Philosophical Studies 168 (2014): 629-644
DOI: 10.1007/s11098-013-0154-2

A Better World

Ryan Preston-Roedder
UNC Chapel Hill

ABSTRACT: A number of moral philosophers have endorsed instances of the following curious argument: it would be better if a certain moral theory were true; therefore, we have reason to believe that the theory is true. In other words, the mere truth of the theory—quite apart from the results of our believing it or acting in accord with it—would make for a better world than the truth of its rivals, and this fact provides evidence of the theory's truth. This form of argument may seem to be an obvious non-starter. After all, the fact that the truth of some empirical claim, say, the claim that there is an afterlife, would be desirable does not, by itself, give us any reason to believe it. But I argue that, when it is properly understood, this form of argument—which I call the better world argument—is valid in moral philosophy. I develop and defend a version of the argument that rests on the view that the correct moral theory cannot exhibit a certain form of self-defeat—a form that, as far as I know, has not been discussed in the literature. I also identify two promising applications of this form of argument. The first is a defense of permissions to promote one's own private aims, rather than promote the greater good, and the second, an argument against the possibility of moral dilemmas.

I

In "Personal Rights and Public Space", Thomas Nagel (1995, pp. 86-93) presents the following defense of individual moral rights: If we have rights, say, to freedom of religious expression and protection against police torture, then we have a certain desirable moral status. But if we do not have rights, then we have some other, less valuable status instead. So we would all be better off if we had rights. And this fact somehow gives us reason to believe that we do, in fact, have them. To be clear, the argument does not rest on the claim that *believing* that we have rights, or acting in accord with this belief, would be desirable or have desirable results. Rather, Nagel's claim is that if we had rights, this *moral* fact would be good by itself, quite apart from people's actions and attitudes, and that this provides some evidence that we actually have rights.

This is, to put it mildly, a puzzling form of argument: if the truth of a moral theory would, by itself, "make for a better world" than the truth of its rivals, this gives us some reason to believe that the theory is, in fact, true (Nagel 1995, p. 92). But Nagel is not the only one who has found it persuasive: Frances Kamm (1996, Ch. 10) and Warren Quinn (1993, pp. 149-74) also appeal to versions of this argument to defend rights.[1] Michael Slote (1985, Ch. 2) seems, on one reading, to use this form of argument to defend the view that we are often permitted to promote our own private aims, rather than promote the greater good. Margaret Urban Walker (1991) implicitly appeals to this form of argument to defend the view that there is moral luck, that there are cases in which our moral assessment of someone rightly depends on factors that are out of her control. And David Enoch (2009) claims that although arguments of this sort are absurd when they concern non-moral matters of

[1] Kamm (2007, Chs. 1, 5, and 8) provides a detailed characterization of this moral status, which she calls "inviolability", together with an account of its desirability.

fact – consider moving from the claim that it would be better if there were an afterlife to the conclusion that there is, in fact, an afterlife – they are not so obviously out of place in moral theory. In short, a number of perceptive moral philosophers have endorsed arguments of this sort; though, in most cases, either the underlying form of argument is too obscure to be adequately assessed or it is made precise in a way that casts doubt on its validity.

Some have argued that certain instances of this form of argument are valid, provided that we accept a metaethical view according to which moral facts are somehow determined by the right sort of procedure.[2] For example, contractualists claim that we act rightly when and because we act in accord with principles that any reasonable person would agree to adopt. So a contractualist might argue that any reasonable person would treat the fact that it would be better if some moral principle were true as grounds for adopting that principle. Or to take another example, someone who claims that moral facts depend on God's will might argue that God would treat the fact that it would be better if some principle were true as a reason to make it true. In each case, the fact that the truth of a candidate moral principle would make for a better world provides some evidence of its truth.

But I want to consider a very different way of making sense of this form of argument – one that does not depend on these metaethical assumptions, which many reject. I will defend a formulation of the argument that rests, not on the view that moral truths are determined by a procedure, but rather on the view that morality cannot exhibit a certain type of self-defeat, a type of self-defeat that, as far as I know, has not been dis-

[2] See Enoch (2009, pg. 223). Others have made this point in conversation.

cussed in the literature.[3] To be clear, I aim to show only that a certain *form* of argument – namely, one that treats the fact that it would be better if a moral theory were true as a defeasible reason to believe the theory – makes sense when it is properly formulated, despite the fact that it seems initially to be a non-starter. I will not try to show that any particular application of the argument, say, Nagel's defense of rights or Walker's defense of moral luck, succeeds, since that would require a more detailed discussion of rights or of moral luck than I can offer here. Nevertheless, I do want to take a crucial step toward determining what a successful application of this argument would look like. To that end, I will discuss what I take to be the main objection that any application of this argument has to face, and I will identify features that an application has to have in order to address this objection.

Provided that this curious form of argument, which I call the *better world argument*, makes sense, we stand to gain a good deal by understanding it more clearly. Most obviously, we may broaden our understanding of how to justify moral principles, and so, become better able to determine which principles are justified. But beyond this, most of us care a great deal about making our world a better place, at least, in certain respects and in certain kinds of circumstances. Recognizing that this form of argument makes sense, and why it makes sense, promises to deepen our understanding of the ways in which morality is relevant to this fundamentally important concern. If the better world argument is valid, morality does not just demand that we adopt attitudes and perform actions that have good

[3] I noted above that several authors appeal to this form of argument in order to defend some moral principle, but I do not claim that any of these authors understand this form of argument in exactly the way I do here. Most instances of the better world argument in the literature merely gesture at the underlying form of argument – my aim is both to make this underlying form of argument more precise and to show that it is valid in certain limited contexts.

effects. It also makes the world a better place in a more direct way – at least, in certain limited respects – through the structure and content of its principles.

II

I can best describe the formulation of the better world argument that I will defend if I begin by briefly discussing another, more familiar argument that treats the fact that a moral theory is self-defeating in a certain respect as an objection to the theory, namely, Derek Parfit's (1984, pp. 95-110; 2011, pp. 301-8) argument that Common-Sense Morality is directly collectively self-defeating, and so, should be revised. A moral theory is directly collectively self-defeating – for ease of expression, I will generally drop the modifier "directly" – just in case it has the following feature: if everyone successfully followed the theory's principles, each person's substantive moral aims would certainly be *worse* achieved than they would be if everyone violated these principles in some way.

I will discuss some of our moral aims in detail below, but it will help to make some general remarks here. The moral aims that a moral theory assigns to us are the aims that, according to the theory, it is important for us to achieve. In addition to assigning to each person the merely formal aim of acting rightly, each theory also assigns substantive aims to her. These may take very different forms from one theory to the next, or even within a single theory. One such aim may be that the best possible outcomes occur, or that outcomes that are good for particular people – say, for one's own children – occur. Another may be that one respect people's autonomy, or that one refrain from performing certain base acts. Achieving these substantive aims is part of the point of observing moral principles; that is, the fact that acting rightly serves such aims helps explain why acting rightly is worth caring about, why it is not just a matter of following pointless rules. Parfit argues that Common-Sense Morality assigns such aims to us, but requires us to act in ways

that, taken together, would cause these aims to be worse achieved than they would be if everyone acted in some other way instead. Parfit argues, in other words, that Common-Sense Morality fails in its own terms, and that this constitutes a powerful objection to the theory.

Parfit's argument rests on his discussion of cases like the following (1984, pp. 84 and 98; 2011, pp. 303-4). Imagine a poor fishing village in which many people earn their living by fishing separately on some large lake. Due to overfishing, the lake's stock of fish has declined, and the fishermen have to struggle just to catch enough fish to keep their children well fed. It is true of each fisherman that, no matter what the others do, if he catches fish without substantial restraint, his own children will be slightly better fed – and as a result, slightly better off – than they will be if he restricts his catch instead. But the lake's stock of fish is so poor that if *everyone* catches fish without restraint, the stock will be depleted, and everyone's children will be severely malnourished. By contrast, if *no one* catches fish without restraint – or in other words, if everyone restricts his catch – the stock will be better maintained, and none of the children in the village will become severely malnourished.

Common-Sense Morality assigns to each person the aim that her own children flourish in certain respects, including being well nourished and avoiding the gnawing pain of severe hunger. And it states that each person is required to promote this aim by doing whatever will promote her children's flourishing in the relevant respects, provided that this does not involve violating anyone's rights or making certain kinds of personal sacrifices. Parfit claims that, because people in this village live in such dire conditions, and because the fishermen have to struggle to catch enough fish to keep their children well fed, this requirement directs each fisherman to catch fish without substantial restraint, and thereby ensure that his own children are slightly better off. To be clear, it does not matter, for my purposes, whether or not Common-Sense Morality actually has this impli-

cation. The important point is that *if* it has this implication, this is a serious problem for the theory. Suppose, for the sake of argument, that Common-Sense Morality does imply that each fisherman is required to fish without substantial restraint. In that case, if everyone in the village observed Common-Sense Morality's requirements, then everyone's children would be much worse off – and so, everyone's substantive moral aims would be much worse achieved – than they would be if everyone violated the requirements by restricting his catch. So Common-Sense Morality is collectively self-defeating: everyone's observing the theory's requirements would, together, undermine the substantive aims that the theory assigns to each person.

If Common-Sense Morality is collectively self-defeating, as Parfit argues, then Parfit seems right to claim that this is a serious objection to the theory. More precisely, this is a serious objection, provided that morality is concerned with the class of cases in which this form of self-defeat arises, and there is some way to revise the theory that would eliminate this sort of self-defeat. Both of these conditions obtain. First, morality is largely a system of principles for the general regulation of behavior – it is centrally concerned with what we do collectively. So the fact that we would together undermine our moral aims if we all observed some moral theory's requirements is a morally significant fact about that theory. Second, Parfit describes a revision of Common-Sense Morality that would eliminate this sort of self-defeat: he suggests that we revise it to state that each fisherman is required to restrict his catch, provided that enough others do so as well. So we have some reason to believe that Common-Sense Morality, as Parfit understands it, should be revised.

The argument from the claim that it would be better if a certain moral theory were true to the conclusion that it is, in fact, true can also be formulated in a way that rests on the view that morality cannot be self-defeating in a certain respect. So understood, the argument is analogous to Parfit's argument

against Common-Sense Morality. But it concerns a form of self-defeat that differs from the one Parfit describes – a form of self-defeat that, as far as I know, has not yet appeared in the literature at all. The argument I will discuss also relies on the claim that bringing about a better world is among our substantive moral aims. We cannot plausibly reject this claim altogether, though there may be more than one plausible way to interpret it: some may judge that morality assigns us the aim that impersonally good outcomes occur, others that it assigns the aim that people be well off, and so on. The point is that morality assigns us the aim of bringing about a better world in *some* respect.[4] The better world argument, as I will now formulate it, concerns moral theories the truth of which would be better, in that very respect, than the truth of other, rival theories.

Suppose it would be better in the relevant respect if the principles that make up moral theory M were true. To be clear, the truth of these principles would be good in itself – quite apart from the effects of people's believing the principles or acting in accord with them, and no matter what the world happens to be like. In that case, the truth of M would, by itself, serve the moral aim of bringing about a better world, an aim that all plausible moral theories assign to us. By contrast, we would all be worse off if any of M's rivals, which comprise different principles, were true instead. So the truth of any rival theory would make for a worse world – and, other things equal, render our

[4] In any case, it seems that any formulation of the better world argument that has any hope of succeeding will assume that morality assigns this aim to us. It would be odd to claim that bringing about a better world is so important that morality tends to include principles whose very content serves this aim, and nevertheless to deny that bringing about a better world is important enough for morality to assign this aim to *us*. So the assumption that bringing about a better world, in some sense, is among our moral aims is, at the very least, an assumption that any plausible formulation of the better world argument will make.

substantive moral aims worse achieved – than the truth of M.[5] In other words, any rival to M that makes plausible claims about the content of our moral aims is also self-defeating in the following sense: it both assigns certain substantive aims to us and includes principles the mere truth of which would render those very aims worse achieved than the truth of other, rival principles. But as I will argue below, the correct moral theory cannot undermine itself in this sense, unless eliminating the self-defeat carries other, comparable costs. So we have some reason to reject the rival theories and accept M, which avoids this form of self-defeat. This means that the fact that it would be better if M, rather than one of its rivals, were true *does* give us some reason to believe M is true.[6]

I can better explain how this form of argument works if I sketch two of its more promising applications. The first, which is adapted from Michael Slote (1985, Ch. 2), is an argument for the view that we are often permitted to promote our own private aims, rather than promote the greater good.[7] Some moral theories, including Common-Sense Morality, state that we are often permitted, say, to marry people we love, pursue careers we are passionate about, or merely take up hobbies we enjoy, even if we could do more good, impersonally considered, by doing something else instead. By contrast, other theories, includ-

---

[5] Other things are *not* equal if the truth of a rival theory would undermine the aim of bringing about a better world, but nevertheless promote other, comparable moral aims. In that case, the truth of the rival theory might undermine our moral aims in one important respect, but promote our moral aims, all things considered.

[6] I am setting aside the possibility that *no* moral theory is true. The argument is addressed to those who grant that there is a correct moral theory and who wish to identify it.

[7] Because Slote does not clearly describe the underlying structure of his argument, it is not clear to what extent this adaptation departs from what he has in mind.

ing Act Consequentialism, state that we are always, or almost always, required to do whatever will produce the best available results, impersonally considered, even when we wish to promote some other, non-optimal aim.[8] If a theory of the former sort were true, we would all have a form of autonomy that Slote (1985, Ch. 2) calls "moral autonomy".[9] That is, each person would have a kind of freedom to exercise her capacity to choose what sort of life she lives, to shape her life in accord with her own conception of what is worthwhile. To be clear, having moral autonomy is not just a matter of being able to pursue some particular life plan without impediment. Rather, someone has this form of autonomy insofar as she is able to choose, among a range of alternatives, whatever life plan she prefers, without bearing the grave costs associated with acting wrongly. Two such costs seem especially important. First, acting rightly enables a person to live in light of her recognition that, although her own interests loom large from her point of view, there is an important sense in which no one is any more or less significant than anyone else. A decent person cares about living in this way for its own sake, but acting wrongly undermines this aim. Second, people tend to care both about securing others' approval and about meriting such approval, each for its own sake. But when someone acts wrongly, she undermines these aims, whether by bringing it about that people blame her or, at the very least, by making it appropriate for them to do so.

---

[8] Act Consequentialism is the best known theory of this sort, but it is not the only such theory. Kagan (1989, pg. 8) and Unger (1996, pp. 149-50) defend versions of the view that we are always required, not to choose the act – whatever it is – that will produce the best available results, but rather, to choose the act – among those acts that are not ruled out by constraints or special obligations – that will produce the best results.

[9] My characterization of moral autonomy draws on Shiffrin's (1991) illuminating discussion.

By contrast, if a theory of the second sort – according to which we are always, or almost always, required to do what will produce the best results – were true, we would all be deprived of this form of autonomy. Of course, we might retain merely physical opportunities to exercise our capacities for choice and self-creation, but we would not be able to exercise these capacities in any substantial way without incurring the costs associated with wrongdoing. So we would not possess the kind of autonomy that I described, at least, not to any substantial degree.

Suppose – plausibly, in my view – that having moral autonomy is highly desirable for its own sake.[10] In that case, the truth of a theory of the former sort, according to which we all have substantial moral autonomy, would make for a better world, other things equal, than the truth of a rival theory, according to which no one has it.[11] So, other things equal, any rival theory

---

[10] Kagan (1989, pp. 336-8) argues that the intuition that moral autonomy is valuable rests on a mistaken assumption that this form of autonomy is analogous to freedom from physical or psychological coercion. But Shiffrin points out that moral autonomy *is* similar to these other forms of freedom in the following important respect: when someone is deprived of moral autonomy, she is deprived of "her opportunity to exercise her capacity for choice without thereby jeopardizing something else she has a right or reasonable expectation to have or enjoy" (1991, pg. 252).

[11] It may be that the type of theory the truth of which would bring about the best outcome, on balance, is one that permits us to devote substantial attention to our deepest commitments, but nevertheless requires us to do far more to promote the greater good than Common-Sense Morality requires. To offer a crude characterization, such a theory might, say, require everyone who has income to spare after she has met her family's basic needs to donate one third of this income to help the world's poorest people, but permit her to use the remainder to promote her own aims. If such a theory were true, we would all be better off in one important respect, because we would have a substantial domain of moral autonomy. But the truth of such a theory might make for a better world in another respect as well: it

that grants, as it should, that bringing about a better world is among our substantive moral aims is self-defeating in the sense I described. For example, Act Consequentialism assigns a single ultimate aim to us, namely, that outcomes be as good as possible, and it states that we are always required to promote this aim by doing whatever will produce the best available outcome. But as I just argued, our merely being subject to such a requirement would, by itself, prevent at least one highly desirable outcome from being realized: it would severely restrict our range of morally permissible actions and thereby prevent us from having moral autonomy. So, other things equal, the truth of Act Consequentialism would actually make the world *worse* – and thereby render the ultimate substantive aim that that theory assigns to us worse achieved – than the truth of certain other theories according to which we are often permitted to promote our own, non-optimal aims. So, other things equal, Act Consequentialism is self-defeating – it fails in its own terms – and we have some reason to reject it. More generally, any set of moral principles that leaves us little room to pursue our own conceptions of the good life comes at a price and bears a burden of justification.[12]

---

may be that conscientious people would recognize the stringency of their requirement to promote the greater good and would supply the resources needed to eliminate the world's worst preventable evils.

[12] An anonymous reviewer pointed out that this argument may seem to have implausibly strong implications: The argument rests on the claim that we would be better off if we were often permitted to pursue our own private aims, because we would have a certain valuable form of autonomy. And it moves from this claim to the conclusion that we have reason to believe that we are, in fact, so permitted. But, provided that more autonomy is better than less, the argument may seem to imply that we have an even stronger reason to believe that all things are permitted. But the argument does not have this implication. First, how much moral autonomy people have is not the only factor that determines how well off they are. If all things were permitted, people would presumably tend to believe that all things were

The second application of the better world argument is loosely adapted from an argument that Geoffrey Sayre-McCord develops in "A Moral Argument Against Moral Dilemmas" (unpublished manuscript). Unlike the defense of permissions to promote non-optimal aims that I just described, this second application does not essentially concern requirements to produce good results. Rather, it concerns, in part, second-order principles that determine whether and how certain kinds of conflicts among moral principles may be permissibly resolved. More precisely, this second application concerns cases in which someone has a strong pro tanto obligation to do one thing and a comparable pro tanto obligation to do some other, incompatible thing, but neither decisively outweighs the other. One famous example is that of Sartre's student, who must choose between going to England to join the Free French Forces – thereby opposing the Nazi occupation of France – and remaining in France to care for his ailing mother (1975, pp. 354-5). Some moral theories state that in all such cases, the person ought to act in accord with one of her pro tanto obligations, but is permitted to act in accord with either. Others state that in at least some of these cases, the person faces a moral dilemma, in which she ought to act in accord with each pro tanto obligation, even though she cannot act in accord with both. In other words, according to the second type of theory, there are cases in which a person cannot avoid committing a serious moral wrong, no matter what she does.

It may be that we would all be better off if moral dilemmas could *not* occur, because the possibility of such dilemmas would

---

permitted, and as a result, social life would be intolerable. This evil would far outweigh the good of having our moral autonomy increased to its upper limit. Second, there may be cases in which the aim that good outcomes occur gets trumped or outweighed by other substantive moral aims, for example, the aim that one respect other people's rational agency. In such cases, our moral autonomy may be limited by principles that serve these other aims.

make our circumstances less fair or more tragic.[13] In that case, other things equal, the truth of the first type of moral theory, according to which anyone who faces the kind of practical conflict I described is permitted to adopt either option, would make for a better world than any rival theory, according to which at least some of these people ought to adopt each option, even though they cannot adopt both. So other things equal, any rival theory that grants, as it should, that bringing about a better world is one of our substantive moral aims is self-defeating in the sense I described. It assigns certain aims to us but includes principles the truth of which would render these aims worse achieved than the truth of alternative principles, namely, principles that rule out the possibility of dilemmas. But morality cannot be self-defeating in this sense, unless eliminating the self-defeat carries other, comparable costs. So it follows that we have some reason to reject the rival theories, and instead accept a theory according to which dilemmas cannot occur.[14]

---

[13] Sayre-McCord claims, in the manuscript cited above, that any theory according to which we may face moral dilemmas is itself unfair in the sense that it makes demands that are impossible to carry out, and Hare argues that dilemmas are tragic because someone who faces a dilemma is "like a rat in an insoluble maze" (1981, pg. 32).

[14] To be clear, the argument against moral dilemmas that Sayre-McCord develops is importantly different from the argument I just described, though the two arguments are superficially similar. Sayre-McCord does not argue that theories that allow for the possibility of moral dilemmas are self-defeating in any way, or that the truth of such a theory would make for a better world. Rather, he argues that the correct moral theory, whatever that turns out to be, merits our allegiance. Other things equal, a theory that rules out the possibility of moral dilemmas is fairer, and so, more worthy of our allegiance, than a theory that allows for such dilemmas. So we have some reason to believe that the former sort of theory is true. A second difference is that Sayre-McCord does not claim that a world in which dilemmas

The distinctive claim on which the better world argument rests – that the fact that a moral theory is self-defeating in the sense I described constitutes an objection to the theory – seems right on reflection. In fact, the charge that a theory is self-defeating in this sense is at least as serious as the charge that it is self-defeating in Parfit's sense. In each case, the theory fails in its own terms: its principles somehow undermine the aims that, according to that very theory, it is important for us to achieve. And in each case, to insist that we observe the theory's principles, even though they undermine our moral aims, is to engage in a form of rule-worship.

The difference between these two forms of self-defeat concerns the route by which the theory's principles undermine its aims. And, if anything, this difference makes the form of self-defeat I describe *more* objectionable than the one Parfit describes. Parfit labels the form of self-defeat that he discusses *direct* collective self-defeat, but that modifier is misleading in the present context. When a theory is self-defeating in Parfit's sense, the route by which its principles undermine our moral aims is not, strictly speaking, direct; rather, it is mediated by our actions. The principles require us to act in ways that, together, would cause our moral aims to be worse achieved than they would be if we violated these principles and acted in accord with other, rival principles instead. By contrast, when a theory is self-defeating in the sense I described, the route is truly direct, or in any case, more direct than the route Parfit describes. The truth of the theory's principles would, by itself, render our moral aims worse achieved than the truth of other, rival principles, quite apart from people's attitudes and actions.

But the mere fact that the kinds of principles with which I am concerned undermine our moral aims in this more direct way – in a way that is not mediated by our attitudes or actions – does

---

can occur is more tragic than a world in which they cannot. Rather, his argument focuses entirely on fairness.

not make the frustration of those aims any less objectionable, nor does it somehow shield the principles from moral criticism. To the contrary, moral principles the mere truth of which would undermine our moral aims – no matter what we believed, how we acted, or what the world happened to be like – seem to me *more* objectionable than principles the general observance of which would, given how the world happens to be, cause these aims to be undermined. After all, the conflict between a theory's principles and the aims the theory assigns us seems both more pronounced and more closely tied to the content of the principles when the truth of the principles would undermine our aims in the more direct – or in other words, unmediated – way.

The fact that moral theories that are self-defeating in the sense I described fail in their own terms also marks an important distinction between this form of self-defeat and another, superficially similar characteristic that some theories possess: our *believing* that the theory's principles are true would cause the aims that the theory assigns us to be worse achieved than they would be if we believed rival principles instead.[15] For example, as I said above, Act Consequentialism assigns us the aim that outcomes be as good as possible, and it states that we are always required to do what will produce the best possible results, even when this involves, say, neglecting people or projects we care about or harming some people in order to help others. But people who believe that they are so required may become, in some sense, estranged from the commitments that make their own lives worthwhile, and they may do pointless harm because they wrongly believe that the harm is necessary to produce some good result.

[15] Parfit (1984, Ch. 1) explains how this characteristic might arise, assesses its significance, and argues, persuasively, that the fact that a theory has this characteristic does not, by itself, show that the theory fails in its own terms.

So it may be that our believing that Act Consequentialism's single ultimate principle is true would cause the single ultimate aim that that theory assigns us to be worse achieved than it would be if we believed other, rival principles, namely, principles that permit us to devote considerable energy and attention to our own commitments and generally prohibit us from doing certain kinds of harm. But this does not, by itself, show that Act Consequentialism fails in its own terms. Consequentialists have long claimed that the requirement to do what will produce the best results is a standard for the rightness and wrongness of actions, not a basis for decision-making.[16] So, if our believing that certain non-consequentialist principles are true is necessary to produce the best available outcome, then this requirement directs us to get ourselves and others to believe these other principles, if possible. In other words, once we take into account the act consequentialist requirement's implications for our moral beliefs, we can see that it directs us to do what, among our available options, would best achieve the aim that Act Consequentialism assigns us. Of course, the fact that our believing that this principle is true would undermine our moral aims may be objectionable on *other* grounds, for example, because it means that the principle is not suited for public acknowledgement, and the correct moral principles must be so suited (Baier 1958, pg. 195; Rawls 1999, pg. 115). But my point is simply that the considerations that give us reason to reject theories that are self-defeating in the sense that I described do not also give us reason to reject theories that have this other, apparently similar characteristic.

III

[16] Sidgwick claims that "the doctrine that Universal Happiness is the ultimate standard must not be understood to imply that Universal Benevolence is the only right or always best motive of action … [I]t is not necessary that the end which gives the criterion of rightness should always be the end at which we consciously aim" (1907, pg. 413).

I have argued that, although arguments from the claim that it would be better if some moral theory were true to the conclusion that the theory *is* true may seem puzzling at first, they can be formulated in a way that makes sense, whether or not moral facts are determined by a certain kind of procedure. But even so, instances of this form of argument face further challenges. Some instances have to address objections to the claim that the truth of the theory to be defended would be better, all things considered, than the truth of its rivals. For example, the defense of moral rights with which I began states that the truth of a theory according to which we have rights would make us better off by giving us a desirable moral status, namely, the status of people who cannot be permissibly harmed in order to benefit others. But one might object that the truth of a rival theory, according to which we do not have rights, would *also* give us a desirable moral status, namely, that of people who are so morally important that others may be permissibly harmed in order to benefit them (Kagan 1991, pp. 919-20). So, to show that this defense of rights succeeds, one would have to show that the former status is more desirable than the latter.

And some instances of this form of argument have to address objections to the claim that the truth of the theory to be defended would better serve our moral aims, on balance, than the truth of its rivals. In other words, they have to rule out the possibility that, although the truth of the theory to be defended would make for a better world, it would also render other, comparable moral aims worse achieved than the truth of some rival theory. For example, the argument against moral dilemmas that I described above states that the truth of a theory according to which dilemmas cannot occur would make our circumstances fairer or less tragic, and so, make for a better world, than the truth of rival theories, according to which dilemmas are possible. But even if we grant, for the sake of argument, that this is right, it may be that the truth of some rival theory would better serve some other moral aim. In that case, to show that the argument succeeds, we would have to

show that realizing a better world is morally more important than achieving this other aim.

The appropriate responses to such objections may vary considerably from one instance of the argument to the next, and in many cases, presenting these responses would require detailed discussions of issues that lie beyond the scope of this essay. So I will set these sorts of objections aside and turn to another, more fundamental objection to the better world argument, one that any successful instance of that argument has to address.

The main general objection to this form of argument is that it proves too much.[17] Typically, arguments that treat the fact that it would be better if some claim were true as a reason to believe that the claim is true are obviously invalid. For example, as I noted above and many have pointed out, such arguments clearly fail when they concern non-moral matters of fact: the fact that it would be better, say, if there were an afterlife or if global poverty would soon be eliminated does not give us any reason at all to believe these claims (Enoch 2009, pg. 222; McNaughton and Rawling 1998, pp. 51-2; Nagel 1995, pg. 92). But such arguments seem equally absurd in some cases in which they concern *moral* judgments. Consider the following argument: We all experience pain, and many of us experience more pain than pleasure over the course of our lives. So we would all be better off if pain were not bad for us – the truth of a moral theory according to which pain is not bad would make for a better world than the truth of a theory according to which it is. So pain is not bad for us. Again, the argument gives us no reason to accept its conclusion. So, in order to defend the claim that the better world argument is worth taking seriously, I have to identify a morally significant difference between the

[17] To be clear, this is the main general objection to *any* formulation of the better world argument, whether that formulation rests on the view that the correct moral theory cannot be self-defeating or on the view that moral facts are determined by some kind of procedure.

kinds of cases in which this reasoning is valid and the kinds of cases I just described, in which it is not.

The difference between applications of the argument that concern moral judgments and those that concern non-moral matters is straightforward. The better world argument does not rest on some more general view that always treats the fact that it would be better if some claim were true as a reason to believe the claim, no matter what the content of the claim happens to be. Rather, as I formulated the argument, it rests on the view that morality cannot undermine the aims that it assigns us, together with the view that morality assigns us the aim of bringing about a better world. But these features of morality have no analogues, for example, in the empirical world. There is simply no relevant sense in which nature aims at what is best. So, although this form of argument is sometimes valid in moral philosophy, we should not expect it to be valid when it concerns, for example, empirical matters.

By contrast, the difference between the valid instances of the better world argument and the parody argument that I presented above is more complicated to describe. Roughly, the valid instances assume that the truth of a moral theory can make for a better world in virtue of that theory's conception of the right, while the parody argument assumes that the truth of a theory can make for a better world in virtue of the theory's conception of the good. Somewhat more precisely, the valid instances rest on the view that the truth of principles of right and wrong behavior – principles that determine how we are permitted to live our lives, and how others are permitted to treat us – can, by itself, shape our circumstances in ways that seem desirable or undesirable on reflection. For example, the truth of such principles might render our circumstances liberating or constraining, fair or unfair, ennobling or demeaning, or the like. But the parody argument rests on the view that the truth of principles concerning the goodness and badness of things can make for a better world simply because these principles stipu-

late that some things have greater value than they seem, on reflection, to possess.[18] I can best clarify this distinction and account for its importance if I begin by explaining more fully why the parody argument fails.

Both the valid arguments and the parody argument rest on the view that morality tends to include principles the truth of which would, by itself, be better than the truth of rival principles, where "principle" refers to any general claim about what is right or wrong, good or bad, or worthy or unworthy of esteem. But the parody argument interprets this view in a problematic way. It assumes that one of the conditions under which the truth of a moral principle would make for a better world is that the principle simply assigns more value to things, or assigns less dis-value to them, than its rivals do. Such a principle might assign great value to common, mild pleasures, like the pleasure of scratching an itch, or it might state that apparently grave evils, like suffering and death, are not bad for us. To be clear, the parody argument assumes that when a principle assigns great value to things, the truth of the principle would be desirable whether or not that assignment of value has an antecedent rationale, and whether or not it is intuitively plausible. After all, neither of these conditions holds for the claim that pain is not bad, but the argument nevertheless assumes that the truth of a principle that simply stipulates – out of the blue, as it were – that pain is not bad would make for a better world.

But that assumption is false. If a moral principle arbitrarily assigns greater value to something than it seems to have on reflection, the truth of that principle would not thereby make the world better in any sense worth caring about. To be clear, since the satisfaction of our substantive moral aims *is* worth caring about – as I said above, the fact that acting rightly serves these aims helps explain why acting rightly is not just a matter of

[18] I am grateful to David Enoch and to an anonymous reviewer for comments that helped me clarify this point.

following pointless rules – this means that the truth of such a principle would not, by itself, make the world better in any sense that serves our moral aims. Of course, the truth of a principle that arbitrarily assigns great value to something might enable us to *say* that some apparent evil, for example, getting the flu, is not bad for us. But the mere truth of the principle would not, by itself, make flu symptoms any easier to bear. Nor would it make the attitudes or actions of someone who tried to avoid getting the flu, say, by getting vaccinated, any less intelligible. Nor would it make the attitudes or actions of someone who seemed indifferent to her risk of getting the flu any less puzzling. In short, the truth of such a principle might enable us to make favorable judgments about certain apparent evils, but only by doing violence to moral language, that is, by severing any link between our value judgments and the things we care about. It would not make our initial, *un*favorable judgments any less apt or intuitive.[19]

Clearly, valid instances of the better world argument cannot appeal to this interpretation of the view that the truth of certain principles would, by itself, be desirable. Recall that the better world argument treats the judgment that the truth of some theory would make for a better world, *and thereby serve our moral aims*, as grounds for the conclusion that the theory's rivals are self-defeating, and should be rejected. If this argument rested, as the parody argument does, on the view that the theory's truth would be desirable simply because the theory assigns greater value to things than they seem, on reflection, to possess, then it would not give us any reason to judge that the theory's truth would serve our moral aims. Nor would it give us grounds for objecting to the theory's rivals.

---

[19] Enoch (2009) offers a tentative defense of the better world argument that focuses, in a different way, on the theory of the good, but Rob van Someren Greve (2011) shows that accepting Enoch's defense would commit us to accepting a parody argument that is relevantly similar to the one I just described.

In order to identify an alternative interpretation of the view that the truth of a moral principle can make for a better world, I will return to the two applications of the better world argument that I described in the previous section. These applications rest on the view that the truth of principles that state that certain behaviors are right or wrong can, by itself, shape the circumstances in which we act, sometimes in ways that make these circumstances better or worse. To illustrate, the first of these applications states that the truth of a theory that includes permissions to promote our own private aims would give us freedom to shape our own lives, without incurring certain grave costs, and thereby make us better off. In other words, the truth of such a theory would be desirable because it would, by itself, prevent us from being boxed in by the ever-present risk of committing a serious wrong. The second application states that the truth of a theory whose principles rule out the possibility of cases in which someone cannot avoid acting wrongly, no matter what she does, would make for a better world by making our circumstances fairer or less tragic than they would be if such cases could occur.

Unlike the parody argument, these applications of the better world argument leave our theory of the good intact. Of course, both applications rely on claims about the goodness and badness of things: the first relies on the claim that being free to shape the content of one's own life, without incurring the costs associated with acting wrongly, is desirable, and the second relies on the claim that living in a world that is tragic or unfair is *un*desirable. But they take for granted our considered judgments about these matters. In other words, these applications rest on the view that the truth of a moral theory can be desirable, not because the theory arbitrarily assigns values to things, but rather – and this is the crucial point – because it includes principles of right and wrong the truth of which would give our circumstances some further feature that seems desirable on reflection. When we judge that the truth of a theory would make for a better world in this way, we preserve the link between our

judgments about what is desirable or undesirable and the things we care about. So we can plausibly claim both that the truth of such a theory would serve our moral aims and that any rival theory that correctly identifies our moral aims is self-defeating. And we can accept the applications of the better world argument that rest on this view of the route by which the truth of moral principles can make for a better world, without committing ourselves to accepting the parody argument.[20]

IV

I have argued that although the better world argument may seem puzzling – if not obviously invalid – at first glance, it can be formulated in a way that makes sense. This form of argument rests on the familiar and deeply plausible view that, other things equal, the correct moral theory cannot be self-defeating, or in other words, it cannot fail in its own terms. But the argument employs this view in a novel way. When a theory is self-defeating in the now familiar sense that Parfit describes in

---

[20] An anonymous reviewer suggested that we may be able to construct a version of the parody argument that focuses on conceptions of the right. For example, provided that it would be bad if people acted wrongly, and better if they did not, the truth of a theory according to which all of the things that people tend to do are morally permissible would make for a better world than any rival theory, according to which one or more of these things, say, telling lies or littering, is *im*permissible. But this argument does not pose a problem for the version of the better word argument that I defend. This new parody argument states that we have reason to reject certain moral theories because the truth of these theories would make for a worse world. But the truth of these theories would make for a worse world because people would act wrongly, or in other words, because they would fail to follow the theories' principles. Such theories need not fail in their own terms. By contrast, the better world argument states that we have reason to reject certain theories because they fail in their own terms. So accepting the better world argument does not commit one to accepting this suggested parody argument.

his critique of Common-Sense Morality, our *successfully following* the theory's principles would serve our moral aims less well than our violating these principles and following other, rival principles instead. In other words, the theory fails in its own terms because the actions or attitudes of people who followed its principles would somehow undermine the aims that, according to that very theory, it is important for them to achieve.

But the main insight of the better world argument is that the truth of certain moral principles would undermine our moral aims in a more direct way, quite apart from the actions and attitudes of people who followed those principles. For example, it may be that the truth of principles that always require us to produce the best available results would be bad for us, by itself, because it would deprive us of a desirable form of autonomy that we would have if we were often permitted to promote our private aims. And it may be that the truth of principles that allow for the possibility of moral dilemmas would make our circumstances less fair or more tragic, and therefore worse, than they would be if such dilemmas could not occur. In that case, on the intuitively plausible assumption that bringing about a better world is among our moral aims, the truth of any moral theory that includes such principles would, by itself, undermine our moral aims, at least in this one respect. In other words, any such theory is self-defeating – it fails in its own terms in at least one respect. But the correct moral theory cannot fail in its own terms unless eliminating this defect carries other, comparable costs. So we have some reason to reject any theory that includes these principles.

As I said above, one would have to do more work to use the better world argument to show that, all things considered, these or any other particular theories should be rejected. But if the arguments that I presented above succeed, the project of seeking out and defending successful instances of this form of argument is well worth pursuing.

Acknowledgements I am grateful to David Enoch, Geoffrey Sayre-McCord, and an anonymous reviewer for very helpful comments.

References

Baier, K. (1958). *The Moral Point of View*. Ithaca: Cornell University Press.

Enoch, D. (2009). Wouldn't It Be Nice If p, Therefore p (for a moral p). *Utilitas*, 21, 222-4.

Hare, R.M. (1981). *Moral Thinking: Its Levels, Methods and Point*. Oxford: Oxford University Press.

Kagan, S. (1989). *The Limits of Morality*. Oxford: Oxford University Press.

_____. (1991). Responses to My Critics. *Philosophy and Phenomenological Research*, 51, 919-28.

Kamm, F.M. (1996). *Morality, Mortality: Volume II: Rights, Duties, and Status*. Oxford: Oxford University Press.

_____. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

McNaughton, D. and Rawling, P. (1998). On Defending Deontology. *Ratio*, 11, 37-54.

Nagel, T. (1995). Personal Rights and Public Space. *Philosophy and Public Affairs*, 24, 83-107.

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

_____. (2011). *On What Matters*, Vol. 1. Oxford: Oxford University Press.

Quinn, W. (1993). Actions, intentions and consequences: The Doctrine of Doing and Allowing. In *Morality and Action*, (pp. 149-174). Cambridge: Cambridge University Press.

Rawls, J. *A Theory of Justice*, Revised ed. Cambridge, MA: Harvard University Press.

Sayre-McCord, G. (unpublished manuscript). A Moral Argument Against Moral Dilemmas.

Slote, M. (1985). *Common-Sense Morality and Consequentialism.* London: Routledge and Kegan Paul.

Sartre, J. (1975). Existentialism is a Humanism, in *Existentialism from Dostoevsky to Sartre*, Revised and Expanded Edition, ed. Walter Kaufman. New York: Penguin Group.

Shiffrin, S. (1991). Moral Autonomy and Agent-Centered Options. *Analysis*, 51, 244-254.

Sidgwick, H. (1907). *The Methods of Ethics*, Seventh ed. London: Macmillan.

Unger, P. (1996). *Living High and Letting Die: Our Illusion of Innocence*. Oxford: Oxford University Press.

van Someren Greve, R. (2011). Wishful Thinking in Moral Theorizing: Comment on Enoch. *Utilitas*, 23, 447-50.

Walker, M. (1991). Moral Luck and the Virtues of Impure Agency. *Metaphilosophy*, Vol. 22, Nos. 1 and 2, 14-21.