Must Kantian Contractualism and Rule Consequentialism Converge?

Derek Parfit’s *On What Matters* argues that three prominent theories in normative ethics converge.[[1]](#footnote--1) These three theories are Kantian ethics suitably revised, Scanlon’s Contractualism suitably revised, and Rule Consequentialism. Parfit’s explicit argument for the convergence of Kantian Contractualism and Rule Consequentialism is as important as any argument in normative ethics. This argument will be the main focus of this paper. The final section of the paper suggests that *On What Matters* also contains an *implicit* argument for normative convergence.

**Section 1: Kant’s Universal Law Formula**

Parfit carefully considers the strengths and defects of many different possible formulations of Kant’s categorical imperative, including many that go beyond Kant’s own formulations. Parfit thinks that all the formulations he considers are Kantian in spirit, and that one of them is preeminently plausible. Here I bypass disputes about whether Parfit’s interpretations of Kant are accurate. Likewise, I bypass disputes about whether there are some ideas in Kant that are more helpful than anything in the Kantian theory that Parfit develops.

Parfit attacks one after another of Kant’s formulations by showing that they have very implausible implications. A formulation has very implausible implications if the formulation implies that a kind of act is morally wrong and yet, clearly, this kind of act is *not* morally wrong. Likewise, a formulation has very implausible implications if the formulation implies that a kind of act is morally permissible and yet, clearly, this kind of act is morally *im*permissible. Whenever Parfit finds that a formulation of the categorical imperative has very implausible implications, he rejects that formulation as implausible.

Kant’s Universal Law Formula of the categorical imperative tells each to act on maxims that he or she can will to be universal laws. Had Kant recognized the difference between the question ‘what could *the agent* rationally will to be universally accepted?’ and the question ‘what could *everyone* rationally will to be universally accepted?’ he would have chosen the question about what everyone could rationally will (*OWM*, vol. 1, p. 340). After all, Kant thought that his various formulations of the categorical imperative agree with one another. And his other formulations of the categorical imperative make clear that he was interested in not merely what the individual agent could rationally will but what everyone could rationally will.

Parfit’s preferred revision of Kantianism is Kantian Contractualism, the theory that everyone ought to follow the principles whose universal acceptance *everyone* could rationally will. This theory obviously has a foundational principle: that everyone ought to follow the principles whose universal acceptance everyone could rationally will. The foundational principle prescribes compliance with other principles that pass a certain test. To help prevent confusion between the foundational principle and the other principles, I will call these other principles *rules*.

Sometimes Kantian Contractualism gets expressed, not as an explicit prescription, but as a proposition about which acts are wrong and why. Kantian Contractualism’s account of wrong acts is that acts are wrong if and because they are forbidden by rules whose universal acceptance everyone could rationally will. The theory also holds that acts are permissible if and because they are permitted by rules whose universal acceptance everyone could rationally will, and acts are morally required if and because they are required by rules whose universal acceptance everyone could rationally will.

**Section 2: Reasons for Willing**

What constitutes rational willing? This is not the place to get caught up in complications that arise when what appear to the agent to be good reasons are in fact not good reasons. So let us suppose the agent knows all the relevant reason-giving facts, knows all the reasons, and knows the comparative weight or importance of the reasons. At least when all these conditions hold, what the agent can *rationally* will depends entirely on what there really is sufficient normative reason for the agent to will.

A common view has been that normative reasons for action derive from the agent’s desires, or (in Bernard Williams’s terminology) from the agent’s ‘subjective motivational set’.[[2]](#footnote-0) Since people’s desires vary widely, so do their normative reasons, according to this ‘internalist’ view. Indeed, this view must admit that normative reasons can be just as fickle, random, and even downright perverse as ‘subjective motivational sets’ can be.

Rejecting desire-based views of practical reasons, Parfit holds that practical reasons are grounded in the features of things that make these things worth choosing, pursuing, promoting, protecting, cherishing, commending, admiring, etc. Practical reasons depend on the features of things that merit such pro-attitudes, not on the ‘internal’, subjective motivational sets of the agent (*OWM*, vol. 1, chs. 2–6).

Parfit distinguishes four kinds of reasons that might be thought relevant to what an agent could rationally will:

1. Self-interested reasons for the agent to will what is beneficial to herself
2. Agent-relative altruistic reasons for the agent to will what is beneficial to people specially connected to her (e.g., her friends and family)
3. Impartial reasons
4. ‘Deontic’ reasons for the agent to will what is morally right

If the categorical imperative is the foundational principle of morality, then deontic reasons to will what is morally right cannot play any role in determining what the categorical imperative requires or permits. In terms of the structure of Kantian normative ethics, deontic reasons must come from, or be outputs of, the categorical imperative rather than serve as inputs to it. Consider this picture of the categorical imperative’s ambition:

|  |
| --- |
| Rightness (as output)  ↑ |
| Categorical Imperative |

But if we say that what we have sufficient reason to will depends on our reasons to will what is right, the picture is:

|  |
| --- |
| Categorical Imperative (as output)  ↑ |
| Rightness |

Clearly we would be offering a uselessly circular account of moral rightness if we said that what makes right acts right is that everyone has sufficient reason to will rules that require those acts and then we conceded that these sufficient reasons come from the rightness of those acts.

Since, in determining what is right, reasons for willing cannot come from rightness, we are left with self-interested reasons, agent-relative altruistic reasons, and impartial reasons.

Parfit introduced the agent-relative/agent-neutral terminology, though the distinction is much older.[[3]](#footnote-1) He does not use this terminology in *On What Matters*. However, he employs the distinction even if he eschews the terminology.

He writes,

We have *self-interested* reasons to care about our own well-being, and *altruistic* reasons to care about the well-being of other people. … We can have strong reasons to care about the well-being of certain other people, such as our close relatives and other people whom we love. Like self-interested reasons, these altruistic reasons are *person-relative* or *partial* in the sense that these are reasons to be specially concerned about the well-being of people who are in certain ways *related to us*. (*OWM*, vol. 1, p. 41)

Person-relative altruistic reasons share with self-interested reasons the property of being, in Parfit’s earlier terminology, *agent-relative*.

Parfit explains what he means by an impartial point of view:

We have an impartial point of view when we are considering possible events that would affect or involve people all of whom are strangers to us. When our actual point of view is not impartial, we can think about possible events from an imagined impartial point of view. We can do that by imagining possible events that are relevantly similar, except that the people involved are all strangers to us. (*OWM*, vol. 1, pp. 40–41; see also vol. 2, p. 150)

So if impartial reasons are the reasons we would have if we considered things from an impartial point of view, and if an impartial point of view is one that ignores any special connections we have to certain people, then impartial reasons are, in Parfit’s earlier terminology, agent-neutral reasons.

I wrote above that Parfit does not use the agent-relative/agent-neutral terminology in *On What Matters*. Avoiding new philosophical terminology is wise where there is already perfectly familiar terminology for the very same thing. *On What Matters* coins the term ‘person-relative’ to mean what ‘agent-relative’ has come to mean. That seems to me no loss. But it does seem to me a loss that *On What Matters* uses the term ‘impartial’ to mean what ‘agent-neutral’ has come to mean. The term ‘impartial’ already has lots of different meanings. Most relevantly, ‘impartial’ is often used to mean ‘caring equally about everyone’. Indeed, it is so hard not to think of this meaning that there is the danger that we will at least sometimes forget that Parfit means by ‘impartiality’ mere ‘agent-neutrality’.

Having distinguished self-interested reasons, agent-relative altruistic reasons, and agent-neutral (impartial) reasons, let us now consider how these reasons determine which rules can be rationally willed for everyone to accept.

Each agent has self-interested reason to reject the universal acceptance of every set of moral rules other than the set whose universal acceptance would be most beneficial to herself. Hence, were self-interested reasons the only kind of reasons, there might well be no rules everyone had sufficient reason to will. Even if there were at least one set of rules that everyone had sufficient *self-interested* reason to will, that set would probably provide very little protection or aid to the powerless. The lack of protection for and aid to the powerless renders this set of rules intuitively inadequate. This is the old complaint against Hobbesian contractualism that it contracts morality too far (Parfit, *OWM*, vol. 1, pp. 345–46).

When choosing rules for everyone to accept, everyone has most impartial reason to will the rules whose universal acceptance would be impartially best. What would be impartially best, according to Parfit, is whatever impartial reasons favour on balance.[[4]](#footnote-2) On this buck-passing account of ‘impartially best’, ‘everyone has most impartial reason to will the rules whose universal acceptance would be impartially best’ equates to ‘everyone has most impartial reason to will the rules whose universal acceptance is favoured by impartial reasons’. Let us use the term ‘optimific rules’ to refer to the rules whose universal acceptance is favoured by impartial reasons.

Universal acceptance of the optimific rules might be disadvantageous to you compared to universal acceptance of some other rules. If universal acceptance of the optimific rules would be disadvantageous to you, you have *self-interested* reason *not* to will the universal acceptance of them. However, Parfit proposes that the *impartial* reason you have to will the universal acceptance of optimific rules is *sufficient* even ifyou have opposed self-interested reason.

Just as optimific rules might be disadvantageous to you, they might be disadvantageous to your friends and family. So we need to ask whether, when one is deciding which rules to will that everyone accept, impartial reasons are sufficient to win out over *agent-relative altruistic* reasons? And, even if impartial reasons are sufficient to win that contest, are they also sufficient to win out over the *combination* of self-interested reasons and agent-relative altruistic ones? Parfit argues that the answer is yes: impartial reason to will the universal acceptance of the optimific rules is always sufficient for willing the optimific rules (*OWM*, vol. 1, pp. 380–89).[[5]](#footnote-3)

**Section 3: Rule Consequentialism**

Rule Consequentialism assesses rules in terms of their consequences and assesses acts in terms of the rules with the impartially best consequences. Again, ‘impartially best’ can be understood in buck-passing fashion as ‘most favoured by impartial reasons’.

There are different versions of Rule Consequentialism. One version holds that what makes an act morally permissible or morally impermissible is whether the act is permitted or prohibited by the set of rules *whose acceptance* *by the agent* would produce the impartially best consequences. Such a view might be called Individualized Rule Consequentialism. A different version of Rule Consequentialism holds that what makes an act morally permissible is that the act is permitted by the code of rules *whose more or less universal acceptance* would produce the best consequences, impartially considered.[[6]](#footnote-4) This view might be called Collective Rule Consequentialism.

Individualized Rule Consequentialism rejects the idea that moral rules should be the same for everyone. In contrast, Collective Rule Consequentialism embraces this idea—as do Kantianism and Contractualism. To the detriment of Individualized Rule Consequentialism, the idea that moral rules should be the same for everyone strikes many people as axiomatic.

In case idea needs defense, however, consider the common assumption that a code of moral rules must be suitable for public acknowledgement and explicit use in interpersonal justification. And a moral code must be suitable for being taught to each new generation as a group. Once this code has been internalized by a generation, the members of this generation will typically have mutual moral expectations and shared moral reactions. Now, moral rules could not be suitable for public acknowledgement, group teaching, mutual expectations, shared reactions, and explicit use in interpersonal justification if those rules were, as Individualized Rule Consequentialism allows, tailored differently for each individual.

I have nothing more to say here about Individualized Rule Consequentialism. Hereafter, therefore, I will refer to Collective Rule Consequentialism more briefly as Rule Consequentialism.

I have formulated Rule Consequentialism in terms of rules whose more or less universal acceptance would be best. The acceptance of rules is conceived of as combining belief that the rules are justified and dispositions to act and react in accordance with them. These dispositions must be strong enough to guarantee corresponding behavior and attitude, except perhaps in rare extreme circumstances.

Compliance with rules is very often the most important consequence of rule acceptance. However, it isn’t the only important consequence. Suppose you and I accept different rules that happen to coincide in the acts they induce from us. Suppose the rule you accept makes you happy, whereas the rule I accept depresses me.[[7]](#footnote-5) Here acceptance of rules has consequences beyond the acts of compliance with the rules. Such consequences must be taken into account.

Also crucially important are the ‘internalization costs’ of rules. The internalization costs are the psychological and cognitive costs involved in getting new generations to learn a rule and to become highly motivated to comply with it. I assume that such costs are higher to the extent that there are more rules or more complicated ones to internalize and to the extent that the rules require regular large self-sacrifices.

So Rule Consequentialism assesses rules in terms of three kinds of consequences. There are the good and bad consequences caused by people’s complying with these rules. There are whatever benefits or costs result from rule acceptance but are not consequences of compliance with the rules. And there are the costs of getting new generations to learn the rules and to become strongly motivated to comply with them.

In my statement of Rule Consequentialism, I added the qualifier ‘more or less’ to ‘universal acceptance’. Formulating Rule Consequentialism in terms of *perfectly* universal acceptance presents a very serious problem. Perfectly universal acceptance of a good set of moral rules would eliminate various problems altogether. For example, it would eliminate the problem of how to deal with amoralists: in a world where absolutely everyone internalizes good moral rules, there are no amoralists. Perfectly universal acceptance of a good set of moral rules would also eliminate the problem of how to treat people misled by bad moral rules. And it would eliminate the problem of how to treat people who fail to have enough moral motivation even if they do have the right moral beliefs. But a plausible moral theory must provide plausible rules for dealing with (e.g. deterring or punishing) amoralists, the morally misguided, and those with insufficient moral motivation. To provide a rationale for including rules about how to deal with such problems, Rule Consequentialism should be formulated in terms of some level of acceptance short of perfect universal acceptance.

A difficulty, however, is that any particular level of acceptance—e.g., acceptance by 90% of everyone—seems unacceptably arbitrary. Why 90% instead of 85% or 75% or 60% or any other particular level? Parfit’s answer is to imagine that the code of rules whose universal acceptance would produce the impartially best consequences would include rules specifying what to do when only 1% of everyone else accepts this set of rules, what to do when 2% of everyone else accepts this set of rules, what to do when 3% of everyone else accepts this set of rules, and so on for every other level of acceptance of this set of rules (*OWM*, vol. 1, pp. 317–19).[[8]](#footnote-6)

Because we are taking into account internalization costs and these rise as a code becomes more complicated, there are fairly severe limits to how complicated an optimific code will be. These limits on complexity are infringed, I believe, by the proposal Parfit makes. Despite the arbitrariness of formulating Rule Consequentialism in terms of acceptance by 90% of each new generation, I think this formulation superior to one that embraces different rules for every different level of social acceptance.

But the question of how to deal with the real world’s merely partial acceptance of optimific rules can be set aside here. It is not a question central to Parfit’s main line of argument in *On What Matters*.

The most familiar way of assessing rules is in terms of resulting aggregate well-being, i.e., the sum of everyone’s personal good, calculated by giving any benefit or harm to one individual the same weight as the same size benefit or harm to anyone else. To assess acts in terms of rules that are selected solely for the resulting aggregate well-being is to be a Rule Utilitarian.

Selecting rules solely on the basis of the resulting aggregate well-being runs into the objection that, while aggregate well-being certainly matters, the pattern of its distribution also matters. The objection might be, for example, that some aggregate well-being is worth sacrificing for the sake of improving the plight of the worst-off. For illustration, compare two possible codes of moral rules. Suppose the consequences of acceptance of the first code would be very high aggregate well-being but also very low well-being for the worst-off. Suppose acceptance of the second code would yield somewhat less aggregate well-being but would leave the worst-off less badly off. If we think the consequences of the second code would be better, then we might reject Rule Utiliarianism in favour of Prioritarian Rule Consequentialism. Prioritarian Rule Consequentialism assesses acts in terms of the rules whose more or less universal acceptance would produce the greatest aggregate good calculated by giving somewhat greater importance to benefits and burdens to the worse off than to the same size benefits and burdens to the better off.

We have seen that Rule Utilitarianism focuses purely on the aggregate well-being resulting from acceptance of rules. Prioritarian Rule Consequentialism puts extra weight on benefits for the worst-off. Other kinds of Rule Consequentialism take other values to be relevant to the assessment of rules. Such possibilities will concern us later. But for the moment focus on Prioritarian Rule Consequentialism.

The rules that Prioritarian Rule Consequentialism endorses specify:

1. Prohibitions on certain kinds of typically harmful actions, such as killing or injuring non-aggressors or threatening to do so, stealing or destroying others’ property, breaking promises, and telling lies (these prohibitions apply even where doing one of these actions would produce somewhat greater impartial good, and even when violating the prohibition would not harm anyone)
2. Special obligations to give some degree of priority to those with whom one has special connections when one is deciding how to allocate one’s own time, energy, attention, and other resources
3. A general duty of beneficence to do good for others and prevent disasters

The prohibitions on certain kinds of typically harmful actions are needed in order to ‘protect persons, property, and promises’ as H. L. A. Hart put it.[[9]](#footnote-7) If persons weren’t protected, then there would be the nightmare that Hobbes described, where each had to invest virtually all his or her attention and energy in self-protection. If property were not protected, each would have to devote any remaining time and energy to protecting his or her own provisions and no one would have the security necessary for extensive investment in developing his or her possessions. If promises weren’t protected, then perhaps the most important instrument for assuring others and making diachronic cooperation rational would be lost.

Special (agent-relative) obligations to friends and family are apposite when one is deciding how to allocate one’s own time, attention, energy, money, or other resources. In such contexts, some degree of partiality towards one’s friends and family is required and more (though not infinitely more) is allowed. Such partiality is *allowed* because agent-relative partiality is integral to friendship and love, which have immense intrinsic and instrumental value.

Could people be raised in such a way that they care not only equally but also very deeply about every single individual in the world? In fantasy, yes. In the real world, no. Maybe in Heaven everyone intensely loves everyone else. But our moral rules should be designed for us real human beings, not for angels. The most relevant difference here between angels and us is that angels have vastly bigger hearts. We real human beings couldn’t intensify our concern for every other individual to the level of our typical concern for our child, or partner, or parent. The only realistic way to equalize our concern for everyone would be to suppress severely our concern for our friends and family.

Would a world in which everyone had a fairly lukewarm concern for everyone else be as happy as a world in which everyone cares intensely about some others but not as much about each of the rest? If people’s concern for friends and family were suppressed sufficiently far, the world would be lonelier and less happy. A world where people have intense concern for some others but not equal concern for all others would be a happier one than a world where people had equal concern for all but intense concern for none (*OWM*, vol. 1, pp. 385, 404–406). So Rule Consequentialism endorses rules that allow a degree of partiality.

Presumably, one’s affection for one’s friends and family will normally induce partiality toward them in one’s decision making. But in some people such affection is often weak, and even people who normally feel strong affection for friends and family sometimes get tired or exasperated. So, for family and friends reliably to get the attention they need, affection needs buttressing by special obligations of partiality when one is deciding how to allocate one’s own time, attention, energy, money, or other resources. Therefore, agent-relative partiality when one is deciding how to allocate one’s own time, attention, energy, money, or other resources would be not only allowed but also required. (Let me reiterate that the level of required partiality is limited, as is the level of allowed partiality.)

The benefits of there being a general (agent-neutral) requirement of beneficence are too obvious to list. The same is true of the benefits of there being a requirement to prevent disaster. These general requirements concern not just benefits to those with whom one has special connections, such as friends and family, but also benefits to strangers.

Rule Consequentialists contend that such requirements are sometimes but not always strong enough to override all the agent’s other projects and commitments. There would be high costs involved in trying to get more demanding rules of beneficence internalized. At some point, the costs of getting more demanding rules of beneficence internalized would outweigh the benefits resulting from the increase in beneficent acts done by those who had internalized these more demanding rules.

In short, it seems that the rules whose acceptance by more or less everyone would be favoured by impartial reasons would include familiar prohibitions, familiar special obligations, and a general duty to do good for others and prevent disasters. If such rules really are part of the optimific code, then Rule Consequentialism mandates these rules.

**Section 4: Parfit’s Explicit Argument for Convergence**

Parfit’s argument from Kantian Contractualism to Rule Consequentialism can be analysed as follows (*OWM*, pp. 378–79; cf. p. 400).

1. Kantian Contractualism holds that everyone ought to follow the rules whose universal acceptance everyone could rationally will, or choose.
2. Everyone could rationally choose whatever they would have sufficient reason to choose.
3. There are some rules whose universal acceptance is favored by the strongest impartial (agent-neutral) reasons.
4. No one’s self-interested or other partial reasons would outweigh their impartial (agent-neutral) reasons to choose these rules.
5. So, everyone would have sufficient reason to choose the rules whose universal acceptance is favoured by the strongest impartial (agent-neutral) reasons.
6. No one would have sufficient reason to choose any rules other than the ones whose universal acceptance is favoured by the strongest impartial (agent-neutral) reasons.
7. The combination of (B), (E) and (F) entails that the only rules that everyone would have sufficient reason to choose, and could therefore rationally choose, are the rules whose universal acceptance is favored by the strongest impartial (agent-neutral) reasons.
8. The combination of (A) and (G) entails that Kantian Contractualism implies that everyone ought to follow the rules whose universal acceptance is favored by the strongest impartial (agent-neutral) reasons.
9. Rule Consequentialism holds that the rules everyone ought to follow are the ones whose universal acceptance is favored by the strongest impartial (agent-neutral) reasons.
10. The combination of (H) and (I) entails that Kantian Contractualism and Rule Consequentialism converge on the same rules.

Parfit writes that the premise most likely to be questioned in this argument is (D), and he goes on to consider the strongest arguments against that premise. I am not going to discuss the debate about (D), because I want to focus on a different worry about his argument.

**Section 5: Impartiality as Agent-Neutrality**

Impartiality enters Parfit’s argument in premise (C) and plays at least some role in every step through (I). Given that Parfit is trying to present an argument that will attract very wide agreement, he understandably defines impartiality in a thin way that is neutral on what impartial reasons there are. The question is whether conceiving of impartiality merely as agent-neutrality is adequate.

As I explained above, for Parfit, impartial reasons are the reasons accessible to one when one considers situations from an agent-neutral point of view—a point of view that ignores any special connections one has with the people affected or involved. If one fact about the situation is that one of the potential beneficiaries or victims in the situation is oneself, this is a fact one must ignore completely when one thinks *impartially* about the situation. If another fact about the situation is the agent-relative consideration that some of the potential beneficiaries or victims are one’s friends, again this fact is irrelevant when one thinks *impartially* about the situation. Agent-relative considerations are not impartial ones.

Parfit’s definition of impariality does not entail that there are impartial reasons to care *equally* about everyone’s well-being (*OWM*, vol. 1, p. 41). To see that someone could take only agent-neutral reasons into consideration without giving everyone’s well-being equal weight, consider the following example. Suppose Jill knows that she and her family and friends are all of average wealth, average talent, and average well-being. Although Jill and her family and friends are not badly off, she prioritizes benefits to the worse-off. In other words, she is a Prioritarian in that she thinks that gains in the personal good of the worse-off matter more than the same size gains in the personal good of the better-off. So Jill wants the worst-off to benefit beyond whatever level would maximize aggregate well-being.

Compare Jill with Jack, who cares most about the well-being of the most beautiful celebrities. Jack is neither beautiful nor a celebrity, and none of his friends or family are celebrities or beautiful. Since beautiful celebrities have no special connection with him, his concern for the well-being of the most beautiful celebrities is agent-neutral, not agent-relative. Furthermore, he does not think that increasing the well-being of beautiful celebrities somehow maximizes aggregate well-being, much less the well-being of the worst-off.

Jack and Jill are here motivated entirely by agent-neutral concern. Thus, if impartiality is merely agent-neutrality, both Jack and Jill are completely impartial and equally so. And yet neither Jack nor Jill is giving overriding weight to aggregate well-being calculated by giving the same weight to benefits or harms to any one individual as is given to the same size benefits or harms to anyone else. Jack thinks the well-being of the most beautiful celebrities is at least somewhat more important than the well-being of others. Jill thinks that benefits to the worst-off are more important than somewhat larger benefits to the better-off. So, if impartiality is mere agent-neutrality, then, as we have just seen, impartiality does not entail equal concern for everyone.

The example about Jack also illustrates that some kinds of agent-neutral assessment are silly. This raises a difficulty for identifying impartiality with agent-neutrality. One of the things Kantianism, most forms of Contractualism, and familiar forms of Rule Consequentialism share is an aspiration to identify a way of *impartially* selecting or justifying a code of moral rules. For that aspiration to be intelligible, impartial selection and justification has to be attractive rather than silly. However, if impartiality is understood as agent-neutrality, then, since the content of some agent-neutral concerns are not attractive, the mere fact that selection of rules is agent-neutral will have no justificatory force.

One advantage of *impartially* evaluating possible moral rules is supposed to be the prospect that this will produce social agreement (even if this agreement is only idealized and hypothetical). But, if every agent-neutral evaluation counts as an impartial evaluation, then people’s evaluating things impartially does not guarantee convergence in their evaluations. Jack’s agent-neutral concern is for the well-being of the most beautiful celebrities; Jill’s agent-neutral concern is for the well-being of the worst-off. Though both Jack and Jill are focused on agent-neutral considerations, they are focused on very different ones and are very unlikely to agree about which moral rules are best. (What makes it unlikely that they will agree is the low probability that the most beautiful celebrities happen to be the worst-off.)

Although Jack’s agent-neutral priority for beautiful celebrities is silly, Jill’s agent-neutral Prioritarianism is a serious view, worth careful consideration. And there are other serious rivals. Often Parfit’s arguments are non-committal about which agent-neutral values are most important. Often the most natural assumption would be that he has utilitarian values in mind. However, at various points he acknowledges that other agent-neutral considerations could generate impartial reasons. For example, he writes,

On some value-based objective theories, there are some things that are worth doing, and some other aims that are worth achieving, in ways that do not depend, or depend only, on their contributions to anyone’s well-being. Scanlon’s examples are ‘friendship, other valuable personal relations, and the achievement of various forms of excellence, such as in art or science’. These we can call *perfectionist aims*.

On such views, it would be in itself good in the impartial-reasons-implying sense if we and others had these valuable personal relations, and achieved these other forms of excellence. The optimific principles might require us to try to achieve some perfectionist aims, and to help other people do the same. Since these are views about how it would be best for things to go, these claims could not give us reasons to reject the optimific principles. (*OWM*, vol. 1, p. 389)

And,

Nor should we assume that principles are optimific only if their acceptance would on the whole best promote everyone’s well-being. The goodness of outcomes may in part depend on other facts, such as facts about how benefits and burdens are distributed between different people, or facts that are not even about people’s well-being. If everyone could rationally choose that everyone accepts some autonomy-protecting principle, this might be one of the principles whose acceptance would make things go best, even if this principle’s acceptance would not on the whole best promote everyone’s well-being. (*OWM*, vol. 2, pp. 150–51)

In the first of these passages, Parfit leaves open the possibility that optimific rules are the ones whose acceptance by everyone would promote perfectionist considerations. In the second passage, he leaves open the possibility that optimific rules are the ones whose acceptance by everyone would protect autonomy. Much more often, Parfit has in mind the Prioritarian idea that the optimific rules are the ones whose acceptance by everyone would produce the greatest aggregate good calculated by attaching somewhat more value to benefits going to the worse off (*OWM*, vol. 2, pp. 196–201, 206, 248, 254, 751).

It is true that some people agent-neutrally assess possible moral rules in terms of whether the more or less universal acceptance of these rules would maximize perfectionist values. Other people agent-neutrally assess possible moral rules in terms of whether autonomy would be protected. Other people agent-neutrally assess possible moral rules in Prioritarian terms. Other people do Rule-Utilitarian assessment. And these are not the only popular agent-neutral approaches.

What is important here about the different forms of agent-neutral assessment is that, because they disagree over agent-neutral reasons, they might well disagree about which rules are the ones that everyone has sufficient agent-neutral reason to will that everyone accept. This point will apply if one set of agent-neutral, or impartial, reasons are guiding a Kantian Contractualist and another set guiding a Rule Consequentialist. This Kantian Contractualist and this Rule Consequentialist are both being guided by agent-neutral reasons, but not to the same destination. Likewise, if one Kantian Contractualist is guided by one set of agent-neutral reasons and another Kantian Contractualist is guided by a different set, they might end up with different rules. The same is true for two different Rule Consequentialists being guided by two different sets of agent-neutral reasons. Unless they are guided by the same agent-neutral reasons, why expect convergence?

One possible reply to the point I have just made is to accept it but then insist that Parfit has nevertheless established a very remarkable conclusion. He has shown that, as long as both Kantian Contractualists and Rule Consequentialists take into account only the *right* agent-neutral considerations, i.e., the *real* agent-neutral reasons, Kantian Contractualists and Rule Consequentialists will converge on the same moral rules. I agree that is an important conclusion.

Indeed, a wider conclusion beckons. If both Kantian Contractualists and Rule Consequentialists take into account only the *same* agent-neutral considerations, then, whether or not these are the right ones, Kantian Contractualists and Rule Consequentialists will converge on the same moral rules. This is also an important conclusion. But appreciating it should not obscure the point that for realistic hope that Kantian Contractualism and Rule Consequentialism to converge on the same set of moral rules, it is not enough that impartiality be identified with agent-neutrality. There must be convergence on which agent-neutral considerations are paramount.

**Section 6: The Reflective-Equilibrium Argument for Convergence**

I think that a great deal of what makes Parfit’s extended argument for convergence compelling is that we are reassured along the way that his Kantian Contractualism and Rule Consequentialism end up with intuitively plausible implications about moral prohibitions, permissions, and requirements. If these theories endorse intuitively plausible prohibitions, permissions, and requirements, then the theories gain enormous support from the plausibility of their implications. And if they each endorse *the same* intuitively plausible prohibitions, permissions, and requirements, then we have convergence.

Least you think I am projecting onto Parfit a commitment to searching for reflective equilibrium that is not really there in his work, note that Parfit quotes approvingly Scanlon’s remark about the method of searching for reflective equilibrium: ‘this method, properly understood, is … the best way of making up one’s mind about moral matters … Indeed, it is the only defensible method: apparent alternatives to it are illusory.’ (*OWM*, vol. 1, p. 367). And later Parfit writes,

For some moral theory to succeed, it must have plausible implications. The Triple Theory [Kantian Contractualism, Scanlonian Contractualism, and Rule Consequentialism] has many such implications. But after we have worked out what this theory implies, and we have carefully considered all of the relevant facts and arguments, this theory might conflict with our intuitive beliefs about the wrongness of certain acts. If there are many such conflicts, or these intuitive beliefs are very strong, we could then justifiably reject this theory. If instead these conflicts are significantly less deep, or less common, we could justifiably follow this theory in revising some of our intuitive moral beliefs. (*OWM*, vol. 1, p. 415)

Why does searching for reflective equilibrium in ethics militate in favour of versions of theories that converge on the same set of rules? There are some confident moral convictions that nearly all of us share. Searching for reflective equilibrium between our moral convictions and our moral theories will push us to favour whatever versions of our theories have implications that cohere with these convictions. Hence, what seem to us the best versions of Kantianism Contractualism and Rule Consequentialism will have very similar implications.[[10]](#footnote-8)

Of course different theories disagree in many ways—not just in their foundational principles but also in their structure, practicality, ambition, etc. But an ambition shared by Kantianism Contractualism and Rule Consequentialism is to select or justify moral rules impartially. As long as impartiality is more than mere agent-neutrality, this ambition is intriguing, especially when paired with a commitment to finding a theory in reflective equilibrium with our confident moral convictions. Success is not guaranteed. Neither is failure.[[11]](#footnote-9)

1. Derek Parfit, *On What Matters*, 2 volumes, Oxford University Press, 2011. From now on in this paper, I will cite this work as *OWM*. [↑](#footnote-ref--1)
2. Williams, ‘Internal and External Reasons’, in Ross Harrison (ed.) *Rational Action*, Cambridge University Press, 1979; reprinted in Williams, *Moral Luck*, Cambridge University Press, 1981 [↑](#footnote-ref-0)
3. Parfit, *Reasons and Persons*, Clarendon Press, 1984, p. 143; for the distinction though not this terminology, see Thomas Nagel, *The Possibility of Altruism*, Princeton University Press, 1970, pp. 47–48, 90–96. [↑](#footnote-ref-1)
4. See *OWM*, vol. 1, pp. 38–42. Here Parfit is accepting a ‘buck-passing’ view of goodness and bestness. For earlier statements of the buck-passing view, see Parfit’s ‘Reason and Motivation’, *Proceedings of the Aristotelian Society Supplementary Volume*, 71: 99–130, p. 124; and T. M. Scanlon, *What We Owe To Each Other*, Harvard University Press, 1998, pp. 96–8. My own views about the buck-passing view appear in P. Stratton-Lake and B. Hooker, ‘Scanlon versus Moore on Goodness’, in Terry Horgan and Mark Timmons, (eds), *Metaethics After Moore*, (Oxford: Clarendon Press, 2006), pp. 149–68. [↑](#footnote-ref-2)
5. Parfit’s argument is focuses on the battle between self-interested and altruistic *reasons*. Richard Brandt has a very similar argument, though Brandt’s is driven by self-interested and altruistic *preferences*. See Brandt’s *Facts, Values, and Morality* (New York: Cambridge University Press, 1996, pp. 58, 240). [↑](#footnote-ref-3)
6. I defend Rule Consequentialism in *Ideal Code, Real World: A Rule-consequentialist Theory of Morality*, Clarendon Press, 2000; ‘Reply to Arneson and McIntyre’, *Philosophical Issues* 15 (2005), pp. 264–81; ‘Rule-consequentialism and Internal Consistency: A Reply to Card’, *Utilitas* 19 (2007), pp. 514­–9; ‘Act-consequentialism versus Rule-consequentialism’, *Politeia* 24 (2008), pp. 75–85; and ‘Publicity in Morality: Reply to Katarzyna de Lazari-Radek and Peter Singer’, *Ratio* 23 (2010), pp. 111–17. [↑](#footnote-ref-4)
7. Shelly Kagan, *Normative Ethics*, Westview Press, 1998, pp. 227–34; ‘Evaluative Focal Points’, in Brad Hooker, Elinor Mason, and Dale E. Miller, *Morality, Rules, and Consequences*, Edinburgh University Press, 2000, p. 139. [↑](#footnote-ref-5)
8. Compare Michael Ridge, ‘Introducing Variable-Rate Rule Utilitarianism’, *Philosophical Quarterly* 56 (2006), pp. 242–53. [↑](#footnote-ref-6)
9. H. L. A. Hart, *The Concept of Law*, Clarendon Press, 1961, p. 197. [↑](#footnote-ref-7)
10. I think this is as true for some other moral theories as it is for the ones that Parfit discusses. Ross-style pluralism and Douglas Portmore’s Agent-Relative Act Consequentialism are cases in point (W. D. Ross, *The Right and the Good*, Clarendon Press, 1930, ch. 2; Douglas Portmore, *Commonsense Consequentialism*, Oxford University Press, 2011). [↑](#footnote-ref-8)
11. I am grateful for Derek Parfit’s wonderfully helpful written comments on the penultimate draft of this paper. Earlier drafts were presented at the British Postgraduate Philosophy Association’s 2012 Conference, at the University of Reading’s Philosophy Department, at the Arizona Normative Ethics Workshop, and at Adam Cureton’s Workshop on Social Norms at the University of Tennessee. I am grateful for comments received on those occasions from Sarah Broadie, James Kirkpatrick, Philip Stratton-Lake, Alex Gregory, David Oderberg, David Owens, James Stazicker, Nat Hansen, Emma Borg, Charlotte Newey, Richard Rowland, Geraldine Ng, Sarah Stroud, Paul Hurley, Shelly Kagan, Tom Hurka, Holly Goldman, Jussi Suikkanen, David Schmidtz, Mark van Roojen, Brad Cokelet, Earle Conee, Mark Timmons, Adam Cureton, Tom Hill, Markus Kohl, Henry Richardson, and Jon Garthoff. [↑](#footnote-ref-9)