Derek Parfit[1]

Derek Parfit is a British philosopher who has made major contributions to the study of ethics, practical reason, and the metaphysics of persons. Born in China in 1942, and educated at Oxford, he is now a Senior Research Fellow at All Souls College, Oxford. He is also a regular Visiting Professor at Rutgers University, Harvard University, and New York University.

Parfit's international reputation was already established in the early 1970s through a series of articles on personal identity. In his magnum opus, *Reasons and Persons* (1984), he presents his fullest account of this theme, as well as a wide-ranging exploration of rationality and morality. This work is recognized by many readers as the most important work in moral philosophy written since the early part of the twentieth century. It ranks in importance alongside the work that is its main inspiration, Henry Sidgwick's *Methods of Ethics* of 1874. *Reasons and Persons* set the agenda for many of the central debates in contemporary moral philosophy. It defined the terms for current discussions of personal identity and its moral significance, rational attitudes toward the past and the future, obligations to future generations, alternative conceptions of wellbeing, and the general structure of value. This book also served to initiate a number of important discussions by revealing new problems, some of which we will be discussing below.

Reason and Persons was followed by a number articles, many of which have played an similar agenda-setting role. These have ranged from contributions to social and political philosophy, such as his work on the value of equality, to his writings on philosophical cosmology, concerning the question of why the universe exists at all and has the orderly structure that it exhibits.[2] He has now nearly completed a second book, tentatively entitled *Climbing the Mountain*, which concerns moral theory. Though this book is still forthcoming, it has been widely circulated in draft form.

---

[1] I am indebted to Derek Parfit and to Larry Temkin for very helpful comments on an earlier draft of this paper.

[2] For the former, see "Equality or Priority?"; and for the latter, see "Why Anything? Why This?"

1: The Fact of Reasons

Though Parfit's writings are broad in scope, to a large extent they are unified by the central theme of reasons. He is concerned with the reasons that bear on the question of how we should act, and on the question of what we should care about. This theme will therefore be the organizing principle of what follows: after discussing Parfit's general conception of reasons, we will turn to a discussion of prudential reasons (reasons of self-interest), then to reasons of beneficence (reasons to help others), and then Parfit's recent work on the structure of moral reasons.

The reasons Parfit is concerned with are called *normative practical* reasons. They are *practical* since they bear on practical questions, and they are *normative* since they concern the question of what we *ought* to do, or *ought* to care about, rather than on the question of what we *in fact* do, or care about, or are motivated to do. An agent may fail to do what she ought to do, or she may fail to care about what she ought to care about, and so an agent may fail to be sufficiently motivated by her normative reasons. More generally, the *normative* force of reasons, or their force in favoring certain actions or concerns, must be distinguished from the *motivational* force of reasons, or their efficacy in motivating agents to act or to care. Parfit argues that there is a strong trend among philosophers to conflate, or to collapse the distinction between, normative force and motivational force, and that many of the central arguments in ethics and metaethics from Hume and Kant to the present day have involved such a conflation.[3] If we lose sight of the distinction between normative and motivational force, then the question of how we should act is reduced to the question of how we are motivated to act, or of how we would be motivated to act under specified circumstances, and so ethics is reduced to a branch of psychology. And this, according to Parfit, is a serious misunderstanding of the object of ethical inquiry.

Even when the conceptual distinction between normative and motivational force is recognized, it is often held that the two are very closely connected. According to the dominant approach to understanding practical reasons, which is represented by what

[3] See "Normativity."

Parfit calls *desire-based theories*, an agent's normative reason for or against an action always consist in a fact concerning how this action would fulfill or frustrate the agent's actual or counterfactual present desires, and the motivational force of such a reason in explained in terms of the strength of the corresponding desire. On the simplest desire-based theory, the desires that determine what an agent has most reason to do at a given time are the ultimate desires she actually has at that time. According to this theory, nothing is by nature worthy or unworthy of desire, and so every consistent set of desires is on an equal footing, none being more rational than any other. An agent has reason to act in some way just in case doing so would promote the satisfaction of her desires *whatever they may be*. Parfit first criticizes such theories in *Reasons and Persons*, where he argues that certain patterns of desire are inherently irrational. One example of an irrational pattern of desire is "future Tuesday indifference," which consists in currently being indifferent to the prospect of painful experiences one may undergo on future Tuesdays, while desiring to avoid painful experiences on every other day of the week (*R&P*, pp. 123-24). In *Climbing the Mountain*, Parfit discusses what he regards as the extremely implausible implications of the simple desire-based theory. This theory implies, for example, that if, at some particular time, one desires to drink sulfuric acid, and one has no desire to avoid the harmful consequences of doing so, then one is rationally required to do so, even if one is certain that one will regret having done so for the remainder of one's (possibly shortened) life.

In order to avoid such implications, many philosophers have adopted a more complex desire-based theory, according to which the desires one has reason to fulfill are not one's actual present desires, but the desires one would have if one knew and had carefully considered all the relevant facts. In particular, on this theory, the ends that one has non-instrumental reason to promote are not the ends that one currently desires for their own sake, but rather the ends that one would desire for their own sake if one had considered all the relevant facts. This theory, Parfit argues, is untenable. For desire-based theories must claim that facts about the objects of our desire can't give us reason to desire these objects as final ends. Therefore it must claim that the ultimate desires we would have were we to consider all the facts would be no more supported by reasons than our actual desires. But if these hypothetical desires are no more supported by reasons

that our actual desires, then there can be no grounds for asserting that it is these hypothetical desires, rather than our actual desires, that are the source of our reasons for action.

Instead of holding a desire-based theory of practical reasons, Parfit holds a *value-based* theory, according to which there are reasons for ultimate desires, namely facts about the objects of these desires that give us reason to desire them. And he holds that our reasons to promote an outcome are provided not by the fact that this outcome would satisfy our desires, but rather in the very same features of this outcome that give us reason to desire it. Thus, what gives us reason to want to avoid being tortured in the future, and to act in such a way as to prevent ourselves from being tortured in the future, is the fact that being tortured would be extremely painful. Since this fact is independent of our present desires, these reasons to not depend our currently having any desires which would be frustrated by being tortured in the future.

The question remains as to what we have reason to desire for its own sake. One answer to this question is that one's ultimate aim should be to maximize one's own well-being, or to ensure that one's life as a whole go as well as possible. This answer, which we shall discuss presently, is the target of many of Parfit's best-known arguments.

2: Prudential Reasons and Personal Identity

According to the *self-interest theory* of practical reason, all one has reason to care about for its own sake is one's own well-being, and what one has most reason to do is whatever would most promote one's well-being. Whenever there is anything else we should care about or promote, this is ultimately to be explained in terms of its contribution to our own well-being. While the desire-based theories of practical reason of the kind we discussed in the previous section are currently dominant among philosophers, Parfit holds that the self-interest theory has been the dominant theory of rationality among people in general for over two thousand years. These two theories are often conflated, since it is sometimes believed that each agent's fundamental desire is that her life as a whole go as well as possible, or that her own welfare be maximized. On this assumption, the action that most promotes the satisfaction of one's current desires always

coincides with the action that would make one's life go best as a whole. But this assumption is false. For people often care more about the nearer future than about the more distant future, and so many people would prefer a life that is better in the short run, but worse on the whole, to a life that is worse in the short run, but better on the whole. The desire-based theory implies that such agents ought rationally to act in ways that make their lives go worse on the whole. Further, the desire-based theory implies that if today I desire some outcome, and I know that tomorrow my desires will change and I will desire some opposing outcome, then I will be rationally required today to promote an outcome while recognizing that tomorrow I will be rationally required to try to prevent this outcome.

But to the self-interest theorist, these implications are unacceptable. On her view, if we will ever have reason to care about some event or outcome, we already have this reason now. The force of a reason to promote an outcome, she insists, is transmitted over time, and its strength is not affected by the distance of this outcomes from the present. And so our concern for how well we fare at future times must not be affected by the distance of these times from the present.

In part II of *Reasons and Persons*, Parfit argues that in making these claims, the self-interest theory occupies an unstable position between two alternative theories. On one side there are what we may call *fully relativistic* theories of reasons, like the simple desire-based theory, according to which what one has reason to care about, and to promote if one can, depends both on who we are and on where we are situated in time. In other words, practical reasons vary both from agent to agent and from time to time. On the other side there are *fully non-relativistic* theories, according to which what one has reason to care about and promote varies neither across agents nor across times. (An example of such a fully non-relativistic theory is rational consequentialism, according to which there is a single rational aim valid for everyone, namely that the history of the world go as well as possible as evaluated from an impartial point of view.) According to the self-interest theory, what one has reason to care about and promote varies from agent to agent (since each agent should be concerned with *his own* well-being) but it does not vary from time to time (since each agent should always have the aim of making his life as

a whole go as well as possible). In order to defend this middle position between these opposing kinds of theory, the self-interest theories must show that there is a principled reason for treating agents and times differently, and hence for requiring a partial attitude toward agents but an impartial attitude toward times. She must show, in other words, that differences among persons have a rational significance that differences among times lack.

One argument, made by Sidgwick, Rawls, and Nozick, is that any supposed requirement to be impartial with respect to persons fails to do justice to the separateness of persons.[4] In Sidgwick's words, "it would be contrary to common sense to deny that the distinction between any one individual and any other is real and fundamental" and hence to deny that this distinction should be "taken as fundamental in determining the ultimate end of rational conduct."[5] By contrast, it might be claimed that the passage of time is merely a subjective illusion, and so the distinction between the nearer and further future should not be taken as fundamental in determining this ultimate end. Parfit argues, however, that no such metaphysical defense of the self-interest theory can succeed.

For one thing, the most plausible version of the self-interest theory is not supported by any viable conception of the metaphysics of time. If one holds that the passage of time is an illusion and that this fact imposes constraints upon what patterns of concern can be rational, then the natural inference to draw is not merely that we should be impartial toward all *future* times, but that we should be impartial toward *all* times, including past times. But if we were impartial toward all times, then, other things being equal, we would have no preference for a situation in which a painful ordeal has occurred in the past over a situation in which this ordeal has yet to occur. But most of us have this preference: if we had amnesia, and could not remember the events of yesterday, and we knew that either we underwent a painful ordeal yesterday, or else this ordeal has yet to occur and is scheduled for tomorrow, most of us would be relieved if we discovered that the ordeal occurred yesterday. And most of us do not regard this bias as irrational. Thus, while a bias in favor of the nearer future over the further future may be irrational, it

---

[4] See Henry Sidgwick, *Methods of Ethics*, London: Macmillan, 1874, p. 498; John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1971), sections 5-6; and Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), pp. 32-33.

[5] Sidgwick, *op cit*.

appears that a bias in favor of the future over the past is not. But if the self-interest theory is to be defended on the basis of the metaphysical view that the passage of time is an illusion, then this theory must claim, counter-intuitively, that these biases are equally irrational.

Parfit argues that the self-interest theory, in addition to lacking support from any plausible conception of the metaphysics of time, is undermined by every defensible conception of the metaphysics of persons. Indeed, Parfit's best known argument againt the self-interest theory is found in his discussion of personal identity in part III of *Reasons and Persons*. Here Parfit argues that on any defensible conception of personal identity, there are possible cases in which what we have ultimate reason to care about is not our own welfare.

In order to show that what we have reason to care about need not be our own future, Parfit employs a famous thought-experiment, which is one of the many ingenious thought-experiments to be found in his writings. He first notes that as we normally think about personal identity, the part of the body that matters for retaining personal identity is the brain, so that if one's brain were transplanted into another body, one would continue t exist with in this new body. He further notes that as we normally think about personal identity, a person could survive an injury in which much of his brain is destroyed, so long as enough of his brain survives in order for her to retain most of her beliefs, intentions, and other psychological characteristics. Thus, as we normally think about personal identity, one could survive an operation in which half of one's brain is destroyed, and the other half is transplanted into another body.

Now consider two cases. In the first case, called *Single Transplantation*, Van Cleve's brain is cloven in half, and that the left half of his brain is transplanted into another body, though the right half is destroyed. Let us assume that most of Van Cleve's memories and other mental states are encoded in both halves of his brain, so that the preservation of either half of his brain is sufficient for him to retain psychological continuity. In this case, we would normally think that Van Cleve survives the operation, and lives on in the body to which the left half of his brain was transplanted.

Now consider *Double Tranplantation*. As in the case of single transplantation we just considered, the left half of Van Cleve's brain is preserved and transplanted into someone else's body. But in this case, the right half is also preserved, and is transplanted into someone else's body. Assume, further, that prior to the operation, both halves of Van Cleve's brain are nearly identical psychologically, since nearly all of Van Cleve's memories and other mental states are similarly encoded in each. Thus, after the operation, there will be one person (or one entity that appears to be a person), who has the left half of Van Cleve's brain, and who has most of Van Cleve's psychological characteristics, whom we may call "Lefty", and another person, or apparent person, who has the right half of Van Cleve's brain, and who likewise has most of Van Cleve's psychological characteristics, whom we may call "Righty." Suppose, finally, that after the operation, Lefty and Righty never interact. Does Van Cleve survive this operation? In other words, is there anyone who exists after this operation, and who is numerically identical with Van Cleve? It seems that there are five answers we could give to the question:

(i) Van Cleve is the same person as Lefty, but not the same person as Righty.

(ii) Van Cleve is the same person as Righty, but not the same person as Lefty.

(iii) Van Cleve is the same person as Lefty, and Van Cleve is the same person as Righty.

(iv) Van Cleve survives the operation as a divided person, of which Lefty and Righty are the both parts.

(v) Van Cleve does not survive the operation.

It seems that we should reject (i) and (ii), since in the case described there does not appear to be anything relevantly different between Van Cleve's relation to Lefty and his relation to Righty. Moreover, since Lefty is not the same person as Righty, they can't each be the same person as Van Cleve, and so we should reject (iii). Further, since Lefty and Righty are each persons, or at least each would be a person in the absence of the other, and since we are assuming that the two do not interact after the operation, there is

strong reason so reject the view that Lefty and Righty together constitute a single person. Therefore we should reject (iv). Hence, there are only two remaining alternatives. One is to adopt the fifth answer and assert that Van Cleve does not surive the operation. And the other is to reject every determinate answer and conclude that there is no fact of the matter concerning relations of identity between Van Cleve and those who exists after the operation. In either case, we cannot affirm that Van Cleve survives the operation of double transplantation.

Although we cannot affirm that Van Cleve survives the operation, we should affirm that being divided into two persons is *as good as* survival, or at least it is not nearly as bad as ordinary death. Surely the preservation of both halves of one's brain can't be significantly worse than the preservation of only one; this hardly seems like a case in which a double success would amount to a failure. Similarly, although we cannot affirm that Van Cleve is identical with either of the people who result from the operation, we can affirm that he has reason to be concerned about the welfare of these persons for its own sake. For whatever reason Van Cleve has, in Single Tranplantation, to be concerned about the welfare of the person who will have the left half of his brain cannot be negated the fact that, in Double Tranplantation, the right half of his brain will also be successfully transplanted.

But the self-interest theorist cannot make these claims, so long as she affirms that Van Cleve survives claims Single Transplantation, but does not affirm that Van Cleve survives Double Transplantation. For then, in Single Transplantation, she must affirm that Van Cleve has reason to care, for its own sake, about the person who will have the left half of his brain, but she cannot affirm this in the case of Double Transplantation. And this is an implausible position

One option open to the self-interest theorist is to deny that Van Cleve survives even in the case of Single Transplantation. So far we have been assuming that in Single Transplantation, Van Cleve survives because he retains enough of his brain to preserve most of his memories and other psychological characteristics. But one might adopt an alternative theory of personal identity according to which this is not enough for survival.

One might therefore say that there is no asymmetry between the attitudes Van Cleve ought to have in the two transplantation cases toward the person who will have his brain after the operation: in both cases he should recognize that this person is not him, and so he has no reason to care about this person's welfare for its own sake. However, even if one regarded this as a tenable position, it would not solve the general problem Parfit raises. For Parfit argues that on any plausible theory of personal identity, there will be some thought-experiment involving division, analogous to the transplantation thought-experiment we have been considering, in which the self-interest theory has similarly counterintuitive implications.[6]

Thus, Parfit concludes, we should reject the self-interest theory. Though we may have special reason to care about the future person with whom we are identical, our reason cannot plausibly be said to derive from the fact that this person will be *us*, for if this were the case, then we would lack this reason in cases of division. Since, in both the case of Single Transplantation and in the case of Double Transplantation, Van Cleve has special reason to be concerned about the person who has the left side of his brain after the operation, it seems that the relation that explains his special reason for concern must be a relation that obtains in both cases. And this relation, as we have seen, does not appear to be the relation of identity. Rather, it is the relation of *psychological continuity*.

Parfit defines psychological continuity in terms of psychological connections, where these are the sorts of relations that exist between an earlier experience and a later memory of this experience, or between an earlier intention and a later fulfillment of this intention. In response to the charge that definitions of personal identity in terms of such relations as remembering and intending are circular as these relations presuppose personal identity, Parfit, following Shoemaker's lead,[7] introduces the relations of *q-remembering* and *q-intending*, relations which are similar to those of remembering and intending but that are defined without presupposing personal identity. If we define a *person-stage* as a stage in the life of a person, then we may say that two person-stages are *strongly connected* just in case there are enough psychological connections between

---

[6] See "Experiences, Subjects, and Conceptual Schemes."

[7] Sidney Shoemaker, "Persons and Their Pasts," *American Philosophical Quarterly* 7, 1970.

them. And two person-stages are *psychologically continuous* just in case they both belong to a sequence of person-stages such that each person-stage belonging to this sequence is strongly connected to the preceding one.

The relevant relations in which Van Cleve stands to the person who will have the left half of his brain in both the single and the double transplantation cases are the relations of psychological continuity and connectedness. In both cases it is these relations, Parfit argues, that explain Van Cleve's special reason for concern. And it is also these relations, and not the relation of personal identity, that explain our own special reason to be concerned about our future welfare. And what is most important, Parfit argues, is the relation of psychological connectendess. Since we are connected to the future person-stages making up our lives to differing degrees, it can be rational, *pace* the self-interest theory, to be concerned about them to differing degrees.

Some have claimed that we cannot coherently deny the importance of personal identity. A prominent example is Christine Korsgaard, who gives an argument of the following form.[8] Any relation that we must necessarily take into account whenever we are deliberating is important from the practical point of view. But the relation of personal identity is such a relation. For when an agent deliberates, she is asking how *she* is to act, and the alternatives among which she is choosing always lie at some distance in the future. Hence, she must regard the actions that will be performed by an agent at some future time as *her* actions, which means that she must regard herself as identical with an agent who will exist in the future. But if any relation that we must take account of in practical reasoning is important from the practical point of view, and if the relation of personal identity is one such relation, then it follows that the relation of personal identity is important from the practical point of view. Thus, the practical importance of the relation of personal identity derives not, as Sidgwick suggested, from its being metaphysically real and fundamental, but rather from its being a necessary presupposition of the practical point of view.

---

[8] In "Personal Identity and the Unity of Agency: A Kantian Response to Parfit." *Philosophy & Public Affairs* (Spring 1989), 18(2):101-132.

But there is an obvious reply to this argument. Granted, in all actual cases of deliberation, we are deciding how *we* shall act in the future. But this is simply because there are no actual cases of fission. Suppose that Van Cleve knows that he will undergo Double Tranplantation, and that the body into which the left half of his brain will be transplanted is in a hospital in which there is a dangerous gas leak. Suppose that after the operation occurs, Lefty will have no time to plan his escape, and will only be able to leave the building alive if he takes immediate and appropriate action. If, prior to the operation, Van Cleve is given a map of the hospital, it seems that he could and should consult this map and deliberate concerning how to escape. But in so doing, he would be deciding not how *he* shall escape from the building, but rather how Lefty shall escape form the building. And the conclusions of such deliberation would be q-intentions whose objects are the actions of Lefty. It seems, therefore, that in cases of division, one can deliberate concerning the actions of an agent with whom one is not identical, and with whom one does not take oneself to be identical. And so it appears that, contrary to Korsgaard, the concept of personal identity over time does not play an ineliminable role in practical reasoning.

Moreover, if Korsgaard is right that claims about the practical importance of a relation can be justified in virtue not of its metaphysical status but rather of its ineliminable role in practical reasoning, then this will strengthen rather than undermine Parfit's position. For as the division case illustrates, the fundamental distinction we must draw among future actions in the context of practical reasoning is not a distinction between actions that we may perform and actions that others may perform, but rather between actions that are up to us, or that we can cause to occur by q-intending that they occur, and actions that are not up to us in this sense. But this is a question of psychological connectedness, not of identity. Thus, what must be presupposed from the practical point of view is not personal identity, put psychological connectedness, which is precisely the relation to which Parfit thinks we should give most weight.

In Parfit's view, we can coherently regard the relation of personal identity as having no significance in relation to the question of how we ought to act. Indeed, this is how we ought to regard this relation. Parfit holds, with the Buddha, that when we free ourselves

from the strangle-hold of the concept of personal identity, then we can abandon the illusion that in order to act rationally we must act selfishly, and we can recognize that very often, what we have most reason to do is to act in such a way as to benefit others, even at the expense of our own well-being.

3: Reasons of Beneficence

According to Parfit, our reasons to benefit others, or *reasons of beneficence*, are among our most important moral reasons. Thus any adequate moral theory must recognize such reasons, and must also specify their content so that we can determine whether our reasons of beneficence favor one course of action or another. Parfit shows, however, in part IV of *Reasons and Persons*, that this is no easy task, since all the prima facie candidate theories of beneficence have unacceptable implications.[9] This part of *Reasons and Persons*, though initially overshadowed by the part on personal identity, is increasingly becoming recognized for its fundamental importance.

One candidate conception of our reasons of beneficence includes the following claims:

(i) We have a greater reason of beneficence to choose outcome A than to choose outcome B just in case, on the whole, A would be better for people than B.

(ii) Unless there is someone whose level of welfare is higher in outcome A than in outcome B, A is not better for people than outcome B.

In cases where the same people will exist regardless of what we choose, or in what Parfit calls *same people choices*, (i) and (ii) have fairly plausible implications. But in cases where who will come to exist depends on how we act, these claims can have very implausible implications, for they fail to solve what Parfit calls the *Non-Identity Problem*.

Suppose we are choosing between two policies, *conservation*, in which we conserve our resources so that they are available for future generations, and *Depletion*, in which we

[9] Some of the arguments from this part of *Reasons and Persons* are developed further and strengthened in "Overpopulation and the Quality of Life."

consume these resources in the near future. Suppose that Depletion would have slightly better consequences for some people who are alive now, and that it would not have worse consequences for anyone will be alive over the next two centuries. Suppose, however, that at all times later than two hundred years from now, the prevailing level of welfare will be much higher if we choose Conservation than Depletion. Suppose, further, that our choice between these two alternatives will have very wide-ranging implications, significantly affecting the daily lives of everyone in the population. On this supposition, Parfit argues that we can reasonably assume that in the population of those affected by our decision, *who* will exist at times later than two hundred years in the future will depend on which of these policies we choose now, and that there is no one in this population who will exist more than two hundred years from now regardless of which of these policies we choose.

In this case, it seems clear that, on the whole, people will be better off if we choose conservation rather than depletion, and that we thus have greater reason of beneficence to choose conservation. But on the conception of beneficence we are now considering, we cannot draw this conclusion. For since there is no one in the affected population who will exist more than two hundred years from now independently of which alternative we choose, there is no one whose level of welfare would be greater if we choose conservation than if we choose depletion. Hence the view under consideration implies that we do not have greater reason of beneficence to choose conservation.

Thus, a common conception of beneficence runs into problems when faced with choices in which who will exist depends on how we act. Moreover, such cases present problems for a great many positions in moral philosophy. Parfit argues that they present serious problems for the moral theories of Gauthier, Harman, Mackie, Rawls, and Scanlon, among others (*R&P*, p. 523).

Any adequate moral theory must explain how an action can be wrong, and specifically wrong from the point of view of beneficence, even if there is no one for whom its outcome would be worse than any available alternative; or in other words, any adequate moral theory must solve the Non-Identity Problem. One obvious solution to

this problem is to conclude that the outcome that is best from the point of view of beneficence (and hence best simpliciter, all else being equal), is the outcome in which the total sum of human welfare or utility is greatest. Call this the *Impersonal Total Principle*. This principle implies that, other things being equal, conservation is preferable to depletion, since it would result in a greater sum of human welfare. Thus the Impersonal Total Principle gives the right answer in the case we have been considering.

Further problems arise, however, if we consider situations in which our choices will affect not only *who* will live, but also *how many* people will live. For the sum total of utility in a population can be increased either by increasing the average level of welfare in the population, or by adding people whose level of welfare is above the zero-level (the level below which lives cease to be worth living). Thus, one population can involve a greater sum total of welfare than a second population even if, on average, people are much better off in the second population, so long as the first population involves a sufficiently large number of people, and so long as everyone in this population has a life that is worth living. Hence the Impersonal Total Principle therefore what Parfit calls the *Repugnant Conclusion*:

> For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence would be better, if other things are equal, even though its members have lives that are barely worth living (*R&P*, p. 388).

To avoid this conclusion we can move to the *Impersonal Average Principle*, according to which the outcome that is best from the point of view of beneficence is the outcome in which people's lives go best on average. This principle avoids the Repugnant Conclusion, since the average level of welfare is clearly higher in a population in which everyone has a very high quality of life than in a population, however large, in which everyone lives a life that is barely worth living. But the Average Principle has its own problematic implications. The worst of these arise in cases in which one must choose the lesser of evils. Suppose we are choosing between Hell A, which consists in a population of a billion innocent people, all of whom experience extreme agony throughout their

lives, and Hell B, which consists in these same billion people undergoing this same degree of agony, plus an additional billion innocent people who likewise experience extreme agony throughout their lives, but to a slightly lesser degree. The Average Principle implies that since the average level of welfare is slightly higher in Hell B than Hell A, our reasons of beneficence favor the choice Hell B.

We can avoid this conclusion, while at the same time avoiding the Repugnant Conclusion, if we suppose that there is an asymmetry between the positive value of lives that are worth living, and the disvalue of lives that are wretched, or not worth living. We might hold that, contrary to the *Impersonal Average Principle*, numbers matter, so that by adding people with positive levels welfare (people whose lives are not worth living) we improve an outcome, regardless of whether we increase the average level of welfare, and that by adding people with negative levels of welfare we make an outcome worse. And yet—and this is where the asymmetry enters—we may hold that there is a limit to how much we can improve a situation by adding people at any given positive level of welfare, but not limit to how much we can make an outcome worse by adding people at any given negative level of welfare. By assuming that there is a limit in the first case, we avoid the Repugnant Conclusion, and by assuming that there is no limit in the second case, we avoid the conclusion that Hell B is preferable to Hell A. However, we now face another unacceptable implication. For now our view implies that if we begin with a population of ten billion people, all but one of whom has an absolutely wonderful life, but one of whom has a life that is not worth living, and we then progressively multiply this population, retaining the proportion between those with wonderful lives and those with lives not worth living, then the disvalue of the tiny fraction of bad lives will come to swamp the positive value of the wonderful lives, so that we eventually reach a world that is worse than a world in which no one exists at all. Parfit calls this the *Absurd Conclusion*.[10]

Thus, in attempting to formulate an adequate principle of beneficence, we seem to be caught between the Scylla of the Repugnant Conclusion and the Charybdis of the Absurd conclusion. Naturally, Parfit considers ways in which we might attempt to navigate a course between them. We might distinguish between three kinds of lives: bad

[10] For a more precise characterization of this conclusion, see *Reasons and Persons*, pp. 410-411.

lives (lives that are not worth living), good lives (lives that are well above the level at which they cease to be worth living), and mediocre lives (lives that are only marginally above the level at which they cease to be worth living). And we might hold that while the positive values of good lives and the disvalues of bad lives should be added up in a similar manner, the positive values of mediocre lives should be added up differently. One solution is to say that while there is no limit to the value or disvalue of additional good or bad lives, there is an upper limit to the value of additional mediocre lives. By placing a limit on the value of additional mediocre lives, we avoid the Repugnant Conclusion, and by placing no limit on the value of additional good lives, we avoid the Absurd Conclusion. Call this the *non-lexical solution*.[11] An alternative solution is to say that while there is no upper limit to the value of additional lives of any kind, the value or disvalues contributed by good and bad lives infinitely outweighs, and hence always takes precedence over, the value contributed by mediocre lives, so that the only significance of mediocre lives is to break ties between outcomes that are equally good with respect to good and bad lives. Since, on this view, the value of good lives always has precedence over the value of mediocre lives, we avoid the Repugnant Conclusion, and since the disvalue of bad lives does not always take precedence over the value of good lives, we avoid the Absurd Conclusion. Call this the *lexical solution*.

Parfit argues, however, that these solutions are unsatisfactory. He demonstrates that while they enable us to avoid the dilemma between the Repugnant Conclusion and the Absurd Conclusion, they leave us with a dilemma between a variant of the Repugnant Conclusion and a variant of the Absurd Conclusion. And since the variants of these conclusions are nearly as counterintuitive as the original conclusions, these solutions, Parfit argues, remain unacceptable.

And there are, I believe, further reasons for rejecting these two solutions, in addition to those given by Parfit. First, there is strong reason to reject the view that value of good lives infinitely outweighs the value of mediocre lives. For given any good life, *G*, there is some possible mediocre life, *M*, such that *G* and *M* can be connected by a

[11] This corresponds to what Parfit calls the "appeal to the valueless level." See *Reasons and Persons*, pp. 412-414.

chain of possible lives wherein no two successive lives differ from one another significantly in any important respect. And if two lives do not differ from one another significantly in any important respect, then neither of these lives will be infinitely outweighed in value by the other. And if two lives, G and M, are connected by a finite chain of possible lives, such that no life belonging to this chain infinitely outweighs the next life in value, then it follows that the value of life G cannot infinitely outweigh the value of life M.[12] In other words, the value of good lives cannot infinitely outweigh the value of mediocre lives. And so we should reject the lexical solution.

But there is also strong reason to reject the non-lexical solution. For it has the implausible implication that we should give more weight to improving the lives of the better off than to improving the lives of the worse off. Let P_1 and P_2 be two populations of equal size such that everyone in P_1 has a level of welfare that is at the dividing line between good lives and mediocre lives, and everyone in P_2 has a level of welfare that is slightly below this dividing line. Formally speaking, if we let g represent the minimum level of welfare for a good life, then we can say that everyone in P_1 has a level of welfare of g, and that everyone on P2 has a level of welfare of $g - \Delta$. Now suppose we have two options: we can either raise the level of welfare of everyone in the better-off population, P1, by a margin of d, so that they all attain a level of welfare of g + d, or we can raise the level of welfare of everyone in the worse-off population, P2, by this same margin, so that they all attain a level of welfare of g. Intuitively, if either alternative is better than the other, then it's the alternative of improving the lives of those in P2, since they are worse off to begin with. But if we adopt the non-lexical solution, we must accept the counterintuitive implication that, so long as the two populations are large enough, it would be better to improve the lives of the better off people (P1) than to improve the lives of an equal number of worse off people by an equal margin

This conclusion follows because on the non-lexical view, it is true of both population P1 and P2 that, as we increase its size, we increase the amount of good we could do by improving the lives of everyone in it by a margin of d. However, on this

[12] Ruth Chang presents a related argument in her introduction to *Incommensurability, Incomparability, and Practical Reason* (Cambridge, Mass: Harvard University Press, 1998).

view, since initially, people in P2 have mediocre lives, as we increase the size of this population, the amount of good we could do by improving the lives of everyone in it by d approaches an upper limit. But since people in P1 initially have good lives, this view implies that as we increase the size of this population, there is no upper limit to how much good we could do by improving the lives of everyone in this population by d. Therefore, if we make the two populations large enough, there will come a point where it will be better to improve the lives of those in the better-off population, P_1 by a margin of Δ than to improve the lives of those in the worse-off population, P_1 by this same margin.[13]

Parfit's explorations of our reasons of beneficence demonstrate a great difficulty of moral theory. The problem is not that there are too many plausible alternative moral theories, and hence that there is too much room for reasonable disagreement. The problem is rather that there is no moral theory that appears to be plausible. For any plausible moral theory would need to account for our reasons of beneficence, and every account of such reasons that has yet been offered has intolerable implications.

4: Impartial Reasons and Morality

Reasons of beneficence are what we may call *teleological* reasons, in the sense that they are reasons to promote ends. And they are also *impartial* reasons, in the sense that if anyone has a reason of beneficence to desire some end, and to promote this end given the opportunity, then everyone has this reason to do so. But while reasons of beneficence are teleological and impartial, they are not the only reasons of this kind, since there are other social, cultural, and ecological ends that we have impartial reason to promote for their own sake. Moral theories differ, however, concerning the moral significance they attribute to such impartial teleological reasons. According to consequentialist moral theories, such reasons are absolutely fundamental, as they are the basis for all moral requirements. Such reasons have not traditionally played a central role in the kinds of moral theories that are the main rivals to consequentialism, such as contractualism and

[13] In "Equality or Priority," Parfit presents and defends a view, called *prioritarianism*, according to which we ought to give priority to the welfare of the worse off. Since the non-lexical solution has the opposite implication, we may call it *antiprioritarian*.

Kantianism. Contractualists and Kantians attempt to ground moral obligations not in impartially valuable ends, but rather in terms of principles that could be rationally chosen or rationally willed, and the rationality of the choice in question is in turn understood without reference to impartial teleological reasons. Parfit argues that no adequate moral theory can be grounded in this manner. If moral obligations are to be derived from principles that we could rationally choose, then the rationality of this choice must be understood in terms of *all* the relevant reasons, including reasons of the impartial, teleological variety. And Parfit goes on to argue that when the Kantian and contractualist theories are formulated in this way, then Kantians, contractualists and utilitarians will all converge on theories that are equivalent from the practical point of view. Thus, while the proponents of Kantianism and contractualism may have intended their theories to support the rejection of consequentialist principles, the best versions of their theories in fact constitute the strongest defense of such principles. Thus, in Parfit's view, the Kantians, contractualists and consequentialists have all been climbing the same mountain from different sides.[14]

Redo next paragraph.

In Kant's view, the fundamental moral principle, or the *categorical imperative*, can be given a variety of formulations, but all of these are equivalent. Parfit shows, however, that on any reasonable assumptions, Kant's various formulations of the categorical imperative are not in fact equivalent, and that some of these formulations could not possibly serve as the fundamental principle of morality. According to the best known formulation, the Formula of Universal Law, one acts rightly just in case one acts on a maxim that one could will to be a universal law, where "could" here means "without incoherence." Parfit shows that this fails to rule out many impermissible actions. Consider the following maxim "if one is white and one is able to enslave a black person, do so." Acting on this maxim would clearly be wrong. But a white person could, without incoherence, will that this maxim be a universal law, or in other words that white people enslave black people whenever possible. For even if there is a kind of rational

[14] Parfit's arguments for this conclusion receive their fullest presentation is in *Climbing the Mountain*, but they were first sketched in "What We Can Rationally Will."

incoherence involved in willing one's own enslavement, there does not seem to be any rational incoherence involved in willing the enslavement of someone else. The problem with the maxim under consideration is not that it couldn't be willed as a universal law by *anyone*, but rather that it could not be willed as a universal law by *everyone*, and in particular, that it couldn't be so willed by blacks.

A better candidate for the fundamental principle of morality is thus the following: one acts rightly just in case one acts on a maxim that everyone could coherently will to be a universal law. But even this is too permissive, since far too many immoral maxims could be willed by everyone to be universal laws without incoherence—there is no contradiction, for example, in universalizing a maxim of causing as much pain as possible. Hence, according to Parfit, the best formulation of a principle of universal law concerns not what everyone could will *coherently*, but rather what everyone could will rationally, in the sense of *having sufficient reason* to will. The best formulation, Parfit argues, can be stated as follows: one acts rightly just in case one acts on principles whose universal acceptance everyone would have sufficient reason to will, or to choose. Parfit calls this the Kantian Contractualist Formula, since it bases the rightness or wrongness of an action on principles that all agents could rationally agree to. Parfit argues that the Kantian Contractualist Formula represents not only the best version of Kantianism, but also the best version of contractualism.[15]

The Kantian Contractualist Formula presupposes that there are principles whose universal acceptance each of us would have sufficient reason to will, or to choose, were we in a position to choose the principles that are to be accepted by everyone. But whether there are any such principles depends on what our reasons are, and on the strength of these reasons. Suppose, for example, that our only reasons are prudential reasons. In this case, it is unlikely that there would be any principles whose universal acceptance everyone would have sufficient reason to choose, since everyone would have decisive reason to choose the universal acceptance of principles that would be optimal in relation to *her own* interests, and it is unlikely that any principles would be optimal in relation to everyone's interests. Suppose, however, that apart from any prudential or

[15] See chapter 13 of *Climbing the Mountain*.

other partial reasons we may have, we also have impartial teleological reason to choose outcomes that are best from a point of view that is valid for everyone. And suppose, further, that it is always rationally permissible (though perhaps not rationally obligatory) to give significant weight to these impartial reasons. In this case, Parfit argues, there will be principles whose universal acceptance everyone would have sufficient reason to will. And these will be precisely those principles whose universal acceptance would have the best consequences from an impartial point of view; that is, these will be the *rule-consequentialist* principles. For these rule-consequentialist principles are the ones that each agent would have strongest impartial reason to choose, and these impartial reasons would in each case constitute sufficient, though perhaps not decisive, reason for the agent in question to choose these principles. But if the acceptance of these principles would not make things go best from an impartial point of view, then there will always be someone who has decisive reason not to choose their universal acceptance. Thus, the only principles that everyone has sufficient reason to choose that everyone accept are the rule-consequentialist principles. And so it follows from Kantian Contractualism that one acts rightly just in case one acts on rule-consequentialist principles.

There are, however, strong objections to rule consequentialist principles. Therefore, if the best versions of Kantianism and of contractualism imply that we act rightly just in case we act on such principles, these objections will count equally against Kantianism and contractualism. Indeed, one of the strongest objections to rule-consequentialist principles can be found in chapter 12 of *Climbing the Mountain*. The problem is that there are principles whose universal acceptance would make things go best or equal-best, but that it would be clearly immoral to act on. Consider the following principle: "never use violence, unless some other people have used aggressive violence, in which case kill as many people as possible." This principle might well be a principle whose universal acceptance would make things go as well as possible, and hence a principle whose universal acceptance everyone would have sufficient reason to choose. For if everyone followed this principle, then no one would ever use violence. But to follow this principle in the actual world, where there will always be others who have used aggressive violence, would involve killing as many people as possible.

To avoid this problem, Parfit claims, we must revise rule consequentialism, so that it states that we act rightly just in case we act on principles whose acceptance *by any number of people* would make things go best. We must similarly revise Kantian contractualism, so that it states that we act rightly just in case we act on principles whose acceptance *by any number of people* everyone would have sufficient reason to choose. That is, in order to act rightly, we must act on principles whose acceptance we could rationally will not only in a situation in which we are choosing the principles to be acted on by everyone, but also in a situation in which we are choosing principles to be acted on by any smaller number of people. And according to Parfit, when consequentialism and Kantian Contractualism are reformulated in this way, they once again converge.

It is doubtful, however, that there are sufficiently many principles satisfying the descriptions in these revised formulation. That is, it is doubtful that in every choice situation there is some principle one could act on whose acceptance by any number of people would make things go best, or whose acceptance by any number of people everyone could rationally will. Consider, for example, the following two rules:

P1: Make a reasonable effort to benefit the poor, but give significant priority to the interests of the near and dear.

P2: Act in such a way as to maximally benefit humanity as a whole, without favoring anyone's interests over anyone else's.

If our choice concerned what principle would be followed by only a single individual, then we may have stronger impartial reason to choose that she accept P2 rather than P1, since in the actual world, there are countless desperately poor people who would benefit far more from her accepting P2 than from her accepting P1, and this benefit would, from an impartial point of view, more than outweigh any loss to the agent in question, or to her near and dear, that would result from her accepting P2. But if we were in a position to choose the principle to be accepted by everyone, then we might have stronger impartial reason to choose P1 than P2. For regardless of whether everyone accepts P1 or P2, poverty will be eliminated or nearly eliminated. But if everyone were to accept P2, then no one could have close personal relations with the near and dear, and this would

arguably be a significant, uncompensated loss. Thus, it seems that the rule that we would have strongest impersonal reason to choose that one person accept differs from the rule we would have strongest impartial reason to choose that everyone accept.[16]

Parfit suggests that we can solve this problem, and arrive at principles whose acceptance by any number of people would be optimal, if we allow for conditional principles of the form "Do A, unless the number or proportion of A-doers is or will be below some threshold, in which case do B."[17] Thus, in the present case, the relevant conditional principle might be the following:

> P3: Make a reasonable effort to benefit the poor, while giving significant priority to the interests of the near and dear, unless there are insufficiently many people who make such a reasonable effort to benefit the poor, in which case act in such a way as to maximally benefit humanity as a whole."

But would the acceptance by everyone of P3 be as good as the acceptance by everyone of P2? In both cases, everyone would give significant priority to the near and dear. But in the former case, this priority would be conditional. That, if everyone accepted P3, they would be willing to sacrifice the interests of the near and dear if insufficiently many people made a reasonable effort to benefit the poor. And it may be that personal relationships that involved this kind of conditional commitment would be less valuable than relationships involving unconditional commitment.

And so we appear to be faced with a dilemma: if we say that the moral principles are those whose universal acceptance would make things go best, then we get the result that too many alternative principles count as moral, many of which are clearly dreadful. And if, on the other hand, we say that the moral principles are the ones acceptance by any number of people would make things go best, then we may get the result that principles count as moral, or at least that too few principles count as moral.

---

[16] Michael Ridge argues very forcelly for this conclusion in "Climb Every Mountain?" forthcoming in *Ratio*.

[17] See *Climbing the Mountain*, Chapter 12.

We can solve this problem by defining the relevant principles in terms of *compliance* rather than acceptance. For a person might happen to comply with a rule, in the sense that all his actions happen to be in accordance with this rule, without his accepting or being guided by this rule. Hence, we might define the moral principles as the principles compliance with which by any subset of people would make things go best. For conditional principles such as P3 might well be principles *compliance* with which by any number of people would make things go best, or equal-best, even if their universal acceptance would not make things go best. After all, in a world in which sufficiently many people make a reasonable effort to benefit the poor, P1 and P3 make the same prescriptions, and so anyone who complies with P1 will also comply with P3. Thus, universal compliance with P3 would be just as good as universal compliance with P1. If a principle has the feature that compliance with it by any number or people would make things go best, then we can call this principle *adaptable*.[18]

Thus, we can avoid the dilemma indicated above if we reformulate consequentialism so that it states that an act is right just in case it accords with *adaptable* principles. And we should similarly revise the Kantian Contractualist Formula so that it states than an act is right just in case it accords with principles compliance with which by any subset of people everyone could rationally will. If Parfit's arguments are sound, the resulting formulations will be equivalent: both formulations will permit just those actions that conform with adaptable principles. But if we make these revisions, then we come very close to adopting act-consequentialism, since it can be shown that any action that is permissible according to adaptable principles must be an action that makes things go best, and so it must also be permissible according to act-consequentialist principles. We may, therefore, be faced with the conclusion that in their best formulations, consequentialism, Kantianism and contractualism converge on a view that in many ways resembles act consequentialism. If this is so, then Parfit's thesis that the three main schools or moral thought converge will still be vindicated. But the summit at which they converge as they climb the moral mountain will turn out to be very far from two of the three base camps from which they began their ascent.

---

[18] See Donald Regan, *Utilitarianism and Cooperation*, (Oxford: Oxford University Press, 1980).

5: Conclusion

Parfit's works have been tremendously influential. Their significance lies not only in the ideas they present, but equally in the manner in which these ideas are presented. His works contain a clarity of prose, a rigor of argumentation, a thoroughness in the exploration of theoretical alternatives, an ingenuity and imaginativeness in the construction of examples, and a breadth of argumentative strategies that had never before been seen in moral philosophy. Countless readers of Parfit, including many of today's leading ethicists, have found in his works a revelation of how moral philosophy can fruitfully be done, and of how undeniable progress in moral philosophy can be made.

Parfit begins *Reasons and Persons* with the following epigraph from Nietzsche: "… all the daring of the lover of knowledge is permitted again; the sea, our sea, lies open again; perhaps there has never been such an 'open sea'."[19] We can only expect that much of the future progress in moral philosophy, like much of its recent progress, will be made in the exploration the open sea that Parfit's writings have revealed.

Works by Parfit

"Personal identity," *The Philosophical Review*, Vol. 80, No. 1, 3-27, 1971.

Reasons and Persons, Oxford: Oxford University Press, 1984.

"Overpopulation and the Quality of Life," in *Applied Ethics*, edited by Peter Singer, Oxford University Press, 1986.

"Equality or Priority?" delivered as the Lindley Lecture at the University of Kansas, 21, November 1991. Reprinted in *The Ideal of Equality*, edited by Matthew Clayton and Andrew Williams, MacMillan Press Ltd, and St. Martin's Press, Inc., 2000.

"The Unimportance of Identity," in *Identity*, edited by H. Harris, Oxford University Press, 1995.

"Reasons and Motivation," *Proceedings of the Aristotelian Society*, Supplementary Volume, 1997.

[19] This quotation is from *The Gay Science*, section 343.

"Why Anything? Why This?" *The London Review of Books*, 22 January and 5 February, 1998.

"Experiences, Subjects, and Conceptual Schemes," *Philosophical Topics* 26, No.s 1 & 2, Spring & Fall 1999.

"Rationality and Reasons," in *Exploring Practical Philosophy*, edited by Dan Egonsson, et al Ashgate 2001.

"What We Could Rationally Will," *The Tanner Lectures on Human Values*, Salt Lake City: University of Utah Press, 2004), pp. 285-369.

"Normativity," in *Oxford Studies in Metaethics*, Volume 1, edited by Russ Shafer-Landau, Oxford: Oxford University Press, 2006.