational aim: that his life go, for him, as well as possible" (p. 4). This theory is not a moral theory, and Parfit does not treat it as providing source material for his unified theory. It might seem odd, then, that Parfit would devote a substantial fraction of Part One to this theory. It turns out, however, that the various aspects of self-defeat that Parfit is concerned to examine in connection with consequentialism and common-sense morality also arise in connection with the Self-interest Theory. Thus, Parfit's discussion of this theory serves mainly – at least in Part One – to acquaint the reader with self-defeat-related concepts and lines of reasoning whose real importance materializes in connection with those moral theories. (Note, however, that the Self-interest Theory serves a further purpose outside of Part One, playing a more central role in Part Two.)

How the Self-interest Theory is indirectly self-defeating

Parfit observes that an obvious way in which the Self-interest Theory is self-defeating is that when people attempt to make their lives go as well as possible, they often fail, due to false beliefs – for example, a thief's overconfidence about being able to steal something without getting caught. Parfit quickly judges that "These cases are not worth discussing" and that this type of self-defeat is "no objection" to the theory because the problem is caused by the agent's incompetent implementation of the theory rather than anything internal to the theory itself (p. 5).

The scenarios that Parfit takes more seriously involve not just isolated acts in which a person attempts to make her life go as well as possible, but an underlying disposition toward performing such acts. One might expect Parfit to focus on the disposition to always try to make one's life go as well as possible, and he does mention this disposition (pp. 5–6). But he claims that we should also include, as pertinent to assessing the Self-interest Theory, many acts that people perform while "acting on a more particular desire" than the desire to make one's life go as well as possible. For example, Nina might help Oren make a cake because she loves him, not because she has any opinion about how that activity will affect her life. In such a case, as long as Nina does not regard her act as making her life go worse, we should (according to Parfit) count her decision-making as aligning with the Self-interest Theory. So, the disposition Parfit elects to focus on is that of being "never self-denying" (p. 6). A person with this disposition might sometimes be inattentive to the goal of making her life go well and might sometimes lack beliefs about her acts' promotion of that goal, but she will (and this is the defining trait of the disposition) never do what she does believe will be worse for her (p. 6).

Parfit presents several cases showing that if a person is never self-denying, then her life can go worse than if she had a non-self-interested disposition. Here is one such case:

> Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger, and the only other driver near. I manage to stop you, and I offer you a great reward if you rescue me. I cannot reward you now, but I promise to do so when we reach my home. Suppose next that I am transparent, unable to deceive others. I cannot lie convincingly. Either a blush, or my tone of voice, always gives me away. Suppose, finally, that I know myself to be never self-denying. If you drive me to my home, it would be worse for me if I gave you the promised reward. Since I know that I never do what will be worse for me, I know that I shall break my promise. Given my inability to lie convincingly, you know this too. You do not believe my promise, and therefore leave me stranded in the desert. This happens to me because I am never self-denying.
>
> *p.* 7

Parfit's example might seem rather peculiar and far-fetched, but it is an instance of a large class of cases that exhibit a well-established and widely applicable concept known as the *paradox of happiness* (though Parfit does not use this term). This is the initially surprising – but generally acknowledged – fact that "regarding happiness as the sole ultimately valuable end or objective, and acting accordingly, often results in less happiness than results from regarding other goods as ultimately valuable (and acting accordingly)."[3] Parfit's example illustrates the paradox perfectly: if he (upon having been driven home) were able to regard keeping his promise as more important than advancing his happiness, then he (earlier, in the desert) could sincerely make that promise and he would thereby advance his happiness more than he actually can, given that he is never self-denying. Hence, being never self-denying is worse for him than having some other disposition (i.e., one that allowed occasional self-denying acts, such as keeping certain promises) would be.

According to Parfit, in cases such as these the Self-interest Theory is "indirectly individually self-defeating" (p. 7). As Parfit defines this term, it applies to a theory T (here the letter "T" is a variable) when the following is true: "if someone tries to achieve his T-given aims [such as making one's life go as well as possible], these aims will be, on the whole, worse achieved" (p. 5). This is Parfit's first major conclusion concerning the self-defeating character of any of the theories he discusses.

The failure of several objections to the Self-interest Theory

Having shown that the Self-interest Theory is indirectly individually self-defeating, Parfit then asks whether this means the theory "fail[s] in its own terms," or whether it "condemn[s] itself" (p. 7). To answer these questions, Parfit notes that the theory implies many things about what people have reason to do, and one of these is the following:

> what each person has most reason to do is to cause himself to have, or to allow himself to keep, any of the best possible sets of motives [which, for Parfit, include dispositions], in self-interested terms. These are the sets of motives of which the following is true. There is no other possible set of motives of which it is true that, if this person had these motives, this would be better for him.
>
> *p. 8*

Because of the paradox of happiness, being never self-denying is "for many and perhaps most people" (p. 17) definitely *not* among the best possible sets of motives. Consequently, Parfit explains, the Self-interest Theory does not tell these people to have that disposition. Thus, when people suffer the ill effects of having that disposition, they are violating, not complying with, the Self-interest Theory. So, the theory is not failing in its own terms (p. 11).[4]

Parfit goes on to consider other questions about the Self-interest Theory, such as whether there might be other grounds for saying that it fails in its own terms (p. 11), or whether it is objectionable because it directs people to be disposed to sometimes act irrationally (p. 12), or whether it is objectionable because it implies that we cannot entirely avoid acting irrationally (p. 16), or whether it is objectionable because it might sometimes tell agents to believe a revised version of itself (p. 19), or whether it would be objectionable if it were *self-effacing*, in the sense of telling "everyone to cause himself to believe some other theory" (p. 24), or whether it is objectionable because the outcome would be bad for a group of people if they believed it and were unable to change their beliefs or dispositions (p. 43), or whether it is objectionable because it treats acting rationally as a mere means (p. 45).

Parfit rejects all of these. Note, however, that for our purposes, as we keep in mind the trajectory of Part One as a whole, Parfit's defence of the Self-interest Theory against these objections has less to do with his support for the Self-interest Theory per se than with his desire to discredit the logic of these objections. For he will want to entertain and dismiss them yet again in his discussion of consequentialism, in order to establish the latter theory as a worthy contributor to his unified theory.

Practical dilemmas

The Self-interest Theory figures prominently in one other major topic of Part One: situations that Parfit calls "practical dilemmas." These include not only traditional, two-person prisoner's dilemmas, but many-person variants as well. The latter, like the former, are characterized by the fact that "it is certain that, if each rather than none of us does what will be better for himself, this will be worse for everyone" (p. 59). Parfit holds that many-person cases occur more frequently than, and matter more than, two-person cases: "Though we can seldom know that we face a Two-Person Prisoner's Dilemma, we can very often know that we face Many-Person Versions. And these have great practical importance. The rare Two-Person Case is important only as a model for the Many-Person Versions" (p. 59). Parfit gives about a dozen examples of such many-person dilemmas (pp. 61–62). One example whose resonance has perilously increased in the decades since the publication of *Reasons and Persons* pertains to climate change: It is better for each person if he or she uses as much energy as is optimal for his or her lifestyle, but if everyone makes this same decision, the result is worse for each person than if everyone exercises some restraint.

Many-person dilemmas are critical to Parfit's discussion of common-sense morality, as we will see below. But they also aid Parfit in exploring one further objection to the Self-interest Theory that (like the objections mentioned above) Parfit dismisses as unsuccessful. This objection is based on another kind of self-defeat: direct collective self-defeat. In defining this notion Parfit gives several criteria of increasing precision, but the essential idea is that compliance with the theory by each of a group of people ultimately frustrates their achievement of their aims (p. 55). For example, the energy-consumption example just mentioned shows that the Self-interest Theory is directly collectively self-defeating because compliance with the theory by each person results in energy consumption of such a magnitude that the result for each person is worse than if they had each exercised some restraint, in defiance of the Self-interest Theory.

Parfit grants that the Self-interest Theory is directly collectively self-defeating (p. 88). But, as with the indirect individual self-defeat discussed above, he holds that this does not underwrite a successful objection to the theory. This is because the theory is not concerned with the collective level – it is only concerned with the individual level. Consequently, its collective self-defeat does not entail that it either fails in its own terms or condemns itself (p. 92).

Consequentialism

What consequentialism is

In contemporary moral philosophy, consequentialism is generally taken to refer to the principle that acting rightly is a matter of choosing the act that leads to the best possible outcome.[5] In *Reasons and Persons*, Parfit discusses a theory that he calls "Consequentialism," whose central claim is that "There is one ultimate moral aim: that outcomes be as good as possible" (p. 24). Another tenet he attributes to this theory echoes the general consequentialist principle mentioned above: "What each of us ought to do is whatever would make the outcome best" (p. 24). This theory forms half of the basis for the unified theory that he proposes at the end of Part One. Accordingly, as indicated above, he both defends this theory against several objections and argues that it should be modified in a particular way.

Before turning to those matters, it is worth pausing briefly to observe that the way Parfit states the central claims of the Self-interest Theory and Consequentialism imply that they operate in different domains, rather than being rivals of each other. The central claim of the Self-interest Theory refers to the supremely rational ultimate aim, while the central claim of Consequentialism refers to the ultimate moral aim. Although Parfit writes that moral theories and theories about rationality both give answers to the question "What do we have most reason to do?" (p. 3) and touches on the opposition of morality and self-interest (p. 88), he does not position the Self-interest Theory and Consequentialism against each other. In fact, the most salient connection between them, in Parfit's discussion, is that they are both targets of largely the same set of unsuccessful objections.

How consequentialism is indirectly self-defeating

Above I claimed that the main purpose of Parfit's discussion of the Self-interest Theory in Part One is to acquaint the reader with self-defeat-related concepts and lines of reasoning whose real importance materializes in connection with the moral theories he discusses. One basis for this claim is that Parfit's discussion of the Self-interest Theory is closely echoed by his discussion of Consequentialism, as he acknowledges at the beginning of the latter discussion (p. 24). Let us review those concepts and lines of reasoning as they arise in relation to Consequentialism.

Paralleling his consideration of the dispositions that we should associate with the Self-interest Theory, Parfit identifies a disposition to associate with Consequentialism: that of being a "pure do-gooder" – someone who is disposed to "always try to do whatever would make the outcome as good as possible" (p. 27). Parfit then argues that if everyone were a pure do-gooder, then the outcome would probably be worse than if some people had some other dispositions. The essence of his argument runs as follows:

> Most of our happiness comes from having, and acting upon, certain strong desires. These include the desires that are involved in loving certain other people, the desire to work well, and many of the strong desires on which we act when we are not working. To become pure do-gooders, we would have to act against or even to suppress most of these desires. It is likely that this would enormously reduce the sum of happiness.
>
> *p. 27*

According to Parfit, in cases such as these Consequentialism is "indirectly collectively self-defeating" (p. 28). This term applies to a theory T (again, using that letter as a variable) "when it is true that, if several people try to achieve their T-given aims, these aims will be worse achieved" (p. 27).

The failure of several objections to Consequentialism

As with the Self-interest Theory, Parfit argues that Consequentialism is not discredited by the scenarios that show that it is indirectly self-defeating. Just as the Self-interest Theory requires us to have the motives that will make our lives go best (not necessarily the disposition of being never self-denying), Consequentialism requires us to have the motives that will have the best consequences generally (p. 26). Assuming these do not include being a pure do-gooder, Consequentialism "tells us that it would be wrong to cause ourselves to be, or to remain, pure do-gooders." Thus, Consequentialism does not fail in its own terms or condemn itself (p. 28).[6]

Parfit goes on to consider a series of questions about Consequentialism that closely follow the further questions about the Self-interest Theory that he considered. These include whether Consequentialism is objectionable because it directs people to be disposed to sometimes act wrongly (p. 32), or whether it is objectionable because it implies that we cannot entirely avoid acting wrongly (p. 36), or whether it is objectionable because it tells us to cause ourselves to do what it claims is wrong (p. 37), or whether it would be objectionable if it were self-effacing (in the sense specified in the discussion of the Self-interest Theory) or *esoteric*, in the sense of "telling those who believe it not to enlighten the ignorant majority" who do not believe it (p. 41), or whether it is objectionable because the outcome would be bad for a group of people if they all believed it and were unable to change their beliefs or dispositions (p. 43), or whether it is objectionable because it treats acting morally as a mere means (p. 45). Although these issues have spawned further discussion,[7] the key point to appreciate for our purposes is that Parfit is maintaining the viability of Consequentialism because it forms half of the basis for the unified theory that he proposes at the end of Part One.

Almost everything about Part One that I have described so far is from the first of its five chapters, which occupies nearly half of it. (The only exception is the material pertaining to practical dilemmas, which is from Chapter 2.) In the preface to *Reasons and Persons*, Parfit writes that the first chapter is the only one in which he does not "try to challenge what we assume." Instead, in that chapter, "I cannot avoid repeating what has been shown to be true." Consequently, he says, it is "dreary" (p. x).

Practical dilemmas

In my overview of Parfit's discussion of the Self-interest Theory, I mentioned that practical dilemmas show that that theory is directly collectively self-defeating, but that Parfit holds this to be acceptable for that theory since it is concerned with the individual level, not the collective level. No such reply would be available to excuse any direct collective self-defeat that Consequentialism could be shown to exhibit, since the collective level is precisely where Consequentialism is meant to operate (since it is a moral theory). But Parfit nips this objection in the bud: he claims that Consequentialism cannot be directly collectively self-defeating because compliance with the theory by each person in a group of people will definitely result in the best possible outcome – that follows directly from what the theory requires (p. 54).

Having stated this simple and apparently decisive point, however, Parfit goes on to investigate a question that poses a different threat to Consequentialism: In practical dilemmas that involve such large numbers of people that "any single altruistic choice would make no difference," can we "explain why we should contribute by appealing to the consequences of our acts" – i.e., without appealing to any nonconsequentialist considerations (p. 67)? If we cannot, then Consequentialism might not imply that we should contribute in such cases. Depending on how obvious we take it to be that we *should* contribute in such cases, this fact about Consequentialism (if it turns out to be a fact) might be taken to indicate that Consequentialism is seriously flawed as a moral theory.

Mistakes in moral mathematics

According to Parfit, we can indeed appeal to the consequences of our acts in order to explain why we should contribute, if we avoid what he calls "mistakes in moral mathematics" (p. 67). He discusses a total of five such alleged mistakes. The first two concern how to credit or blame people for the combined effects of their individual acts (pp. 67–73), and Parfit's claims here have generated some discussion.[8] The third mistake is to ignore possible events that are highly unlikely, such as the possibility that one's vote will determine the outcome of a presidential election or the possibility that one of many nuclear-reactor components will fail and cause a catastrophe. Parfit writes that such possibilities must not be ignored in decision-making, though of course it is appropriate to discount them according to how improbable they are (pp. 73–75).

However, for the trajectory of Part One as a whole, the most important mistakes are the fourth and fifth ones. The fourth mistake is to regard very small effects as morally insignificant (p. 75). For example, in the energy-consumption example mentioned above, each consumer might think that because the effects of his or her consumption decisions on other people are very small, those effects are morally insignificant, and thus provide no reason for him or her to exercise restraint. The fifth mistake is similar to the fourth, but refers to *imperceptible* effects instead of *very small* effects (p. 75). Parfit declines to address the fourth mistake separately and, instead, argues against the fifth mistake at length. This suffices to address the fourth mistake if one thinks – as seems plausible – that if *imperceptible* effects matter in the moral assessment of an act, then surely any *perceptible* (even if very small) effects also matter in such an assessment.

Parfit's argument for the moral significance of imperceptible effects includes several artfully constructed hypothetical examples. Probably the most important of these is "The Harmless Torturers" (p. 80). In this case, there are a thousand torturers and a thousand victims. Each victim is connected to a machine that causes pain, where the intensity of the pain is determined by the position of a dial on the machine. The dial has a minimal setting of 0, which corresponds to mild pain, and can be advanced by a thousand increments. Every one-increment turn of the dial is imperceptible to the victim, but the maximal setting of 1,000 is extremely painful. There is a button that simultaneously advances each of the thousand dials by one increment. One day, all of the dials are initially set at 0. During the day, each of the thousand torturers presses the button once. By the end of the day, each victim is in severe pain. (This example closely resembles an example created by Jonathan Glover and M. J. Scott-Taggart,[9] which Parfit cites. He writes that his discussion of the five mistakes, and "especially" his Harmless Torturers example, "derives entirely from the stimulus of this brilliant example" (p. 511, n. 44).)

Parfit judges that "the torturers are clearly acting wrongly" (p. 80), and he supports this conclusion with a disjunctive argument hinging on whether it is, or is not, possible for a person's pain to worsen imperceptibly (which Parfit acknowledges is debatable). If this is possible, then the argument for the wrongness of the actions of the torturers is straightforward: each torturer causes a lot of pain, even though no individual victim perceives the worsening of pain inflicted by any individual torturer (p. 80). On the other hand, if it is not possible for a person's pain to worsen imperceptibly, then although no torturer harms anyone, we can conclude that each acts wrongly because "they together impose great suffering" (p. 80). Either way, by the original construction of the example, the effects of each torturer's act are imperceptible, so Parfit concludes from this example that imperceptible effects can be morally significant. As mentioned above, this suffices to show that small effects are morally significant, too.

I mentioned at the beginning of this section that Parfit's discussion of these "mistakes in moral mathematics" provides the key to explaining why we should contribute in many-person practical dilemmas. We can now see how it does this, by considering once again the energy-consumption example. If Parfit is correct to hold that small and even imperceptible effects of acts can be morally significant, then the greenhouse-gas effects of even a single act of energy consumption cannot justifiably be ignored in an accounting of the consequences of that act. Such an accounting would include those effects and would supply the elements of a consequentialist argument in favour of exercising some restraint. Thus, Consequentialism is saved from implying the unpalatable verdict that it is morally permissible for a person to consume energy unrestrainedly.

Common-sense morality

The third and final theory that plays a major role in Part One is common-sense morality. Parfit argues that this theory is self-defeating in a way that should persuade even its proponents to embrace a small but meaningful modification of it – leading toward the unified theory that he proposes at the end of Part One.

What common-sense morality is

In ordinary parlance (at least among moral philosophers), common-sense morality is understood as differing from consequentialism in virtue of designating some particular kinds of conduct as being especially bad – if not prohibited absolutely, then at least having a strong presumption of wrongness and therefore being justifiable only in exceptional circumstances. Such kinds of conduct might include lying, stealing, breaking promises, racial discrimination, religious persecution, torture and murder. According to common-sense morality, the badness of these kinds of conduct is not just a matter of their tending to have bad consequences, though that of course does matter greatly. In this way, common-sense morality contrasts with consequentialism since the latter assesses conduct strictly through the lens of its consequences, and therefore does not designate some kinds of conduct as being especially bad in and of themselves.

Although Parfit acknowledges this element of what is generally called common-sense morality (p. 112), he focuses on a different, but perhaps no less integral, component of that moral outlook in order to characterize the theory he calls Common-Sense Morality. He writes that "Common-Sense Morality largely consists in" the following moral belief:

> Most of us believe that there are certain people to whom we have special obligations. These are the people to whom we stand in certain relations— such as our children, parents, friends, benefactors, pupils, patients, clients, colleagues, members of our own trade union, those whom we represent, or our fellow-citizens. We believe that we ought to save these people from certain kinds of harm, and ought to try to give them certain kinds of benefit.
>
> *p. 95*

Part of the content that Parfit sees these special obligations as having is that they can require a person to save a child (or a parent, friend, etc.) from some harm even at the cost of saving a stranger from a larger harm. For example, Common-Sense Morality might well require that a parent save her child from a broken arm even if she knows that she is thereby forgoing the opportunity to save a stranger from a different injury that she knows will be more serious. Parfit adds, however, that even according to Common-Sense Morality, "This priority is not absolute. I ought not to save my child from a cut or bruise rather than saving a stranger's life." Nevertheless, there is some priority: according to Common-Sense Morality, "I ought to save my child from some harm rather than saving a stranger from a *somewhat* greater harm" (p. 95).

It is useful to situate Parfit's characterization of Common-Sense Morality in the context of his distinction, which has proven influential, between agent-neutral theories and agent-relative theories (p. 27). Consequentialism is an agent-neutral theory because it assigns the same moral aims to all agents: that outcomes be as good as possible. In contrast, Common-Sense Morality is an agent-relative theory because the aims it assigns to an agent (and their relative weights) are relative to that agent's values, relationships, or other attributes. For example, in some situations, Common-Sense Morality might assign Parfit the (primary) aim of saving his child from a particular harm even at the cost of some other person incurring a somewhat greater harm, while it might assign the other person's parent the (primary) aim of averting that harm even at the cost of Parfit's child incurring a harm.

Parfit's characterization of Common-Sense Morality is clearly rather idiosyncratic and schematic, focusing on the structural feature of agent-relative obligations (especially ones that diverge from producing the best possible outcomes) rather than the kind of substantive rules and values that we should want children to be taught by their parents and teachers. The reason for this eccentric focus is that this structural feature is the essential ingredient in Parfit's argument that Common-Sense Morality is self-defeating, which is a key claim in his argument for his unified theory.

How Common-Sense Morality is directly self-defeating

To set the stage for Parfit's comments on the self-defeat of Common-Sense Morality, let us recall his self-defeat-related conclusions about the previous two theories. Both the Self-interest Theory and Consequentialism are indirectly self-defeating at the levels with which they are concerned – the individual level and the collective level, respectively. But indirect self-defeat is not a serious flaw. As for direct self-defeat (which can be a serious flaw), the Self-interest Theory is directly self-defeating only at the collective level, but that is a non-issue since the Self-interest Theory is not concerned with the collective level. And by the logic of Consequentialism, it cannot be directly self-defeating at the collective level. So, each theory avoids direct self-defeat at the level where that attribute would be a serious flaw.

Matters are different with Common-Sense Morality: Parfit argues that at the level with which it is concerned – the collective level – it is directly self-defeating. He shows this by presenting several hypothetical examples that are ingenious variations on the traditional case of the prisoner's dilemma. Several of these examples are what he calls "Parent's Dilemmas" (p. 96), and the simplest one has the following form (though I have rewritten it to make some of its key features explicit).

> Quinn has a son, and so does Ramona. (But there are no relationships of special obligations between the two families.) Each child is vulnerable to two harms, a small one and a large one. Quinn can either save her son from the small harm or save Ramona's son from the large harm. Similarly, Ramona can either save her son from the small harm or save Quinn's son from the large harm. The harms are close enough in size that, according to Common-Sense Morality, each parent's obligation to save her own son from the harm threatening him that she can avert (the small harm) outweighs her obligation to save the other boy from the harm threatening him that she can avert (the large harm).

Parfit also presents Parent's Dilemmas involving the conferral of benefits rather than the blocking of harms (pp. 96–97), but in all cases, the upshot is the same: Common-Sense Morality requires the parents to prioritize their own children rather than avert the most harm or do the most good. The result is that every child suffers a larger harm than was necessary in the circumstances, or receives a smaller benefit than was possible in the circumstances.

To see how this example shows that Common-Sense Morality is directly collectively self-defeating, recall the definition of that notion from the discussion of the Self-interest Theory and practical dilemmas: the essential idea is that compliance with the theory by each of a group of people ultimately frustrates their achievement of their aims (p. 55). In the example of Quinn and Ramona, Common-Sense Morality holds that each of them has a special obligation to shield her own son from harm (as well as an ordinary obligation to shield other people from harm). We may consider the safety of their children their aim, for the purposes of applying the definition of self-defeat just given. And when Quinn and Ramona comply with Common-Sense Morality, this aim is achieved worse than was necessary in the circumstances. Here, then, is a case in which Common-Sense Morality frustrates the achievement of the aim it tells people to pursue. Thus, it "is here directly collectively self-defeating" (p. 99).

Parfit concedes that two-person parent's dilemmas do not arise often. But he claims that situations often arise in which there are not just two parents, but many parents, with each parent facing a choice like the ones in the parent's dilemmas, such that if each parent gives priority to his or her child, then all of the children end up faring worse than they needed to. This is, for example, the problem of public goods (p. 98). Parfit adds that "Similar remarks apply to all similar obligations—such as those to pupils, patients, clients, or constituents. With all such obligations, there are countless many-person versions like my three Parent's Dilemmas" (p. 98; see also p. 102).

One feature of these cases that justifies the centrality that Parfit claims for them is that they arise even when a person is willing to justify his or her behaviour in moral terms – not merely in self-interested terms. Consider, for example, a person who is planning to buy a huge, gas-guzzling SUV, imposing harms on others that exceed the benefits he will secure for himself (relative to, say, a more fuel-efficient car). Although we cannot accuse the person of acting irrationally in terms of his own self-interest, we might think that if only he could be persuaded that he must also account for his behaviour in moral terms, then he would acknowledge that he is behaving wrongly. However, the same person might adapt some remarks that Parfit discusses (p. 100; see also p. 62) and say, "Oh, if it were just me, I would be glad to buy a smaller car. But I also have to think about my children, and their safety and comfort. And my obligations to them are stronger than my obligations to people in general. Thus, all things considered, the moral reasons ultimately point in the direction of the bigger car." Thus, the shift from a self-interested point of view to a moral point of view is not enough to get people to make decisions that, collectively, result in the maximal achievement of their moral aims. Some moral points of view, such as that of Common-Sense Morality, exhibit the same kind of collective self-defeat that we saw in the case of the Self-interest Theory. But whereas the Self-interest Theory can shrug off collective self-defeat as being a non-issue because it is only concerned with the individual level, collective self-defeat is a serious flaw for Common-Sense Morality or any other moral theory.[10]

Revising Common-Sense Morality

The collective self-defeat of Common-Sense Morality leads Parfit to propose revising that theory in certain ways. He proposes three distinct revisions, but they all basically boil down to the idea that people should deviate from the prescriptions of Common-Sense Morality when those prescriptions frustrate the achievement of the aims of the theory. For example, in situations of the kind we have been considering, people should do what will cause those aims to be better achieved, not worse achieved (pp. 100–103). To understand the precise character of this idea, recall that the aims in question are the aims that Common-Sense Morality prescribes for people. Thus, the idea is not that people should feel free to ignore the demands of Common-Sense Morality (or morality in general) in order to more fully achieve their self-interest. Rather, the idea is that the aims of Common-Sense Morality are worth promoting, and that people should promote them more thoughtfully than the dictates of Common-Sense Morality prescribe. For example, a person should be willing to contribute to a public good that will benefit his or her child, even if it would be better for that child for the parent to decline to contribute and devote the same resources to that child individually.

After specifying this revised version of Common-Sense Morality, Parfit gives several arguments for the claim that proponents of Common-Sense Morality should find these revisions appealing. One argument begins by emphasizing the inherent senselessness of accepting that one's moral theory is directly collectively self-defeating. After all, the *direct* self-defeat of Common-Sense Morality means that the self-defeat stems not from some agents' failing to act in accordance with the theory, but from all agents' successfully acting in accordance with the theory (p. 103). And on most views about the nature of morality, a moral theory – unlike a theory of individual rationality such as the Self-interest Theory – purports to offer a code for *collectively* achieving our moral aims (p. 106; see also p. 113). Thus, the direct collective self-defeat of a particular theory is an absurdity that any proponent of that theory should want to remedy. Since the revisions Parfit proposes remedy Common-Sense Morality's self-defeat while leaving the rest of its content intact, proponents of Common-Sense Morality should regard his revisions as costless improvements (p. 106).

A second argument proceeds independently of the concept of self-defeat. Parfit invites us to imagine that in one of his parent's dilemmas, two additional conditions hold: (1) the parents can communicate and (2) each parent can best advance the interests of his or her child by making a conditional promise to the other parents that if they refrain from giving priority to their own children, then he or she will refrain from doing that, too. (For example, Quinn and Ramona could each promise to save the other's son from the larger harm, rather than save her own son from the smaller harm.) In situations in which these two conditions hold, a parent contemplating making the promise would not be contemplating *deviating* from Common-Sense Morality. Instead, making the promise would be *required* by Common-Sense Morality: by hypothesis, that is each parent's best option for advancing the interests of his or her child. And once the promise has been made, even Common-Sense Morality would prescribe that it be kept, simply as an instance of its moral regard for promise-keeping generally. In a sense, then, Common-Sense Morality can be seen as endorsing exactly the kinds of revisions to it that Parfit proposes (p. 107).

Parfit acknowledges that communication among the parties is not always possible. But he claims that people can fulfill the aims of Common-Sense Morality more fully when they can communicate than when they cannot. This provides a reason for any supporter of the aims of Common-Sense Morality to take the communication scenario as a reference point when thinking about how it would be desirable for people in such situations to act, even if communication is not possible. Thus, the impossibility of communication does not undermine the basic thrust of the argument just given: any supporter of the aims of Common-Sense Morality would wish for people to act as if they had made, and were intent on complying with, the promises described above. So, we see again "a sense in which [Common-Sense Morality] itself tells us to accept this revised version of itself" (p. 108).[11]

The unified theory of morality

As I mentioned above, Part One culminates in Parfit's proposing a new theory of morality. He prepares the reader for his statement of this theory by commenting on how his discussions of Consequentialism and Common-Sense Morality reduce the "distance" or "disagreement" between those theories (p. 111). First, recall his discussion of Consequentialism – specifically, his discussion of its indirect collective self-defeat. One lesson of that discussion is that instead of being pure do-gooders, people should have strong desires to benefit their families and friends, and to do their work well, which would result in their having strong desires to benefit their pupils, patients, clients and other people that they have special professional relationships with. Additionally, people should have strong aversions to committing certain kinds of acts, such as murder and deception. All of these desires will impel people to be disposed to act as Common-Sense Morality requires (p. 112).

Second, recall Parfit's discussion of Common-Sense Morality. As we saw just above, Parfit argues that any adherent of Common-Sense Morality should embrace the revisions he proposes. He writes that the revised theory "is Consequentialist, giving to all of us common moral aims" (p. 111) – here Parfit may be claiming that the revised theory is agent-neutral. He then says that moving from Common-Sense Morality to the revised theory "*reduces the disagreement* between Common-Sense Morality and Consequentialism" (p. 111, emphasis added).[12]

With Consequentialism and Common-Sense Morality brought closer together in this way, Parfit conjectures that "We might be able to develop a theory that includes and combines revised versions of both" of them. "Call this the *Unified Theory*" (p. 112). Parfit actually says very little about the content of this theory, writing that developing it "would take at least a book" (p. 113). Instead, he emphasizes a major obstacle that the project would face: the differences in the moral judgements that follow from Consequentialism and Common-Sense Morality. Although Consequentialism can be brought closer to Common-Sense Morality by way of dispositions to act in ways that Common-Sense Morality prescribes, the two theories still have different contents. Thus, there will be cases in which complying with one theory means violating the other, and vice versa. In articulating the Unified Theory, "our greatest task would be to reconcile these conflicting beliefs" (p. 113). (Strangely, Parfit says nothing here about the extent to which the unification project could be furthered by the revisions to Common-Sense Morality that he recommended. Recall that he argued that they eradicate much if not all of the content of Common-Sense Morality that cannot be reconciled with Consequentialism.)

Let me conclude this overview of Part One by describing a conversation that took place on 20 February 1865. On that date, John Stuart Mill spent time with John Russell, whose son Bertrand would be born seven years later and receive the honour of having Mill as his godfather. The conversation turned to the topic of moral progress, and Russell's diary reports the following of Mill:

> It did one great good to hear him & raised one into a hopeful state of mind. He said the wish & intention to do good was good in itself—and he said the great thing was to consider one's opponents as one's allies; as people climbing the hill on the other side.[13]

Parfit quotes Mill's "climbing the hill on the other side" phrase in the penultimate paragraph of Part One, and characterizes his Unified Theory in that spirit (p. 114). Consequentialism and Common-Sense Morality, rather than being seen as rivals, can be seen as the outlooks of people who are on different sides of a mountain, but who share the goal of reaching the summit and seeing the whole truth of morality.

This metaphor remained a touchstone for Parfit. More than a decade later, he wrote a manuscript called *Climbing the Mountain*,[14] which, although never published, was circulated widely among moral philosophers. In 2011, much of this manuscript appeared as – or was superseded by – Parfit's two-volume treatise *On What Matters*. There Parfit writes that although Kantians, Contractualists and Consequentialists are often seen as disagreeing deeply with one another, that perception is a mistake: "These people are climbing the same mountain on different sides."[15] More important than the metaphor, one of the main claims of *On What Matters* is the convergence of the three views just mentioned. Thus, the quest for a unified theory that we find in Part One of *Reasons and Persons* is significant not only because of the impact of that book, but because it is the first major articulation of what would prove to be one of the major themes of Parfit's cumulative body of work.

Notes

[1] I would like to thank Chris Heathwood, Brian Hedden and Andrea Sauchelli for helpful comments on an earlier version of this chapter.

[2] Parenthetical in-text page references are to Parfit (1984/87).

[3] Eggleston (2013a: 3794).

[4] For criticisms of this defence of the Self-interest Theory, see Dancy (1997: 4–11) and Adams (1997: 255–256).

[5] Shafer-Landau (2018: 122).

[6] For a criticism of this defence of Consequentialism and Parfit's reply, see Gruzalski (1986a: 771–777) and Parfit (1986: 865, section 2). For further criticism, see Dancy (1997: 11–16).

[7] See, e.g., de Lazari-Radek and Singer (2010) and Eggleston (2013b).

[8] See, e.g., Jackson (1997), Eggleston (2000), Eggleston (2003) and Petersson (2004).

[9] Glover and Scott-Taggart (1975: 174–175).

[10] For criticism of this reasoning and Parfit's reply, see Kuflik (1986) and Parfit (1986: 849–854 and 865–867). For further criticism, see Gruzalski (1986b: 150–151) and Adams (1997: 256–261).

[11] For criticism of the inferences Parfit draws from the collective self-defeat of Common-Sense Morality, see Mendola (1986).

[12] For criticism of Parfit's claim that his proposed revisions bring Common-Sense Morality closer to Consequentialism, see Gruzalski (1986b: 145–150).

[13] Russell and Russell (1937: 373).

[14] MacFarquhar (2011: 50).

[15] Parfit (2011: vol. I, p. 419; see also vol. II, p. 259).

Bibliography

Adams, R. M. 1997. "Should Ethics Be More Impersonal?" in J. Dancy (ed.), *Reading Parfit* (Oxford: Blackwell Publishers), pp. 251–289.

Dancy, J. 1997. "Parfit and Indirectly Self-defeating Theories" in J. Dancy (ed.), *Reading Parfit* (Oxford: Blackwell Publishers), pp. 1–23.

de Lazari-Radek, K., and P. Singer. 2010. "Secrecy in Consequentialism: A Defence of Esoteric Morality." *Ratio* vol. 23, no. 1 (March), pp. 34–58.

Eggleston, B. 2000. "Should Consequentialists Make Parfit's Second Mistake? A Refutation of Jackson." *Australasian Journal of Philosophy* vol. 78, no. 1 (March), pp. 1–15.

Eggleston, B. 2003. "Does Participation Matter? An Inconsistency in Parfit's Moral Mathematics." *Utilitas* vol. 15, no. 1 (March), pp. 92–105.

Eggleston, B. 2013a. "Paradox of Happiness" in *The International Encyclopedia of Ethics*, edited by H. LaFollette (Blackwell Publishing Ltd., 2013), pp. 3794–3799. DOI: 10.1002/9781444367072.wbiee202.

Eggleston, B. 2013b. "Rejecting the Publicity Condition: The Inevitability of Esoteric Morality." *The Philosophical Quarterly* vol. 63, no. 250 (January), pp. 29–57.

Glover, J., and M. J. Scott-Taggart. 1975. "It Makes No Difference Whether or Not I Do It." *Proceedings of the Aristotelian Society, Supplementary Volumes*, vol. 49, pp. 171–209.

Gruzalski, B. 1986a. "Parfit's Impact on Utilitarianism." *Ethics* vol. 96, no. 4 (July), pp. 760–783.

Gruzalski, B. 1986b. "Parfit's Unified Theory of Morality." *Philosophical Studies* vol. 50, no. 1 (July), pp. 143–152.

Jackson, F. 1997. "Which Effects?" in J. Dancy (ed.), *Reading Parfit* (Oxford: Blackwell Publishers), pp. 42–53.

Kuflik, A. 1986. "A Defense of Common-Sense Morality." *Ethics* vol. 96, no. 4 (July), pp. 784–803.

MacFarquhar, L. 2011. "How to Be Good." *The New Yorker*, 5 September, p. 42–53.

Mendola, J. 1986. "Parfit on Directly Collectively Self-Defeating Moral Theories." *Philosophical Studies* vol. 50, no. 1 (July), pp. 153–166.

Parfit, D. 1984/87. *Reasons and Persons*, "Reprinted with further corrections 1987" edition (Oxford: Clarendon Press).

Parfit, D. 1986. "Comments." *Ethics* vol. 96, no. 4 (July), pp. 832–872.

Parfit, D. 2011. *On What Matters* (Oxford: Clarendon Press).

Petersson, B. 2004. "The Second Mistake in Moral Mathematics is not about the Worth of Mere Participation." *Utilitas* vol. 16, no. 3 (November), pp. 288–315.

Russell, B., and P. Russell. 1937. *The Amberley Papers: The Letters and Diaries of Bertrand Russell's Parents*. New York: W.W. Norton & Company, Inc.

Shafer-Landau, R. 2018. *The Fundamentals of Ethics*, 4th edition. Oxford: Oxford University Press.