JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XIII · NUMBER 3
July 2018

ARTICLES

DISCUSSION

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
http://www.jesp.org

The *Journal of Ethics and Social Philosophy* (ISSN 1559-3061) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge and is available to all readers under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license. Funding for the journal has been made possible through the generous commitment of the Gould School of Law and the Dornsife College of Letters, Arts, and Sciences at the University of Southern California.

The *Journal of Ethics and Social Philosophy* aspires to be the leading venue for the best new work in the fields that it covers, and it is governed by a correspondingly high editorial standard. The journal welcomes submissions of articles in any of these and related fields of research. The journal is interested in work in the history of ethics that bears directly on topics of contemporary interest, but does not consider articles of purely historical interest. It is the view of the associate editors that the journal's high standard does not preclude publishing work that is critical in nature, provided that it is constructive, well-argued, current, and of sufficiently general interest.

Journal of Ethics and Social Philosophy
VOL. 13, NO. 3 · JULY 2018
https://doi.org/10.26556/jesp.v13i3.259

PARFIT, CONVERGENCE, AND UNDERDETERMINATION

Marius Baumann

One especially persistent concern for moral realism takes its force from the observation that there are widespread and deep disagreements when it comes to moral issues. How could there be a truth of the matter in a field so pervaded by often fundamental dissent? This worry is at the root of what has been aptly called the *argument from moral disagreement*. Recently, Derek Parfit has provided us with a highly remarkable defense of moral realism against this argument. In his 2011 *On What Matters*, Parfit aims to show that the best versions of three of the most important families of moral theories—namely Kantianism, consequentialism, and contractualism—actually agree about what is right and wrong, forbidden, mandatory, or permissible. This, he argues, strengthens the case for moral realism vis-à-vis the argument from disagreement. If our best moral theories turn out to be in agreement about what is right and wrong, we have reason to be more optimistic about the prospect of truth in ethics.[1]

Many commentators have already taken issue with Parfit's sometimes idiosyncratic reading of the authors he engages with and expressed doubts that he can achieve the promised convergence. I outline a new challenge that only sets in at a stage where Parfit considers his argument to be already won. As I see it, Parfit might have shown that moral theories can indeed agree in their extension, i.e., the set of deontic verdicts they yield for every particular act—past, present, and future.[2] However, he has not shown that the more theoretical disagreements, relating to the explanation of *why* certain acts are right or wrong, can also be settled. Instead, we are left with a case of extensionally equivalent yet theoretically incompatible theories. Interestingly enough, as Dietrich and List have recently observed, a structurally analogous situation has been discussed for some time now in the philosophy of science under the name of the *under-*

1 For better readability, I will often use "right or wrong" as a short form for all the deontic verdicts.

2 At least in this world. I discuss the modal strength of this claim below.

determination of theory by evidence.[3] Following Pierre Duhem and W.V. Quine, philosophers of science have debated the possibility of there being rival scientific theories that can account for exactly the same evidence, but do so by giving explanations that are themselves incompatible. More remarkably still, this idea traditionally figures in arguments for *anti-realist* positions, a fact that has not yet attracted enough attention when it comes to the analogy to ethics.

My aim is to outline how we can reinterpret Parfit's findings in a similar anti-realist vein and thus question the effectiveness of his project to vindicate moral realism. Here is how I will proceed. I start by giving some background to the notions of moral realism and the argument from disagreement. I then outline Parfit's convergence argument and point out its main flaw, which is that it fails to address the remaining explanatory disagreements. Next, I turn to the philosophy of science for some inspiration. I introduce the idea of underdetermination and sketch how it figures in an anti-realist argument, which I subsequently adapt to the realm of ethics. Finally, I discuss three possible realist rejoinders, ultimately arguing that none of them does the trick for Parfit.

1. MORAL REALISM AND THE ARGUMENT FROM DISAGREEMENT

Let me set the stage for Parfit's reasoning in *On What Matters* (OWM) by introducing two crucial notions: *moral realism* and the *argument from disagreement*.

1.1. Moral Realism: A Thin Definition and Two Additional Components

Many of our moral claims convey the impression of stating truths about the world. If I say that your lying to your parents was wrong, or that the supreme principle of morality is the categorical imperative, or that an action is forbidden because it produces less total utility than an alternative, I seem to be stating propositions that purport to be true.

Moral realists take these claims at face value and hold that at least some of them are actually true. Sayre-McCord accordingly identifies two universally shared commitments of moral realism:

a. Moral claims, when literally construed, have truth values.
b. At least some moral claims actually are true.[4]

3 I came up with the analogy independently of Dietrich and List and was only recently made aware of their paper. To the best of my knowledge, Dietrich and List, "What Matters and How It Matters," includes the first explicit mention of the analogy.

4 Compare, not in exact wording but in spirit, Sayre-McCord, "The Many Moral Realisms" and "Moral Realism."

We can call this the *thin* definition of moral realism.[5] It is obviously no more than the lowest common denominator of the different views that go by the name of moral realism. Most realists will want to add other components to their preferred version. Two options are especially relevant to our case. First, one might prefer a metaphysically thicker version that also entails something about the ontological status of moral facts or properties. For example, one might want to argue that the aforementioned moral claims can only be true if they refer to some moral property that exists independently of any natural properties. Yet this preference for a metaphysically committed view is not shared by all realists and, as it happens, is not shared by Parfit, who advocates for what he calls a *non-metaphysical* version of realism that entails no ontological claims about moral facts or properties, a position that has attracted much attention lately.[6]

Second, definitions of realism often comprise an epistemic component, relating to our ways of recognizing moral truths. That epistemic component comes in different strengths. Some realists rest comfortably pointing out solely that we do in principle have the required abilities to find out about some of the moral truths. Others take a stronger stance and line out to what extent we have already succeeded, or would under sufficient conditions be succeeding, in this undertaking. Parfit, it will turn out, subscribes to a rather strong variety of the epistemic component. This has crucial implications for what means are available to him in answering the challenge I am about to outline.

Thus Parfit subscribes to an epistemic but non-metaphysical understanding of moral realism. That means that the thin definition given above does *not* encode

5 I hesitate to call it the *minimal* definition, since that has other connotations. The quest to give a definition of realism has been rendered more complicated by the emergence of what has been called *minimalism* about truth or factuality, a position that allows non-realist-leaning philosophers to avail themselves of talk about moral facts or moral truths. People have consequently doubted whether the talk of moral truth (or facts or properties, for that matter) is any longer the distinguishing characteristic of realism; see Dreier, "Metaethics and the Problem of Creeping Minimalism," for a succinct statement of that view. Sayre-McCord attempts to exclude such positions by adding the proviso that we have to construe those claims *literally*. I put these issues aside for the rest of the paper.

6 See Parfit, *On What Matters*, 2:486. Other proponents of that view include Dworkin, *Justice for Hedgehogs*, and Scanlon, *Being Realistic about Reasons*. More precisely, Parfit is a realist about reasons, which is a distinct position that adds to the former definition of realism what has sometimes been called *reasons fundamentalism*, i.e., the view that we can analyze all normative notions in terms of reasons, which are themselves not further analyzable. I will neglect any complications that might arise from this specific view, on the assumption that all of the problems I am about to bring out could in principle be restated in terms of reasons. I will also neglect terminological subtleties, such as Parfit's preference for the label *cognitivism* as opposed to *realism*. If we understand realism in the sense just outlined, there should be no misunderstandings.

his complete view. However, it will prove useful to keep it in mind and think of the other components as *additions* that are characteristic of specific forms of realism. On the one hand, this will enable us to pinpoint more precisely where the problems with Parfit's line of argument lie. On the other hand, if Sayre-McCord is right, his definition picks out what is, *mutatis mutandis*, common to *all realisms*, not just the moral variety.[7] It should thus be convenient as a working definition when comparing discussions of realism in different domains of inquiry.

1.2. The Argument from Disagreement

One classic argument against moral realism originates from the simple observation that, when it comes to morality, we encounter wide and strong disagreements. Considering how common and deeply divisive these disagreements often seem, realism faces a problem. Why should we expect there to be a fact of the matter in a field that is so pervaded by disagreements, as is the case in morality? Seeking to account for this, it would seem that we do better without the idea of as yet to be detected moral truths. Instead, other accounts, having to do with, e.g., the cultural or evolutionary origins of our moral beliefs, readily suggest themselves. These, or so it is suggested, do a better job of explaining our disagreements, and they go against the assumption of moral truths. Thus, the argument is commonly presented in the form of an *inference to the best explanation*.[8]

As it stands, the argument obviously needs improvement. The mere fact that people differ in their moral outlooks is an observation that is no less indisputable than it is inconclusive with regard to its metaethical upshots. A multitude of factors can explain the disagreements between people without challenging the assumption that there is after all a truth to be found. People might not know the relevant nonmoral facts or understand what moral claims are supported by those facts; they might be under some sort of distorting influence; or they might be using different concepts, thus merely talking past each other.[9]

What might make the case worse for the moral realist, though, is disagreement persisting between those people who have most thoroughly contemplated the issue. This, or so it is commonly assumed, is the case in normative ethics. Disagreement about the assessment of moral questions has continued to exist

7 See Sayre-McCord, "The Many Moral Realisms," 5.

8 As, for example, Mackie does in one of the defining modern statements of the argument, where he rather misleadingly calls it the *argument from relativity*, seemingly presupposing what would have to be proven (*Ethics*). Tersman points out that there are at least two alternative ways in which the argument from disagreement can be construed (*Moral Disagreement*, xiii). We do not have to go into this in more detail, though, since Parfit clearly does not have Tersman's alternatives in mind.

9 See Parfit, *On What Matters*, 2:552–63, for an extended discussion of these distorting factors.

between proponents of the most prominent and widely held traditions of moral theorizing up until today. Such is the way in which Parfit seems to think of the problem. He worries considerably about being in disagreement with other experts whom he considers to be his *epistemic peers*:

> [People who] may be responding to the same evidence, their judgement in other cases may have been as reliable as ours, and they may not be more likely to have been misled.[10]

Indeed, Parfit fears that such disagreement might threaten our view that anything matters at all.[11] Yet, even if we tone down the rhetoric somewhat, it should be clear why this kind of disagreement is more threatening to realism, *provided* that it incorporates an epistemic component.[12] Granted that one can blame distorting influences for many disagreements between laymen, it would strain the realist's credibility to claim that we are able to detect moral truths, only to find out that even our best, most worked-out theories disagree about a considerable amount of nonperipheral verdicts.

2. Parfit's Convergence Argument and the Remaining Explanatory Disagreements

This awkward situation is what Parfit has set out to change. In volume one of *OWM*, Parfit argues, among many other things, that the best versions of three of the most important families of moral theories arrive at the same conclusions about what matters.

2.1. Convergence and the Vindication of Realism

Parfit's reasoning for this surprising conclusion proceeds via a very detailed and

10 Parfit, *On What Matters*, 2:428.

11 See Parfit, *On What Matters*, 2:426–30.

12 That is on the assumption that one only needs to bother about people disagreeing if one holds that there is a way for us to find out about moral facts. However, as Tersman (*From Scepticism to Anti-Realism*) suggests, this might be too much of a concession to the realist. Proponents of the argument from disagreement can try to strengthen their case by insisting that absent the ability to come to know any moral truths, we are not entitled to think that there are any in the first place. Thus, showing that there are indeed such deep disagreements would suffice to challenge the realist position even if that position comes down to only the metaphysical component. I am sympathetic to that line of reasoning as an attempt to broaden the appeal of the argument beyond Parfit's specific form of realism. However, since it is not strictly needed for the overall argument, considering Parfit clearly does subscribe to an epistemic definition of realism, I will not incur the additional requirement of defending it.

intricate succession of arguments that I will not be able to restate in detail. The main structure is something like this: Parfit identifies, through a rigorous analysis of problems and objections, what he sees as the best versions of Kantianism, consequentialism, and contractualism. He is happy to acknowledge that his main interest is not in staying true to every detail of the original theories, but rather in searching for the best possible forms those theories could take.[13] In a most remarkable move, Parfit then construes what he calls the *Kantian argument for rule consequentialism*. The argument is supposed to show that those principles that everyone can rationally will are simply those that, if universally accepted, would make things go best. Since the former formulation is Parfit's preferred version of Kantianism and the latter amounts to the best version of (rule) consequentialism, Kantianism therefore implies consequentialism. To top things off, the only principles that everyone can rationally accept are also highly likely to be the ones that no one can reasonably reject, granting compatibility with (the best version of) contractualism as well. Taking all those moves together, we can dub this the *convergence argument*.

The argument relies heavily on substantial views about reasons and rationality. In particular, Parfit is of the opinion that everyone has reasons to want the best outcomes as they would be seen from an impartial point of view (short: the optimific outcomes), and that these reasons are at least not decisively outweighed by non-optimific considerations.[14] This is what ultimately allows him to argue that Kantians would indeed choose the same principles as consequentialists. That claim has already attracted considerable criticism.[15] Much ink has also been spilled on whether Parfit's portrayal of the authors he engages with is a fair and unbiased one, and whether his argument for convergence goes through.[16]

However, these issues are not my concern here. Parfit is convinced that his preferred versions are not too far off to still be considered representatives of the three moral traditions, and I am willing to grant this as well as the convergence claim. How is this supposed to strengthen the case for moral realism, then? As we have seen, the argument from disagreement tries to capitalize on the fact that

13 Compare Parfit, *On What Matters*, 1:338–39 and 369–70.

14 Parfit, *On What Matters*, 1:377–79.

15 See, for example, Otsuka, "The Kantian Argument for Consequentialism," and Setiya, review of *On What Matters*.

16 Compare Herman, "A Mismatch of Methods"; Scanlon, "How I Am Not a Kantian"; and Larmore, "Morals and Metaphysics" for the former, and Ross, "Should Kantians Be Consequentialists?"; Wolf, "Hiking the Range"; and Darwall, "Agreement Matters" for the latter critique. What is most interesting is that representatives of *all three* of the major traditions have objected that their view gets misrepresented, and that convergence is only achieved because of an underlying bias toward one of the other frameworks.

we disagree about our moral verdicts all the time. But if Parfit is correct, this is not so for our best theories. The three main traditions agree when it comes to their main principles. Parfit does not tell us in detail which principles these are. Yet, based on his examples, we can assume that what is meant is a set of principles that specify the deontic status of classes of acts in certain circumstances.[17] I will refer to these as the *deontic principles*. Since, very plausibly, verdicts in particular cases follow directly from such principles, the theories must also agree on the former. In other words, they turn out to be *deontically equivalent*. That is, they lead to exactly the same sets of verdicts about which particular acts are right or wrong, mandatory, allowed, or forbidden.[18] Since these theories are also the best versions of our most important traditions, at least one of which most experts have thought to be true, this should make us confident that the verdicts we arrive at are also the correct ones.[19] Ethics therefore turns out to be in no worse condition than many areas of inquiry where we are more confident that there is a truth of the matter. We may find much disagreement at first sight, but if we turn to our best theories, those disagreements vanish.

2.2. The Remaining Explanatory Disagreements

The depth and length at which Parfit follows through his argument for the convergence of moral theories is impressive. Nevertheless, I think that, even if the argument is successful, that is not enough to vindicate moral realism. To begin to understand why, let us first turn to one of the most famous passages in *OWM*. After outlining the convergence argument, Parfit ends volume one on a memorable note:

17 Examples Parfit discusses, and repeatedly modifies, include the *Consent Principle*—"It is wrong to treat anyone in any way to which this person could not rationally consent" (*On What Matters*, 1:181)—and the *Mere Means Principle*—"It is wrong to treat anyone merely as a means" (1:212).

18 When putting forward this claim, Parfit does not explicitly specify its modal strength, that is, whether equivalence holds for all worlds or just some subset. However, later comments suggest a strong reading, e.g., when he states: "Fundamental normative truths are not about how the actual world happens to be" (*On What Matters*, 2:489). He goes on to state that pain would be bad and rational beings would have reasons to achieve their rational aims in any possible world. As Nebel points out, this view is a direct upshot of Parfit's more general metaethical convictions ("A Counterexample to Parfit's Rule Consequentialism," 1). Nonnatural properties cannot be discovered by empirical means, yet if they were merely conditional, that is how they would have to be discovered on his view.

19 The point is aptly expressed by Ridge, who submits that "these three traditions (Kantian, contractualist, and consequentialist) would surely be on any reasonable person's short list of the most promising moral theories yet developed" ("Climb Every Mountain?" 60).

> It has been widely believed that there are such deep disagreements between Kantians, Contractualists, and Consequentialists. That, I have argued, is not true. These people are climbing the same mountain on different sides.[20]

The metaphor of the mountain has come to be seen as the epitome of Parfit's view.[21] At the same time, it strikes me as particularly telling of the fundamental problem with his project. I take the metaphor to signify that when the different theorists reach the summit, i.e., when they have perfected their theories, they will notice that they agree on the same principles. Thus, there are indeed no disagreements remaining on this level. But the metaphor also betrays something else. It suggests that the theorists take different roads to the summit. This immediately prompts an additional question: Why do these differences not matter? Why is it not important how we get to the mountain?

To put it less metaphorically, I think that Parfit's argument falls short of its aim because what has been shown is, at most, that different theories can indeed lead to the same deontic principles and verdicts. Parfit thinks that he has therefore settled all the relevant conflicts. But this, or so I will argue, is not so. Though deontic equivalence might have been proven, there remain differences when it comes to those parts of the theories that go beyond the mere production of deontic verdicts.

This point has recently been stressed by Suikkanen, who accords with Parfit that the different traditions (he only considers Kantianism and consequentialism) can agree about right and wrong actions. However, he goes on to ask why Kantians and consequentialists nevertheless disagree, and I think his answer is exactly on the right track:

> Why do Kantians and consequentialists then disagree despite this? Perhaps the best way to understand why they still disagree is to think that they have different views about what makes the intuitively right acts right. Consequentialists claim that the acts which we all believe to be right are right because they bring about the best outcomes. In contrast, Kantians claim that these acts are right because the relevant maxims for them can be willed to be universal laws. Consequentialists and Kantians give competing explanations for why certain acts are right even if they can agree on which acts are right.[22]

20 Parfit, *On What Matters*, 1:419.

21 Indeed, the manuscript that was widely circulated before the publication of *OWM* was titled "Climbing the Mountain."

22 Suikkanen, *This Is Ethics*, 104.

What Suikkanen is bringing to our attention here is that moral theories are not only in the business of producing the correct particular deontic verdicts. They also seek to explain *why* certain acts are right or wrong (obligatory, forbidden, or allowed). This attempt at explanation I take to be at the very heart of moral theorizing. Indeed, ethicists often seem to be more interested in getting the explanation right than in fine-tuning their theory's output.

What kind of explanation does Suikkanen refer to here? Certainly not the *metaethical* one, having occupied philosophers since Harman, of whether purported moral facts figure in our best explanations of our moral beliefs.[23] Instead, it is a more generic notion of explanation inherent to normative ethics, pertaining to the reasons we are giving within ethical discourse and practice. Different ways of understanding this kind of explanation present themselves and we have to be careful here since not all accounts fit well within Parfit's framework. For example, Berker has recently proposed that moral explanations are quite naturally amenable to an analysis in terms of the notion of *grounding*.[24] On such an analysis, the alternative traditions of moral theorizing make claims as to what is fundamentally prior in ethics. Importantly, that priority is neither epistemic nor causal or conceptual.[25] Instead, according to that view, moral explanations identify the ultimate metaphysical grounds for an act's rightness or wrongness.

The problem with such an analysis is that Parfit, as we have seen, defends a non-metaphysical version of moral realism.[26] Presumably, this also includes explanatory claims. In order to stay within Parfit's framework, it is thus better to adopt a non-metaphysical understanding of moral explanation. I suggest that the most straightforward is a semantical one. Seen this way, moral explanations consist of truth-apt sentences (or propositions) about why acts are right or wrong. This is in accordance with Parfit's remarks on other moral claims. Despite having no ontological implications, Parfit informs us, such claims can nevertheless be true in the strongest sense.[27] Thus, just as we do not make any ontological commitments when attributing to acts the property of rightness, saying that some feature makes an act right or wrong does not entail any ontological consequences.

How do these different explanations relate to one another? We can begin by observing that they use markedly different language. Kantians typically refer to such concepts as autonomy, good will, humanity as an end in itself, etc. Con-

23 Harman, *The Nature of Morality*.

24 Berker, "The Unity of Grounding."

25 Compare Berker, "The Unity of Grounding," 10–13.

26 Compare Parfit, *On What Matters*, 2:486.

27 Parfit, *On What Matters*, 2:479.

tractualists use the ideas of a (hypothetical) agreement, some kind of initial position, and cooperation between rational (self-interested) agents, among others. Finally, the notions that consequentialists emphasize include that of an (actual or foreseeable) outcome, its overall utility, and its ranking on a scale of goodness. Most ethicists think that the differences in these explanations reach deeper than the terminological surface. They agree with Suikkanen that the different traditions give *competing* explanations.

This might be mistaken, though. There is a possibility that the different traditions are merely *notational variants*. When putting forward seemingly incompatible claims about what makes acts right, Kantians, consequentialists, and contractualists would merely be using different terminology. A better understanding of these claims might indicate a way of translating one explanation into the terms of the other and demonstrate that there is no real incompatibility after all. These matters are not to be conclusively decided here. Since a translation of predicates could be rather complicated and surprising, we cannot dismiss that possibility out of hand. However, I contend that this is not what is called for at this stage. First, it is important to note that the notational variants view clearly goes against the orthodox picture. There is no obvious way to translate the explanatory notions used by one of these frameworks into those used by its rivals. Standard textbooks conceive of the different traditions as rival, incompatible frameworks.[28] The onus is therefore on the critic of the standard view to show that the three traditions are in fact not giving competing explanations, despite it obviously seeming like they do. Parfit certainly has not shown this.

Second, there is good additional reason to think that it could not be shown. It seems that it is constitutive of the moral traditions that they give explanations of a certain kind. A consequentialist theory that does not maintain that the rightness of an act is an upshot of its outcomes alone is (obviously) not a consequentialist theory. What is more, a Kantian theory that would accept this *same claim* about the rightness of an act is not a Kantian theory. Thus, there is a very strong sense that these theories give mutually exclusive explanations *and* could not cease to do so without losing a constitutive feature. This finds its expression in the fact that the different traditions are sometimes even *defined* in opposition to each other.[29] Pending arguments to the contrary, we thus have very strong

28 For recent examples, compare Tännsjö, *Understanding Ethics*; Driver, *Ethics*; and Deigh, *An Introduction to Ethics*.

29 For instance, in "Deontological Ethics," Alexander and Moore use consequentialism as a foil to define deontology.

reason to consider these alternative theories to be logically incompatible when it comes to their explanations.[30]

If this is correct, Parfit is not justified in claiming that all differences among the alternative traditions have been resolved. Though Kantians, contractualists, and consequentialists might agree on a set of principles (and thus on the deontic verdicts that follow from them), they put forward different, mutually exclusive explanations for why those are the correct principles. This leads us to a puzzling result. On the one hand, alternative traditions of moral theorizing seem to be able to lead to the same deontic verdicts. On the other hand, differences remain as to how those theories explain the deontic status of acts. In other words, those theories are *deontically equivalent while at the same time being logically incompatible*. How are we able to explain this?

3. UNDERDETERMINATION AND REALISM

Remarkably, in the philosophy of science a similar phenomenon has attracted a lot of attention for quite some time now: the underdetermination of theory by evidence. This analogy between science and ethics has recently been pointed out by Dietrich and List.[31] Since it takes a prominent role in our argument, it will be useful to provide some background to the phenomenon in science before considering the analogy.

3.1. Underdetermination in Science

Following the work of Duhem and Quine, philosophers of science have debated a highly influential, though nonetheless controversial, idea: *the underdetermination of theory by evidence*. Duhem and Quine both held that sometimes alternative scientific theories are able to accommodate exactly the same observable evidence, while simultaneously making radically different assertions about unobservables.[32] These theories thus give mutually exclusive explanations for the

30 Since this is quite a mouthful, I will also speak of *theoretical* or, more specifically, *explanatory incompatibility* when I want to highlight that the incompatibility is not regarding the deontic content. This does not designate a special kind of theoretical/explanatory incompatibility but is rather a short form for logical incompatibility concerning the theoretical/explanatory claims.

31 Dietrich and List, "What Matters and How It Matters."

32 Duhem's *The Aim and Structure of Physical Theory* (*La Théorie Physique: Son Objet et sa Structure*) is widely regarded as the *locus classicus* for the underdetermination thesis. Quine's "Two Dogmas of Empiricism" introduced the idea to the analytic tradition, while his own views have kept changing over time. (Compare Quine, "Two Dogmas of Empiricism," "On

same evidence. They are, in other words, *empirically equivalent while at the same time being logically incompatible*.[33]

The standard examples to illustrate underdetermination are mostly drawn from the history of science, especially physics. Famously, at the time of Copernicus, the choice between his new theory and its Ptolemaic alternative was underdetermined by observational data. Likewise, not that long ago, the relevant observations seemed to be compatible with both a corpuscular and a wave theory of light.[34] For our purposes, it would take us too far off track to consider them more closely. However, the basic idea can be illustrated by means of everyday phenomena. Ladyman invites us to consider the case of some person at a train station waiting for a delayed train.[35] While waiting, this person develops several hypotheses on why the train might be delayed, such as problems with the engine, a staff shortage, or a signal failure. All the hypotheses are compatible with the data available to the person at that time, but the explanations are incompatible (barring the case of multiple causation). Of course, such a case of underdetermination is not very interesting for the obvious reason that it is a consequence of the epistemic situation of a specific person at a specific time. Other people probably already know the correct explanation and the person herself could find it out easily with some research. What makes for a philosophically more interesting case are situations in which the whole scientific community, given all the data available, cannot decide among the different theories. This, or so it is claimed by proponents of underdetermination, has been (and still is) the case with at least some of our scientific theories.

Not much more is common ground besides this. Duhem and Quine already held quite different views about the nature and scope of underdetermination. Whereas Duhem argued for a moderate, restricted thesis that was informed by detailed historical examples, Quine envisioned a more radical version that

---

the Reasons for Indeterminacy of Translation," "On Empirically Equivalent Systems of the World," and "Three Indeterminacies.")

33 At a later stage, Quine renounces his claim of *logical* incompatibility between the rival theories ("On Empirically Equivalent Systems of the World"). Instead, he now holds that the incompatibility between the empirically equivalent theories merely consists of us not being able to find a way to reconcile them by way of reconstruing their predicates. That is, the theories are not logically incompatible; they are simply using different predicates that we have not yet managed to translate into each other. For reasons I explain in note 42, anti-realists should resist this move and insist on there being logical incompatibility.

34 Both these cases have already served as examples of underdetermination to Duhem. More recently, Stanford, *Exceeding Our Grasp,* and Bonk, *Underdetermination*, have discussed a range of further cases at length.

35 See Ladyman, *Understanding Philosophy of Science*, 162–63.

pertains to the totality of our knowledge, and for whose justification he relied more heavily on general epistemological, logical, and linguistic considerations.[36] Many authors have adopted Quine's wider understanding, so that the thesis that is most often discussed today looks something like this:

> *Underdetermination of Scientific Theory* (UST): There are alternatives to even our best scientific theories that can account for exactly the same evidence while containing incompatible propositions.[37]

As might be expected, considering how long philosophers have grappled with the idea of underdetermination, a fair amount of criticism has also been directed at it. These criticisms can be roughly divided into two classes. On the one hand, philosophers have questioned whether underdetermination is in fact a sufficiently widespread phenomenon to justify a thesis like UST. Thus, it has been pointed out that, for most of the time, the discussion has focused on a very small set of examples, which are mostly from a subset of physics, and which are not representative of science *per se*.[38] This kind of criticism, though of great relevance for the debate in science, should not bother us too much, and for a simple reason. As we shall shortly see, I am not advancing a claim to the effect that *because* there is underdetermination in science we should also expect to find it in ethics.

On the other hand, there are also more general lines of criticism that cast doubt on the mere possibility of rendering the idea of underdetermination plausible (or coherent) at all.[39] Such criticism is more dangerous to our purpose because it might generalize to any form of underdetermination in any area of inquiry. I address some of these objections below. For now, I will simply continue on the assumption that underdetermination can be rendered a coherent idea.

36 For a similar assessment of what the differences between Duhem and Quine amount to, see Pietsch, "Defending Underdetermination or Why the Historical Perspective Makes a Difference."

37 A very concise overview of the different versions that have been proposed can be found in Park, "Philosophical Responses to Underdetermination in Science."

38 See Kitcher, *Real Realism*, and Stanford, *Exceeding Our Grasp*.

39 These include the claims that underdetermination is a nonstarter because the distinction between observables and unobservables cannot be upheld (Maxwell, "The Ontological Status of Theoretical Entities"), that most versions of the underdetermination thesis are based on a wrong-headed equation of the logical consequences of a theory with its evidential support (Boyd, "Realism, Underdetermination, and a Causal Theory of Evidence"; Laudan, "Demystifying Underdetermination"), and that the notion of empirical equivalence cannot be rendered precise in a way that would support the underdetermination thesis (Laudan and Leplin, "Empirical Equivalence and Underdetermination," and Worral, "Underdetermination, Realism and Empirical Equivalence").

3.2. *The Anti-Realist Argument from Underdetermination*

On that note, let us turn to the more pressing issue for our purpose, which is the upshot of underdetermination for the realism debate. Why is underdetermination considered to be dangerous for realists? At first glance, it merely poses a problem for theory choice. Although we know that we are facing rival theories, since they contain incompatible propositions, our choice between them is rendered indeterminate by the fact that they can account for exactly the same evidence. Yet, in addition, many philosophers have also held that underdetermination poses a specific problem for scientific realism.[40]

To get a firm grasp of what the danger to scientific realism amounts to, we can start with a thin definition of scientific realism modeled on Sayre-McCord's definition of that doctrine for the moral realm:

a. Scientific claims, when literally construed, have truth values.
b. At least some scientific claims actually are true.

Note again that, according to such a thin definition, there is neither a metaphysical nor an epistemological component to realism. However, as is the case in ethics, most scientific realists will add such additional components to their preferred view.[41] It is important to see that whether underdetermination poses a threat to a realist view depends on which component that is and how strong it is formulated.

Before I can go on spelling that out, we first need a formulation of the basic idea behind the anti-realist argument on the table. So as not to complicate matters too much, I will outline an intuitive version of the argument and postpone discussion of its complications to a later stage. Hence, the argument could read something like this:

P1. If two scientific theories (ST) can account for exactly the same evidence, it is equally reasonable to believe either of them.
P2. If it is equally reasonable to believe either of two ST, we have no reason to attribute truth to one but not the other.

40 Indeed, the argument from underdetermination is sometimes seen as one of two main arguments against scientific realism, together with the so-called *pessimistic metainduction*. Compare Bortolotti, *An Introduction to the Philosophy of Science*, and Stanford, "Underdetermination of Scientific Theory."

41 Compare Newton-Smith and Lukes, "The Underdetermination of Theory by Data," and Bortolotti, *An Introduction to the Philosophy of Science*. Most scientific realists would probably not even see the point in defending the doctrine if described so thinly.

P3. If two ST contain incompatible propositions, they cannot both be true.[42]

P4. If two ST cannot both be true, and we have no reason to attribute truth to one but not the other, then none of them should be considered true.

UST. There are alternatives to even our best scientific theories that can account for exactly the same evidence while containing incompatible propositions.

C. Therefore, even our best scientific theories should not be considered true.

Most readers will probably have immediate reservations about one or more of the premises. I must ask them to bear with me for a little longer. Here, I only want to draw attention to how the argument relates to the different versions of realism that we identified in the beginning.

First, note that the argument, as it stands, does not directly challenge scientific realism as we defined it. On our definition, scientific realism is a position about scientific claims, not theories. However, we can reasonably argue that, since our scientific claims follow from our best theories and the latter should not be considered true, we also should not believe the former. Still, there is a further problem. The argument is evidently of the *skeptical* variety due to its focus on the reasonableness of belief. It states that we should not consider the theories (or their claims) true because, based on the evidence, we have no reason to prefer one theory and we also know that they cannot both be true. But insofar as this poses a problem for realism, it presupposes that there is an epistemic component to that position.[43] Moreover, the strength of the argument is inversely proportional to the strength the realist ascribes to the epistemic component. The stronger the realist insists on us being able to find out about scientific truths, the

42 Here is the reason why anti-realists need to insist on logical incompatibility between the rival theories. If incompatibility comes down to nothing more than our practical inability to reconstrue predicates, we would have no reason to believe P3. That being said, the burden is not automatically on the anti-realist to prove logical incompatibility. If there is a strong initial indication that there is logical incompatibility between the theories, it is on the opponent of the argument to show that the seeming incompatibility can be resolved.

43 But recall note 12. The anti-realist can try and strengthen the argument by charging that the lack of any ability to attain knowledge about moral truths renders baseless the idea that there are such truths. The argument, in combination with this claim, would thus also pertain to forms of realism that include only a metaphysical component. This surely needs to be taken into consideration when assessing the strength of the argument beyond our present focus on Parfit.

more damning the argument will prove if it indeed succeeds in establishing that we cannot find out about those truths. This will shortly become very important.

Second, note that the argument, as it stands, is deliberately formulated in terms of the truth of propositions and theories. It does not depend on any metaphysical assumptions. This gets overlooked easily because philosophers of science are naturally tempted to express the idea of underdetermination in terms of unobservable objects, thus in ontological terms. But it is commonly acknowledged that the underlying problem is a broader epistemological one.[44] We can therefore resist this tendency and formulate the challenge solely in terms of the truth of (scientific) theories and propositions. This opens up the possibility of transferring it to other domains. More specifically, it opens up the possibility of transferring it to debates about forms of realism that are ontologically noncommittal.

3.3. *Adapting the Argument to the Realm of Ethics*

If we are to believe anti-realists in the scientific domain, empirically equivalent theories that contain mutually exclusive propositions pose a problem for scientific realism. But how does this relate to our case in ethics? This is where Dietrich and List come into play.[45] Drawing on decision-theoretic work, they propose a new formal framework for the classification of moral theories. In that context, they introduce what they call the *reason-based representation* of moral theories. Echoing Suikkanen, this framework distinguishes between two dimensions of moral theories:

> Reason-based representations encode not only a theory's *action-guiding recommendations* (that is, how we should act, according to the theory), but also the *reasons* behind those recommendations (that is, why we should act in that way).[46]

Dietrich and List's primary use of their framework is for a formal taxonomy of moral theories, which need not occupy us here. However, it also helps to drive home an important analogy. As they point out themselves, their framework "shed[s] light on an important but still underappreciated phenomenon . . . : the *underdetermination of moral theory by deontic content*."[47] In their view, just as scientific theories can be underdetermined by the empirical evidence, moral

44 See Ladyman, *Understanding Philosophy of Science*; Stanford, "Underdetermination of Scientific Theory."

45 Dietrich and List, "What Matters and How It Matters."

46 Dietrich and List, "What Matters and How It Matters," 422.

47 Dietrich and List, "What Matters and How It Matters," 422.

theories can be underdetermined by their deontic content. By formally distinguishing between the two dimensions, they show that it is at least theoretically possible that theories would differ when it comes to the second dimension but not the first. Moral theories can give different explanations of what makes acts right or wrong but nevertheless agree about the deontic status of those acts. This is, in a structural way, similar to what has been discussed in the underdetermination debate in the philosophy of science.

The analogy relies on treating both the data of scientific theories as well as the deontic verdicts of our moral theories as *extensions* of those theories. Dietrich and List do not flesh out in detail what this amounts to. However, it seems clear that it does not mean that the particular verdicts are epistemically on a par with the data of scientific theories *in every respect*. First, it is not being claimed that *particular* verdicts are prior in the sense that they have an initially higher credibility than, e.g., mid- or highest-level principles. What the analogy presupposes is only that particular verdicts have to be accounted for by moral theories, just as the data has to be accounted for by scientific theories. Second, it is not being claimed that particular *verdicts* are the ultimate ground of epistemic justification in ethics. Instead, what is proposed here is compatible with the view that our moral verdicts are themselves grounded in, for example, (non-doxastic) intuitions of the same content.[48]

More importantly for our present case, by ascribing particular verdicts the role of evidence in ethics, we do not presuppose anything that Parfit does not already acknowledge. We will shortly contemplate whether particular verdicts are the *only* evidence in ethics. But unless we deny that the particular moral verdicts play any role whatsoever in theory choice in ethics, the analogy stands. Parfit definitely does not deny this, since otherwise he would not have to worry about alternative theories arriving at different deontic verdicts in the first place.[49]

Though Dietrich and List's main focus is on the so-called *consequentializing* debate, they do comment on Parfit's project in OWM. They contend that their rea-

48 Indeed, talk of the *data* of moral theories has always had particularly wide currency in intuitionist theorizing. Compare Ross, *The Right and the Good*, 41, for a classical statement of that view, and Audi, "Intuition, Inference, and Rational Disagreement in Ethics," 476, for a more recent one.

49 One might object to the analogy on a different ground. Perhaps one does not think that moral theories are supposed to account for the moral verdicts that we *actually*, at this time, hold. Instead, they tell us what verdicts we *should* be holding. But this objection misses the point. My aim is to show that Parfit's solution gets him into problems. For this purpose, it is sufficient to point out that Parfit gives the verdicts that we do actually hold a central role in deciding which theories are correct. Indeed, much of volume one of OWM is a constant refinement of the different theories in order to make them compatible with specific cases.

son-based representation supports at least the possibility of different theorists climbing the same mountain on different sides, while remaining uncommitted on whether this is indeed the case with Parfit's preferred theories.[50] However, they remain largely silent when it comes to the metaethical consequences of this observation.[51] Yet, if the proposed analogy stands, it is easy to see how we can construct a structurally analogous anti-realist argument for the moral realm as was conceived for the scientific realm. We only have to substitute *moral theories* for *scientific theories* in the argument above to see this:

P1′. If two moral theories (MT) can account for exactly the same evidence, it is equally reasonable to believe either of them.
P2′. If it is equally reasonable to believe either of two MT, we have no reason to attribute truth to one but not the other.
P3′. If two MT contain incompatible propositions, they cannot both be true.
P4′. If two MT cannot both be true, and we have no reason to attribute truth to one but not the other, then none of them should be considered true.
UMT. There are alternatives to even our best moral theories that can account for exactly the same evidence while containing incompatible propositions.
C. Therefore, even our best moral theories should not be considered true.

Ethics might thus face a similar problem as science. We find theories with the same extension that nevertheless contain propositions that cannot be true at the same time. If we are unable to choose between those theories, that makes it difficult for us to believe the claims they put forward. Instead of helping realism by refuting the argument from disagreement, Parfit might have laid the ground for a new anti-realist challenge.

4. THREE POSSIBLE REALIST REJOINDERS

So far, I have only presented a very basic version of the anti-realist argument and made a suggestion as to how to transfer it to the moral realm. To add some more depth to this, I am going to discuss three possible realist rejoinders, which take

50 See Dietrich and List, "What Matters and How It Matters," 451.

51 They mention how a parallel view to scientific instrumentalism in ethics might clash with their reason-based representation (Dietrich and List, "What Matters and How It Matters," 425–26). However, they do not pursue this line of thinking any further.

their inspiration from similar objections in the scientific debate. They bear on P1, P2, and P3, respectively.[52] Even though I will ultimately argue that Parfit is not in a position to make use of any of them, it should also become clear that the situation is much more complicated than I have been able to convey so far. Thinking in terms of underdetermination is apt to raise some intricate new issues.

4.1. Additional Evidence and Theoretical Virtues

P1 invites two kinds of criticism. First, according to the proposed analogy, it is the theories' *deontic* verdicts alone that take the place of empirical evidence. However, it might sensibly be objected, moral theories also generate non-deontic verdicts, such as *axiological* ones. For example, they might also yield verdicts on the goodness and badness of acts, of states of affairs, or of the characters of agents. Yet if one theory were to yield much more plausible verdicts of such a kind than any of its rivals, these might well tip the balance in its favor. The overall argument would thus fail because it relies on too restricted an account of what constitutes the evidence in ethics.

There is one swift reply available to this objection if we focus on the present context only. Parfit thinks that his convergence argument helps realism by settling the deontic quarrels. However, if differences remain about other verdicts, the proposed convergence is much less effective in countering the argument from disagreement. Moreover, if the rest of our observations are correct, the theories are incompatible and we still need to decide among them. Even if there remain axiological judgments to decide the case, we can no longer do so on the basis of deontic ones. There is, in sum, *less* of a basis to distinguish between the theories if the convergence argument proves successful. Settling the deontic conflicts thus makes it *more* difficult to adjudicate between the incompatible theories and thereby more difficult to claim that we can know some truths in ethics. Hence Parfit's convergence argument does not advance the case for realism.

However, if we look beyond Parfit and ask whether options remain for other realists to counter the argument from underdetermination, the objection gains more traction. Perhaps the correct deontic verdicts can indeed be accounted for by incompatible theories, yet that still leaves open the possibility of choosing among them on the basis of (a class of) other verdicts they yield. Since the aim of this paper is not to defend a general version of the underdetermination argument, I will not be able to counter this objection in detail. However, I want to

52 I will not consider objections to P4, which I take to be the least controversial. I am also not claiming that these are the only possible objections, though I do think that they are among the most important ones.

at least hint at two possible routes that the anti-realist's reply can take. One is to insist that deontic verdicts or intuitions enjoy a privileged status. This might be the case because our intuitions about them are more firm, conferring a lesser status to our intuitions about, e.g., axiology. Anti-realists thus needed to defend a moral epistemology that vindicates the primacy of deontic verdicts. Alternatively, it might have something to do with the purpose of moral theories themselves. If moral theories are, in an important sense, *practical*, deontic verdicts will have a privileged status because action-guiding verdicts follow from them, not from axiological ones.

Another route is to agree in principle that other evidence can play a role in theory choice, while maintaining that it is insufficient to actually tip the balance in the present case. The anti-realist would thus have to show that, if there are differences regarding axiological verdicts, none of those speak decisively in favor of one of the competing theories. Maybe the rival theories each get equally important subsets of those other verdicts correct. Bringing in additional evidence does not guarantee that we will be able to choose.

The second kind of criticism takes issue not so much with what we put in place of the empirical evidence, but with the idea that the *empirical evidence alone* can adjudicate among the rival theories. We have seen that the anti-realist tries to exploit the fact that different theories can account for exactly the same evidence, to argue that it is equally reasonable to accept any of them. But, so it has been argued in the scientific case, this is misleading, at least if we consider the evidence to narrowly consist of the empirical data. Many critics have objected to what they see as an impoverished understanding of scientific methodology underlying the anti-realist argument, to the effect that it identifies the deductively deducible consequences of a theory with its evidential support. To put it in a catchphrase, they claim that empirical equivalence does not entail evidential equivalence.[53] Instead, theories exhibit additional *theoretical virtues* by which we compare them, for example simplicity, predictive fruitfulness, or non-*ad-hoc*-ness. Such theoretical virtues might well be brought in to decide between theories that are empirically equivalent.

Anti-realists have reacted to this objection in basically two ways. One way is to deny that theoretical virtues are relevant at all to the truth of a theory. Such virtues, it is argued, give us a pragmatic criterion for which theory to use, but whether a theory is, e.g., simple does not have anything to do with it being true.[54]

53 Critics of underdetermination that have contributed to driving home this point include Boyd, "Realism, Underdetermination, and a Causal Theory of Evidence"; and Laudan, "Demystifying Underdetermination."

54 Van Fraassen has very famously defended this line of argument (*The Scientific Image*).

The second way for an anti-realist to react is by pointing out difficulties for the project of deciding between theories on the basis of such additional criteria. Tulodziecki argues that we might face very difficult problems of weighing different virtues when they are exhibited by rival theories, even problems of incommensurability.[55] The debate about these issues remains open. Yet, at least in principle, it is easy to see how such considerations could be brought into play in the moral case as well. So far I have relatively uncritically assumed that deontic (and possibly axiological) equivalence suffices to get the anti-realist argument flying. However, moral theories exhibit theoretical virtues as well, and those might tip the balance in favor of one of the rival traditions.[56] I am not going to take sides in this dispute. For what it is worth, I do think that, no matter whether theoretical virtues are indeed relevant to the question of truth, further investigation into them would be of value in ethics. They have not received as much attention yet as they deserve. In ethics, we generally take for granted that the different theoretical traditions arrive at very different conclusions about deontic verdicts. This renders it unnecessary to search for further grounds on which to decide among them. One especially interesting question that Parfit's convergence argument opens up is whether theoretical virtues might not break the tie among the rival traditions.[57]

However, if we narrow down the question again to whether there is a strategy that suits Parfit specifically, I am fairly sure that this is not it. The reason for that is that it would jeopardize what could be called his *reconciliatory project*. Parfit is explicit that rather than proposing a new moral theory, he wants to learn from existing ones.[58] His is not a reductionist project but one of reconciliation among the major moral traditions. Hence, all three mountaineers reach the summit. But by referring to additional theoretical virtues we would, if successful, decide which of the deontically equivalent theories is the correct one after all.[59] The whole point of this strategy is to reach a decision *for one* of the rival theories. If theoretical verdicts could decide the case, we would no longer have reason to believe in the three traditions of moral theories, but only in one of them. The price that the discussed strategy takes for saving Parfit's project from the argument

55 Tulodziecki, "Epistemic Equivalence and Epistemic Incapacitation."

56 Hooker offers an influential line of reasoning for rule consequentialism along these lines (*Ideal Code, Real World*).

57 Carrier suggests that underdetermination in science fulfills a similar function, in that it can serve as a test-tube to lay open the nonempirical virtues that play a role in theory choice ("Underdetermination as an Epistemological Test Tube").

58 Compare Parfit, *On What Matters*, 1:174.

59 The same point holds for bringing in additional axiological evidence.

from underdetermination would thus be to give up on reconciliation. Since I take the reconciliatory project to be very dear to Parfit's heart, I do not think that he would want to opt for this strategy.

This points to a conflict at the heart of Parfit's project. Parfit wants to have his cake and eat it, too. He wants to get rid of the disagreements that threaten realism, but he does not want to make up his mind about which moral theory is the correct one. Thus he undertakes much effort to clear away any deontic disagreements, making it more difficult in the process to choose among the theories. After the convergence argument has gone through, it is harder (if possible at all) to find the correct theory for lack of a deontic basis on which to decide. However, this does not seem to bother Parfit, which is baffling. Indeed, if we take heed of the lessons from the philosophy of science, it looks like we end up with a picture that is more congenial to anti-realist views. Faced with mutually exclusive theories that are extensionally equivalent, the typical realist reaction is to look for other criteria to decide the case. It is anti-realists who have pointed out problems for this proposal and who are generally much more comfortable with different, noncompatible explanatory frameworks. Thus, Parfit's whole project of reacting to disagreements by showing that different theories can account equally well for them has a distinctly anti-realist ring to it.

4.2. What We Know and What Is True

Except, what if we do not need to bother about being able to find the correct theory? The second objection goes right at the epistemic component we identified as one of the possible additions to the thin version of realism. P2 draws a direct connection between the reasonability of some of our beliefs about a subject matter and our attribution of truth to that same subject matter. This, a realist might reply, seems an overly hasty inference. There is a perfectly acceptable alternative. We could simply claim that we have not, or will never have, the evidence to adjudicate among some theories, but that we are nevertheless justified in believing that only one of them is true. That is, we could hold on to the conviction that there are facts, and correspondingly possible reasons, inaccessible to us, that decide which theory is true.[60] Would Parfit want to opt for a similar strategy in the moral case? That is, could he hold that, in the case of underdetermination, we are ignorant about which of the theories is correct, while at the same time insisting that one of them is?

60 There is much literature on this in the philosophy of science. A highly illuminating discussion is between Newton-Smith and Lukes ("The Underdetermination of Theory by Data") and Bergström ("Underdetermination and Realism") on the so-called *ignorance response* to underdetermination.

I have already pointed out that the fact that Parfit takes disagreements as seriously as he does is a strong indication that he subscribes to a strong epistemic understanding of realism, which suggests that he would not want to argue along these lines.[61] But we need not turn to such indirect clues. As it turns out, such an answer is not compatible with Parfit's explicitly stated views on moral epistemology. Witness the following passage:

> If we had strong reason to believe that, even in ideal conditions, we and others would have deeply conflicting normative beliefs, it would be hard to defend the view that we have the intuitive ability to recognize some normative truths. We would have to believe that, when we disagree with others, it is only we who can recognize such truths. But if many other people, even in ideal conditions, could not recognize such truths, we could not rationally believe that we have this ability. How could *we* be so special? And if none of us could recognize such normative truths, we could not rationally believe that there *are* any such truths.[62]

As these remarks show, it would not be in the spirit of Parfit's moral epistemology to opt for a solution along these lines. Importantly, it is not Parfit's commitment to the claim that we have intuitive abilities to recognize some normative truths that renders his position vulnerable to the skeptical attack. Intuitionists can consistently hold on to the idea that we have an intuitive ability to recognize truths while admitting that this does not guarantee that we will be able to agree about them. Rather, it is Parfit's conviction that if it turns out that we do in fact disagree substantially about some of our deep moral beliefs, we are no longer

61 Whether this is itself a reasonable conviction is an open question. Many commentators have stressed that Parfit's fear of disagreements, which seems to underlie those epistemological claims, might be overstated. Several have also pointed out that there is a discrepancy between Parfit's heightened uneasiness with disagreements about normative concerns and his more confident reaction to metaethical disagreements. See Larmore, "Morals and Metaphysics"; Darwall, "Agreement Matters"; and Smith, review of Derek Parfit, *On What Matters*, vol. 2.

62 Parfit, *On What Matters*, 2:546. The passage might seem out of place, since Parfit is here talking about our ability to recognize normative truths, whereas my objection is premised on him being committed to our ability to recognize explanatory truths (what makes acts right). I will attend to this worry shortly. However, there are other passages where Parfit makes similar suggestions that are not restricted to normative disagreements, such as the following: "Such disagreements give us reasons to doubt that we are the people whose beliefs are true. These disagreements may also give us reasons to doubt that any of the conflicting beliefs are true. Perhaps none of us is right, because our questions have no answers" (2:427–28). In this passage, Parfit is referring to disagreements about "what it would *be* for things to matter, and about whether anything *could* matter" (2:427). Thus, it does not seem to me that my proposed reading is an unfair one.

entitled to trust in these abilities. Since on his view there are no other ways to recognize moral truths, Parfit is convinced that we would then no longer have reason to believe in such truths at all.

Of course, Parfit is not committed to the view that we must already have found out about those truths, since he adds the proviso that we only have to be able to do so *in ideal conditions*. But this does not change the picture. Parfit acknowledges that the different traditions come to the same set of deontic verdicts; indeed, that is the whole point of his convergence argument. Since the deontic consequences of those traditions are thus identical, there could be no way, even in ideal conditions, of distinguishing among them on that basis. Moreover, if my reply to the first objection is correct, Parfit would not want to appeal to any other criteria, either. Yet there are nevertheless deep disagreements between the alternative traditions when it comes to their explanatory content. We thus have strong reasons to belief that we and others would, even in ideal conditions, disagree *at least insofar* as we cannot adjudicate among our theories' incompatible explanatory claims.

4.3. Do the Explanatory Disagreements Matter?

This leads us to the final objection. We have seen that the problem for the realist arises because, although it might be possible to put the disagreements about our *deontic* principles (and with that about all specific cases) to rest, other disagreements persist regarding explanation. P3 states that two theories cannot both be true if such disagreements persist. However, to the realist, maybe this is not what we were looking for in the first place. What if we were simply to claim that we are only concerned about the deontic disagreements and not the explanatory disagreements? That is, what if we were to claim that our position only entails that there are truths about our deontic principles and verdicts, but not about why these principles and verdicts are the correct ones? Parfit might in such a spirit restrict the importance of our coming to find moral truths to the deontic realm. Since there are no remaining differences there, his line of reasoning would then be unaffected by the anti-realist argument.

The textual basis in OWM on this issue is rather thin. At least one passage strongly suggests that Parfit might consider such a solution to the problem. After outlining in great detail how distorting influences are responsible for many of our disagreements about deontic verdicts, he makes the following claims:

> Some other moral disagreements are not about *which* acts are wrong, but about *why* these acts are wrong, or what *makes* them wrong. Different answers are given by different systematic theories, such as those developed

> by Kantians, Contractualists, and Consequentialists. Such disagreements do not directly challenge the view that we are able to recognize some moral truths. In defending this view, it is enough to defend the claim that, in ideal conditions, there would be sufficient agreement about which acts are wrong. Though we also have intuitive beliefs about why many acts are wrong, and about the plausibility of different systematic theories, we would expect there to be more disagreement about these other questions. As I have also argued, however, when the most plausible systematic theories are developed further, as they need to be, these theories cease to conflict. If that is true, these theoretical wars would end.[63]

Parfit does not explain why we would expect there to be more disagreements about explanatory questions. Also, if I am correct, he has not shown that the theories would cease to conflict. Nevertheless, the passage suggests that he does consider some kinds of disagreements, such as the ones about explanation, not to be of equal importance to the realist's case.

I am not sure how much importance to accord to this passage. Yet even if this turned out to be Parfit's preferred solution, we need to ask whether it is a tenable one. Two reasons speak strongly against it. First, the solution seems *ad hoc*. Parfit is of course right that the explanatory disagreements do not logically contradict the position that there are some moral truths. However, at least at first sight, the explanations we ordinarily put forward for why certain acts are right or wrong seem to be meant to be taken at face value. This generalizes to our theories. Typically, our moral theories are not only trying to enumerate and systematize our particular verdicts, but also have an explanatory aspiration. Barring further arguments to the contrary, we should take these explanations at face value, too. At the very least, the burden of proof is on the realist to prove otherwise.

The impression of *ad-hoc*-ness is reinforced by the fact that Parfit does consider some disagreements other than deontic ones to be threatening to realists:

> Disagreements are deepest when we are considering, not the wrongness of particular acts, but the nature of morality and moral reasoning and what is implied by different views about these questions. If we and others hold conflicting views, and we have no reason to believe that *we* are the people who are more likely to be right, that should at least make us doubt our view. It may also give us reasons to doubt that any of us could be right.[64]

63 Parfit, *On What Matters*, 2:554

64 Parfit, *On What Matters*, 1:418–19.

Thus, according to this passage, there are also deep disagreements about other aspects of morality and those seem to give rise to the same worries as the ones about deontic verdicts. The fact that Parfit is talking about the nature of morality and moral reasoning here suggests that he is thinking of metaethical differences, not of the explanative ones that I have in mind. But the general point stands nevertheless. If such disagreements are indeed a threat, what makes our explanatory disagreements so special that they are not? Why should we not be bothered by the fact that we cannot find out which explanation of the deontic status of our acts is correct?[65]

Comparison with the scientific case hints at an additional, deeper reason. There is a corresponding discussion in the philosophy of science. It has been asked whether one can be a realist about the observable and remain agonistic about unobservables. Yet, strikingly, this move is generally taken by *anti*-realists. Thus Van Fraassen, who has famously argued that science does not aim for more than empirically adequate theories, is also very explicit about his own constructivist empiricism being an alternative to scientific realism.[66] What he denies is that the entities our theoretical concepts refer to in order to explain the data have the same entitlement to be taken at face value. Most scientific realists accept this challenge. They seem to think that a realist position that deserves its name cannot restrict its realism to the claims that theories make about what we can all readily observe, but also has to apply to the more theoretical claims. The pressing realist questions are in the end not about the data itself, but about the further claims that we make to account for them.

Does this generalize to ethics? I am not sure. There is no obvious reason why the criteria for realism should be the same over all domains of inquiry. Instead, realism might come in domain-specific variations. Maybe in ethics we can be satisfied with knowing which acts are right or wrong without knowing why this is so. But again, the realist would at least have to give us an explanation for the asymmetry with the scientific case. Why is the glass taken to be half full in ethics, whereas it would in similar cases be considered half empty in science? Barring further arguments, the idea of restricting the aim of our moral theories to deontic adequacy looks suspiciously similar to the one of restricting the aim of

65 Note again that our definition of realism does not commit a realist to hold that there is a fact of the matter to every moral question. However, to exclude a whole class of statements (those about moral explanation) from being truth-apt seems arbitrary. Usually, realists argue that when we are not able to find out the truth about some moral question this is explainable by citing phenomena such as vagueness. It would yet have to be shown why explanatory statements summarily suffer from such an impairment.

66 Compare Van Fraassen, *The Scientific Image*.

our scientific theories to empirical adequacy, which is, after all, an anti-realist suggestion.[67]

5. CONCLUDING REMARKS

Let me finish by taking stock and tying up some loose ends. Rather than challenging Parfit's interpretation of his favorite authors, or the tenability of his convergence argument, I have tried to point out a problem that only arises when his initial argument has gone through. For all I know, Parfit might indeed have shown that the different traditions (or at least some plausible versions of these traditions) agree on what matters, i.e., they might be deontically equivalent. But this, or so I have argued, is not enough to vindicate moral realism. The theories still differ in the explanations they give us for why those acts are right or wrong. I have described this as a case of underdetermination, taking inspiration from the philosophy of science. If that is an accurate description, Parfit's convergence argument might backfire and lay the ground for a similar anti-realist argument for the moral realm as has been conceived for the scientific realm.

Importantly, my remarks do not amount to a general argument against realist positions in ethics. Parfit's convergence argument, on which the underdetermination claim crucially depends, is far from being universally accepted and I have done nothing to defend it further. Even if the convergence argument is granted, and my suspicions about the remaining explanatory disagreements prove to be correct, several potentially successful options remain to block the anti-realist argument. I have tried to argue the case that at least three of the most promising options are not open to Parfit. However, whether the argument could be broadened to pertain also to realist views that do not share Parfit's conciliatory spirit or his moral epistemology remains to be seen.

In addition, we have only discussed underdetermination among three specific theories. Parfit thinks that those are the best versions of the most important traditions and, if he is right, the result would thus be of great importance. Yet we have to expect that other authors will have their own favorites. To provide a more general argument, we would need to show underdetermination to be a more pervasive phenomenon. Philosophers of science often advance more ambitious claims to the effect that *any* given theory is underdetermined. Going

67 Some might think that Parfit's reasons fundamentalism provides a way out: if reasons are the fundamental normative notion, we might well be satisfied with knowing *which* our reasons are. But this merely pushes back the question; we can still ask *why* these are our reasons.

forward, it would be of great interest to learn the prospects for such a pervasive kind of underdetermination in the realm of ethics.[68]

I want to conclude with a more general observation. I think that the issues discussed have their root in the way Parfit conceives of the problem of disagreement. I submit that when doing metaethics, many (if not most) philosophers prefer to discuss their pre-theoretical notions and views on morality, and neglect (or put aside) the results of normative ethical theorizing. If we look at, e.g., the metaethical realism debate, it is noticeable that this debate does not build on the findings, preliminary as they may be, of normative ethical theorizing. Instead, an argument like that from disagreement is mostly posed in terms of disagreements between laymen, or between different cultures. This is in stark contrast to the philosophy of science. In the philosophy of science, it is scientific theories that get most of the attention, and it is to such theories that arguments like the underdetermination argument or the pessimistic metainduction refer. Parfit's way of thinking of the challenge in terms of disagreements among theories reveals a closer connection between these two domains. This allows us to adapt some questions to the moral realm that are being discussed in the scientific literature. And it might ultimately suggest that the whole project of reconciling rival theories in order to vindicate moral realism is not the most promising idea after all.[69]

University of Bern
marius.baumann@philo.unibe.ch

68 A promising starting point for such investigations could be the project referred to as the *consequentializing* of moral theories. Proponents of that project try to establish the claim that there is a deontically equivalent consequentialist alternative to any (minimally plausible) non-consequentialist theory. See, for example, Dreier, "Structures of Normative Theories," and more recently Portmore, *Commonsense Consequentialism*. If this were to be successful, and there did remain explanatory differences between those theories, it would suggest a more pervasive form of underdetermination.

69 I would like to thank: Claus Beisbart for invaluable help and encouragement; members of the Reflective Equilibrium Group (especially Georg Brun and Tanja Rechnitzer) for several rounds of constructive criticism; Monika Betzler, Jamie Dreier, Gerhard Ernst, Brad Hooker, Christian List, Philip Stratton-Lake, Jussi Suikkanen, Folke Tersman, and Silvan Wittwer for insightful suggestions; and, finally, audiences at Edinburgh, Erlangen, and Utrecht for stimulating discussions, as well as the Swiss National Science Foundation for generous funding.

REFERENCES

Alexander, Larry, and Michael Moore. "Deontological Ethics." *Stanford Encyclopedia of Philosophy* (Spring 2015). https://plato.stanford.edu/entries/ethics-deontological.

Audi, Robert. "Intuition, Inference, and Rational Disagreement in Ethics." *Ethical Theory and Moral Practice* 11, no. 5 (November 2008): 475–92.

Bergström, Lars. "Underdetermination and Realism." *Erkenntnis* 21, no. 3 (November 1984): 349–65.

Berker, Selim. "The Unity of Grounding." *Mind* (forthcoming).

Bonk, Thomas. *Underdetermination: An Essay on Evidence and the Limits of Natural Knowledge*. Dordrecht: Springer, 2008.

Bortolotti, Lisa. *An Introduction to the Philosophy of Science*. Malden, MA: Polity Press, 2008.

Boyd, Richard N. "Realism, Underdetermination, and a Causal Theory of Evidence." *Noûs* 7, no. 1 (March 1973): 1–12.

Carrier, Martin. "Underdetermination as an Epistemological Test Tube: Expounding Hidden Values of the Scientific Community." *Synthese* 180, no. 2 (May 2011): 189–204.

Darwall, Stephen. "Agreement Matters: Critical Notice of Derek Parfit, *On What Matters*." *Philosophical Review* 123, no. 1 (January 2014): 79–105.

Deigh, John. *An Introduction to Ethics*. Cambridge: Cambridge University Press, 2010.

Dietrich, Franz, and Christian List. "What Matters and How It Matters: A Choice-Theoretic Representation of Moral Theories." *The Philosophical Review* 126, no. 4 (October 2017): 421–79.

Dreier, James. "Metaethics and the Problem of Creeping Minimalism." *Philosophical Perspectives* 18, no. 1 (December 2004): 23–44.

———. "Structures of Normative Theories." *The Monist* 76, no. 1 (January 1993): 22–40.

Driver, Julia. *Ethics: The Fundamentals*. Chichester, UK: Wiley-Blackwell, 2007.

Duhem, Pierre. *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press, 1954.

Dworkin, Ronald. *Justice for Hedgehogs*. Cambridge, MA: Belknap Press of Harvard University Press, 2011.

Harman, Gilbert. *The Nature of Morality: An Introduction to Ethics*. Oxford: Oxford University Press, 1977.

Herman, Barbara. "A Mismatch of Methods." In Parfit, *On What Matters*, 2:83–115.

Hooker, Brad. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Oxford University Press, 2000.

Kitcher, Philip. "Real Realism: The Galilean Strategy." *Philosophical Review* 110, no. 2 (April 2001): 151–97.

Ladyman, James. *Understanding Philosophy of Science*. Abingdon, UK: Routledge, 2002.

Larmore, Charles. "Morals and Metaphysics." *European Journal of Philosophy* 21, no. 4 (December 2013): 665–75.

Laudan, Larry. "Demystifying Underdetermination." In *Scientific Theories*, vol. 14, edited by C. Wade Savage, 267–97. Minneapolis: University of Minnesota Press, 1990.

Laudan, Larry, and Jarrett Leplin. "Empirical Equivalence and Underdetermination." *Journal of Philosophy* 88, no. 9 (January 1991): 449–72.

Mackie, J. L. *Ethics: Inventing Right and Wrong*. London: Penguin Books, 1977.

Maxwell, Grover. "The Ontological Status of Theoretical Entities." In *Scientific Explanation, Space, and Time: Minnesota Studies in the Philosophy of Science*, vol. 3, edited by Herbert Feigl and Grover Maxwell, 181–92. Minneapolis: University of Minnesota Press, 1962.

Nebel, Jacob. "A Counterexample to Parfit's Rule Consequentialism." *Journal of Ethics and Social Philosophy* 6, no. 2 (July 2012): 1–10.

Newton-Smith, W., and Steven Lukes. "The Underdetermination of Theory by Data." *Proceedings of the Aristotelian Society* 52 (1978): 71–91.

Otsuka, Michael. "The Kantian Argument for Consequentialism." *Ratio* 22, no. 1 (March 2009): 41–58.

Parfit, Derek. *On What Matters*. 2 vols. Oxford: Oxford University Press, 2011.

Park, Seungbae. "Philosophical Responses to Underdetermination in Science." *Journal for General Philosophy of Science* 40, no. 1 (July 2009): 115–24.

Pietsch, Wolfgang. "Defending Underdetermination or Why the Historical Perspective Makes a Difference." In *EPSA Philosophy of Science: Amsterdam 2009*, edited by Henk W. de Regt, Stephan Hartmann, and Samir Okasha, 303–13. Dordrecht: Springer, 2012.

Portmore, Douglas W. *Commonsense Consequentialism*. New York: Oxford University Press, 2011.

Quine, W. V. "On Empirically Equivalent Systems of the World." *Erkenntnis* 9, no. 3 (November 1975): 313–28.

———. "On the Reasons for Indeterminacy of Translation." *Journal of Philosophy* 67, no. 6 (March 1970): 178–83.

———. "Three Indeterminacies." In *Perspectives on Quine*, edited by Robert B. Barrett and Roger F. Gibson, 1–16. Chichester, UK: Wiley-Blackwell, 1990.

———. "Two Dogmas of Empiricism." *Philosophical Review* 60, no. 1 (January 1951): 20–43.

Ridge, Michael. "Climb Every Mountain?" *Ratio* 22, no. 1 (March 2009): 59–77.

Ross, Jacob. "Should Kantians Be Consequentialists?" *Ratio* 22, no. 1 (March 2009): 126–35.

Ross, W. D. *The Right and the Good*. Oxford: Clarendon Press, 2002.

Sayre-McCord, Geoffrey. "The Many Moral Realisms." *Southern Journal of Philosophy* 24, no. S1 (Spring 1986): 1–22.

———. "Moral Realism." *Stanford Encyclopedia of Philosophy* (Spring 2015). https://plato.stanford.edu/entries/moral-realism.

Scanlon, T. M. *Being Realistic about Reasons*. Oxford: Oxford University Press, 2014.

———. "How I Am Not a Kantian." In Parfit, *On What Matters*, 2:116–39.

Setiya, Kieran. Review of *On What Matters*, by Derek Parfit. *Mind* 120, no. 480 (October 2011): 1281–88.

Smith, William. Review of *On What Matters*, vol. 2, by Derek Parfit. *Journal of Moral Philosophy* 11, no. 2 (2014): 241–44.

Stanford, Kyle. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press, 2006.

———. "Underdetermination of Scientific Theory." *Stanford Encyclopedia of Philosophy* (Spring 2016). https://plato.stanford.edu/entries/scientific-underdetermination.

Suikkanen, Jussi. *This Is Ethics: An Introduction*. Chichester, UK: Wiley-Blackwell, 2014.

Tännsjö, Torbjörn. *Understanding Ethics: An Introduction to Moral Theory*. Edinburgh: Edinburgh University Press, 2002.

Tersman, Folke. "From Scepticism to Anti-Realism." Unpublished manuscript, 2018.

———. *Moral Disagreement*. Cambridge: Cambridge University Press, 2006.

Tulodziecki, Dana. "Epistemic Equivalence and Epistemic Incapacitation." *British Journal for the Philosophy of Science* 63, no. 2 (June 2012): 313–28.

Van Fraassen, Bas C. *The Scientific Image*. Oxford: Oxford University Press, 1980.

Wolf, Susan. "Hiking the Range." In Parfit, *On What Matters*, 2:33–57.

Worrall, John. "Underdetermination, Realism and Empirical Equivalence." *Synthese* 180, no. 2 (2011): 157–72.

Journal of Ethics and Social Philosophy
VOL. 13, NO. 3 · JULY 2018
https://doi.org/10.26556/jesp.v13i3.249

HELPING THE REBELS

Massimo Renzo

1. REVOLUTION AND INTERVENTION

IT IS A STRIKING FEATURE of revolutionary wars that they often fail to meet the requirement of having a reasonable chance of success—despite otherwise meeting the traditional *jus ad bellum* principles—unless other states or international institutions militarily intervene to support the insurgents. Thus, the very permissibility of waging such wars, which are necessary to end some of the most tyrannical regimes we are familiar with, often depends on military intervention being permissible. In this respect, the permissibility of intervention becomes a precondition of the permissibility of rebellion against tyranny. The latter might not be permissible if the former is not.

On the other hand, intervention tends to significantly increase the length of revolutions and civil wars.[1] This is partly for the obvious reason that when intervention takes place weapons take longer to run out, and partly because intervening parties tend to feel the costs of these wars (both economic and human) less than locals, and thus have less of an incentive to end hostilities. Indeed, their interest is often to escalate the conflict when the side they support faces defeat (as in the Syrian case).

As these quick remarks illustrate, identifying the conditions for the permissibility of military intervention in support of attempts to rebel against authoritarian regimes has a central role in any account of revolution. And yet the question has received surprisingly little attention in the most recent debate on just war. The problem was addressed in the late 1970s by Michael Walzer and a group of philosophers who engaged with the arguments of his seminal *Just and Unjust Wars*.[2] But it has rarely been addressed since, despite the fact that its urgency

1 Regan, "Third-Party Interventions and the Duration of Intrastate Conflicts," 55–73.

2 Walzer, *Just and Unjust Wars* and "The Moral Standing of States"; Doppelt, "Walzer's Theory of Morality in International Relations"; Wasserstrom, review of *Just and Unjust Wars*"; Luban, "Just War and Human Rights"; Beitz, *Political Theory and International Relations*; Tesón, *Humanitarian Intervention*.

has been highlighted, once again, by the wave of revolutions that swept the Arab world beginning in 2010, most notably in Libya.

One important exception to this glaring lacuna in the current philosophical debate is the work of Allen Buchanan, who has addressed this question in a pair of recent papers that together outline an ambitious account of the ethics of revolution and its implications for the ethics of intervention.[3] Buchanan's account is bold and yet sophisticated. It is bold in that it advances a number of theses that will no doubt strike the reader as highly controversial; it is sophisticated in that it rests on a nuanced account of the dynamics that characterize the rise and development of revolutions and, more importantly, of the constraints that the right to political self-determination places on intervention. The notion of political self-determination also plays a crucial role in Walzer's account of the relationship between the permissibility of rebelling and the permissibility of military intervention, but while his critics have invariably criticized Walzer's account, not much has been done by philosophers working on revolution and intervention to replace it with a more plausible one.[4] One of the merits of Buchanan's account is that it takes on this important task.

Buchanan argues that, despite the importance of political self-determination, military humanitarian intervention may be permissible, at least in some cases, without the consent of the rebelling population that the intervention intends to benefit. Indeed, given certain structural features of the way revolutions typically unfold, there are often reasons to disregard the consent of the population oppressed and intervene *before* the revolution starts.[5] More controversially, he argues that the aims of the intervention need not be limited to overthrowing the unjust regime. Military force may also be permissibly employed to nullify the democratic constitutional choice of the newly liberated population and impose a particular form of democratic government, if doing so is necessary to guarantee the conditions for the future exercise of the right of self-determination

3 Buchanan, "The Ethics of Revolution and Its Implications for the Ethics of Intervention," 291–323, and "Self-Determination, Revolution, and Intervention." (Some of the themes explored in the first article were anticipated in his "Revolutionary Motivation and Rationality," which focuses on Marx's theory of revolution). Two other recent contributions are Finlay, "Reform Intervention and Democratic Revolution," and Dobos, *Insurrection and Intervention*.

4 An exception is Charles Beitz, who discusses political self-determination at length both in his *Political Theory and International Relations* and in "The Moral Standing of States Revisited." However, Beitz's discussion does not focus specifically on revolutions, as Buchanan's does.

5 Buchanan, "The Ethics of Revolution and Its Implications for the Ethics of Intervention."

(and if the population will be able, through constitutional means, to replace the imposed democratic government with a different one).[6]

In this paper, I further elaborate Buchanan's account of political self-determination and argue that once correctly understood, the sort of picture of political self-determination he operates with (which seems to me roughly correct) places tighter constraints on intervention than he allows. Thus, his bold conclusions should be resisted.

2. BUCHANAN'S ACCOUNT

Typically, those who have addressed the question of the permissibility of humanitarian intervention have framed it in terms of a tension between the moral demand to protect human rights and the moral demand to respect political self-determination. In the same way in which individuals have a right against others interfering with their own agency against their will in order to protect them from harm or to make them better off, political communities are said to have a right against others interfering with their own agency against their will in order to protect them from harm or to make them better off. So understood, the objection to humanitarian intervention ultimately has an anti-paternalistic foundation.

Michael Walzer famously defends this view. He argues that humanitarian intervention is permissible only in cases of supreme humanitarian emergency, such as massacre, enslavement, or mass deportation. Any intervention to bring down tyrannical regimes that do not engage in this sort of widespread or systematic violation of human rights would be impermissible since it would constitute an unjustified form of interference with the right to self-determination of the community in question. In this case, revolution would be permissible. The members of the community are "as free not to fight as they are free to rebel. But that freedom does not easily transfer to foreign states or armies and become a right of invasion or intervention; above all, it does not transfer at the initiative of the foreigners."[7]

In response, Buchanan distinguishes between a negative and a positive component of political self-determination: the latter refers to the right of a political

6 Buchanan, "Self-Determination, Revolution, and Intervention."

7 Walzer, "The Moral Standing of States," 223; see also Walzer, *Just and Unjust Wars*, 89–91. Walzer mentions two further "rules of disregard." Intervention is permissible (a) when a particular state includes more than one political community and some of them are trying to secede, or (b) when another state has already intervened in a civil war to support one of the factions and the effects of this earlier intervention need to be neutralized. These two exceptions are less important for the purposes of my discussion, so I will bracket them here.

community to govern itself through the exercise of its own autonomous agency; the former refers to the right of the political community not to be subject to external interference.[8] The problem with Walzer's view, Buchanan argues, is that it assigns paramount importance to negative self-determination, ignoring the fact that its value ultimately resides in protecting positive self-determination.[9] But it is a mistake to think that all it takes to preserve the self-determination of a political community is ensuring that the community is not interfered with by others, for not being determined by others is not equivalent to being self-determining.[10] There will be cases in which a political community enjoys negative self-determination, in that it is not interfered with, and yet it fails to exercise positive self-determination because it lacks what it takes to govern itself through the exercise of its autonomous agency.

I am not sure it is correct to say that Walzer ignores the value of positive self-determination, since his argument for noninterference (i.e., negative self-determination) is precisely that the process of positive self-determination "has value even if it is not always pretty, and even if its outcome does not conform to philosophical standards of political and social justice."[11] The problem with Walzer's view is not that it focuses on negative self-determination, ignoring positive self-determination. The problem is that Walzer operates with an implausible conception of positive self-determination, according to which the internal balance of power generated within a political community around certain institutions, no matter how authoritarian, constitutes a genuine expression of the will of the community.[12]

8 On this distinction, see also Cassese, *Self-Determination of Peoples*, 5–12; Beitz, *Political Theory and International Relations*, 92–93; Patten, "Self-Determination for National Minorities," 120–44 (though these authors prefer the label "internal/external self-determination").

9 Buchanan argues that it is doubtful that negative self-determination has any value other than that of protecting positive self-determination ("Self-Determination, Revolution, and Intervention," 452), but this seems too strong. To see this point, consider for a moment the value of self-determination as it applies to personal as opposed to collective agency. Suppose I enjoy external self-determination (freedom from being subject to external interference) but not internal self-determination (the capacity to govern myself in light of values and goals I have autonomously chosen), whereas you enjoy neither. There seems to be a sense in which my condition is preferable to yours. True, my life lacks self-direction, as does yours. But at least I am not someone else's puppet. Neither of us is the author of his or her own life, but your condition seems worse than mine, because in addition to being unable to form and pursue your own goals, you are being used to serve someone else's goals. I do not suffer this further wrong. The same point applies to collective self-determination.

10 Buchanan, "Self-Determination, Revolution, and Intervention," 451.

11 Walzer, "The Moral Standing of States," 232.

12 Walzer, *Just and Unjust Wars*, 87–91, and "The Moral Standing of States," 230–34. Buchanan

Indeed, the reason why Walzer allows intervention in cases of massacre, enslavement, or mass deportation is that only in these cases, within his own view, can the conditions for positive self-determination be said to break down. For "when a government turns savagely upon its own people, we must doubt the very existence of a political community to which the idea of self-determination might apply." [13] As we will see, this is precisely the move that Buchanan makes in order to conclude that respect for political self-determination does not always require refraining from humanitarian intervention that has not been consented to: when the conditions for self-determination are not in place in a given political community, intervention cannot be impermissible on the grounds that it would violate the community's self-determination.[14] The problem with Walzer's view is that it offers an implausible view of the conditions under which self-determination is not in place, because it rests on an implausible conception of what positive self-determination consists in.[15]

Buchanan does not provide a fully developed account of political self-determination, but the model he operates with, according to which political self-determination requires some sort of "group agency," is much more plausible than Walzer's.[16] In his words, "group agency requires more than that certain political outcomes be the result of activities of members of the group: they must be the result of the exercise of agency by the group, which in turn requires that the group be organized in such a way that it can be said that the group can decide and act. In other words, self-determination, where this means determination of political outcomes by the group—as distinct from those outcomes being caused

convincingly rejects this view ("The Ethics of Revolution and Its Implications for the Ethics of Intervention," 316), drawing on some of the arguments offered by Walzer's critics, as mentioned in note 1 above. A helpful discussion of this objection can also be found in Finlay, "Reform Intervention and Democratic Revolution."

13 Walzer, *Just and Unjust Wars*, 101.

14 Charles Beitz offers a similar argument in "The Moral Standing of States Revisited," 341.

15 The other two "rules of disregard" introduced by Walzer also support the conclusion that for him negative self-determination is ultimately valuable insofar as it protects positive self-determination. The reasons why intervention is permissible to help a community that is trying to secede from a multinational state is that there is no "fit between the government and the community," and thus the former cannot be said to constitute an expression of the will of the latter. The reasons why intervention to defend a faction in a civil war is permissible when the enemy faction is already receiving some outside help is that the second intervention counterbalances the effects of the first, preventing it from unduly affecting the internal balance of forces, which for Walzer constitutes a genuine expression of political self-determination.

16 Henceforth, I will use "political self-determination" to denote what Buchanan calls "positive self-determination."

by aggregated actions of individuals—requires that the group be an agent, not just that the individual members are agents. It must make sense to say that the group acts, and this requires a degree of organization—a structure or process that coordinates the actions of the individual members in such a way as to justify the claim that there is a collective agent."[17]

But if political self-determination requires group agency, and if group agency requires that the individual members of the political community coordinate their action in certain ways—say, by voting in free elections, supporting certain political leaders, or engaging in some other form of collective deliberation—then we should conclude that, when a country is run by an authoritarian regime in which only a minority has the power to determine how the political community will act, then the community in question is not really self-determining.[18] For in this case, how the community acts, far from being the result of the exercise of agency by the whole group, is determined by what the minority in power wants. Thus, political self-determination is undermined not only in cases of massacre, enslavement, or mass deportation, but also whenever authoritarian regimes perpetrate violations of human rights that, while not widespread or systematic, are sufficiently serious to prevent the sort of group agency that Buchanan is talking about. (Members of political communities are unable to take part in processes of collective deliberation when their basic human rights, such as the right to life or the right not to be tortured, are constantly threatened and when they lack the capacity to engage in minimal forms of political participation.)

Suppose now that, while serious, the violations in question do not prevent the political community from being able to exercise its group agency. Even so, Buchanan argues, the permissibility of intervention is not conditional on consent to it having been secured from the oppressed population. This is for two reasons. First, given that tyrannical regimes typically curtail important freedoms—such as freedom of speech, association, and political participation—serious epistemic obstacles will afflict any attempt to ascertain that consent has been given under these circumstances. (The regime can hardly be expected to organize a referendum to enable the population to deliberate whether to accept help in overthrowing it.)[19] Second, even when consent is somehow given (or

17 Buchanan, "Self-Determination, Revolution, and Intervention," 450–51.

18 I do not take a position here on the specific kind of group agency that political self-determination requires. For the purposes of this paper, I simply rely on the general model based on the notion of "group agency" that Buchanan operates with. Two recent accounts of political self-determination compatible with Buchanan's model can be found in Stilz, "The Value of Self-Determination"; Moore, *A Political Theory of Territory*.

19 Buchanan, "The Ethics of Revolution and Its Implications for the Ethics of Intervention," 317–18.

refused), there is reason to suspect that it is the product of coercion or manipulation by the "aspiring revolutionary leaderships" (ARL) that started the revolution, rather than a genuine expression of the will of the population to receive (or refuse) help via military intervention.[20] This is because, as Buchanan's illuminating discussion shows, coercion and manipulation are typically the most effective ways (sometimes the only ways) in which the ARL can mobilize the masses, overcoming the formidable coordination problems that beset any attempt to start a revolution.[21]

In light of these problems, Buchanan's conclusion is that often the best way to respect the autonomy of a population subject to a tyrannical regime is, somewhat counterintuively, to intervene early, without its consent, before the ARL has a chance to take control of the revolution and coerce or manipulate the rest of the population into consenting (or refusing to do so) according to the ARL's own preferences. This intervention would not be subject to the charge of unjustified paternalism, Buchanan argues, because its main aim would not be to stop the human rights violations or bring down the unjust regime, but rather to establish the conditions under which valid consent to the intervention could be formulated and communicated by the population. For example, the intervener could "impose a ceasefire, physically separate the two sides, and then investigate the attitudes of the population toward the revolutionary struggle under conditions in which they can be freely expressed. . . . In intervening for this reason, it would not . . . be intervening to support the revolution, but rather to help create conditions under which it could determine whether to support the revolution."[22]

This is a powerful battery of arguments. In the rest of the paper I consider them in turn.

20 Buchanan, "The Ethics of Revolution and Its Implications for the Ethics of Intervention," 318. For the same reasons, Buchanan argues, the permissibility of intervention is not conditional on the revolution being supported by widespread popular participation. Like consent, participation can be the product of manipulation or coercion. On the other end, lack of participation might be the product of the significant costs associated with raising against the regime, rather than reflecting genuine aversion to the revolutionary cause (Buchanan, "The Ethics of Revolution and Its Implications for the Ethics of Intervention," 315–17).

21 Buchanan, "The Ethics of Revolution and Its Implications for the Ethics of Intervention," 309–14. See also Buchanan, "Revolutionary Motivation and Rationality."

22 Buchanan, "The Ethics of Revolution and Its Implications for the Ethics of Intervention," 321.

3. How Political Self-Determination Constrains the Permissibility of Intervention

Buchanan is certainly right that, insofar as political communities run by authoritarian regimes lack the capacity to exercise the sort of group agency required by political self-determination, any intervention aimed at restoring that capacity cannot be said to be interfering with an exercise of their political self-determination. In these cases, humanitarian intervention is not conditional on the community in question having consented to it for the simple reason that to the extent that it lacks the capacity for group agency, the community can neither give nor withhold consent. However, Buchanan makes a further claim—namely that in these cases intervention "is not a case of lack of proper regard for self-determination, and no violation of the right of self-determination has occurred."[23] This further claim is, I contend, too strong.

Suppose that the democratic government of country *Y* is replaced at some point by a regime so authoritarian that *Y*'s political community can no longer be said to be able to exercise the sort of group agency required by political self-determination. There are nonetheless some constraints on what may be permissibly done to *Y*, based on the fact that *Y* retains a right to self-determination. *Contra* Buchanan, this right can be violated despite the fact that *Y* cannot currently exercise it.

To see this point, consider what we might call "personal self-determination," i.e., the capacity that individuals possess to deliberate so that their actions can be said to be an expression of their autonomous agency. Suppose that I deprive you of the capacity to exercise your personal self-determination—for example, I drug you or hypnotize you, so that you cannot form and execute the intentions required to act and shape your life as you wish. It is certainly true that if a bystander intervenes to rescue you, she would not be interfering with an exercise of your self-determining agency. To the extent that you are under the effect of the drug or the hypnosis, you cannot formulate and act upon the intentions required for such agency to be in place. But is it true that there are no demands that your right to personal self-determination places on a bystander who could help you?

Suppose that the only way she could stop me is to kill me, but the bystander knows that you would not want me to die. You would prefer to suffer the terrible fate I have imposed on you rather than being the reason why I am killed (say because you are a committed pacifist or because you know I am about to find a cure for a disease that afflicts someone you love); or perhaps you would want

23 Buchanan, "Self-Determination, Revolution, and Intervention," 461; see also 455–56.

to be rescued, but not by the bystander (say, because the bystander would do so in a way you find immoral or because she has severely wronged you in the past). We can imagine cases in which, if you were to be rescued by the bystander, the life that you would be living thereafter would be less valuable to you, less close to the plan of life you had been autonomously pursuing up until the moment of my attack, than the one you would be living if you were to be rescued by the bystander. This provides the bystander with some reasons, though not necessarily conclusive reasons, not to kill me. And these reasons are ultimately grounded in your right to decide how to shape your life and what should happen to you.

The fact that you are momentarily incapable of exercising your self-determining agency does not undermine your right to decide what happens to you. That right persists in virtue of the fact that, although currently unable to exercise your capacity to act as a self-determining agent, you retain that capacity. You are still an autonomous agent, despite the fact that your capacity has been momentarily impaired.[24]

What is worth stressing here is that respecting the way in which you currently exercise your personal agency by consenting is not the only way in which we can respect your right to self-determination, though it is typically the best way, when available. There are other ways in which we can do that. We can respect your right to self-determination by acting in a way that conforms to

a. how you have previously exercised your self-determining agency (suppose in the past you wrote a detailed account of how you would like others to act, should your self-determining agency be disabled because you are in a coma or drugged), or
b. how we have reason to believe you would want to exercise your self-determining agency in light of sufficiently reliable evidence available to us.[25]

The same is true in the case of political self-determination. There might be cases of intervention that would violate *Y*'s right to self-determination, despite the fact that the intervention in question would not interfere with any current exercises

24 Or, perhaps more precisely, in virtue of the fact that you are sufficiently connected (in terms of whichever properties ground personal identity) to the entity that had that capacity before my attack and to the one that will gain that capacity again after your rescue. Insofar as that identity persists, the right also persists. For classic discussions of the properties that explain the persistence of your identity in this sort of case, see Parfit, *Reasons and Persons*; McMahan, *The Ethics of Killing*.

25 I believe something like this view of personal self-determination is ultimately what underlies Parfit's discussion of the different forms of consent (actual consent, past consent, and retroactive endorsement) in *On What Matters*, vol. 1.

of *Y*'s self-determining agency. This is because, although *Y* is currently unable to act as a self-determining agent, the intervention might be incompatible with

a. previous exercises of *Y*'s self-determining agency that are still binding on intervening parties, or
b. what we can reasonably expect *Y* to want in light of the goals and preferences *Y* has autonomously set for itself in the past. Those goals retain their normative force as an expression of *Y*'s self-determining agency, even if *Y* currently lacks the capacity to pursue them.

For example, *Y* might have previously signed a treaty by which it consented to receive help from certain parties but not others (say, former allies but not former enemies), or in certain forms but not others (say, through the institution of no-fly zones, but not through air raids), should military intervention on *Y*'s territory be necessary.[26] When this is the case, respect for *Y*'s self-determination counts as a reason against intervention by any of the parties *Y* refused to be helped by, or by any interveners that would employ methods that *Y* has previously objected to. Similarly, if military intervention, or military intervention of a certain kind, would be at odds with some of *Y*'s autonomously chosen goals (perhaps *Y* is a community of committed pacifists, or perhaps the members of *Y* aspire to realize the Millian/Walzerian ideal that a political community should earn its own freedom by fighting, rather than having its freedom handed to it by someone else), respect for *Y*'s self-determination would count as a reason, though not necessarily a decisive reason, against intervention.

The problem with Buchanan's account is that it focuses on respect for actual consent, given at the time of intervention, as the only way to discharge the duty to respect *Y*'s political self-determination. This account, however, is too narrow because, as we have seen, we can also respect *Y*'s political self-determination by respecting

a. its *past consent*, i.e., by treating *Y* in the way *Y* previously asked to be treated (by giving or refusing to give actual consent), should the current conditions materialize, and
b. its *presumed consent*, i.e., by treating *Y* as we can reasonably expect *Y* to want to be treated in light of its values and preferences.[27]

26 For example, with the Treaty of Guarantee, signed in 1960, Cyprus authorized Greece, Turkey, and the UK to intervene in its territory, should that become necessary to restore the status quo established by the treaty.

27 Interestingly, Buchanan elsewhere considers the possibility of resorting to past consent. See Buchanan and Keohane, "Precommitment Regimes for Intervention." On past consent, see

Thus, while Buchanan is right that humanitarian intervention cannot be conditional on *Y* consenting to it at the time of intervention, if *Y* is unable to formulate or communicate consent at that time, the permissibility of intervention is nonetheless conditional on it being compatible with *Y*'s right to self-determination understood more broadly along the lines I have suggested. Even if *Y* lacks the capacity to exercise its group agency at the time of the intervention, or to communicate its decision to consent, its right to self-determination can be violated if the intervention goes against previous decisions autonomously made by *Y* or against what we can reasonably presume *Y* to will in light of previous exercises of its political self-determination.[28]

Stressing these further dimensions of political self-determination is important, not only because it provides a more nuanced account of the constraints that this notion places on intervention, but also because it enables us to address the two worries raised by Buchanan in relation to the reliability of actual consent, given at the time of intervention, as an epistemic proxy for what *Y*'s population truly wants. Relying on past consent or presumed consent is the best way to ensure that intervention respects the autonomous preferences of *Y*'s population when its actual consent cannot be secured at the time of the intervention, either because the regime prevents any reliable way to express it or because the ARL's efforts suggest that the validity of *Y*'s consent might be invalidated by coercion or manipulation.

Finally, focusing on the conceptual resources provided by the richer notion of political self-determination I have outlined enables us to revisit one of the most important insights of Buchanan's analysis—namely his conclusion that there might be circumstances in which the best way to respect *Y*'s political self-determination is to intervene early, without its consent, in order to establish the conditions under which *Y* can formulate and communicate valid consent without being subject to coercion or manipulation by the ARL. We can now see more clearly why this view, while tempting at first, fails to take seriously Walzer's claim that what is objectionable about intervention is that it removes from *Y*'s control the decision about whether to rise in arms.[29]

True, the sort of intervention that Buchanan invokes is different from the one that Walzer discusses, insofar as it does not aim to take down the regime or even

---

Parfit, *On What Matters*, 1:195. On presumed consent see Fabre, *Cosmopolitan War*, 155; Lazar, "Authorization and the Morality of War."

28 I further develop this argument in my manuscript, "Revolution and Intervention." The previous paragraph draws on that paper.

29 Walzer, *Moral Standing of States*, 224. See also Finlay, "Reform Intervention and Democratic Revolution," 575.

simply to stop the human rights violations, but rather to place *Y* in a position to formulate and communicate its autonomous decision about whether to accept military help. Still, there is an important decision that is taken out of *Y*'s hands—namely whether *this* sort of intervention should take place. For this decision is based entirely on the intervener's assessment, rather than *Y*'s, of whether the good effect produced by the intervention (i.e., reducing the risk of coercion or manipulation by the ARL) is worth the costs imposed by it. This would not be the case however, if the decision to intervene was guided by respect for the broader notion of political self-determination I have outlined above. The intervener could then rely on *Y*'s past or presumed consent in deciding what to do. If the limited kind of intervention described by Buchanan was ruled out in light of *Y*'s previous autonomous decisions or in light of what we can reasonably presume *Y* to want in these circumstances, this would give the intervener some reasons, although not necessarily conclusive reasons, to refrain from intervening.

4. How Political Self-Determination Constrains the Scope of Intervention

So far, I have addressed Buchanan's answer to the question of the conditions under which military intervention to depose an authoritarian regime and bring back the conditions for political self-determination in a given political community would be permissible, despite the fact that consent from the community has not been secured. But Buchanan's more controversial thesis concerns what we might call the *scope* of humanitarian intervention, i.e., the goals that the intervening state may legitimately pursue once it has deposed the authoritarian regime.

Most writers on humanitarian intervention agree that the intervening party would be permitted to assist with the process of rebuilding the political institutions of the newly liberated country, preventing any threats that might afflict this process. Indeed, some have argued that the intervening party has a duty, rather than a mere liberty, to do so (in line with the Responsibility to Protect doctrine).[30] Buchanan argues that the scope of intervention is even broader, and includes the permission to "nullify the *democratic* constitutional choice of a newly liberated population, if that choice can reasonably be expected permanently to undercut the conditions for future exercises of the right of self-determination."[31] Indeed, according to him, it is permissible not only to nullify the result of the democratic process, but also to "impose a particular form of democratic gov-

30 Pattison, *Humanitarian Intervention and the Responsibility to Protect*; Fabre, *Cosmopolitan War*, 187–92.

31 Buchanan, "Self-Determination, Revolution, and Intervention," 449.

ernment on a newly liberated population, if (as a contingent matter) it is the only feasible form of government that will ensure the conditions for the future exercises of the right of self-determination, and if the imposed political structure allows for the population, through constitutional means, later to discard it in favor of another one."[32]

This position will strike many as overly permissive, but I see the force of it. For like Buchanan, I believe that a plausible justification for humanitarian intervention must ultimately be grounded not only in the moral demand to prevent human rights violations, but also in the moral demand to protect the right of the community in question to exercise its political self-determination.[33] And like Buchanan, I also believe that political self-determination is ultimately grounded in the existence of a particular interaction between the members of the political community, which makes it apt to regard how the community acts as an expression of its collective agency, rather than as an aggregation of instances of individual agency. Thus, I share his concern for the importance of protecting the conditions under which this interaction can take place.[34] However, I believe we should resist Buchanan's conclusion that a particular form of democratic government may be permissibly imposed on the newly liberated population if the constitutional arrangement they have chosen undermines the conditions for future exercises of the right to political self-determination by other members of the same community.

To see why, consider again the nature of political self-determination. We have seen that political self-determination requires the existence of a particular relationship between the agency of the political community and the agency of its individual members. The members of the community must interact in a certain way so that it makes sense to regard how the group acts as an expression of the unified agency of the community. And it makes sense to do so insofar as the way in which the group acts somehow bears the mark of the agency of its members, in virtue of the fact that they have engaged in the relevant sort of group agency. This

32 Buchanan, "Self-Determination, Revolution, and Intervention," 449.

33 It is worth mentioning that this is a minority position in the most recent debate on humanitarian intervention, where many deny that political self-determination can place any significant constraint on intervention. According to philosophers like Fernando Tesón, Andrew Altman and Christopher Wellman, or Jeff McMahan, the only necessary condition for the permissibility of humanitarian intervention is the fulfillment of traditional *jus ad bellum* principles, particularly proportionality. Tesón, "The Liberal Case for Humanitarian Intervention," 106–7; Altman and Wellman, *A Liberal Theory of International Justice*, 109; McMahan, "Humanitarian Intervention, Consent, and Proportionality," 52.

34 I offer an account of how respect for political self-determination constrains the permissibility of humanitarian intervention and revolution in "Revolution and Intervention."

is why respecting the choices of the group is ultimately a way of respecting the autonomous agency of the members of the community: the inputs that generate the conduct of the group are produced by its members *and only by its members*.[35]

But things would change drastically if a third party were to impose a particular form of government. This imposition would undermine the process just described, as the members of the political community would have to determine the way in which they exercise their collective agency by responding to an alien input. And an extremely significant input indeed, since it shapes how the very basis of political life in the community in question is to be organized. When this is the case, the way in which the group acts no longer reflects what the political community has autonomously decided, because the process of collective deliberation is now shaped to a significant extent by the will of the intervening party.

To the extent that the functioning of the government constitutes the main framework within which the inputs of the members of the political community are combined, it is hard to see how the community in question could be genuinely self-determining in this condition. The way in which the community acts is now determined to a significant extent by an alien entity, since the very way in which the process of collective deliberation is structured has been decided by the intervening party, rather than by the community itself.

But what is the alternative in those cases where the form of government chosen by the intervening party would be the only one capable of ensuring the conditions for the future exercises of the right of self-determination of its members? Are we forced here to accept Buchanan's conclusion, if we value political self-determination? I do not think we are. In those cases, respecting political self-determination requires sacrificing the adoption of a system that would *ensure* that the right of self-determination not be restricted in the future for one that does not provide such assurance. The community in question should be given the chance to set up the institutions that it has autonomously chosen, around which its members can arrange their collective deliberation by interacting in the way required by the process of political self-determination; and it should be given this chance even if there is a risk that in the future its choice might lead it to violate some of its members' right to self-determination.

It would be ideal, of course, if the community selected a constitutional arrangement that ruled out this risk, and the intervening party is permitted to of-

35 I articulate my own formulation of this idea in two unpublished manuscripts: "Political Self-Determination and Wars of National Defence" and "Why Colonialism Is Wrong." In the former, I argue that we can regard the agency of a political community as an expression of the agency of its members, even if (a) typically only few members, if any, can make a difference as to how the community will act, and (b) the way in which the community acts does not align with the personal preferences of each member.

fer incentives to this end, including negative incentives, such as increased trade barriers. If these incentives are unsuccessful and if the risk that the chosen constitutional arrangement would lead to the permanent disenfranchisement of a minority is too high, I agree with Buchanan that intervention to nullify that constitutional decision might be permissible.[36] But even in that case, the intervening party would not be permitted to impose a new constitutional arrangement that has not been chosen by the political community in question. For any decision produced within that constitutional arrangement would not constitute the expression of the community's will. Rather it would be to a significant extent the expression of the will of the intervening party.[37]

Here too the way in which we think about the value of personal self-determination supports my conclusion. Suppose that given his professed values, as well as his previous conduct, Alex is likely to choose a life of crime and harm others. We normally think that, while it is permissible to try and dissuade him from doing so, threaten him with hard treatment, and even physically constrain him (under certain conditions), we are not permitted to manipulate his deliberative process in a way that bypasses his autonomous agency. It would be impermissible, for example, to brainwash him, hypnotize him, or subject him to the "Ludovico technique," so that he will refrain from engaging in harmful conduct. This is a case in which respecting the autonomy of moral agents comes at a cost: the risk that Alex will go on and harm someone. While we may offer incentives to him, including the threat of inflicting significant harm, to prevent him from doing so (the criminal law offers negative incentives of this sort), we may not manipulate the way in which he autonomously deliberates. And this is true even if we assume (a) that those harmed by Alex will be unable to exercise their own personal self-determination, and (b) that it falls outside the scope of Alex's personal self-determination to act in a way that will undermine his victims' personal self-determination in this way.

The same holds for collective self-determination. Imposing a particular constitutional arrangement on a newly liberated country, as suggested by Buchanan,

36 This is because, like Buchanan ("Self-Determination, Revolution, and Intervention," 459, 462), I believe that there are limits to the right of self-determination of political communities. Disenfranchisement and serious forms of discrimination clearly fall outside the scope of how political communities are permitted to exercise their self-determination.

37 On the other hand, the intervening party is permitted, possibly required, to ensure that functioning institutions are created before leaving. Leaving too soon typically leads to unstable regimes and new humanitarian emergencies, which in turn require further military intervention. East Timor is a case in point. After the peacekeeping mission left in 2005, violence quickly resurfaced and a new intervention was needed only a year later. See Stromseth, Wippman, and Brooks, *Can Might Make Rights?*

would be a way of manipulating its autonomous agency. If we did that, the way in which the community deliberates and acts once the new political order is in place could no longer be considered a genuine expression of the way in which its members have exercised their agency as a political community; for a crucially important input in its exercise of collective agency would be generated by the intervening state. When this is the case, the self-determining agency of the community in question is undermined at its roots.

5. CONCLUSION

The question of the permissibility of military intervention in support of attempts to rebel against authoritarian regimes has a central role in any account of revolution, and yet the question has received scant attention in the contemporary debate. In his most recent work, Buchanan has begun to address this gap in the literature. Relying on a sophisticated account of the limits that political self-determination places on intervention, he has defended two controversial views. First, when the injustice suffered by a given political community is serious enough to undermine its capacity for group agency, nonconsensual military intervention does not violate the right to political self-determination of the community in question, and is thus permissible, provided that traditional *jus ad bellum* principles are fulfilled. Second, when intervention takes place, its aims need not be limited to overthrowing the unjust regime. The intervening party may employ military force to nullify the democratic constitutional choice of the newly liberated population and impose a particular form of democratic government, if this is necessary to guarantee the conditions for future exercises of the right of self-determination.

I have suggested that both views should be rejected. The first one should be rejected because respecting the right to political self-determination of political communities requires respecting not only their actual consent, but also their past consent and their presumed consent. Intervention might be incompatible with respecting the right to political self-determination of its intended beneficiaries, despite the fact that at the time of the intervention they lack the capacity to exercise their group agency, if it either goes against previous decisions they have autonomously made or goes against what we can reasonably expect them to want in light of goals and preferences they autonomously set for themselves.

The second view should be rejected because, in imposing a particular form of democratic government, the intervening party would be shaping the very way in which the newly liberated political community will exercise its collective deliberation. Because of this, the decisions taken by the new government will re-

flect, at least in part, the will of the intervening party instead of being a genuine expression of the will of its people. Taking this option off the table is the price to pay for taking seriously the capacity of political communities to act autonomously and be self-determining agents, the price to pay to truly respect their right to political self-determination.[38]

King's College London
massimo.renzo@kcl.ac.uk

REFERENCES

Altman, Andrew, and Christopher Heath Wellman. *A Liberal Theory of International Justice.* Oxford: Oxford University Press, 2011.

Beitz, Charles R. "The Moral Standing of States Revisited." *Ethics and International Affairs* 23, no. 4 (Winter 2009): 325–47.

———. *Political Theory and International Relations*. Princeton: Princeton University Press, 1999.

Buchanan, Allen. "The Ethics of Revolution and Its Implications for the Ethics of Intervention." *Philosophy and Public Affairs* 41, no. 4 (Fall 2013): 291–323.

———. "Revolutionary Motivation and Rationality." *Philosophy and Public Affairs* 9, no. 1 (Autumn 1979): 59–82.

———. "Self-Determination, Revolution, and Intervention." *Ethics* 126, no. 2 (January 2015): 447–73.

Buchanan, Allen, and Robert O. Keohane. "Precommitment Regimes for Intervention: Supplementing the Security Council." *Ethics and International Affairs* 25, no. 1 (Spring 2011): 41–63.

Cassese, Antonio. *Self-Determination of Peoples: A Legal Reappraisal*. Cambridge: Cambridge University Press, 1995.

Dobos, Ned. *Insurrection and Intervention: The Two Faces of Sovereignty*. Cambridge: Cambridge University Press, 2011.

Doppelt, Gerald. "Walzer's Theory of Morality in International Relations." *Philosophy and Public Affairs* 8, no. 1 (Autumn 1978): 3–26.

Fabre, Cécile. *Cosmopolitan War*. Oxford: Oxford University Press, 2012.

Finlay, Christopher J. "Reform Intervention and Democratic Revolution." *European Journal of International Relations* 13, no. 4 (December 2007): 555–81.

Lazar, Seth. "Authorization and The Morality of War." *Australasian Journal of Philosophy* 94, no. 2 (2016): 211–26.

38 Thanks to Allen Buchanan and Laura Valentini for helpful comments.

Luban, David. "Just War and Human Rights." *Philosophy and Public Affairs* 9, no. 2 (Winter 1980): 160–81.

McMahan, Jeff. *The Ethics of Killing*. Oxford: Oxford University Press, 2002.

———. "Humanitarian Intervention, Consent, and Proportionality." In *Ethics and Humanity*, edited by Ann N. Davis, Richard Keshen, and Jeff McMahan, 44–74. New York: Oxford University Press, 2010.

Moore, Margaret. *A Political Theory of Territory*. New York: Oxford University Press, 2015.

Patten, Alan. "Self-Determination for National Minorities." In *The Theory of Self-Determination*, edited by Fernando R. Tesón, 120–44. Cambridge: Cambridge University Press, 2016.

Pattison, James. *Humanitarian Intervention and the Responsibility to Protect: Who Should Intervene?* Oxford: Oxford University Press, 2010.

Parfit, Derek. *On What Matters*, 2 vols. Oxford: Oxford University Press, 2011.

———. *Reasons and Persons*. Oxford: Clarendon Press, 1984.

Regan, Patrick M. "Third-Party Interventions and the Duration of Intrastate Conflicts." *Journal of Conflict Resolution* 46, no. 1 (February 2002): 55–73.

Renzo, Massimo. "Political Self-Determination and Wars of National Defence." Unpublished manuscript.

———. "Revolution and Intervention." Unpublished manuscript.

———. "Why Colonialism Is Wrong." Unpublished manuscript.

Stilz, Anna. "The Value of Self-Determination." In *Oxford Studies in Political Philosophy*, vol. 2, edited by David Sobel, Peter Vallentyne, and Steven Wall, 98–127. New York: Oxford University Press, 2016.

Stromseth, Jane, David Wippman, and Rosa Brooks. *Can Might Make Rights? Building the Rule of Law After Military Interventions*. Cambridge: Cambridge University Press, 2006.

Tesón, Fernando R. *Humanitarian Intervention: An Inquiry into Law and Morality*. Ardsley, NY: Transnational Publishers, 2005.

———. "The Liberal Case for Humanitarian Intervention." In *Humanitarian Intervention: Ethical, Legal, and Political Dilemmas*, edited by J. L. Holzgrefe and Robert O. Keohane, 93–129. Cambridge: Cambridge University Press, 2003.

Walzer, Michael. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, 4th ed. New York: Basic Books, 2006.

———. "The Moral Standing of States: A Response to Four Critics." *Philosophy and Public Affairs* 9, no. 3 (Spring 1980): 209–29.

Wasserstrom, Richard. Review of *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, by Michael Walzer. *Harvard Law Review* 92, no. 2 (December 1978): 536–45.

Journal of Ethics and Social Philosophy
VOL. 13, NO. 3 · JULY 2018
https://doi.org/10.26556/jesp.v13i3.323

ON *EX ANTE* CONTRACTUALISM

Korbinian Rüger

CONTRACTUALISM is a claims-based model of moral rightness. It is the view, brought forward most notably by T. M. Scanlon, that an action is right if and only if it is justifiable to all. An action is justifiable to all just when it is licensed by a principle that cannot be reasonably rejected by any single individual.[1] Further, a principle can only be reasonably rejected for personal reasons. Contractualism thus construed excludes impersonal reasons derived from, for example, the overall value of an outcome. It thereby denies the permissibility of interpersonal aggregation of harms and benefits to determine which action is right. In situations in which individuals have competing claims to be helped, we always ought to pursue the policy that satisfies the single strongest individual claim, or, in converse, minimizes the strongest individual complaint against it, by following "the principle whose implications are most acceptable to the person to whom it is least acceptable."[2]

This implication of contractualism clearly demarcates the view from thoroughly aggregative theories like utilitarianism. I here understand utilitarianism as standard act utilitarianism, where we always ought to pursue the action that will lead to the greatest (expected) sum total of well-being. The difference between the two rival theories becomes apparent in cases like:

> *Death versus Headaches:* We can either save Ann from a terminal illness or prevent any number of different people from suffering a mild headache.

By virtue of what Ann stands to lose, her claim to be saved from death is clearly greater than any other individual claim to be spared a headache. Under contractualism we therefore ought to save her. This is the case irrespective of how many people stand to suffer a headache. Under utilitarianism, on the other hand, our answer will depend on the number of people that we could spare the headache. For some number of people, the benefits derived from the spared headache will *in sum* outweigh the benefit to Ann if we choose to save her. Contractualism

1 See Scanlon, *What We Owe to Each Other*, 189–248.

2 Kumar, "Risking and Wronging," 31.

demands what many people take to be the obviously correct choice in *Death versus Headaches*.[3]

This is straightforward in hypothetical situations of absolute certainty like the above. The approach, however, is less clear about situations in which we do not yet know the outcomes our choices will lead to. These cases, however, are much more common. With Barbara Fried, one could even say that

> in the real world, no conduct, judged ex ante, is certain to harm others. This is true even of harms that are intended.... If I point a gun at your head and pull the trigger, I am overwhelmingly likely to kill or seriously injure you, but I am not certain to do so. The gun could misfire, I could have forgotten to load it, [etc.].[4]

So rather than occurring with certainty, most harms result from risks that have been imposed on people or have not been eliminated. It is therefore imperative for contractualists to offer an account of how their theory deals with risk.

Johann Frick has developed such an account: *ex ante* contractualism.[5] In brief, *ex ante* contractualism holds that in situations involving risk we ought to act in accordance with principles that license the action that satisfies the strongest individual claim, where those claims are a function of the expected value that a given policy gives each person *ex ante*. It thus offers an alternative to the *ex post* reasoning employed by other contractualists, most notably Scanlon himself.[6]

I here challenge Frick's version of *ex ante* contractualism on contractualist grounds.[7] My argument proceeds as follows. In the first section I distinguish between *ex ante* and *ex post* contractualism in more detail. In the second section I argue that adopting *ex ante* contractualism would have far-reaching implications that contractualists would find very hard to accept. I show that *ex ante* contractualism in fact includes an implicit appeal to the interpersonal aggregation of

3 See, e.g., Voorhoeve, "Why One Should Count Only Claims with Which One Can Sympathize."

4 Fried, "Can Contractualism Save Us from Aggregation?" 50.

5 Frick, "Contractualism and Social Risk." See also Frick, "Treatment versus Prevention in the Fight against HIV/AIDS and the Problem of Identified versus Statistical Lives." Unless noted otherwise, henceforth "*ex ante* contractualism" refers to Frick's version of the view.

6 See Scanlon, *What We Owe to Each Other*, 189–248. See also Reibetanz-Moreau, "Contractualism and Aggregation." Note that Scanlon has since changed his position, crediting an earlier version of Frick's article. See Scanlon, "Reply to Zofia Stemplowska." For a critique of the *ex post* approach, see Ashford, "The Demandingness of Scanlon's Contractualism."

7 My argument is specifically addressed at Frick's way of developing *ex ante* contractualism. It is possible that there is another way of developing the view to which my argument does not apply. I do not pursue this possibility here.

harms and benefits. In the third section I show that Frick's argument for the principled priority of identified over unidentified lives, another troubling implication of *ex ante* contractualism, is unsound. In the fourth and final section I briefly comment on a possible pluralistic approach to get around some of the defects of *ex ante* contractualism. I conclude that, to deal with uncertainty, contractualists should not adopt *ex ante* contractualism, at least not Frick's version. Rather, they should adopt a suitably amended *ex post* approach.

1. *EX ANTE* AND *EX POST* CONTRACTUALISM

Let me introduce the *ex post* approach and then contrast it with the *ex ante* approach by way of one of Frick's examples.

> *Mass Vaccination*: One million children are threatened by a virus, which will kill all of them if we do nothing. We must choose between producing one of three vaccines:
>
> - Vaccine 1 is certain to save every child's life. However, if a child receives Vaccine 1, the virus will permanently paralyze one of the child's legs.
> - Vaccine 2 gives every child a 99.9 percent chance of surviving the virus completely unharmed. However, for every child there is a corresponding 0.1 percent chance that Vaccine 2 will be completely ineffective. (Assume that the outcomes for different children are probabilistically independent.) Call the children who end up dying the *luckless children*.
> - Vaccine 3 is sure to allow 999,000 children to survive the virus completely unharmed. However, because of a known particularity in their genotype, Vaccine 3 is certain to be completely ineffective for 1,000 identified *doomed children*.[8]

First consider a choice between only Vaccines 1 and 3 ($V1$ and $V3$). Here we are not dealing with uncertainty and it is straightforward what contractualism recommends. If we choose $V1$, no single child will have a complaint that is as strong as the individual complaints of the doomed children if we choose $V3$.[9]

8 See Frick, "Contractualism and Social Risk," 181–83. Note that Frick presents two distinct cases, in both of which Vaccine 1 is available, but Vaccines 2 and 3 only are available in one case.

9 Frick writes: "The individual burden of becoming paralyzed in one leg, though significant, is not even close to that of losing one's life at a young age" ("Contractualism and Social Risk," 183). Note that this information underspecifies (or even ill specifies) the strength of the individual complaints. If we assume a counterfactual account of harm, the complaints of the doomed children if we pick $V3$ over $V1$ are not complaints against being left to die, where

We therefore ought to choose V_1. If, on the other hand, we consider a choice between V_1 and V_2, we are entering the territory of risk and things are less clear. This is because there are two ways of singling out the relevant complaints that we should take into account. Under one interpretation we look at the outcome that a given vaccine will produce and look at the single strongest complaint any individual will have in that outcome. If we choose V_2, we expect one thousand children to die.[10] Though we do not know how many children exactly will die, it is statistically certain that at least one child will die.[11] Since we are concerned with the single strongest individual complaint, this is all we need to know. Like in V_3, this complaint will be stronger than any complaint under V_1. Again, we ought to choose V_1. This is the *ex post* approach.

According to the *ex ante* approach, on the other hand, the relevant complaints are a function of the expected value an action gives each individual before it is performed. Under this account, a complaint against being subjected to a risk of suffering a harm is the complaint against being subjected to that harm with certainty discounted by the unlikelihood of the harm actually occurring. In *Mass Vaccination* the individual *ex ante* complaints against V_2 are thus only 0.1 percent as strong as a complaint against dying from the virus with certainty. The strongest *ex ante* complaint against V_2 is therefore much smaller than the strongest *ex ante* complaint against V_1, which in turn is smaller than the strongest *ex ante* complaint against V_3.

Accordingly, Frick's account selects V_2 over V_1, V_1 over V_3, and V_2 over V_3. This ensures that in each choice we minimize the strongest *ex ante* complaint. The *ex post* approach on the other hand would choose V_1 over V_2 and V_3, and would likely be indifferent between V_2 and V_3.

Mass Vaccination thus shows how the *ex ante* and *ex post* approaches come apart. According to Frick it also shows why *ex post* contractualism is unattractive. It fails to make a principled distinction between V_2 and V_3. As long as we know that *someone* will die if we pick V_2, and therefore they have a stronger complaint than anyone else if we choose V_1, this is enough for *ex post* contractualism to rule out V_2. It fails to take into account the special predicament the doomed children find themselves in under V_3 as it assimilates their fate to those of the luckless

the alternative would be life in full health, but complaints against being left to die, where the alternative would be life with one paralyzed leg. Such complaints are presumably much weaker. I think this point is overlooked by Frick. Nonetheless it is reasonable to assume that these weaker complaints are still decisively stronger than the complaints of the other (non-doomed) children against a policy that leaves them with one paralyzed leg, where the alternative would be life in full health.

10 $0.001 \times 1{,}000{,}000$.

11 $1 - \left(^{999}/_{1{,}000}\right)^{1{,}000{,}000}$.

children in $V2$. Frick would say that it fails to distinguish between the fact that "we know that someone will die" ($V2$) and the fact that "there is someone whom we know will die" ($V3$).[12]

Because of these alleged shortcomings of *ex post* contractualism, Frick proposes his *ex ante* approach. The main argument for this approach is the *argument from the single-person case*.[13] According to this argument, if we have an option available that is in the best *ex ante* interest of all individuals, we ought to choose it. We ought to adhere to the *ex ante* Pareto principle.

> *Ex Ante Pareto Principle:* If an alternative has higher expected utility for every person than every other alternative, then this alternative should be chosen.[14]

Frick argues that the argument from the single-person case establishes the *ex ante* Pareto principle as a principle of contractualist ethics. We can decompose cases like *Mass Vaccination* into a large number of single-person gambles. Suppose again that we are facing a choice between $V1$ and $V2$ (recall that *ex post* contractualism chooses $V1$). This choice can be broken down into one million single-person cases. Suppose that Ann is one of the affected children and we ask ourselves what we would choose if we were solely motivated by her self-interest. We know that $V1$ will let her survive the virus but leave her with one paralyzed leg and that $V2$ will let her survive the virus completely unharmed with probability 999/1,000 and will lead to her death with probability 1/1,000. Given reasonable assumptions about which level of well-being (or utility) these three possible outcomes would deliver, we can calculate the expected value of both options. Suppose we assume that, for Ann, life with one paralyzed leg is four-fifths as good as life at full health, which we can arbitrarily fix to utility level 10, with death corresponding to 0. The expected utility of $V1$ then is 8, while the expected utility of $V2$ is 9.99.[15] Thus, the expected utility of $V2$ for Ann exceeds that

12 See Frick, "Contractualism and Social Risk," 200, and "Treatment versus Prevention in the Fight against HIV/AIDS and the Problem of Identified versus Statistical Lives," 193. I do not take this distinction to be morally as important as Frick thinks it is. I shall not argue for this claim directly, though. Rather I will show that there are cases where even Frick's own account fails to make the distinction.

13 See Frick, "Contractualism and Social Risk," 186–94, and "Treatment versus Prevention in the Fight against HIV/AIDS and the Problem of Identified versus Statistical Lives," 133. For similar arguments see Dougherty, "Aggregation, Beneficence, and Chance," and Hare, "Should We Wish Well to All?"

14 This formulation is taken from Fleurbaey and Voorhoeve, "Decide as You Would with Full Information!" 114.

15 $4/5 \times 10$ and $999/1{,}000 \times 10 + 1/1{,}000 \times 0$.

of V_1, and, if we are only concerned with her best interest, we ought to choose V_2.[16] This seems to be the right course of action. After all, what other than Ann's best interest would we base our decision on?

But, of course, this reasoning is correct for every single child in *Mass Vaccination*, where the possible outcomes and corresponding odds are exactly the same as in the one-person case. Thus, if we are concerned with every child's best interest, we ought to choose V_2, just like we ought to choose V_2 in the one-person case when we are only concerned with Ann's best interest. The contractualist rationale behind this is that choosing V_2 is the only action that is justifiable to all. Whatever the outcome of choosing V_2, we can offer each child the following justification: "When we had to choose, we did what was in your own best interest." This justification is not available to us if we choose V_1. I confess that I find this argument very seductive. In the following I argue, however, that contractualists ought to reject it and with it the *ex ante* Pareto principle.

2. THE IMPLICATIONS OF *EX ANTE* CONTRACTUALISM

Return to *Mass Vaccination*. Only now suppose that instead of V_3, we have V_3* available. Like V_3, V_3* is sure to allow 999,000 children to survive the virus completely unharmed. However, because of a certain particularity in their genotype, V_3* is certain to be completely ineffective for one thousand *unidentified* doomed children, instead of *identified* doomed children. We can imagine, for example, that we have tested all one million children for that genotype and have found that the vaccine will be ineffective for exactly one thousand of them. However, before we communicated the test results to anyone, our system broke down and we now have no way of assigning the positive results to any particular children.

Given that we chose V_2 over V_3, should we now choose V_2 or V_3*? In order to answer that question we need to investigate whether V_3* is relevantly different from V_3. Only if it is can we justify choosing V_2 over V_3, being indifferent between V_2 and V_3*. If V_3* is not relevantly different, then, given that we chose V_2 over V_3, we also ought to choose V_2 over V_3*. In this section I will argue that, first, we should not judge V_3 and V_3* differently; second, that *ex ante* contractualism, however, is committed to doing so; and, third, that this puts the account in a precarious position.

To me, V_3* seems like V_3 in all important respects. In V_3* as in V_3, we know the exact outcome. We know that exactly one thousand children are going to die and that for them the vaccine was always going to be ineffective. Like with

16 Note that this result will be achieved even if Ann considers life with one paralyzed leg only slightly worse than life at full health.

V_3, these children are doomed to die if we choose V_3*. I therefore cannot see why we should choose V_2 over V_3, but be indifferent between V_2 and V_3*, i.e., prefer to have V_3* rather than V_3 available. To see that this is implausible, suppose that we not only have V_3* but two different vaccines—V_3* and V_3**—available. However, for each of these vaccines it will (very likely) be an entirely different group of one thousand children for whom the vaccine will not work and who will be killed by the virus. Obviously we have no reason to choose any one of these vaccines over the other. Whatever we do, one thousand unknown, doomed children are going to die. The vaccines are equally choice worthy and we should randomize.

Suppose that we settle on V_3*. Before we actually administer the vaccine, however, we learn who the children are for whom V_3* will not do anything (maybe we were able to restore our database for V_3*). Should we now because of that switch to V_3**? I think clearly not. This would be an unnecessary "second lottery" and would arbitrarily favor those children for whom V_3* is ineffective to the disadvantage of those children for whom V_3** is ineffective. Nothing about the vaccines has changed and we said above that they are equally choice worthy. They still are. This, however, is in effect the same situation we face when comparing Frick's V_3 and my V_3*. We therefore ought not to judge V_3 and V_3* differently.

Frick's account, however, is committed to judging V_3 and V_3* differently. It is committed to judging V_3 impermissible, but V_3* (along with V_2) permissible. This is because the argument from the single-person case applies to V_3* as it applies to V_2. Here too it would be in each individual child's best interest to choose the risky vaccine over V_1. The expected value of V_3* for each individual child is the same as V_2.[17]

Though Frick does not consider V_3*, he considers a nearby case. This case is like my V_3*, only here there is a test we could carry out to identify the doomed children, but it would be very expensive. Frick argues that in this case adminis-

17 *Ex ante* contractualists could reply that there is one important difference between V_2 and V_3* that I have overlooked—namely, that while in V_3* it is merely *epistemically* uncertain who will die, in V_2 it is *objectively* (or *physically*) uncertain who will die. Frick, however, carries out his discussion on the assumption that *all* probabilities are merely epistemic. He writes that "when using the terms 'probability' or 'chance' … I assume that we are speaking not about objective indeterminacy at the level of physical reality itself, but about epistemic probability" (Frick, "Contractualism and Social Risk," 182). He furthermore argues, rightly I think, that for the moral assessment of risky policies this distinction makes no difference. (See Frick, "Contractualism and Social Risk," 197–201.) In any case, it is doubtful whether objective probabilities at the physical level even exist. (See, e.g., Lewis, "A Subjectivist's Guide to Objective Chance.") Letting one's moral theory depend on the assumption that they do exist seriously diminishes its attractiveness.

tering the vaccine would be justifiable to all and therefore permissible. In such a case we can say to each child "given justifiable limits on the resources we can be expected to expend in gathering further information about your particular case, [the vaccine] is highly likely to benefit you, and has only a tiny chance of turning out to your disadvantage."[18] If administering the vaccine where it is *very costly* to find out which children will not be helped by it is permissible, then *a fortiori* administering $V3^*$, where it is *impossible* to find out which children carry the problematic gene, must also be permissible. I now argue that this judgment concerning $V3^*$ spells trouble for the account. Consider the following case.

> *Glass Box Villain (Known Victim):* An evil villain has taken twenty-six hostages named Ann, Bob, Carl … and Zeta. He places you in the following diabolic choice situation: he has placed all of them in twenty-six individual glass boxes standing up side by side. The last box is made of regular glass and the other twenty-five boxes are made of extra-heavy glass. You can see that Zeta is placed in the last box. The villain asks you to decide between the following two options: (1) he will either fire a shot at her box or (2) fire twenty-five individual shots at the other boxes. If he fires at Zeta's box, the bullet will not be stopped and Zeta will be killed. If he fires at the twenty-five boxes made of extra-heavy glass, the glass will divert the bullets. However, the glass will crack and the debris will disfigure the twenty-five hostages in a way that permanently leaves them at a well-being level 9.5 on a scale from 0 to 10, where 10 corresponds to a life in full health and 0 to death. If you refuse to decide, the villain will blow up all boxes, killing all twenty-six hostages. How should you decide?

I assume that refusing to decide should be ruled out as an option. Between the two remaining options, it is clear what contractualism tells you to do. You should choose (2). Choosing (1) would kill Zeta only to save twenty-five other people from a relatively minor harm. The complaints of the twenty-five on you are not even close to Zeta's complaint. And since contractualism prohibits you from aggregating the twenty-five weak complaints to outweigh Zeta's strong complaint, you ought to save Zeta's life and let the villain fire at the twenty-five boxes made of extra-heavy glass. Since there is no uncertainty involved, *ex ante* and *ex post* contractualism do not come apart in this case. Consider, however, the following variation of the case.

> *Glass Box Villain (Unknown Victim):* Everything is as before, only now the boxes are opaque and neither you nor the hostages know whether it is

18 Frick, "Contractualism and Social Risk," 194.

Zeta or any of the other twenty-five in the box made of regular glass. How should you decide?

I think if in *Glass Box Villain (Known Victim)* you ought to stop the villain from firing at the last box, then in this case you ought to act in the same way. I cannot possibly see why the fact that the twenty-six boxes are now opaque should change our moral assessment of the case in any way. (Remember that there is nothing about Zeta as a person that should make us favor *her* over the other twenty-five hostages in any way.) However, Frick's *ex ante* contractualism is committed to the view that while in *Glass Box Villain (Known Victim)* you ought save Zeta, in *Glass Box Villain (Unknown Victim)* you ought to let the villain kill the person in the last box.

It is so committed because in *Glass Box Villain (Unknown Victim)* for all you know it could be any of the twenty-six hostages in the last box. In this respect it is parallel to V_3*, above. You have no reason to assume that any of the twenty-six was more likely to end up there than anyone else. As far as you know, for each of them there is a 1/26 chance that they are the one in the last box and a corresponding 25/26 chance that they are among the ones in the boxes made of extra-heavy glass. This means that for each of them if you let the villain fire at the last box, there is a 1/26 chance that they will die and a 25/26 chance that they walk away completely unharmed. If you choose otherwise, on the other hand, for each hostage there is a 1/26 chance that they walk away unharmed (if they are the one in the last box) and a 25/26 chance that they walk away slightly but permanently disfigured.

I have arbitrarily assumed that this disfigurement leaves them at utility level 9.5 out of 10.[19] If we also assume that death leaves the hostages at "utility level" 0, then the choice situation can be represented by the following table.

	S_1		S_2		S_3		⋯	S_{26}	
	A	Others	*B*	Others	*C*	Others	⋯	*Z*	Others
First 25 Boxes	10	9.5	10	9.5	10	9.5	⋯	10	9.5
Last Box	0	10	0	10	0	10		0	10

You have two available actions (again, ignoring the option of doing nothing): "first 25 boxes" and "last box." There are twenty-six equiprobable states of the world (S_1–S_{26}, $p = 1/26$), corresponding to the twenty-six possibilities of who could be the one in the last box, where S_1 corresponds to the state of the world in which Ann is the one in the last box, S_2 to the state in which Bob is the one,

19 If you think that this is too low or too high, then you can adjust the level and change the number of hostages accordingly without affecting the basic structure of the case.

and so on. The table shows the utility levels of the hostages for each of these twenty-six states and the two available actions. For example, if you decide on "first 25 boxes" and Ann is the one in the last box, then she will be left at level 10, corresponding to full health, while the other twenty-five hostages (the "others") will be left at level 9.5. On the other hand, if you decide on "last box" and Ann is the one in that box, she will die ("level 0") and the others will be left unharmed at level 10. We can now calculate the expected utility for each hostage under each of the two available actions. If you choose "first 25 boxes," the expected utility is 9.52.[20] If you choose "last box," the expected utility is 9.62.[21] Since 9.62 is greater than 9.52, if you want to do what is in each of the hostage's best interest, you ought to choose "last box." If you could ask them, they would want you to do so, or if for each of the hostages there was a guardian present who is only motivated by their beloved's interest, they would tell you to do so. Therefore, via the argument from the single-person case, *ex ante* contractualists (and proponents of *ex ante* Pareto in general) are committed to letting the villain kill the person in the last box. Note that here *ex ante* contractualism is so committed although "there is someone whom we know will die." We know that there is a person we will willingly sacrifice—namely the person in the last box.

This fact points to an objection that could be pressed against my exposition: it is not in fact true that every hostage has a 1/26 chance of being the one in the last box.[22] At the time of decision there is a fact of the matter who the person in that box is. From this it follows that it is not actually true that choosing "first 25 boxes" is in the best interest of everyone.

I think this objection will not succeed, at least not for an *ex ante* contractualist of Frick's kind. This is because this same objection could be pressed against someone, like Frick, who distinguishes between V_3 and V_3^*, above, deeming V_3 impermissible and V_3^* permissible. In both V_3^* and V_3 there is a fact of the matter who the children are that are going to die. The only difference is that in V_3^* informational constraints keep us from knowing the identities of these children. The same holds for *Glass Box Villain (Unknown Victim)*. So if one thinks that we should not distinguish between *Glass Box Villain (Unknown Victim)* and *Glass Box Villain (Known Victim)* because in both cases there is a fact of the matter who is in the last box, then by parity of reasoning we also ought not to distinguish between V_3 and V_3^*, because here in both cases there is also a fact of the matter who the one thousand children are for whom the vaccine will do

20 $1/26 \times 10 + 25/26 \times 9.5$.

21 $25/26 \times 10$.

22 This was suggested to me by Jeff McMahan and Tom Sinclair.

nothing.[23] As things stand, *ex ante* contractualists are committed to choose "first 25 boxes" in *Glass Box Villain (Known Victim)* and "last box" in *Glass Box Villain (Unknown Victim)*.

I think this result should worry *ex ante* contractualists, especially since, *qua* contractualists, they would be deeply committed to choose otherwise in *Glass Box Villain (Known Victim)*. The case thus lays bare the implications of the view that its proponents need to accept. These are implications many contractualists, or nonconsequentialists in general for that matter, find hard to stomach. *Glass Box Villain (Unknown Victim)* shows that *ex ante* contractualists need in some cases to be prepared to sacrifice a person's life in order to protect many other people from a relatively minor ailment. This strikes me exactly as the kind of interpersonal aggregation that contractualism set out to avoid in the first place.

Now, *ex ante* contractualism's proponents might be prepared to bite the bullet. They could say that the fact that the number of people affects each individual prospect (holding everything else fixed) is simply directly implied by the way *ex ante* contractualism is defined. Frick calls this "counting the numbers without aggregating."[24] One could thus object to my exposition that I am implicitly assuming what I intend to show—namely that *ex ante* contractualism cannot be correct. For if one instead assumes that *ex ante* prospects are what we should be concerned with in a case like *Glass Box Villain*, then it plainly follows that we should order the villain to fire at the last box. To some extent this objection is warranted, for I *am* assuming that a theory that tells us to let the person in the last box be killed in *Glass Box Villain (Unknown Victim)* should strike contractualists as dubious, if not wrong. The point is that, rather than embracing this "number counting" as a welcome implication of the view, contractualists should be worried about a view that has these implications since it allows the numbers of people on each side of a binary choice to affect what we ought to do, even though the individual benefits and burdens are not affected.

The reason why most contractualists (and other nonconsequentialists) are opposed to interpersonal aggregation is because it violates what, following Rawls, has come to be called the "separateness of persons."[25] According to one very strict version of this thesis the aggregation of harms across different individuals is meaningless since there is no single entity to suffer the aggregate harm. As C. S. Lewis writes:

23 This, of course, is the position I am arguing for. It is however not available to *ex ante* contractualists, as they *want* to distinguish between V_3 and V_3^*.

24 Frick, "Contractualism and Social Risk," 201.

25 See Rawls, *A Theory of Justice*, 167.

> Suppose that I have a toothache of intensity x: and suppose that you, who are seated beside me, also begin to have a toothache of intensity x. You may, if you choose, say that the total amount of pain in the room is now $2x$. But you must remember that no one is suffering $2x$: search all time and space and you will not find that composite pain in anyone's consciousness. There is no such thing as a sum of suffering, for no one suffers it.[26]

This, however, is exactly what *ex ante* contractualists overlook in *Glass Box Villain (Unknown Victim)*. If we let the number of people in the boxes made of extra-heavy glass affect what we believe we ought to do, then we are overlooking the fact that the harm any of the hostages is going to suffer does not increase or decrease with that number.

Might *ex ante* contractualists respond to my argument so far by claiming that there is a principled difference in importance between saving an identified person and saving an unidentified person that I have overlooked? If so, this difference could explain why we should in fact let the villain kill the person in the last box in *Glass Box Villain (Unknown Victim)*, while we should stop him from killing Zeta in *Glass Box Villain (Known Victim)*, as well as explain why we should choose V_2 over V_3, but be indifferent between V_2 and V_3*. In the following section I investigate this possibility.

3. The "Pro Identified Lives Argument"

Many people attach greater importance to saving identified lives than to saving unidentified lives.[27] It is doubtful, however, that this psychological fact is of any moral relevance.[28] I, for one, do not think it is. It will have to be, however, in order to justify *ex ante* contractualism's way of distinguishing between V_3 and V_3*, as well as between *Glass Box Villain (Unknown Victim)* and *Glass Box Villain (Known Victim)*. Luckily for *ex ante* contractualists, Frick offers an ingenious argument to that effect. He argues that correctly applying the *ex ante* contractualist rationale to cases that are "competitive *ex ante*" yields the conclusion that we ought to prioritize identified over unidentified lives. In this section I attempt to show that this argument does not succeed.

In section 1 we saw how *ex ante* contractualism coincides with the *ex ante*

26 Lewis, *The Problem of Pain*, 103–4.

27 See Moore, "Caring for Identified versus Statistical Lives"; Jenni and Loewenstein, "Explaining the 'Identifiable Victim Effect.'"

28 See Schelling, "The Life You Save May Be Your Own"; Brock and Wikler, "Ethical Challenges in Long-Term Funding for HIV/AIDS"; and Otsuka, "Risking Life and Limb: How to Discount Harms by Their Improbability."

Pareto principle in cases in which there are actions that are in the *ex ante* interest of everyone. The principle, however, does not apply in cases that are competitive *ex ante*. Here every action that is in the interest of one group of people comes at a cost to another group of people even at the *ex ante* stage. Take the following example employed by Frick.[29]

> *Miners*: A single miner, Jones, is trapped in a mineshaft and if we don't help him, he will die. The rescue mission, however, would be very costly. These resources could instead be used to make the mine safer for everyone working there in the future. Suppose there are 100 other people working at the mine and with the resources we would have to use on the rescue mission, we know that we could instead reduce their risk of suffering a fatal accident from 3 percent to 1 percent. What should we do?[30]

If we decide to let Jones die and make the mine safer for future workers, we can expect to save two workers' lives in the future instead of saving Jones's life now.[31] Frick argues that *ex post* contractualists here are committed to letting Jones die and saving the two other workers' lives instead.[32] This is because no matter what we do, the strongest individual complaints are equally strong in both cases. These are the complaints of the miners who will die when we could have prevented it, Jones in the one case and the unnamed two miners in the other case. And since under Scanlon's contractualism "numbers break ties" when the strongest complaints are equally strong on both sides, we ought to do what satisfies the greater number of strongest claims.[33]

Again, Frick thinks *ex post* contractualism goes wrong here. He offers his "pro identified lives argument" to show why this is so and takes this argument to provide a principled defense of the claim that we ought to prioritize identified lives over unidentified lives. The argument starts from the premise that, in general, people have a stronger claim to be saved from suffering a harm with

29 See Frick, "Contractualism and Social Risk," 212.

30 Further assume that we know that no one else but these one hundred people will ever work at the mine.

31 100×0.02.

32 See Frick, "Contractualism and Social Risk," 214.

33 Scanlon's "tie-breaking argument," where he draws on an argument by Frances Kamm (see Kamm, *Morality, Mortality*, 101, 114–19; Scanlon, *What We Owe to Each Other*, 229–41) is contested (see, e.g., Otsuka, "Scanlon and the Claims of the Many versus the One" and "Saving Lives, Moral Theory, and the Claims of Individuals"). However, despite the defects of this particular argument, I find it highly plausible that, when deciding between one claim on the one hand and two claims of equal magnitude on the other hand, we ought to satisfy the two claims.

certainty than to be saved from suffering that same harm with some probability $p < 1$.[34] This claim is undoubtedly correct. Suppose we have to decide between saving Ann from certain death or reducing Bob's risk of death from 3 percent to 1 percent. It is clear that we ought to help Ann in this case. Now Bob's claim to have his death risk reduced is identical, Frick continues, to each of the one hundred miners' claims in *Miners*. From this it follows that no individual miner has a stronger complaint than Jones. Coupled with the contractualist ban on interpersonal aggregation, it follows that we ought to minimize the single strongest complaint and save Jones.

Frick claims that, first, this argument provides a principled defense for the privileging of identified over unidentified lives and, second, that it also shows where *ex post* contractualism goes wrong. He claims that *ex post* contractualists are committed to the view that in *Miners* there is someone who has a stronger claim than Bob in the one versus one case. He writes, "somehow, the fact that, if we save [Jones], it is foreseeable that *someone* from the group of 100 will die in a future accident is thought to strengthen the complaint of whoever turns out to be harmed."[35] This, Frick argues, is an implicit appeal to interpersonal aggregation over "different possible worlds."

Regarding the first point: I think that the argument does not provide a principled defense for favoring identified lives in general, but only in a very narrow class of cases like *Miners*. It only provides a defense for favoring an identified person *when and because* that person holds a claim that is stronger than any competing claim. It, for example, does not provide a defense of the type needed to justify the *ex ante* contractualist's choices in *Glass Box Villain (Unknown Victim)*. Here, the dialectic of comparing *ex ante* claims and then satisfying the single strongest claim does not work, since here all *ex ante* claims are equally strong, as we have seen. It thus fails to provide a justification for why it is more important to save Zeta in *Glass Box Villain (Known Victim)*, than to save the unidentified person in the last box in *Glass Box Villain (Unknown Victim)*. This is because the argument does not provide a principled defense for the claim that it is more important to save an identified person rather than an unidentified person *because* that person is identified. Such a defense, however, would be needed to justify *ex ante* contractualism's verdicts in the *Glass Box Villain* cases.[36]

Regarding the second point, first of all, it is not clear that *ex post* contractual-

34 See Frick, "Contractualism and Social Risk," 215. See also Frick, "Treatment versus Prevention in the Fight against HIV/AIDS and the Problem of Identified versus Statistical Lives," 188–91.

35 Frick, "Contractualism and Social Risk," 217.

36 For an attempt at providing an argument to that effect, see Hare, "Should We Wish Well to

ism really is committed to letting Jones die in *Miners*. We do not *know* that two miners will die in the future if we decide to save Jones. Yes, this is the expected outcome, but, of course, it is only one of many different possible outcomes. The chance that *exactly* two miners will die is only around 27 percent.[37] The chance that *at least* two miners will die, so as to tip the scales in favor of letting Jones die, is around 60 percent.[38] No part of *ex post* contractualism commits proponents of the view to disregard these probabilities entirely. Frick assumes that they would take the expected outcome of an action and then simply act as if they knew that that expected outcome would actually eventuate. This, of course, would be a mistake. By doing so they would not be able to differentiate between cases like *Miners* and a case in which we have to decide between saving one person from certain death and saving two different people from certain death. But I do not think that anything commits them to this precarious position. Instead, they could take into account the likelihood of enough miners dying so as to outweigh Jones's claim. I take this to be the most plausible interpretation of *ex post* contractualism.[39] As we have seen, the likelihood of at least two miners dying is only 60 percent. So why should we just assume that *ex post* contractualists would not rescue Jones?

Second, as I have argued before, the main problem with many instances of interpersonal aggregation of harms is that any sum of weaker harms together does not constitute anything meaningful, since there is no one suffering from this aggregate harm. This, however, is not the case in *Miners*. Here, the aggregate of the many trivial harms *is* suffered by a single individual. The more people who work at the mine, the likelier it becomes that *someone* will die as a result of us not making the mine safe. This is a different kind of aggregation. Contrast this with a variation of *Miners*, where we can either save Jones or use the resources to distribute lifelong supplies of aspirin to all future miners who occasionally suffer headaches because of the stuffy air in the mine. This aggregation is more like the kind of aggregation employed by the *ex ante* contractualist in *Glass Box Villain (Unknown Victim)*. Here, as we have seen, the number of people involved has no effect on the harm that the most burdened individual has to suffer. As long as we lack an independent objection against this second, different kind of

All?" 267–71. I am not convinced by Hare's argument. Discussing it here, however, would lead us too far afield.

37 $^{100}/_{2} \times 0.02^{2} \times 0.98^{98}$.

38 Pr (100 deaths) – Pr (0 or 1 deaths).

39 I attempt to fully specify such a view elsewhere. For a similar account, see also Otsuka, "Risking Life and Limb," and Horton, "Aggregation, Complaints, and Risk," 65–66.

aggregation, I do not see why *ex post* contractualists need to be moved by this particular argument.

4. PLURALISM AS A WAY OUT?

Let me now turn to the final problem with *ex ante* contractualism that I want to raise in this essay. This problem is acknowledged by Frick himself. Consider a variation of *Miners*, only now there are one thousand other miners in addition to Jones. Call this *Miners 1,000*. In this case, we would expect twenty miners to die in the future if we save Jones now. Frick submits that *ex ante* contractualism here "goes too far."[40] For him it is clear that given some number of expected deaths (which could be greater or less than twenty) we ought to let Jones die. Frick concedes that this problem for *ex ante* contractualism can only be solved "by scaling back the ambitions of contractualism as a moral theory."[41] He argues that in cases in which his theory is unable to yield the intuitively correct verdicts, it should be assisted by other noncontractualist principles. As a candidate, Frick suggests that we should take into account the effect an action has on people's well-being in general. For example, in *Miners 1,000* we should take into account that there will be "a much greater loss of life" if we save Jones.[42]

This sounds like Frick is suggesting that the contractualist should call utilitarianism to her rescue when her theory fails her intuitions. This *ad hoc* move, however, is available to *ex post* contractualists as well. They, too, can be pluralists about interpersonal morality. Like Frick, they too can say that in some cases their theory needs to be assisted by impersonal concerns to decide what the right course of action is. I see no reason why, *prima facie*, it should seem more plausible to restrict *ex ante* contractualism in such a way than it is to restrict *ex post* contractualism in the same manner. The only difference being that it would be different cases that the theory can deal with "on its own." In *Mass Vaccination*, for example, facing a choice between $V1$ and $V2$, *ex post* contractualists can say the following: "In principle we ought to choose $V1$ here, since this minimizes the largest complaint *ex post*. However, the consequences of doing so in terms of overall well-being are too grave to be ignored. After all, if we do choose $V1$, we will leave one million children with only one functioning leg for the rest of their lives. This overall loss in well-being is much greater than if one thousand children die prematurely." If *ex ante* contractualists can legitimately resort to these

40 Frick, "Contractualism and Social Risk," 219

41 Frick, "Contractualism and Social Risk," 219.

42 Frick, "Contractualism and Social Risk," 222.

impersonal reasons when their theory yields intuitively unattractive implications, the same route should be open for *ex post* contractualists as well.

However, I have misgivings about this ready resort to pluralism. As Frick himself notes, Scanlon's theory itself is already pluralist in a way.[43] He limits his contractualism to the domain of "what we owe to each other." We might call this domain of morality *interpersonal morality* or, following Kamm, *M1*.[44] However, Scanlon deems his theory to exhaust this part of morality. Frick, on the other hand, thinks that contractualism should be assisted by other principles even within this already limited domain. The question then is how valuable contractualism is as a theory above and beyond these other principles. I suspect that it is no longer very valuable. Rather, it seems unacceptably *ad hoc* and gerrymandered to fit a very narrow class of cases. Whenever we look beyond this narrow class of cases and the theory fails to yield the right result, its proponent can resort to pluralism. Leaving the theory open in this way, however, limits its value. It means that the theory has too many free parameters, limiting its predictive power and testability, thereby putting into doubt its value as a standalone moral theory.

Moreover, even granting this pluralistic approach, it will not get *ex ante* contractualism around the implications of the *Glass Box Villain* case in section 2. Utilitarianism here pulls in the same direction as *ex ante* contractualism. All other things equal, the overall aggregate value of an outcome where twenty-five people are spared a 0.5-unit decrease in utility is higher than the value of an outcome where one different person is spared a ten-unit decrease in utility. So even if the misgivings I have with the pluralistic approach are unwarranted, this problem remains.

5. CONCLUSION

In this article I have challenged Johann Frick's *ex ante* contractualism. I argued that adopting the view leads to implications contractualists will find hard to stomach. This has become especially vivid in *Glass Box Villain (Unknown Victim)*, where *ex ante* contractualists are committed to sacrificing one person in order to save twenty-five different people from relatively minor harm. I have argued that this is an instance of the kind of interpersonal aggregation of harms that contractualists sought to avoid in the first place. I also argued that this kind of aggregation is more troublesome than the kind of aggregation Frick accuses *ex post* contractualists of. In connection to this last point I have argued that

43 See Frick, "Contractualism and Social Risk," 220n47.

44 Kamm, *Intricate Ethics*, 455–90.

Frick's argument for the principled priority of identified over unidentified lives also fails, because it can only account for *ex ante* contractualism's verdict in a very narrow class of cases. Finally, I have argued that Frick's resort to pluralism is *ad hoc* and further unable to block some of the unwelcome implications of the view. I conclude that, if there is no other way of developing *ex ante* contractualism that does not run into these problems, contractualists ought to be concerned with the probability that harm could befall someone, rather than with the probability that harm could befall a specific person. For contractualists, a suitably amended *ex post* approach is better equipped to honor this commitment.[45]

Balliol College, University of Oxford
korbinian.rueger@balliol.ox.ac.uk

REFERENCES

Ashford, Elizabeth. "The Demandingness of Scanlon's Contractualism." *Ethics* 113, no. 2 (January 2003): 273–302.

Brock, Dan W., and Daniel Wikler. "Ethical Challenges in Long-Term Funding for HIV/AIDS." *Health Affairs* 28, no. 6 (November/December 2009): 1666–76.

Dougherty, Tom. "Aggregation, Beneficence, and Chance." *Journal of Ethics and Social Philosophy* 7, no. 2 (May 2013): 1–19.

Fleurbaey, Marc, and Alex Voorhoeve. "Decide as You Would with Full Information! An Argument Against *ex ante* Pareto." In *Inequalities in Health: Concepts, Measures, and Ethics*, edited by Nir Eyal, Samia A. Hurst, Ole F. Norheim, and Dan Wikler, 113–29. Oxford: Oxford University Press, 2013.

Frick, Johann. "Contractualism and Social Risk." *Philosophy and Public Affairs* 43, no. 3 (Summer 2015): 175–223.

———. "Treatment versus Prevention in the Fight against HIV/AIDS and the Problem of Identified versus Statistical Lives." In *Identified versus Statistical Lives: An Interdisciplinary Perspective*, edited by I. Glenn Cohen, Norman Daniels, and Nir Eyal, 182–203. Oxford: Oxford University Press, 2015.

Fried, Barbara H. "Can Contractualism Save Us from Aggregation?" *Journal of Ethics* 16, no. 1 (March 2012): 39–66.

45 I thank Ralf Bader, Hilary Greaves, Joe Horton, Kacper Kowalczyk, Benjamin Lange, Jeff McMahan, Johanna Privitera, Tom Sinclair, Bastian Steuwer, and an anonymous reviewer of this journal for helpful comments on earlier versions of this article.

Hare, Caspar. "Should We Wish Well to All?" *Philosophical Review* 125, no. 4 (October 2016): 451–72.

Horton, Joe. "Aggregation, Complaints, and Risk." *Philosophy and Public Affairs* 45, no. 1 (Winter 2017): 54-81.

Jenni, Karen E., and George Loewenstein. "Explaining the 'Identifiable Victim Effect.'" *Journal of Risk and Uncertainty* 14, no. 3 (May/June 1997): 235–57.

Kamm, F. M. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press, 2007.

———. *Morality, Mortality*, vol. 1, *Death and Whom to Save from It*. Oxford: Oxford University Press, 1993.

Kumar, Rahul. "Risking and Wronging." *Philosophy and Public Affairs* 43, no. 1 (Winter 2015): 27–51.

Lewis, C. S. *The Problem of Pain*. New York: Collins/Fontana Books, 1957.

Lewis, David. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, edited by Richard C. Jeffrey, 83–132. Lanham, MD: University Press of America, 1980.

Moore, Randall F. "Caring for Identified versus Statistical Lives: An Evolutionary View of Medical Distributive Justice." *Ethology and Sociobiology* 17, no. 6 (November 1996): 379–401.

Otsuka, Michael. "Risking Life and Limb: How to Discount Harms by Their Improbability." In *Identified versus Statistical Life: An Interdisciplinary Perspective*, edited by I. Glenn Cohen, Norman Daniels, and Nir Eyal, 77–93. Oxford: Oxford University Press, 2015.

———. "Saving Lives, Moral Theory, and the Claims of Individuals." *Philosophy and Public Affairs* 34, no. 2 (March 2006): 109–35.

———. "Scanlon and the Claims of the Many versus the One." *Analysis* 60, no. 3 (July 2000): 288–93.

Rawls, John. *A Theory of Justice*. Oxford: Oxford University Press, 1999.

Reibetanz-Moreau, Sophia. "Contractualism and Aggregation." *Ethics* 108, no. 2 (1998): 296–311.

Scanlon, T. M. "Reply to Zofia Stemplowska." *Journal of Moral Philosophy* 10, no. 4 (2013): 508–14.

———. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press, 1998.

Schelling, Thomas C. "The Life You Save May Be Your Own." In *Problems in Public Expenditure Analysis*, edited by Samuel B. Chase, 127–62. Washington, DC: Brookings Institution, 1968.

Voorhoeve, Alex. "Why One Should Count Only Claims with Which One Can Sympathize." *Public Health Ethics* 10, no. 2 (July 2017): 148–56.

Journal of Ethics and Social Philosophy
VOL. 13, NO. 3 · JULY 2018
https://doi.org/10.26556/jesp.v13i3.367

IS LIBERALISM COMMITTED TO ITS OWN DEMISE?

Hrishikesh Joshi

Are immigration restrictions compatible with liberalism? Recently, Christopher Freiman and Javier Hidalgo have argued that immigration restrictions conflict with the core commitments of liberalism.[1] A society with immigration restrictions in place may well be optimal in some desired respects, but it is not *liberal*, they argue. So if you care about liberalism more deeply than you care about immigration restrictions, you should give up on restrictionism. You cannot hold on to both. I argue here that many restrictions on contractual, economic, and associational liberties seem to be justified by considerations other than liberty—thus the (undischarged) task for Freiman and Hidalgo is to tell us why such restrictions are justified but immigration restrictions are not. Moreover, even if this worry can be addressed, I argue, liberalism is not committed to its own demise in scenarios where there exist large enough numbers of would-be immigrants who accept and endorse illiberal norms in a way that is sufficiently resistant to change. Such a commitment requires thinking of border coercion as violating an *absolute* deontological constraint. This, I contend, is implausible.

1. FREIMAN AND HIDALGO'S ARGUMENT

The argument proceeds as follows. Immigration restrictions involve restricting people's basic liberties. Most fundamentally, they involve restrictions on freedom of movement, which is an important component of basic liberties like freedom of association and freedom of occupation. Primarily this affects would-be migrants. When a would-be migrant is stopped from relocating to another country by threat of coercive force at the border or at the airport, they are thereby forbidden to associate with employers, current and future friends and relatives, etc. In addition, they are stopped from pursuing certain occupational prospects. Immigration restrictions also curtail the freedoms of citizens. Most important-

1 Freiman and Hidalgo, "Liberalism or Immigration Restrictions, but Not Both."

ly, they prevent people from associating with would-be migrants. If you would like to hire somebody who happens to be a citizen of a different country but are unable to procure a work visa for that person, your freedom of association is thereby restricted. The same points hold for friends and relatives whom you would like to associate with on a regular basis in person.

Now, according to liberalism, Freiman and Hidalgo argue, only liberty-based reasons can be adequate for restricting liberty. The state may thus interfere with your freedom of occupation when it comes to your choosing to be a hitman. This is because being a hitman interferes with the basic liberties of others. However, the state may not interfere with your basic liberties for economic or cultural reasons. Thus it may not interfere with your decision to become a painter because you would increase the GDP by a greater amount were you to become something else. Likewise, it may not interfere with your professing Buddhism or teaching Nietzsche if doing so would alter the nation's culture in the long run. Or suppose that Buddhists are having more children on average than non-Buddhists and this is bound to change the culture of the country in the long run. This is not sufficient grounds for the state to interfere with the reproductive liberty of Buddhists within its territory. Of course, such liberty-restricting measures may conceivably arise within a democratic context—the current majority may favor them. Even if they are democratically selected, however, they are not *liberal*. They conflict with liberalism.

Nevertheless, the authors note, the reasons given in favor of immigration restrictions usually appeal to economic or cultural considerations. David Miller, for example, argues that a country's citizens have a right to collective self-determination, and they may thus exclude foreigners so as to promote cultural continuity.[2] Others, for example Stephen Macedo, argue that adverse economic effects on the worst-off members of society are a good reason to limit immigration.[3] Yet, if cultural or economic considerations are not good enough to restrict basic liberty, and if freedom of movement constitutes or is an essential precondition for a basic liberty or liberties, then such arguments proceed from premises that are not consistent with liberalism. In other words, since the offered reasons in favor of immigration restrictions are not liberty-based, liberalism is not consistent with restricting immigration for those reasons.

2 Miller, "Immigration: The Case for Limits" and "Is There a Human Right to Immigrate?"

3 Macedo, "The Moral Dilemma of U.S. Immigration Policy" and "When and Why Should Liberal Democracies Restrict Immigration?"

2. ONLY LIBERTY-BASED REASONS?

While Freiman and Hidalgo take freedom of movement as an important liberty that liberal states protect, they do not specify what it amounts to. Presumably they think a detailed account is unnecessary—there is certainly a sense in which a resident of New York is free to move to Los Angeles in a way that she is not free to move to Vancouver (she will need to go through a visa process). They might be operating under the assumption that this intuitive distinction is all that is required for their argumentative purposes.

The problem is that even liberal states restrict freedom of movement within their borders in important ways. I am not free to enter your property or stay there without your permission. Similarly, the government may decide to disallow the general public from entering a particular national park during caribou-mating season. Yet, intuitively, these types of restrictions are manifestly compatible with liberalism. The question that arises then is: on what conception of freedom of movement will it turn out that immigration restrictions violate liberty but property laws and national parks do not? Further, notice that some people enjoy the freedom of movement to specific areas within liberal polities that others do not. You do not need permission to enter your property, but I do. Rangers or maintenance staff might be allowed to enter the national park during caribou-mating season. So, why are these distinctions unproblematic while the distinction between citizens and legal permanent residents on the one hand and "nonresident aliens" on the other is problematic?

Moreover, it is not obvious that all restrictions on liberty need to have liberty-based reasons according to liberalism. Cigarette taxes restrict your liberty. But the most compelling justifications for such taxes are paternalistic or economic. Occupational licensing laws restrict your freedom of occupation, but the justification for them is the provision of a public good—namely the ability of individuals to trust the medical, legal, and other systems—and to ensure safety standards. Zoning laws restrict your liberty, and the justifications for them often appeal to economic, distributional, and aesthetic considerations. A minimum wage law of $\$x$ per hour limits your freedom of association by forbidding you to hire someone at a rate of less than $\$x$ per hour; in this way, it also restricts would-be employees' freedoms of association and occupation. Similar points can be made about a host of other things. Importantly however, a view according to which liberalism commits us to getting rid of these restrictions is extremely revisionary; hence, relying on such a view would render the authors' argument much less interesting than it appears at first glance. While these are all restrictions on liberty in some sense, one may argue that such measures are not suf-

ficiently drastic to count as violations of *basic liberties*, but on the other hand, immigration restrictions are sufficiently drastic. The task for Freiman and Hidalgo, then, will be to give a characterization of what counts as a violation of basic liberties that is not merely an *ad hoc* construction to support their view. This is brought out, for example, by the fact that minimum-wage laws restrict your freedoms of occupation and association. Freiman and Hidalgo may say that you still enjoy adequate freedom of occupation even if you cannot work for somebody willing to only pay less than $\$x$ per hour. In other words, minimum-wage laws do not interfere with your freedom of occupation *simpliciter*. Rather, they merely impose certain conditions on employment. But notice that the case is similar with immigration—immigration restrictions do not typically restrict anybody's freedom of movement *simpliciter*. Rather, they merely impose the condition that movement into a country's territory must be accompanied by the appropriate visa. Indeed, would-be immigrants are (typically) free to move about within their origin countries as well as any other country that allows them to enter and stay within its territory.[4]

3. IS LIBERALISM COMMITTED TO SUICIDE?

In what follows, I will assume these challenges can be met. Even so, I argue, liberalism can be consistent with (and may even require) immigration restrictions, because there can be strong, forward-looking, liberty-based reasons for some such restrictions.

Consider the two hypothetical countries below:

> LIBERAL DEMOCRACY is a country where people enjoy and support liberal freedoms. There are robust protections for freedom of speech and press.

4 A further problem for Freiman and Hidalgo is that one of the most popular conceptions of basic liberties—the Rawlsian idea—will be of little help in making their case. According to leading Rawls scholar Samuel Freeman, the Rawlsian conception of basic liberties is that they are "an essential social condition for the adequate development and full exercise of the two powers of moral personality over a complete life" (Freeman, *Rawls*, 55). Of the two moral powers, the first is the capacity to "have a rational conception of the good—the power to form, revise, and to rationally pursue a coherent conception of values, as based in a view of what gives life and its pursuits their meaning." The second is the capacity to "understand, apply, and cooperate with others on terms of cooperation that are fair" (Freeman, *Rawls*, 54). The problem with using this view of the basic liberties is that freedom of movement across national borders is typically not needed to develop these powers. You can develop these powers while living in the United States even if you are not free to move to Canada or Brazil or India without meeting residency visa requirements. (This point is developed in Brennan (*Against Democracy*) to argue that political liberties are not basic liberties.)

> Sexual acts between consenting adults of the same sex are not criminalized. There are no public dress codes (with exceptions for nudity). There is robust freedom of religion, and so on. The country has a population of ten million.
>
> THEOCRACY is a democratic country where the overwhelming majority of people do not support liberal freedoms. Blasphemy against the majority religion is punishable by execution. Sexual acts between consenting adults of the same sex are criminalized. Men and women have different laws applicable to them, and the latter are second-class citizens in many ways. Religious minorities are *de facto* persecuted, and defection from the majority religion is officially banned. The country has a population of two hundred million.

Suppose that LIBERAL DEMOCRACY has a GDP per capita that is twenty times higher than that of THEOCRACY, so that many residents of the latter want to move to the former for economic reasons. Economists and social scientists estimate that roughly half the population of THEOCRACY would move to LIBERAL DEMOCRACY within five years if the latter eliminated visa restrictions.

Now, it is a thoroughly empirical question whether, if people from THEOCRACY move to LIBERAL DEMOCRACY in such large numbers, they will keep or change their illiberal norms and beliefs. Let us suppose that norms and cultural beliefs are sticky, and that immigrant communities within LIBERAL DEMOCRACY tend to form homogenous pockets that facilitate and promote the maintenance of antecedent cultural norms. Hence, whether or not immigrants from THEOCRACY to LIBERAL DEMOCRACY will adopt liberal norms and beliefs will partially depend on the numbers that are accepted and the time period over which they are accepted; let us suppose that if one hundred million people were to move within five years, their norms will largely remain intact.

Notice that the question here is not *merely* one of culture. The difference in norms that is relevant here is not merely the difference between language, food, greeting methods, types of holiday celebrations, etc. Rather the difference is between liberal norms and beliefs on the one hand, and illiberal ones on the other.

Let us suppose that if such a movement occurs, various kinds of informal social institutions and norms will shift markedly in the illiberal direction. There will be a "chilling effect" on things like speech, dress, movement, and association. Given the huge population shift, the original residents of LIBERAL DEMOCRACY will become wary of professing atheism or insulting the religion of THEOCRACY. Women might have to change their public behavior in order to avoid severe

harassment. LGBT folk will be pushed to be less open about their sexuality. And so on.

All this can happen without the new immigrants having voting rights. If they are granted voting rights, then given that they outnumber the original population ten to one, they will vote for politicians who will push for laws that resemble the laws of THEOCRACY. Soon enough, LIBERAL DEMOCRACY will no longer be a liberal democracy.

This is a hypothetical scenario, but an important test case for Freiman and Hidalgo's view. Since such a scenario cannot be ruled out *a priori*, and since there are possible worlds in which it is true, we can ask what the demands of liberalism are in this scenario. Is liberalism bound to commit suicide in such cases? That is, does liberalism commit us to policies that would, under certain circumstances, foreseeably eliminate its existence?

Such a consequence seems implausible for two chief reasons. For one, it seems that liberal societies are intrinsically valuable given the relationships between coresidents that they embody. Liberal societies are also instrumentally valuable insofar as they promote certain kinds of cultural and scientific achievements, given the ability of individuals to speak, think, and associate in a relatively free way. They also stand as a model for other, less liberal societies to emulate.

Second, there seem to be liberty-based reasons to restrict the freedom of movement of people from THEOCRACY seeking to migrate to LIBERAL DEMOCRACY—namely that doing so will preserve the liberties that citizens of LIBERAL DEMOCRACY enjoy. The aim of maintaining and promoting the existence of liberal polities is a liberty-based aim.

Freiman and Hidalgo may worry that the kind of reasoning sketched here will also motivate other restrictions on liberty that are intuitively at odds with liberalism. Thus suppose that, within LIBERAL DEMOCRACY, there exists an illiberal minority that is growing in influence and number. Would it be consistent with liberalism for the state to restrict their freedoms of speech and association, or seek to deport this group to a country that will accept them in exchange for aid, for example?

There are two main things to be said in response here. First, some measures are greater violations of individual liberty and autonomy than others, and thus require a greater burden of justification. Sin taxes are restrictions on liberty but stand in need of less justification than restrictions on freedom of speech, for example. Likewise, while granting that border controls involve restricting would-be migrants' liberty, such controls stand in need of less justification than moves to deport long-term residents. Second, some restrictions on liberty, even if they are to be tolerated in principle, are more prone to practical difficulties than oth-

ers. Controlling immigration has less potential for institutional slippery slopes than controlling the speech of even extremely illiberal elements. Restrictions on speech, in other words, are ripe for abuse in a way that legal restrictions on immigration are not.[5]

Freiman and Hidalgo might bite the bullet here and contend that even in the hypothetical situation described above, it is illiberal for LIBERAL DEMOCRACY to seek to impose immigration restrictions. For, they may argue, the demands of liberalism take the form of absolute deontological constraints. Therefore, even if open borders between the two countries will foreseeably end the existence of liberal democracy, so be it. Let justice be done though the heavens fall, as the saying goes.

The problem is that plausible absolute deontological prohibitions are very rare (if they exist at all).[6] Hence, many deontological theorists are not absolutists. W. D. Ross's theory, for example, allows for obligations (*prima facie* duties) to be outweighed by sufficiently weighty considerations.[7] As an illustration, consider the commonly acknowledged deontological prohibition against breaking promises. This might allow us to say that if on the day of a promised lunch, even if I calculate that my staying home would result in +11 utils, whereas my fulfilling the promise would result in +10 utils, I may not break the promise. Yet, if breaking a promise is the only way to stop a murder, it is permissible, and depending on the circumstance even obligatory, to break the promise.[8]

Now, it is highly implausible that restrictions on the freedom to move across international borders are barred by *absolute* deontological prohibitions. Indeed, if stopping one person at the border who does not have the required visa is the only way to stop a nuclear catastrophe from killing one million people, the border stopping is the right thing to do.

But perhaps, more weakly and plausibly, there is a *prima facie* duty not to restrict such freedoms. The task for Freiman and Hidalgo, then, is to argue that the foreseeable end of liberal democracy in the hypothetical scenario is not an evil weighty enough to warrant restricting the freedom of movement for those

5 The challenges I have in mind include, but are not limited to, the sorts of worries raised famously in Mill, *On Liberty and Other Essays*.

6 Immanuel Kant famously thought that you should not lie even to prevent a murder, but I take it that most modern ethicists believe this is quite implausible. However, there is some dispute as to whether he is committed to this; see Korsgaard, "The Right to Lie."

7 Ross, *The Right and the Good*.

8 A further issue is that absolute prohibitions can lead to a proliferation of ethical dilemmas. If there is an absolute deontological prohibition against breaking promises, but the only way you can fulfill promise A is by breaking promise B, what should you do?

in THEOCRACY. Given the intrinsic and instrumental value of liberal societies mentioned above, it is not obvious that they can succeed.

Now, in the hypothetical scenario, I have assumed that LIBERAL DEMOCRACY will definitely come to resemble THEOCRACY in its norms and institutions. But what if, given the best evidence, this is not a certainty, but a mere (sizable) risk? While a consequentialist would just perform a cost-benefit analysis, does weakening the assumption pose a special problem for someone who accepts deontological restrictions against border coercion? Plausibly not: for deontological theories can take risks into account when determining whether some *prima facie* duty is overridden. Indeed, many freedoms are rightly restricted because of the risks involved with respect to the liberties of others—even if the bad outcome is not certain to come about. For example, consider the restrictions on drunk driving, entering airports, pollution, gun ownership, etc.

If this section's argument succeeds, then the question of whether liberalism is committed to open borders turns on the empirical question of whether the actual world sufficiently resembles the hypothetical scenario. I do not wish to delve into that empirical question here. However, if I am right, the connection between the core commitments of liberalism and immigration policy hinge on empirical issues to a much greater degree than the authors appreciate.[9]

University of Michigan, Ann Arbor
joshih@umich.edu

REFERENCES

Brennan, Jason. *Against Democracy*. Princeton: Princeton University Press, 2016.

Freeman, Samuel. *Rawls*. London: Routledge Press, 2007.

Freiman, Christopher, and Javier Hidalgo. "Liberalism or Immigration Restrictions, but Not Both." *Journal of Ethics and Social Philosophy* 10, no. 2 (May 2016).

Korsgaard, Christine. "The Right to Lie: Kant on Dealing with Evil." *Philosophy and Public Affairs* 15, no. 4 (Autumn 1986): 325–49.

Macedo, Stephen. "The Moral Dilemma of U.S. Immigration Policy: Open Borders Versus Social Justice?" In *Debating Immigration*, edited by Carol M. Swain, 63–81. Cambridge: Cambridge University Press, 2007.

———. "When and Why Should Liberal Democracies Restrict Immigration?"

9 Thanks to Jonathan Anomaly and Daniel Jacobson for extensive comments on earlier drafts, and to the editors and an anonymous reviewer at the *Journal of Ethics and Social Philosophy*.

In *Citizenship, Borders, and Human Needs*, edited by Rogers M. Smith, 301–23. Philadelphia: University of Pennsylvania Press, 2011.

Mill, John Stuart. *On Liberty and Other Essays*. Edited by John Gray. Oxford: Oxford University Press, 1991.

Miller, David. "Immigration: The Case for Limits." In *Contemporary Debates in Applied Ethics*, edited by Andrew I. Cohen and Christopher Heath Wellman, 193–206. Malden, MA: Blackwell, 2005.

———. "Is There a Human Right to Immigrate?" In *Migration in Political Theory: The Ethics of Movement and Membership*, edited by Sarah Fine and Lea Ypi, 11–31. Oxford: Oxford University Press, 2016.

Ross, W. D. *The Right and the Good*. Oxford: Oxford University Press, 1930.