JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XII · NUMBER 2
November 2017

OPEN ACCESS FREE FOR AUTHORS AND READERS ISSN 1559-3061

JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XII · NUMBER 2
November 2017

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. 12, NO. 2 · NOVEMBER 2017
https://doi.org/10.26556/jesp.v12i2.223

MORAL UNCERTAINTY ABOUT POPULATION AXIOLOGY

Hilary Greaves and Toby Ord

POPULATION ETHICS is the study of the unique ethical issues that arise when one's actions can change who will come into existence: actions that lead to additional people being born, fewer people being born, or different people being born. The most obvious cases are those of an individual deciding whether to have a child, or of society setting the social policies surrounding procreation. However, issues of population ethics come up much more widely than this. How bad is it if climate change reduces the planet's "carrying capacity"? How important is it to lower the risks of human extinction? How important is it, if at all, that humanity eventually seeks a future beyond Earth, allowing a much greater population?

An important part of any plausible ethical theory, consequentialist or otherwise, is its axiology: its ranking of states of affairs in terms of better and worse overall, or (if cardinal information is also present) its assignment of *values* to states of affairs. The two most famous approaches to population axiology are the Total View and the Average View. The Total View says that the value of a state of affairs is the *sum* of the well-being of everyone in it—past, present, and future. The Average View instead holds that the value is the *average* lifetime well-being of everyone in it. These views agree when the size of the (timeless) population is fixed, but can disagree when comparing larger and smaller populations. Other things being equal, the Total View suggests that the continuation and expansion of humanity are extremely important, while according to the Average View, they are matters of relative indifference.

In *Reasons and Persons*, Parfit showed that the Total View leads to a conclusion many find troubling (the "Repugnant Conclusion"): that for any world, even one with billions of very well-off people, there is a better world (with far more people) in which no individual has a life that is more than barely worth living.[1]

Much of the history of population ethics since then has been an attempt to develop axiologies that avoid the Repugnant Conclusion. However, a series of

1 Parfit, *Reasons and Persons*, pt. 4, ch. 17.

impossibility theorems has shown that the only way to avoid this is to take on other counterintuitive implications, be they formal problems (like cyclic betterness orderings) or substantive problems (like preferring adding people with negative well-being to adding people with positive well-being).[2] In this situation, the reaction of any honest inquirer has to be one of *uncertainty* about population axiology. How, then, are we to decide what to do in the many domains in which our actions may change the population?[3]

One approach would be to press on with the philosophical work, better understand the available options, and attempt to resolve the moral uncertainty. We certainly approve of this approach, but progress will not be instantaneous, and in many cases immediate decisions are required: the question remains of how to decide what to do while we do still have uncertainty.

We could look more carefully at the real-world questions that concern us, and see if there is agreement between the theories we are considering. For example, we might note that, since living standards have improved over the centuries, the Average View might not be indifferent to continued human existence after all. Even if living standards stopped improving now, additional generations at this level would continue to bring up the timeless average. In this way, we might be in a position of knowing which acts are better despite our uncertainty over the underlying evaluative theory (and hence over precisely *why* those acts are better than the alternatives). This scenario certainly simplifies matters when it arises, but not all of the practical questions we face have this convenient feature.

Our problem can be formalized into the question of *axiological uncertainty*: given a set of available options, and credences in each of a set of axiologies that disagree among themselves about the values of those options, how should one choose?

At least when one's relevant moral uncertainty is restricted to the domain of axiology, the answer to this question will involve a rule for identifying one's *effective axiology*: the axiology that one should use for guiding decisions, in whatever way one should generally use an axiology for guiding decisions (maximizing, satisficing, maximizing subject to certain side constraints, or whatever).[4] The

2 Parfit, *Reasons and Persons*, pt. 4, ch. 19; Ng, "What Should We Do about Future Generations?"; Carlson, "Mere Addition and Two Trilemmas of Population Ethics"; Kitcher, "Parfit's Puzzle"; Arrhenius, "An Impossibility Theorem for Welfarist Axiologies" and "Population Ethics," ch. 11.

3 Similar questions occur in the context of group decision-making in the presence of interpersonal disagreement. The approach we will explore in this paper could also be applied in that context.

4 Matters are more complex in the more general case, in which one's normative uncertainty extends to both the axiological and the non-axiological parts of normative theory. It is a sub-

question then becomes: how is one's effective axiology related to the various first-order axiologies in which one has nonzero credence?

The general literature on moral uncertainty suggests four approaches to answering this question. The first approach ignores the agent's credences (and beliefs), and says that the effective axiology is simply the *true* axiology, no matter that the agent is in no position to know which this is.[5] This is a singularly unhelpful answer to people who find themselves in this predicament, but its proponents argue that it is the most one can say.

A second approach says that the effective axiology is the one in which the agent has highest credence. This is the "My Favourite Theory" approach.[6] This approach sounds initially intuitive, but has several deeply unsatisfactory features. (1) It gives very counterintuitive results if there are many theories under consideration and the agent's highest credence is low. For example, if the agent has a credence of 10 percent in her favorite axiology, then this approach to moral uncertainty may lead her to select an option that she is 90 percent sure is much worse, when there was a rival option she was 90 percent sure was much better. (2) It gives the agent no reason to be interested in finding out what the other theories say, even if she has only slightly less credence in them, and thus cannot capture the intuition toward seeking out options that have broader support. (3) It is well defined only relative to some privileged way of individuating theories, but it is unlikely that there is any such privileged individuation.

A third approach appeals to a notion of all-out belief, as opposed to credence: the effective axiology is the one that the agent *believes*. This theory inherits the third of the above problems with the "My Favourite Theory" approach; in addition, in any case involving significant axiological uncertainty, there is unlikely to be any axiology that that agent all-out *believes*, in which case this third approach is simply silent on what one is to do.

This brings us to a fourth approach: to use the same approach to axiological uncertainty that we use for empirical uncertainty, i.e., use an effective axiology that corresponds to the ordering of alternatives according to their *expected value*. This approach ranks options on the basis of the breadth of support across different theories (weighted by how likely those theories are), and also on the basis of

stantive question whether or not, in that general case, anything like an "effective axiology" plays a role in appropriate choice under normative uncertainty. In this paper, we set these more complex issues aside and focus on clarifying the simpler case.

5 Harman, "Does Moral Ignorance Exculpate?"; Weatherson, "Running Risks Morally"; Mason, "Moral Ignorance and Blameworthiness."

6 Gracely, "On the Noncomparability of Judgments Made by Different Ethical Theories"; Lockhart, *Moral Uncertainty and Its Consequences*, 58–59; Gustafsson and Torpman, "In Defence of My Favourite Theory."

how much each theory considers to be at stake. For instance, even if 60 percent of the agent's credence is in theories that judge *A* to be slightly superior to *B*, if the remaining theories find *A* to be vastly worse, this could lower the expected moral value of *A* enough that the effective axiology ranks *B* above *A*.

In this paper, we will focus on this fourth alternative: the "expected moral value" (EMV) approach to axiological uncertainty. In part this is because it is obvious what the other three approaches canvassed above recommend. But it is also because we find EMV to be a very plausible approach to axiological uncertainty (just as its analogue is for empirical uncertainty)—both intrinsically and because the problems for the alternative approaches strike us as serious.

What we will argue is that the EMV approach to axiological uncertainty implies, in a sense that we will make precise, that in certain large-population limits the effective ranking of certain (potentially important) alternative pairs under population-axiological uncertainty coincides with that of the Total View, even if one's credence in the Total View is arbitrarily low, and even if most of the alternative theories generate the opposite ranking of the alternatives under consideration.[7] Readers who start out unsympathetic both to EMV as an approach to moral uncertainty and to the Total View as a first-order population axiology may be inclined to read this as a further *reductio* of EMV; we have some sympathy with this reaction, and we discuss the extent to which it is reasonable in section 8.

The remainder of the paper proceeds as follows. While we seek to analyze the most general case of population-axiological uncertainty that we can, a fully general treatment lies beyond the scope of the present paper: for tractability, we will be restricting attention to axiologies that are in specifiable senses mathematically well behaved. Section 1 flags the restrictions in question.

The biggest challenge for the EMV approach is in determining how the moral stakes on one theory line up with those on another. This is known as the *problem of intertheoretic comparisons*. Section 2 surveys the possible solutions to this problem; our own approach will be neutral between these solutions, requiring rejection only of the skeptical position according to which intertheoretic comparisons are impossible.

Section 3 highlights the fact, crucial to our later analysis, that according to the EMV approach the effective ranking of alternatives depends not only on the

7 Technically: with a Critical Level view, not the Total View itself. We defer discussion of this relative subtlety until section 5.

We do not, of course, claim that there are no situations in which the stakes are much higher on other views than on the Total View, so that it is the Total View that gets "overpowered" on the EMV approach. For some such examples, see Temkin, *Rethinking the Good*, 441–45. Our claim concerns specifically the "large-population limit" constructions we discuss, a class of constructions that seems to us particularly important.

agent's credences in the various possible axiologies, but also on whether some axiologies judge there to be *more at stake* in the decision situation under consideration than other theories do. Existing work on moral uncertainty recognizes the resulting possibility that, in some cases, what one ought to do under uncertainty can reliably track what is recommended by some particular theory even when one's credence in that theory is relatively low. The key theme of our subsequent analysis is that something like this might systematically happen in population ethics. When it does, we say that the theory that carries the day for practical purposes, despite the agent's low credence in that theory, "overpowers" the rival theories.

Section 4 turns to the detailed investigation of the case of population axiology. We analyze three scenarios: (1) adding a single extra person; (2) taking some risky action that improves well-being for presently existing people but increases the risk of human extinction in the near future; (3) making some sacrifice in the well-being of present earthbound humans in order to send expensive missions to seed new human civilizations on other planets. In all three types of case, we identify a precise sense in which, "in the limit of large populations," and for an agent whose credences are split between a specified (but quite wide) range of population axiologies but who has nonzero credence in the Total View, the alternative with the higher expected moral value is the one that is preferred by the Total View, despite the fact that it remains dispreferred by many rival theories.

Section 5 develops one minor refinement to the claims of section 4. The Total View is one member of a more general family of population axiologies, the "Critical Level" family. When the class of population axiologies under consideration also includes other members of this family, in general the axiology that overpowers others in large-population limits is not necessarily the Total View itself, but may be some other member of this family. This refinement, however, is unlikely significantly to alter the practical import of our conclusions. (This section is more technical than the remainder of the paper, and may be skipped by readers who are interested only in the broader features of our argument.)

Section 6 takes on the question of whether, granted that this overpowering occurs in a theoretical large-population limit, the overpowering will actually occur in practice: that is, are the population sizes that are actually involved in empirically realistic versions of our scenarios sufficiently large? The issues here are somewhat complex, both because the relevant empirical parameters are themselves very uncertain, and because the manner in which one settles questions of intertheoretic comparisons will make a difference. However, reasonable back-of-the-envelope calculations suggest that it is at least very plausible that the overpowering we discuss may actually occur.

Section 7 notes that, for very similar reasons, the EMV approach to axiological uncertainty is committed to analogs of some versions of the notorious Repugnant Conclusion. Section 8 takes up the (related) question of whether one might take the overpowering results we have discussed as *reductios* of the EMV approach to moral uncertainty. Section 9 is the conclusion.

1. RESTRICTIONS TO OUR ANALYSIS

In this paper, we use some important simplifying assumptions. First, we restrict our attention to population *axiology*: comparisons of states of affairs (possibly involving different populations) in terms of overall betterness. That is, we are focused on evaluative questions such as whether it would be better to have a larger population so long as the total well-being goes up, rather than directly on deontic questions of what one ought to do or choose. (Similarly, the Total View and Average View that we discuss are not average and total *utilitarianism*, in the sense that they are only theories of the good; they say nothing about whether one *ought* to *maximize* goodness, or instead satisfice, maximize subject to side constraints, or anything else.) Importantly, this does not involve any assumption that axiology is the full moral story. Most approaches to morality, consequentialist or otherwise, hold that considerations of overall betterness are at least *one important part* of the full story, and would thus agree that it is worth working out what that part looks like.[8]

Second, we focus on axiologies that give cardinal values for these comparisons, such that we can ask how many times bigger the value difference between outcomes *A* and *B* is than the value difference between outcomes *C* and *D*. This rules out merely ordinal axiologies, but in practice it includes all the main axiologies under discussion in population ethics.

Third, we set aside theories in which the betterness relation is incomplete or cyclic. While we have some sympathy with theories involving incomplete betterness, they introduce a number of choices for how to fit them into a theory of axiological uncertainty, and substantially complicate the analysis.[9] Unlike the earlier ones, this assumption *is* a moderately large restriction in practice: the approaches of, e.g., Bader, Heyd, and Temkin lie outside the scope of our discussion.[10]

8 The point is made forcefully by Rawls, himself no consequentialist: "All ethical doctrines worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy." Rawls, *A Theory of Justice*, 30.

9 See, e.g., MacAskill, "The Infectiousness of Nihilism."

10 Bader, "Neutrality and Conditional Goodness"; Heyd, "Procreation and Value: Can Ethics

Finally, we set aside theories that violate axiological invariance: the requirement that the value of a state of affairs is independent of which state of affairs is actual. This principle is violated by "actualist" theories.[11] Including such theories in our analysis would be straightforward in principle and would not change our qualitative result, but it would complicate the analysis.

We are thus restricting our attention to theories of population ethics that are mathematically quite well behaved. This is a serious restriction to our analysis: clearly, any fully general treatment of axiological uncertainty will also have to say what one should do when one has nonzero credence (as one plausibly should) in some "badly behaved" theories, and will therefore have to address the deeper problems that are discussed by, e.g., MacAskill.[12] The motivation for our restriction is pragmatic: we have very little idea of how to develop a plausible theory of axiological uncertainty for the fully general case, and in the meantime it seems worth working out what can be said about the more tractable cases.

2. THE PROBLEM OF INTERTHEORETIC COMPARISONS

2.1. Skepticism about Intertheoretic Comparisons?

To construct an effective axiology on the EMV approach, we need to be able to compute, for any pair of alternatives—A and B—whether the difference in expected moral value $\text{EMV}(B) - \text{EMV}(A)$ is positive or negative: the EMV ordering ranks B above A iff this difference is positive. But that requires that we have a meaningful notion of averaging the value differences between A and B according to rival axiologies; this in turn effectively requires that rival axiologies use the *same* scale of possible value differences. How, though, is the value scale postulated by one axiology to be compared to that postulated by another?

Several authors have claimed that no such "intertheoretic comparisons" exist.[13] The source of the worry is that, at least on the face of it, the moral theories themselves do not contain any resources that could determine how the value differences between pairs of alternatives according to one theory compare to those according to a different theory. Suppose, for example, that A and B are alternative possible populations as follows:

Deal with Futurity Problems?"; and Temkin, "Intransitivity and the Mere Addition Paradox" and *Rethinking the Good*.

11 Bigelow and Pargetter, "Morality, Potential Persons and Abortion"; Warren, "Do Potential People Have Moral Rights?"; Arrhenius, "Population Ethics," ch. 10, sec. 3.

12 MacAskill, "The Infectiousness of Nihilism."

13 Hudson, "Subjectivization in Ethics," 224; Gracely, "On the Noncomparability of Judgments Made by Different Ethical Theories"; Broome, *Climate Matters*, 185.

	Average Well-Being	Population Size	Total Well-Being
A	50	4	200
B	25	16	400

In this example, one might naively think, for an agent who has credence one-half in each of the Total View and Average View, that the difference in expected moral value between alternatives *A* and *B* is given by

$$\text{EMV}(B) - \text{EMV}(A) = \tfrac{1}{2} \times (25 - 50) + \tfrac{1}{2} \times (400 - 200) > 0,$$

in which case the effective axiology ranks *B* above *A*. However, if the only facts there are are restricted to what the rival views each *separately* say about (i) the ordering of alternatives and (ii) the ratios of such value differences between alternatives, then we have freedom to rescale each axiology's value function by a *separate* positive linear transformation. We might just as well, for instance, have represented the Average View by means of a value function according to which $V(A) = 50$ million and $V(B) = 25$ million (while still using the values 200 and 400 respectively for the Total View's values); but doing so would, of course, have reversed the result of the above calculation.

If there are no constraints on the scaling of one axiology's value function relative to another's, then the EMV approach to axiological uncertainty is doomed. The subsequent analysis in our paper will require that we have rejected this condition. The relevant facts cannot be restricted to the categories (i) and (ii) in the previous paragraph. Next, we briefly survey the space of remaining possibilities.

2.2. Three Non-Skeptical Approaches

There are three more positive approaches to the issue of intertheoretic comparisons.[14]

The first approach is *content-based*.[15] This approach is available when (as is sometimes, but not always, the case) there is some significant subset of alternatives such that the two theories in question agree on all ratios of value differences regarding pairs of alternatives in the privileged subset. In that case, there may be grounds (based on the content of the theories) for having unit intertheoretic comparisons on the region of overlap; this requirement, together with the existing intratheoretic structure within each theory, then determines the intertheo-

14 Our taxonomy follows MacAskill, "Normative Uncertainty," ch. 4. That chapter also contains a concise survey of the various problems that each approach faces.

15 See, e.g., Ross, "Rejecting Ethical Deflationism," 764–65; Sepielli, "What to Do When You Don't Know What to Do," pts. 4 and 5.

retic comparisons elsewhere. As an example, consider someone whose credence is split between the Total View on the one hand, and a presentist, person-affecting view on the other. The latter view is one way of trying to flesh out the intuition that "we are in favor of making people happy, but neutral about making happy people": on this view, only people who presently exist at the time of the decision count from a moral point of view.[16] There appears to be a natural way of comparing values between these theories, as it seems they agree about the nature of value, but disagree about the bearers of value. One could set the value of a unit of well-being in a person's life according to the Total View to be equal to the value of a unit of well-being in a presently existing person's life according to the presentist theory. The two theories would then agree on the intrinsic value of (say) improving the health or lengthening the life of an already existing person, but the Total View would hold that it is ten times as valuable to improve the lives of ten future people by a given amount than it is to improve the life of one present person by that same amount, while the presentist theory would hold that improving the lives of future persons generates no gain in value at all.

The second approach is *structure based*. This approach seeks a way of normalizing theories against one another that is "purely structural" in the sense that, unlike the first approach just mentioned, it does not attribute any significance to the *content* of an alternative, but utilizes only the ratios of value differences postulated by the theories to be ranked. The most commonly discussed normalization rule in this family is the "zero-one" or "range normalization" method, according to which the value difference between the best and worst alternative is the same for each theory.[17] Cotton-Barratt, MacAskill, and Ord have recently argued for the superiority of an alternative "variance normalisation" approach over others in the structuralist family, in part (but not only) because range normalization is defined only for bounded value functions.[18] One key decision point for such a "structural" approach is whether, for the purpose of a particular choice situation, to normalize the range of values of the options in that choice situation, or to normalize it across a broader set of options, such as all possible options. The former has the formal problem of choice-set dependence, while the latter is difficult to precisely define. Herein lie the disadvantages of the structural approach; its advantage over the content-based approach, meanwhile, is that it remains available

16 Narveson, "Moral Problems of Population," 80.

17 For example, the "principle of equity among moral theories" used in Lockhart, *Moral Uncertainty and Its Consequences*, 84.

18 Cotton-Barratt, MacAskill, and Ord, "Normative Uncertainty, Intertheoretic Comparisons, and Variance Normalisation."

even when comparing theories that are so radically different that the common ground required by the content-based approach does not exist.

The third approach is the "universal scale" approach.[19] This approach does not in itself provide an answer to the question of how to settle intertheoretic comparisons in particular cases, but it does provide a reply to the worry that any such comparisons must be "meaningless." On this approach, individual moral theories (initial appearances perhaps aside) do after all assign moral values to alternatives *on a scale that already has intertheoretic validity*; there are pairs of theories that are genuinely distinct but that agree with one another on all ratios of value differences between alternatives. A particular version of the Total View, for example, might say that the value difference between *A* and *B* (in our above example) is three times as large as that posited by a particular version of the Average View; but different versions of the two views would generate different intertheoretic comparisons. In addition to having credences in the Total View and the Average View as theory families, a rational agent has credences distributed in some particular way among the infinitely many possible particular theories within each family, and these latter credences give rise to this agent's effective views on intertheoretic comparisons.

It is also worth noting the possibility of *subjectivism* about intertheoretic comparisons.[20] This is an analogue of subjectivism about credences: subjective Bayesians hold that each agent is rationally required to have settled (somehow) on some credence function, but that there is a wide range of rationally permissible credence functions, and no rules or guidelines to direct the choice among them. In the context of intertheoretic comparisons, the analogous view holds that each agent is rationally required to have settled (somehow) on some standard of intertheoretic comparisons, but there is a wide range of rationally permissible such standards (including, but certainly not restricted to, the ones that correspond to some reasonably natural content-based or structuralist approach),

19 See MacAskill, "Normative Uncertainty," ch. 4; Riedener, "A Theory of Axiological Uncertainty," sec. 3.4.

20 See, e.g., Ross, "Rejecting Ethical Deflationism," 763–64; Riedener, "A Theory of Axiological Uncertainty." Subjectivism is in the first instance a view about rational permissibility, while the content-based, structure-based, and universal-scale approaches discussed above are views about the metaphysics of intertheoretic comparisons. Subjectivism is naturally understood as a supplement to the universal-scale approach: the content-based and structure-based approaches both (*qua* metaphysical views) imply that there is a *unique metaphysically correct* way of drawing intertheoretic comparisons in any given case, and so would presumably give rise to correspondingly unique rational requirements (in conflict with subjectivism). Note, though, that neither the subjectivist nor the universal-scale advocate needs object to elements of the content-based and structure-based approaches being used to shape agents' beliefs about intertheoretic comparisons.

and no rules or guidelines to direct the choice among them. (Riedener provides a representation theorem for the case of axiological uncertainty, analogous to the theorems of expected utility theory for empirical uncertainty.[21]) The significance is that if subjectivism is true, then there can be intertheoretic comparisons (in the required sense) even in the absence of any defensible general proposal for the grounds of "correctness" for intertheoretic comparisons.

Our subsequent discussion will assume that some such positive view is correct, but (with the exception of section 6) will be largely neutral as to which.

3. The Importance of Relative Stakes

A key tenet of the EMV approach is the idea that, in a particular decision situation, if one moral theory holds that there is a lot at stake while rival theories regard relatively little as being at stake, then one should sway one's ranking of alternatives toward that recommended by the "high-stakes" theory, relative to what one might expect based on one's credences alone. For instance, if one has equal credence in two theories and those two theories disagree as to which of two given alternatives is better, then one should choose according to the theory that regards this particular choice as being higher stakes. For another type of example, sometimes one should follow the dictates of a theory in which one has relatively *low* credence, even when that theory disagrees with all other theories in which one has nonzero credence on the relative ranking of two particular alternatives—if the low-credence theory alone regards the choice between this particular pair of alternatives as being high stakes.[22]

This is, of course, all analogous to the verdicts of ordinary, expected utility theory on cases of empirical uncertainty. One should not accept a gamble according to which one gains £10 if the fair coin lands heads but loses £1,000 if it lands tails, despite the fact that one has equal credences that one would win or lose such a bet. And under at least some circumstances, one should take precautions even against events that one considers to be relatively unlikely: one's credence that one's bike would be stolen on any given day if one neglected to lock it up outside one's office, for instance, is probably less than 5 percent, but still one locks it, since it costs much less to turn the key than it would to lose the bike.

21 Riedener, "A Theory of Axiological Uncertainty."

22 Most (non-skeptical) approaches to intertheoretic comparisons permit such differences in stakes across theories. Exceptions include maximally "narrow" implementations of structural normalization, according to which, for the purpose of comparing two given alternatives, the set of options whose value ranges (or variances, etc.) are to be equalized contains only those two alternatives.

This possibility of one theory's overpowering another within the EMV approach, on grounds of differential stakes and beyond the point that one would expect on grounds of credence alone, has received some limited discussion in the literature on moral certainty. Most obviously, as Ross and MacAskill have both noted, if a "uniform" theory is one according to which every alternative is equally as good as every other alternative, the ranking of alternatives by expected moral value depends only on one's relative credences in nonuniform theories.[23] One's credence, if any, in the uniform theory has no effect. Even if one has credence 0.999, say, in a uniform theory, with the remaining 0.001 credence distributed equally between two nonuniform theories T_1 and T_2, one's EMV ranking of alternatives will be identical to the ranking that one would have if one had credence one-half in each of T_1 and T_2, and zero credence in the uniform theory. In this sense, except in the extreme case of credence 1 in the uniform theory, nonuniform theories overpower uniform theories.

This phenomenon of *total silencing* of one theory by others on grounds of relative stakes is an extreme case. More commonly, but more messily, similar things can occur when one theory judges that the amount at stake is *much less* than other theories judge. For the simplest instance of this, suppose that one starts with two rival theories (T_1 and T_2) and a relatively natural construal of the intertheoretic comparisons between them, but then decides that the version of T_2 in which one actually has nonzero credence is a "hysterical" theory, one that deems *everything* one million times more important than the "natural" version did. (This particular description, of course, makes sense only on the universal-scale approach to intertheoretic comparisons, since any strict content- or structure-based approach would leave no freedom for such "rescaling.") In that case, *for fixed relative credences in* T_1 and T_2, T_2 will now contribute one million times more to the relevant expected value calculations than it did previously, and may thereby overpower T_1. In this simple instance, however, the overpowering is easily avoided simply by having very low (but not necessarily zero) credence in such "hysterical" theories, a move that independently seems quite reasonable.[24]

The project of this paper is to explore a more subtle instantiation of the phenomenon of overpowering via extreme relative stakes, in the specific context of population ethics. Section 4 begins this task by analyzing three scenarios of distinct structures, and considering the results of applying EMV when credences are split between a fairly wide family of population axiologies (subject to the limitations noted in section 1, above).

23 Ross, "Rejecting Ethical Deflationism"; and MacAskill, "The Infectiousness of Nihilism."

24 Ross, "Rejecting Ethical Deflationism," 766.

4. SCENARIOS

4.1. Preliminaries

To understand better how the changes in relative stakes can affect decisions under uncertainty, we explore three hypothetical scenarios, concerning (1) mere additions, (2) extinction risk, and (3) space settlement. A general theme we follow is that, as the scenarios involve more and more people (in a sense that can be made precise on a case-by-case basis), the Total View ascribes the choice a higher relative weight, eventually coming to dominate the ranking of actions according to the EMV view of axiological uncertainty, regardless of one's credence in the Total View (provided only that it is nonzero) and regardless of how the intertheoretic comparisons have been fixed.

We use the following notation. For an arbitrary population X, let $|X|$ be the number of people in X, and let $\overline{X}$ be the average well-being level in X. In this notation, the total well-being in X is $\overline{X}|X|$. For an arbitrary population X and natural number n, write nX for the population that consists of "n copies of X" (that is, for every well-being level w, if X contains exactly m people at well-being level w, then nX contains exactly nm people at well-being level w).

4.2. Axiologies under Consideration

Using the notation above, we can easily compare a number of extant population axiologies.[25] As we shall see, most of these involve calculating the product of some form of an average well-being with some form of the number of people, producing something akin to a total well-being.

In our notation, the Total View and Average View are represented by the following value functions:

Total: $V(X) = \overline{X}|X|$

Average: $V(X) = \overline{X}$

We also consider two types of a Variable Value View, in which there is a kind of diminishing marginal value in creating extra people (hence the value of adding

25 Our list includes every actually advocated theory we are aware of that is both (i) sufficiently precisely specified for us to know what the corresponding value function is, and (ii) consistent with the structural limitations that we laid out in section 2. While we do not explicitly discuss it here, our results also hold for Geometrism (Sider, "Might Theory *X* Be a Theory of Diminishing Marginal Value?")—a theory that was described but never seriously advocated.

a particular life can vary). These are from Hurka, and correspond respectively to his theories "V1" and "V2":[26]

Variable Value I:	$V(X) = \overline{X}g(\lvert X\rvert)$	where g is a strictly increasing and strictly concave function with a horizontal asymptote
Variable Value II:	$V(X) = f(\overline{X})g(\lvert X\rvert)$	where f and g are strictly increasing and strictly concave functions and g has a horizontal asymptote

We then consider two "person-affecting" views which attempt to cash out the intuition that "we are in favor of making people happy, but neutral about making happy people."[27] Presentism is the view that only past and present people matter morally: people who will come into existence in the future are considered to have no moral value at the time a decision is made.[28] Necessitarianism is the view that only people who will exist regardless of the choice one is currently making matter from a moral point of view.[29] Assuming that these theories further take the value of the state of affairs to be the *sum* of the well-being of all people who have moral value, these theories are represented respectively by the following value functions:[30]

Presentism:	$V(X) = \overline{P}\lvert P\rvert$	where P is all people in X who presently exist
Necessitarianism:	$V(X) = \overline{N}\lvert N\rvert$	where N is all people in X who exist in all alternatives

Finally, we eventually also consider the Critical Level family of views that has been defended by Broome and by Blackorby, Bossert, and Donaldson:[31]

Critical Level:	$V(X) = (\overline{X} - a)\lvert X\rvert$	where a is a specific well-being level

26 Hurka, "Value and Population Size," 502–4. Note that Variable Value I is identical to the view Ng ("What Should We Do about Future Generations?") calls "Theory X'."

27 Narveson, "Moral Problems of Population."

28 Arrhenius, "Population Ethics," ch. 10, sec. 1.

29 Singer, *Practical Ethics*, 103–4; Arrhenius, "Population Ethics," ch. 10, sec. 2.

30 Including other versions of the Presentist and/or Necessitarian views would further complicate our analysis, but we are not aware of any extant (or at all plausible) precisification that would alter our qualitative conclusions.

31 Broome, *Weighing Lives*; Blackorby, Bossert, and Donaldson, "Intertemporal Population Ethics."

This theory says that the value of adding an extra person to the world, if it is done in such a way as to leave the well-being levels of others unaffected, is equal to the new person's well-being level *minus* the constant a. Thus, according to this theory, adding an extra person with a well-being level of precisely a is neutral in terms of overall value; adding a person with well-being level $w > a$ is an improvement; and adding a person with well-being level $w < a$ makes things worse, even if the new person has a life worth living (i.e., even if $w > 0$). (The combination "$w > 0$ and $w < a$" is of course possible only if $a > 0$, but advocates of the Critical Level theory generally do propose $a > 0$.)

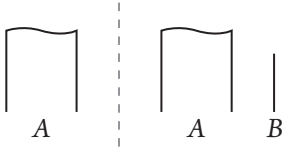
We have listed the Critical Level theory here for completeness, but for ease of exposition, we set it aside until section 5. In the present section, we consider the case in which credences are split between the other theories on the above list.

Note that none of these axiologies is sensitive to how well-being is distributed within a population. However, it is quite easy to tweak them to construct distribution-sensitive versions. For example, if one uses a different form of average (a generalized mean instead of the arithmetic mean), one can end up with prioritarianism.[32] This lets one have total, average, variable-value, and person-affecting versions of prioritarianism. Nothing we say below depends on the type of mean used, so our results apply to all of these theories too.

Note also that the above statements of the respective value functions do not imply that the units of value are directly comparable between the theories. We could apply additional scaling factors to compare them.

4.3. Scenario 1: Adding a Single Person

For our first scenario, suppose that the two populations we seek to compare differ only via the addition of a single person, whose well-being level is above zero but is below the average:



In this and other, similar diagrams, we use a wavy top for the box representing

32 For example, using a geometric mean corresponds to a logarithmic priority function and a root square mean corresponds to the square root priority function. In both cases, these incorporate a Fleurbaey transformation, which takes a particular approach to how prioritarianism should interact with uncertain outcomes. Other approaches to uncertainty can be accommodated, but we will not end up with generalized means in those cases.

a population to mean that the members of the population need not all have the same well-being level—the height is just an average level.

Different axiologies give different verdicts about whether the larger population is better, and by how much. The amount by which the larger population is better can be expressed as the value of the larger population minus the value of the smaller: $V(A \cup B) - V(A)$. The axiologies disagree about whether this expression is positive or negative, and about its magnitude.

In this section, we are particularly interested in what happens for large populations. We formalize this by considering what happens as the size of the population approaches infinity ($|A| \to \infty$) while both the average well-being in A and the well-being of the added "B-person" are kept fixed.[33] Loosely speaking, what happens in this case is that the theories that posit a negative value to adding another person (with below-average well-being) care less and less about this when the base population gets higher (tending toward indifference), while the theory that posits a positive value to adding another person (as long as that person's well-being level is positive) care just as much about this in all cases.

In more detail, here is what our various candidate axiologies have to say about the large-population limit $|A| \to \infty$:

	Value Difference as $\lvert A\rvert \to \infty$	*Explanation*
Total:	$V(A \cup B) - V(A) = \overline{B}$	i.e., $A \cup B$ is better by $\overline{B}$ units
Average:	$V(A \cup B) - V(A) \to 0$	as the averages converge
Variable Value I:	$V(A \cup B) - V(A) \to 0$	as the averages converge and the difference between $g(\lvert A\rvert)$ and $g(\lvert A \cup B\rvert)$ vanishes
Variable Value II:	$V(A \cup B) - V(A) \to 0$	as the averages converge and the difference between $g(\lvert A\rvert)$ and $g(\lvert A \cup B\rvert)$ vanishes
Presentism:	$V(A \cup B) - V(A) = 0$	as the person in B cannot be present at the time of choice so those present have unchanged well-being

33 If we used a distribution-sensitive theory, we would also have to make sure the shape of the distribution of well-being in A was kept roughly the same while the size of the population was scaled up.

Necessitarianism:	$V(A \cup B) - V(A) = 0$	as the necessary people have the same distribution of well-being in both cases

Thus on these views, as the number of people who are guaranteed to exist increases, the value of adding another person is either a fixed positive amount ($\overline{B}$), or tends to zero. The lack of any axiology positing a fixed negative value to adding this additional person has a striking effect on the effective axiology according to the EMV approach: for any fixed set of nonzero credences in these axiologies and any fixed way of drawing intertheoretic comparisons, for a sufficiently large base population the EMV approach ranks adding an extra person with a life worth living above not adding them, even when that lowers the overall average.[34] This is true regardless of how intertheoretic comparisons are performed (provided only that the normalization does not itself vary with base population size), because the ratio of the amount at stake according to the Total View to the amount at stake according to other views approaches infinity.

Interestingly, this goes against a common intuition that such "below-par additions" tend to amount to an overall improvement if the preexisting population is *small*, but make it worse if the preexisting population is *large*.[35] Indeed, it is largely on the grounds of that intuition that "variable value theories" seek to mimic the Total View at small populations but the Average View at large populations.[36] In contrast, we have shown that, under the EMV approach to axiological uncertainty, the result of splitting one's credence either between the Total View and the Average View, or between all of the theories listed above, is *precisely the opposite*: in the above-specified sense, one's effective axiology defers to the Total View when the preexisting population is sufficiently *large*, and is more likely to agree with the Average View when the preexisting population is *small*.

4.4. Scenario 2: Extinction Risk

Suppose we have the option of performing some action that would certainly slightly raise the well-being of the present generation, but that would also generate a nonzero chance of extinction.[37] For the sake of simplicity, let us model

34 Critical Level views might postulate a fixed negative value for the addition of an extra person with positive well-being—that will happen whenever the extra person's well-being, although positive, is below the "critical level." As mentioned above, we defer detailed exploration of Critical Level views to section 5.

35 Hurka, "Value and Population Size."

36 Hurka, "Value and Population Size"; Ng, "What Should We Do about Future Generations?"

37 More precisely: that would slightly raise the chance of extinction. We set aside other sources

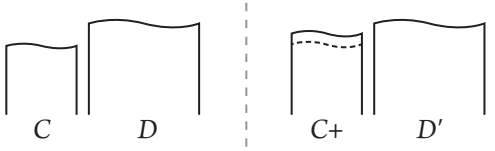
extinction as the nonexistence of any generation after the present one. There are then three possibilities:

(1) We do nothing ("Safe"), in which case past and present people have their "status quo" well-being levels, and there are also future people.

(2) We perform the action ("Risky"), and get away with it: past people are unaffected, present people have a slightly increased well-being level relative to the "status quo," and future people are just as in case (1).

(3) We perform the action ("Risky"), but extinction results: past people are unaffected, present people enjoy the increased well-being level as in case (2), but there are no future people.

We can represent this scenario as follows:



Here C and $C+$ are the same population (representing the past and present people), but with a higher average well-being in C+. The potential future people are represented by D and D'. D' either represents the same population as D or (with a small probability, p) represents an empty population. We shall set this up with well-being averages as follows: $\overline{C} < \overline{C+} < \overline{D} = \overline{D'}$ (i.e., the average well-being in D' *conditional on existence* is equal to the average well-being in D).

In this scenario, the "large-population limit" we consider is that in which the size of the possible future population tends to infinity: $|D| \to \infty$. In that limit, the Total View again overpowers the rival views we are considering, although in this case this happens for a structurally different reason than in the case of the Mere Addition scenario discussed above. In the Extinction Risk case, as $|D| \to \infty$, we have $V_{\text{total}}(\text{Safe}) - V_{\text{total}}(\text{Risky}) \to \infty$, while the value difference according to any axiology that ranks Risky over Safe at most approaches a finite bound. Therefore, the *ratio* of this value difference according to the Total View to the corresponding value difference according to any of the rival views currently under consideration again approaches infinity, so that the Total View again overpowers these rival theories in the large-population limit.

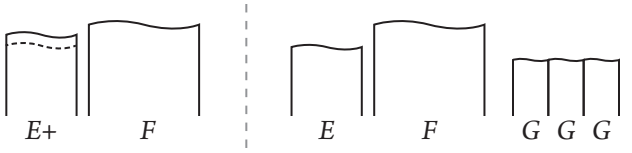
of extinction risk for simplicity of exposition; including it would complicate the detailed expression of our analysis, but would not affect its basic points.

4.5. Scenario 3: Space Settlement

In the future, we may reach a time at which we have the option of settling other planets—potentially, a very large number of other planets. This would involve some well-being cost to the people present at that time, but would dramatically increase the number of people who live in the further future. Living space on Earth is limited, but settling other planets would permit a much larger total population at any given future time—not to mention the fact that our own Sun will eventually die. It is extremely unclear how the average well-being level of those who would thereby live on other planets would compare to that of future Earth dwellers: that would depend on what conditions on the other planets in question turn out to be. Thus, settlement may or may not turn out to be a good move according to the Average View. Given the assumed cost to present people, however, it *is* clear that investing in settlement would be a bad move according to a presentist or necessitarian person-affecting theory. Meanwhile, it is likely to be a good move, and potentially a *very* good move, according to the Total View: for even modest human population sizes on other planets, the increase in total well-being due to the increase in population size is likely to trump the costs of settlement.[38]

A natural model of this scenario is as follows. Let $E+$ denote the population consisting of all past and present lives at the time humanity is deciding whether to settle other planets. If settlement goes ahead, this population is replaced with E, which consists of the same people as $E+$ but with slightly lower average well-being. Let F be the population consisting of all lives *on Earth* after the time of possible settlement. We assume, harmlessly idealizing for the sake of simplicity, that F is unaffected by whether or not the settlement project goes ahead (perhaps because the costs of the settlement project have been borne entirely by the E-people, and there is no further interaction between Earth and the settlements once the latter are established). Let G be a typical settlement population. For the sake of the further analysis, it does not matter how high the well-being in G is, so long as it is positive, but since the theories disagree the most when it is low, we shall illustrate it thus. We might establish several settlements, in which case the aggregate off-Earth population is some constant scaling up nG of G. Our choice is then between the populations $E+ \cup F$ (no settlement) on the one hand, and $E \cup F \cup nG$ (settlement) on the other:

38 Bostrom, "Astronomical Waste."



The "large-population limit" we consider in this case is the limit $n \to \infty$—that is, the limit in which the number of possible settlement inhabitants tends to infinity. For sufficiently large such populations, much as for the Extinction Risk scenario, the Total View favors settlement over non-settlement, *and does so by an amount that increases without bound as n increases*. In contrast, while various other views favor non-settlement over settlement, they do so by at most an amount that remains bounded as n goes to infinity. Therefore, in the limit $n \to \infty$, the Total View overpowers the rival theories that we are considering.[39]

4.6. Further Properties

We have seen that, in all three scenarios, if we spread out credence between the axiologies we have considered, the highest-ranked alternative under the EMV approach to axiological uncertainty is the same as the alternative that is highest-ranked by the Total View, when the number of people involved gets large enough. This eventually happens for any nonzero credence in the Total View, no matter how low.

As well as this, there are two related results. First, the way that the Total View came to overpower its rivals was by having the amount at stake become many times as much as that posited by the other axiologies, without limit. This means that, not only does the alternative recommended by the Total View eventually get an expected moral value that is higher than the other alternative, the difference between these expected moral values of the alternatives grows without bound, and the ratio between them grows without bound. This matters when it comes to empirical uncertainty, because it can mean that even when there is only a *low chance* of the alternatives leading to situations like those in the scenarios (e.g., a choice in which one alternative only slightly increases the chance of space settlement), according to the EMV approach the effective axiology would

39 In a more realistic model, the "settlement" option would correspond not to actually settling other planets, but to having a nontrivial probability of settling other planets (since any settlement attempt we make might fail). This would complicate the model in ways analogous to the treatment of probability in Scenario 2. Here, we stick with the simpler model for the space settlement case for ease of exposition. It is trivial to adjust the calculations we carry out in section 6 to generate a quantitative analysis of the more complicated model.

still agree with the Total View, when the size of the possible population at stake is sufficiently high.

Second, the difference in expected moral value between the alternatives changes *monotonically* as the number of people affected is scaled up. That is, *all* increases in the population improve the relative standing of the alternative that the Total View favors. This implies that, once the alternative that is top ranked by the Total View becomes top ranked in the effective axiology, it remains top ranked as the population is scaled up—there is no oscillation back and forth.

In addition to the restrictions that we noted in section 1, our results are, however, limited in the following two senses.

First, we have shown only that, *of the axiologies we have considered*, all but Critical Level theories are overpowered by the Total View in the specified large-population limits. We have of course not claimed that there is *no possible* population axiology that would not be overpowered in this way. That further claim is clearly false: one mathematically possible (but substantively completely implausible) such axiology, for example, is one we might call the "Reverse Total View", according to which the value of any state of affairs is precisely *minus* the value that the Total View assigns to it.

The more interesting possibility is that there might be some *reasonably plausible* population axiology that we have not considered, and that would (nevertheless) not be overpowered in the cases we have discussed. We have no nonexistence proof here. But it is worth noting what it takes for a theory to avoid overpowering in these cases: the theory must hold, in our scenarios of Extinction Risk and Space Settlement, not only that the alternative favored by the Total View is inferior in the large-population limit, but that the *amount* by which it is inferior grows without bound as the relevant population size increases. In the Extinction Risk case such a theory must, for example, have a preference for Risky over Safe that gets stronger and stronger, without bound, as the size of the threatened possible future population increases. This condition seems difficult to meet; while there may be serious candidate axiologies that we have not considered, we doubt that any of them meet the conditions needed to avoid overpowering. We are aware of only one partial exception, to which we turn in section 5.

Second, so far we have shown only that *in the limit as the relevant population size goes to infinity*, the Total View overpowers the extant rival theories. For practical purposes, these limit results supply a useful heuristic: it is *worth considering the question* of whether actual population sizes are sufficiently large. But nothing that we have said so far takes on the question of whether overpowering will actually occur in practice, rather than only in theory. We address this in section 6.

5. CRITICAL LEVEL AXIOLOGIES

As explained in section 4, the Total View is a member of the Critical Level family of axiologies, corresponding to the special case in which the critical level is zero. The caveat to the overpowering claims we made in section 4 is this: strictly speaking, the theory that "overpowers" others in our large-population limits will usually not be the Total View itself, but some other member of this Critical Level family. Like the Total View, all Critical Level theories have non-diminishing returns to the value of additional people, and thus tend to generate unbounded values in large-population limits.

But what happens in cases in which one has nonzero credence in two different Critical Level axiologies, where the value of the critical level is different? While we omit the details due to lack of space, it can be easily proven that the contributions to the expected moral value made by one's credence in multiple Critical Level theories is just the same as if all that credence was placed in a single Critical Level theory —whose critical level is set to be a weighted average of the individual ones. For example, if one has 40 percent credence in the Total View and 10 percent credence in a Critical Level theory whose critical level is α, then the expected moral value of any option will be exactly the same as if one instead had credence 50 percent in a Critical Level theory whose critical level was $\alpha/5$.

Arguably, however, this modification is unlikely to make very much difference to our qualitative conclusions. For example, in the Extinction Risk and Space Settlement scenarios it is reasonable to suppose that the additional people have well-being greater than the weighted average of plausible critical levels. If so, the combined Critical Level views also push in favor of avoiding extinction risk and settling the cosmos. However, if a scenario envisaged rather mediocre additional lives or if one had a lot of credence in Critical Level theories with a very high bar, then the conclusions could be reversed, with the Critical Level theory's aversion to a large population overpowering any other theories that were in favor of risk reduction or expansion.

6. EMPIRICAL ANALYSIS OF EXISTENTIAL RISK AND SPACE SETTLEMENT

6.1. Preliminaries

What we have argued so far is that, for the three scenarios outlined, *in the limit of large affected populations*, EMV recommends the same alternative as one's effective Critical Level theory, even if one thinks it is overwhelmingly likely that that alternative is the inferior option. But how large does a population have to

be in practice before this happens? In particular, will this overpowering of other theories by the Total View and Critical Level theories ever actually happen in practice, or is it merely a theoretical curiosity?

This question can be answered only by crunching the numbers for plausible estimates of (for instance) the expected remaining life span of humanity (for the Extinction Risk scenario) or the number of future persons who might exist if we succeeded in settling space (for the Space Settlement scenario), the rough size of cost in terms of present well-being that might be associated with lowering extinction risk or settling space (respectively), and the amount by which this sacrifice of present well-being might succeed in reducing extinction risk (in that scenario). Any such estimate is open to significant debate. However, for illustrative purposes, here we sketch how the numbers fall for estimates that we ourselves consider quite reasonable.

To simplify the calculations, we consider the case of an agent who has nonzero credence only in the Total View and a person-affecting theory. The inclusion of other axiologies would be unlikely significantly to alter our qualitative conclusions, but would vastly complicate the analysis.

The calculations in question are, of course, crucially affected by how one draws intertheoretic comparisons between the theories in question. In section 2, we outlined two relatively specific ways of fixing intertheoretic comparisons, drawing respectively on "content" and "structure." Our conclusions will be that on the content-based approach the kind of overpowering we have been discussing is indeed moderately likely to occur in practice and not only in theory; on a structuralist approach matters are more complex, and all bets are off.

6.2. Content-Based Intertheoretic Comparisons

We first assume that the value scales of the Total View and person-affecting theory are normalized against one another according to the natural "content-based" prescription mentioned in section 2: that is, we assume that these theories agree with one another about the value of any given change to the well-being of an already existing person, and merely disagree about whether or not future/non-necessary persons have any axiological significance at all.

In the Extinction Risk scenario: suppose, for instance, that the expected remaining life span of humanity is one million years, that there will on average be an additional seven billion people per century until humanity goes extinct, and that each person lives for one hundred years.[40] Suppose that the amount of

40 For context: the species *Homo sapiens* has already been around for 200,000 years; the average mammalian species lasts for one million to two million years; the average historical frequency of mass extinction events is one per one hundred million years; the heating up

well-being that the present generation would forgo in order to reduce extinction risk amounts to 0.1 percent of each person's lifetime well-being level, and suppose that this sacrifice would reduce the probability of imminent extinction by 1/100,000. Then the amount by which the Total View favors the Safe option over the Risky one is ninety-nine times the amount by which a person-affecting theory favors Risky over Safe. Therefore, provided our agent's credence in the Total View is more than about 1 percent of one's credence in person-affecting theories, under axiological uncertainty (according to EMV, and with the intertheoretic comparisons fixed as stated above) the Total View overpowers the person-affecting theory for the purposes of this particular decision.

The analysis for space settlement has much in common with that of existential risk. If we could settle many new worlds with populations that last many generations and have a good quality of life, it is easy to see how the Total View could assign this a very high value relative to the value of improving the well-being of a single generation. In fact, it seems substantially easier for the Total View to overpower person-affecting views in the Settlement case than in the Extinction Risk case.

Numbers for the Settlement case are even more speculative than for Extinction Risk, but the qualitative conclusions are robust to changing the numbers by a large amount. Let us ask what would happen if we could settle one in a million of the planets in our galaxy (and no planets elsewhere). This would be about 100,000 new planets. We suppose, fairly conservatively, that each settlement would last an average of 200,000 years, that each will have a tenth as many people as Earth did at the time the settlement begins, and that quality of life will only be half as good. Let us suppose that, in order to launch the settlement, present people must sacrifice enough to reduce their quality of life by 10 percent for one hundred years (which we are supposing would be enough to start a cascade of settlements, each of which can settle further, eventually reaching all 100,000 new planets). In this case, the amount by which the Total View favors settling is *one hundred million* times the amount by which a person-affecting theory favors not settling, so that our agent would favor settlement provided only that her credence in the Total View was more than about 1 in 100 million. This is an enormous ratio; for any remotely reasonable relative credences, the Total View would still overpower a person-affecting theory even if the numbers were

of the Sun will dry out Earth in something over one billion years' time. Note that we are interested in humanity's *expected* remaining life span, so that even a small credence in life spans anywhere near the upper end of this range can substantially increase the figure that is relevant for our purposes. While there is room for plenty of debate here, in our view this makes our suggested figure of one million if anything a very conservative estimate.

changed to be much less favorable (e.g., if the settlements only lasted one thousand years and there were only ten of them).

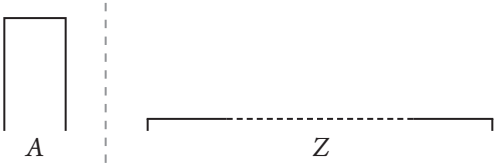
6.3. Structuralist Intertheoretic Comparisons

The back-of-the-envelope calculations of section 6.2 made essential use of the content-based method of fixing intertheoretic comparisons; what, then, of structuralist approaches? One of the key *motivations* of structuralist approaches is to ensure that rival moral theories have (in some sense) "equal say" in decisions when the agent's credence is split equally between the theories in question. This makes overpowering considerably more difficult. It turns out, however, that overpowering can occur, but only on some structuralist approaches and in a more complex set of circumstances.[41]

7. THE EFFECTIVE REPUGNANT CONCLUSION

The Total View notoriously implies:

> *The Repugnant Conclusion*: For any state of affairs *A*, no matter how large the population is and no matter how high people's well-being levels are in *A*, there is a better state of affairs, *Z*, in which no one has a life that is more than barely worth living.



Virtually everyone has at least some degree of pretheoretic intuition that the Repugnant Conclusion is false. Defenders of the Total View argue that this intuition is not in the end to be trusted. For most people, however, avoidance of the Repugnant Conclusion is a very strong desideratum.

Consider now:

> *The Effective Repugnant Conclusion*: For any state of affairs *A*, no matter how large the population is and no matter how high people's well-being

41 For these purposes it matters whether one normalizes by range or by variance, whether one normalizes over only alternatives that are available in the choice at hand or over a wider set of alternatives, and (on the variance-normalization approach) which measure over the set of alternatives is used. Since these details are messy and not especially illuminating, we omit detailed discussion and calculations in the interest of brevity.

> levels are in *A*, there is a state of affairs, *Z*, in which no one has a life that is more than barely worth living, and such that the effective axiology ranks *Z* above *A*.

At first sight, one might suspect that the EMV approach to axiological uncertainty implies the Effective Repugnant Conclusion for reasons similar to those given in section 4 for our three scenario types of primary interest. And, if so, those who think the first-order Repugnant Conclusion is strong evidence against the Total View might well think that the Effective Repugnant Conclusion is strong evidence against the EMV approach.

In reply, three comments are in order. First: In fact the EMV approach does not imply the Effective Repugnant Conclusion, for the reasons given in section 5. The theory that "overpowers" others in large-population limits is not necessarily the Total View, but rather one's effective Critical Level theory. But, as in first-order discussions of Critical Level theories, it is debatable whether this sweetens the pill enough. For the EMV approach to axiological uncertainty *does* imply:

> *The Effective Weak Repugnant Conclusion*: For any state of affairs *A*, no matter how large the population is and no matter how high people's well-being levels are in *A*, there is a state of affairs, *Z*′, in which no one has a life that is more than barely above the effective critical level, and such that the effective axiology ranks *Z*′ above *A*.

How bad this is depends, of course, on how high one's effective critical level is. But an effective critical level that is too high will give rise to further problems, and in any case at least *some* agents will have an effective critical level that is very close to zero (perhaps because their credence in the Total View, conditional on the proposition that some Critical Level theory is true, is high). For those agents, the Effective Weak Repugnant Conclusion is scarcely different in substance from the Effective Repugnant Conclusion. The fact that, strictly speaking, the EMV approach implies "only" the Effective Weak Repugnant Conclusion, and not the Effective Repugnant Conclusion itself, is therefore unlikely to satisfy those who find the Effective Repugnant Conclusion implausible in the first place.

Second: Even the (non-Weak) Effective Repugnant Conclusion is at least *somewhat* more plausible than the first-order Repugnant Conclusion. Granted, the majority of non-Total axiologies rank the *A*-world above the *Z*-world, but they generally hold that the difference in value between any given *A*-world and any *Z*-world is *relatively* modest. In contrast, the Total View holds that sufficiently large *Z*-worlds are *much, much* better than any given *A*-world, by an amount that grows without bound as the size of *Z* increases. The sort of considerations of

relative stakes that we have been considering in this paper, therefore—precisely the considerations that cause EMV to imply some form of Effective Repugnant Conclusion—also serve as an explanation of why an Effective Repugnant Conclusion might be true, even if the first-order Repugnant Conclusion is false. We assume, however, that many of those who find the Repugnant Conclusion implausible in the first place also have recalcitrant intuitions against the Effective Repugnant Conclusion, this consideration notwithstanding.

Third: Section 6 raised the possibility that if intertheoretic comparisons are fixed in a structuralist (variance-normalization) way, then while overpowering is a real theoretical phenomenon in cases of sufficiently large populations, it is an entirely open question whether or not realistic empirical parameters are such that overpowering will actually occur in realistic Extinction Risk and/or Space Settlement cases. There is, however, no hope of avoiding the fact that the EMV approach to axiological uncertainty implies the Effective Weak Repugnant Conclusion via any analogous considerations, since Repugnant Conclusions are and always have been matters of purely *theoretical* large-population limits.

8. *REDUCTIO*?

We have argued that, according to the EMV approach to axiological uncertainty, (i) for three fairly realistic decision scenarios, the Total and Critical Level views overpower other extant axiologies in specified large-population limits; (ii) depending on the details of how intertheoretic comparisons are settled, it is at least somewhat plausible that such overpowering will actually occur with empirically realistic parameter values; and (iii) the Effective Weak Repugnant Conclusion is true.

As with any argument, our arguments themselves are silent on the question of whether the appropriate reaction is to accept their conclusions or reject one or more of their premises. In the present context, the plausible option in this second camp is to take our arguments to be a *reductio* of the EMV approach to axiological uncertainty.[42] In this section, we comment on the degree to which this is a reasonable reaction.

First: *Sometimes* the right reaction to an overpowering result is to read it as a

42 The other option in the *reductio* camp would be a *reductio* of the claim that it is rationally permissible to have nonzero credence in the Total View. Since the Total View is both an extremely natural extension of a plausible fixed-population axiology, and is one of the handful of population axiologies that actually commands the assent of a sizeable minority of the theoretical community, however, this claim of rational permissibility strikes us as considerably more secure than the EMV approach to axiological uncertainty, so that this reaction is implausible.

reductio. It indeed does not seem, for example, that any arbitrarily low credence that it is sufficiently good to set cats on fire for fun should rule one's decisions, when one has credence well over 99.9999 percent that setting cats on fire for fun is extremely bad; so much the worse for any theory of axiological uncertainty that implies otherwise.

Second: The importance of relative stakes notwithstanding, there may be independent pressures to resist evaluative theories with precisely the expected-value structure, in cases involving extremely low probabilities of extremely high stakes. This point applies to empirical, as well as normative, uncertainty. For example, consider the following case (adapted from Bostrom):

> *Pascal's Mugging*: A mugger approaches you. He has no weapon, but exhorts you to hand over your wallet: "In return, I will give you any finite amount of utility that you ask for. I am able to do this because I have secret powers. Now, you might think it is extremely unlikely that I am telling the truth here, but surely you have *nonzero* credence that I am, and if so, you only have to stipulate a sufficiently high utility reward, and then handing over your wallet will have positive expected utility for you."[43]

Expected-utility theory perhaps entails that one is rationally required to hand over the wallet in this case, provided only that one has *nonzero* credence that the mugger is telling the truth. But that seems wrong. If so, the lesson is that expected utility seems to give wrong verdicts *in cases involving extremely high stakes and extremely low probabilities.*

Third: In the empirical case, however, it is *not* plausible to reject expected-utility theory wholesale, in response to the case of Pascal's Mugging. It remains true that expected-utility theory behaves well in general, including in cases that involve very (but not absurdly) low probabilities of very (but not absurdly) high stakes. Expected-utility theory tells a very plausible story, for instance, about why it is rational to buy building insurance for one's home, despite believing that the chance one will ever claim on such insurance is well under 1 percent. If we seek to modify expected-utility theory in response to Pascal's Mugging, therefore, we had better seek a relatively localized modification that mainly affects such *extreme* low-probability/high-stakes cases, not a wholesale rejection of the theory.

Fourth: Given the above comments, the salient question is whether the overpowering results that we have discussed are more like insurance cases on the one hand, or more like Pascal's Mugging (and the above example of setting cats on fire) on the other. The three relatively realistic decision scenarios we have dis-

43 Bostrom, "Pascal's Mugging."

cussed (Mere Addition, Extinction Risk, Space Settlement) are more like insurance cases, and are crucially disanalogous to the example of setting cats on fire. For one thing, one's credence that it is extremely good to set cats on fire should be *extremely* low—well under 0.0001 percent, for instance. But given the state of play in first-order population-axiological theorizing, an honest enquirer should not have such *extremely* low credence in the Total View or Critical Level views (that credence should probably not be less than, say, one percent, however dim a view one is initially inclined to take of the Repugnant Conclusion). For another thing, the recommendations of the Total View and Critical Level views *vis-à-vis the three relatively realistic decision scenarios we have analyzed* are not actively repugnant; at most, they overturn rather mild contrary preferences of other theories or untutored intuitions. (Most people's *pretheoretic* intuition, for instance, is in fact that human extinction would be very bad, while adding extra persons and (relatedly) space settlement strike most people as at worst neutral.)

Fifth: The Effective (Weak) Repugnant Conclusion, though, is a somewhat different story. Unlike the overpowering results for empirically realistic versions of our decision scenarios, the EMV approach entails the Effective Weak Repugnant Conclusion even for an agent who has *arbitrarily* low (but nonzero) credence in the Total View and Critical Level views, and despite the fact that the Repugnant Conclusion strikes most people who are not sympathetic to Totalism as a first-order axiology as *strongly* repugnant. The overpowering result that leads to the Effective Weak Repugnant Conclusion, therefore, may be much more closely analogous to Pascal's Mugging, and hence it is much more plausible to read *this* result as a *reductio* of the EMV approach. Again, however, in the light of the dearth of worked-out, plausible, extant alternatives to the EMV approach, this observation only really motivates seeking a relatively conservative modification of that approach, whose implications are limited to extreme low-probability/high-stakes cases. We should not too hastily conclude, that is, that the relatively mundane overpowering conclusions discussed in the main body of our paper will also be casualties of this modification, any more than contemplation of Pascal's Mugging should incline us to stop insuring our homes.

Some readers, however, will already be inclined to read our main overpowering results as *reductios* of the EMV approach, even without any appeal to any Repugnant Conclusion. While this raises a serious question of what the alternative approach to axiological uncertainty should be, this reaction does not seem unreasonable, and we have not argued against it. For those inclined toward this reaction, we therefore offer the following comments on what our paper has added to the preexisting "overpowering-based case against EMV." Others have previously noted that such overpowering can occur at least when one theory as-

signs an *infinite* value difference to some pair of alternatives, while a rival theory assigns a finite value difference.[44] In that case, any arbitrarily small (but finite) credence in the "hysterical" theory would lead to overpowering. To this basic observation, this paper adds, first, that the same phenomenon can occur with theories that postulate only finite value differences (even for agents who again have *arbitrarily* low credence in the relevant theories), so there is no prospect of avoiding the basic issue by ruling out "infinite value-difference theories" as somehow ill formed. That this phenomenon is in principle *possible* is fairly obvious on reflection. Second, though, we have shown that, in the case of population axiology, such overpowering under EMV is not merely an abstract possibility, but seems fairly likely actually to occur, for reasonable estimates of the relevant empirical parameters and for reasonable credence distributions. So the prospects for avoiding all finite-value-difference overpowering in practice simply by having sufficiently low credence in the "offending" theories also looks fairly dim; if one wants to avoid overpowering, the only escape route in the offing is rejection of the EMV approach to axiological uncertainty.

9. CONCLUSION

It has frequently been observed that, in the context of population ethics in particular, we need to make decisions under conditions of moral uncertainty, including axiological uncertainty. Since even "inaction" is in the relevant sense an action, we are forced to act now, and cannot simply wait until our uncertainty has been resolved.

At the theoretical level, at least one of the serious contenders for the effective axiology under axiological uncertainty is the ranking of alternatives according to their expected moral value (EMV). There has, however, previously been little investigation of what the EMV approach actually recommends, in the case of population-ethics dilemmas. In this paper we have established, for three different decision scenarios, that in an appropriately specified "large-population limit" the alternative that has the higher expected moral value is the one that is preferred by a particular Critical Level theory (where the identification of the critical level is determined by the agent's credences among Critical Level views, including the Total View itself). In this sense, Critical Level views overpower all other extant rival population axiologies *in those large-population limits*. Depending on precisely how one fixes intertheoretic comparisons, there (further) seems to be at least some very real prospect that actual population sizes are large

44 Ross, "Rejecting Ethical Deflationism"; Sepielli, "Along an Imperfectly Lighted Path"; Beckstead, "On the Overwhelming Importance of Shaping the Far Future."

enough for this overpowering to occur in practice, and not only in some counterfactual limit case.

The EMV approach also entails the Effective Weak Repugnant Conclusion, which is likely to strike many people as repugnant. If so, that is a reason to reject the EMV approach to axiological uncertainty in full generality; the Effective Weak Repugnant Conclusion is, structurally speaking, an axiological analogue of Pascal's Mugging. However, this consideration, as in the empirical case, motivates only a relatively conservative modification of expected value theory, and (because of that) is unlikely to provide any sound motivation for rejecting our more mundane overpowering results. One might, however, read those more mundane results as a further reason to reject the EMV approach to axiological uncertainty across the board, and thus to postulate a deep structural difference between empirical and axiological uncertainty.

University of Oxford
hilary.greaves@philosophy.ox.ac.uk
toby.ord@philosophy.ox.ac.uk

REFERENCES

Arrhenius, Gustaf. "An Impossibility Theorem for Welfarist Axiologies." *Economics and Philosophy* 16, no. 2 (October 2000): 247–66.

———. "Population Ethics: The Challenge of Future Generations." Unpublished manuscript.

Bader, Ralf. "Neutrality and Conditional Goodness." Unpublished manuscript.

Beckstead, Nick. "On the Overwhelming Importance of Shaping the Far Future." PhD diss., Rutgers University, 2013.

Bigelow, John, and Robert Pargetter. "Morality, Potential Persons and Abortion." *American Philosophical Quarterly* 25, no. 2 (April 1988): 173–81.

Blackorby, Charles, Walter Bossert, and David Donaldson. "Intertemporal Population Ethics: Critical-Level Utilitarian Principles." *Econometrica* 63, no. 6 (November 1995): 1303–20.

Bostrom, Nick. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15, no. 3 (November 2003): 308–14.

———. "Pascal's Mugging." *Analysis* 69, no. 3 (July 2009): 443–45.

Broome, John. *Climate Matters: Ethics in a Warming World*. New York: W.W. Norton and Company, 2012.

———. *Weighing Lives*. Oxford: Oxford University Press, 2004.

Carlson, Erik. "Mere Addition and Two Trilemmas of Population Ethics." *Economics and Philosophy* 14, no. 2 (October 1998): 283–306.

Cotton-Barratt, Owen, William MacAskill, and Toby Ord. "Normative Uncertainty, Intertheoretic Comparisons, and Variance Normalisation." Unpublished manuscript.

Gracely, Edward J. "On the Noncomparability of Judgments Made by Different Ethical Theories." *Metaphilosophy* 27, no. 3 (July 1996): 327–32.

Gustafsson, Johan E., and Olle Torpman. "In Defence of My Favourite Theory." *Pacific Philosophical Quarterly* 95, no. 2 (June 2014): 159–74.

Harman, Elizabeth. "Does Moral Ignorance Exculpate?" *Ratio* 24 no. 4 (December 2011): 443–68.

Heyd, David. "Procreation and Value: Can Ethics Deal with Futurity Problems?" *Philosophia* 18 nos. 2–3 (July 1988): 151–70.

Hudson, James L. "Subjectivization in Ethics." *American Philosophical Quarterly* 26, no. 3 (July 1989): 221–29.

Hurka, Thomas. "Value and Population Size." *Ethics* 93, no. 3 (April 1983): 496–507.

Kitcher, Philip. "Parfit's Puzzle." *Noûs* 34, no. 4 (December 2000): 550–77.

Lockhart, Ted. *Moral Uncertainty and Its Consequences*. Oxford: Oxford University Press, 2000.

MacAskill, William. "The Infectiousness of Nihilism." *Ethics* 123, no. 3 (April 2013): 508–20.

———. "Normative Uncertainty." DPhil thesis, University of Oxford, 2014.

Mason, Elinor. "Moral Ignorance and Blameworthiness." *Philosophical Studies* 172, no. 11 (November 2015): 3037–57.

Narveson, Jan. "Moral Problems of Population." *Monist* 57, no. 1 (January 1973): 62–86.

Ng, Yew-Kwang. "What Should We Do about Future Generations?" *Economics and Philosophy* 5, no. 2 (October 1989): 235–53.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.

Riedener, Stefan. "A Theory of Axiological Uncertainty." DPhil thesis, University of Oxford, 2015.

Ross, Jacob. "Rejecting Ethical Deflationism." *Ethics* 116, no. 4 (July 2006): 742–68.

Sepielli, A. "Along an Imperfectly Lighted Path: Practical Rationality and Normative Uncertainty." PhD diss., Rutgers University, 2010.

———. "What to Do When You Don't Know What to Do." *Oxford Studies in Metaethics* 4 (2009): 5–28.

Sider, Ted. "Might Theory *X* Be a Theory of Diminishing Marginal Value?" *Analysis* 51, no. 4 (October 1991): 265–71.

Singer, Peter. *Practical Ethics*. Cambridge: Cambridge University Press, 1979.

Temkin, Larry S. "Intransitivity and the Mere Addition Paradox." *Philosophy and Public Affairs* 16, no. 2 (Spring 1987): 138–87.

———. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning.* Oxford: Oxford University Press, 2012.

Warren, Mary Anne. "Do Potential People Have Moral Rights?" *Canadian Journal of Philosophy* 7, no. 2 (June 1977): 275–89.

Weatherson, Brian. "Running Risks Morally." *Philosophical Studies* 167, no. 1 (January 2014): 1–23.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 2 · NOVEMBER 2017
https://doi.org/10.26556/jesp.v12i2.227

THOMSON'S TROLLEY PROBLEM

Peter A. Graham

NO ONE HAS DONE MORE over the past four decades to draw attention to the importance of, and attempt to solve, a particularly vexing problem in ethics—the Trolley Problem—than Judith Jarvis Thomson. Though the problem is originally due to Philippa Foot, Thomson showed how Foot's simple solution would not do and offered some of her own.[1] No solution is uncontroversial and the problem remains a thorn in the side of non-consequentialist moral theory. Recently, however, Thomson has changed her mind about the problem. She no longer thinks she was right to reject Foot's solution to it. I argue that, though illuminating, Thomson's current take on the Trolley Problem is mistaken. I end with a solution to the problem that I find promising.

In sections 1–3, I present Thomson's version of the Trolley Problem (one involving a twist on Foot's original version) and her various responses to it. In sections 4 and 5, I evaluate her various takes on the problem, including her most recent rejection of the problem. In section 6, I offer a diagnosis of the purported data on the basis of which Thomson has mistakenly come to reject the problem. And in section 7, I present and defend my own preferred solution to the Trolley Problem.

1. THE PROBLEM STATED

Foot's version of the Trolley Problem revolves around pairs of cases like these:

> *Big Man*: An out-of-control trolley—the driver is unconscious—is barreling toward five workmen trapped on the track ahead of it. If nothing stops the trolley, the five will be run over and killed. A big man whose weight would stop the trolley before it reaches the five were he to sacrifice his life by jumping in front of it decides not to. However, a thin man can push the big man into the path of the trolley.
>
> *Driver*: An out-of-control trolley is barreling toward five track workers

1 Foot, "The Problem of Abortion and the Doctrine of Double Effect."

> who are trapped on the track ahead of it. If the driver does nothing, the five will be run over and killed. The driver cannot stop the trolley, but he can turn it onto a spur of track to the right, on which there is another trapped track worker who would be run over and killed were he to do so.

Foot took it to be obvious that, though the thin man may not push the big man, the driver may turn the trolley. And the problem she took herself to be addressing was why that was the case. After all, in both cases the agent faces the choice of whether to kill one to save five. Her answer was that the cases were importantly asymmetrical: though the thin man lets five die if he does not kill, the driver kills either way. As Thomson puts it: "If the driver fails to turn his trolley, he does not merely let the five track workmen die; he drives his trolley into them, and thereby kills them."[2] And if, as Foot suggested, the following two claims are true,

(I) killing one is worse than letting five die, and

(II) killing five is worse than killing one,

then the solution to her problem was right at hand. As (I) dictates that pushing the big man is worse than not pushing him, pushing the big man is impermissible. But, as (II) dictates that not turning the trolley is worse than turning it, turning it is permissible. Furthermore, Foot argued, this solution comports with our intuitive conviction that negative moral duties—e.g., the duty *not* to kill—are more stringent than positive moral duties—e.g., the duty *to* rescue.

Thomson argued that Foot's solution fails.[3] She suggested we consider:

> *Bystander*: An out-of-control trolley—the driver is unconscious—is barreling toward five track workers trapped on the track ahead of it. A bystander can either (i) do nothing, in which case the five will die, or (ii) flip a switch to the right, diverting the trolley onto a right-hand spur of track away from the five, thereby killing another track worker who is trapped there.

It is no less morally permissible, Thomson claimed, for the bystander to flip the switch than it is for the driver to turn his trolley in *Driver*. Foot's (I) and (II) offer no help in explaining this. Why may the bystander turn the trolley when he, like the thin man, chooses between killing one and letting five die? What is needed is a solution that accommodates the data in all the cases.

2 Thomson, "The Trolley Problem," 1397.

3 See Thomson, "The Trolley Problem."

2. THOMSON'S SOLUTIONS TO THE PROBLEM

In "The Trolley Problem," Thomson offered a solution—call this her First Solution—according to which the bystander may flip the switch in *Bystander* because were he to do so (1) he makes what was threatening five come to threaten only one and (2) he does so not by any means that constitute an infringement of any right of the one's.[4] If the thin man pushes the big man in *Big Man*, by contrast, though (1) would be true, (2) would not: he would make the threat to the five threaten the one by means that do infringe the one's rights, in particular, the right the one has against him that he not push him. But does the bystander not infringe the one's rights in turning the trolley onto him? In one sense, yes, but in another, no. True, the one has a right that the bystander not kill him and the bystander does kill him. But the turning of the trolley does not *in and of itself* violate any of the one's rights. So, though the bystander does indeed infringe the one's right not to be killed, his doing that which saves the five—turning the trolley—does not, in and of itself, infringe any right of the one's.

Thomson's First Solution reflects a certain picture that many proponents of rights-based moral theories find attractive. There is a standing moral requirement to minimize harm wherever possible. However, people's rights prevent them from being permissibly sacrificed to do so. To use Ronald Dworkin's evocative metaphor, rights trump utilities, and on this picture they do so by being *means-blocking barriers*.[5] Rights act to constrain how we may go about minimizing overall harm by blocking certain causal pathways to that result; only if it is caused *in certain ways* is it permissible to cause harm in order to minimize harm overall. Our moral worth and integrity are in large part constituted by the rights we have and so only by avoiding these barriers do we respect the moral worth and integrity of others.

In *The Realm of Rights*, however, Thomson rejects her First Solution because she now thinks that the bystander's turning of the trolley is indeed an infringement of the one's rights.[6] Having rejected her First Solution, Thomson then offers a Second Solution: the bystander may turn the trolley because all six workmen—the five and the one—belong to a group such that at some point in the past it was in each member's individual interest that the bystander turn the trolley were *Bystander* ever to arise. As track assignments were determined by

4 Thomson also discusses the Trolley Problem in "Killing, Letting Die, and the Trolley Problem"; however, there she gestures at, but does not explicitly defend, a solution to it.

5 For Dworkin's talk of rights trumping utilities, see Dworkin, "Rights as Trumps."

6 She thinks this because she now accepts a principle from which it follows that turning the trolley infringes one of the one's rights. I discuss this in detail in section 5.

lot, each workman's expected utility, prior to that determination, is greater if the bystander flips the switch should *Bystander* arise than if he does not. And it is in virtue of this, she held, that the bystander may flip the switch.

These are Thomson's solutions to the Trolley Problem. I think her First Solution is on the right track and she veered off it with her Second Solution. I say why below. For now, however, we need to look at her third and final take on the problem. This time she does not offer a solution to the problem; instead, she rejects that there is a problem in need of solving at all.

3. THOMSON'S ABOUT-FACE

Thomson has had a change of heart about *Bystander*. She now thinks the bystander's turning the trolley is impermissible. Her argument begins with:

> *Three Options*: Things are as they are in *Bystander* except that the bystander has a third option: (iii) he can flip the switch to the left, diverting the trolley onto a left-hand spur of track on which he himself is trapped, thereby saving the five but killing himself.

Thomson takes it to be obvious that, though it would be permissible for the bystander to choose either (i) or (iii)—forgoing killing someone, even if it means letting five innocent people die, and voluntarily sacrificing oneself in order to save five innocent people from dying are both certainly permissible—it would *not* be permissible for him to choose (ii). What these reflections show, Thomson maintains, is the truth of the following:

> (*): If *A* wants to do a certain good deed, and can pay what doing it would cost, then—other things being equal—*A* may do that good deed only if *A* pays the cost himself."[7]

And, in particular, it follows from (*), she claims, that a certain *ceteris paribus* principle is true:

> *Third Principle*: *A* must not kill *B* to save five if he can instead kill himself to save them.[8]

All of this, she maintains, supports the impermissibility of option (ii) in *Bystander*.

Thomson suggests that if the bystander would not turn the trolley onto himself and pay the cost of his own life to save the five were he instead in *Three Options*, then "there is no way in which he can decently regard himself as entitled

7 Thomson, "Turning the Trolley," 365.

8 Thomson, "Turning the Trolley," 365.

to make someone else pay it [in *Bystander*]."[9] But this could not really be that which underwrites the wrongness of the bystander's turning the trolley. It cannot really be thought that, though it is impermissible for the bystander to turn the trolley in *Bystander*, it would have been permissible had he been just a little bit more of a noble sort. In fact, Thomson agrees. Of the bystander who *would* turn the trolley on himself were he in *Three Options* instead of *Bystander*—"the altruistic bystander"—she says:

> [He] is not entitled to assume that the one workman is equally altruistic, and would therefore consent to the bystander's choosing option (ii). Altruism is by hypothesis not morally required of us. . . . Suppose, then, that the bystander knows that the one workman would not consent, and indeed is not morally required to consent, to his choosing option (ii). The bystander has a permissible alternative, namely choosing option (i)—that is, letting the five die. I think it very plausible therefore that there is no way he can justify to himself or to anyone else his choosing option (ii), and thus he cannot decently regard himself as entitled to choose it.[10]

What matters here, whether the bystander is altruistic or not, then, is that the one does not, and is not required to, consent to being killed and the bystander has a permissible alternative.

If Thomson is correct here, then it is not morally permissible for the bystander to turn the trolley in *Bystander*. As the permissibility of that option was the basis of her rejection of Foot's solution to her problem—that of explaining the data in *Big Man* and *Driver*—Foot's solution to that problem stands. And, as there is no moral difference between *Big Man* and *Bystander* to be explained, Thomson concludes, the Trolley Problem is a "nonproblem."

4. EVALUATING THOMSON'S ABOUT-FACE

For Thomson, what is crucial to the impermissibility of the bystander's taking option (ii) in *Bystander* is that the one neither consents nor is morally required to consent to his taking it and also that the bystander has a permissible alternative. Thomson (implicitly) offers a principle:

> (&): *X* may make *Y* suffer a harm in doing a good deed only if either (a) *Y* consents (or would consent) to *X*'s doing it, (b) *Y* is morally required to consent to *X*'s doing it, or (c) *X* has no permissible alternative to doing it.

9 Thomson, "Turning the Trolley," 366.

10 Thomson, "Turning the Trolley," 367.

Perhaps we should accept it. "Who are you," you might think, "to make someone else suffer a harm she does not and need not acquiesce to in suffering to achieve an end you need not bring about?"

Should we accept (&)? I think not. To see why, consider:

> *Swedes*: Chang finds himself in the central square of an isolated Swedish mountain village. Against the wall are five innocent teenage boys whom the local sheriff is about to have executed as punishment for a revolt by their parents. Because Chang is a visitor, however, the sheriff offers him the "privilege" of executing one of the five, Sven. If Chang accepts, the sheriff will release the four other boys in celebration of the special occasion. If he declines, then the sheriff will execute all five as he had planned. Chang asks Sven if he consents to his shooting him, but, because the sheriff will kill Sven's sister if he does, Sven refuses.[11]

Swedes is a counterexample to (&). Chang may make Sven suffer the harm of death in order to do the good deed of saving the four other boys even though none of (&)'s clauses is satisfied: Sven does not, nor is he morally required to, consent to Chang's killing him, and Chang does have a permissible alternative to killing Sven, viz., he may refrain from killing him.

But maybe consent is a red herring. Perhaps what is crucial to the impermissibility of the bystander's turning the trolley onto the one in *Bystander* is that the one would not be required to bring the harm in question upon himself in order to bring about the good in question had he the option of doing so. Perhaps what Thomson is (implicitly) appealing to, then, is:

> (&′) *X* may make *Y* suffer a harm, *h*, in doing a good deed, *g*, only if either (a) *Y* consents (or would consent) to *X*'s doing so, (b) were *Y* in a position to bring about *g* at the cost of suffering *h* herself, *Y* would be morally required to bring about *g* and suffer *h*, or (c) *X* has no permissible alternative to doing so.

Maybe this will do the trick.

It will not. Consider:

> *Shepherd*: An out-of-control trolley is barreling toward five track workers who are trapped on the track ahead of it. Nothing can stop the trolley before it reaches the five and there are no side spurs of track onto which the trolley can be diverted. A shepherd standing by the side of the track

11 This is a variation on Williams's famous Jim and the Indians case in Smart and Williams, *Utilitarianism*.

> can use his crook to pull the five off the track before the trolley reaches them (option (ii′)). Unfortunately, if he does this, though he will save the five, the trolley will then run over and kill another track worker trapped a hundred yards behind the five. Were he to not pull the five off the track (option (i)), then though the five would die, the one would not be killed, for the weight of the five dead bodies would halt the trolley well in front of him.

Both (i) and (ii′) are clearly permissible. *Shepherd* is thus a counterexample to (&′). Since the shepherd may take option (ii′) and he has a permissible alternative, viz., (i), and the one would not be required to remove the five from the track if he could, (&′) is false.

These cases cast doubt upon (&) and (&′). What is more, Thomson even grants that it is intuitive that the bystander may turn the trolley in *Bystander*; that is why she needs an argument to toss it overboard. But no argument within the vicinity withstands scrutiny. So, the Trolley Problem is not a nonproblem. It needs a solution.

5. RECONSIDERING THOMSON'S REJECTION OF HER FIRST SOLUTION

Though it does need a solution, we need not look far to find the beginnings of one. Thomson's First Solution contains an important moral insight, which in section 7 I will argue is an integral part of the correct solution. But first I will explain why her Second Solution fails and why her reasons for abandoning her First Solution were bad ones.

Thomson's Second Solution is a nonstarter. According to it, the bystander may turn the trolley in *Bystander* because doing so was in the interest of all six track workers prior to when their track positions were determined. But were that the reason, pushing the big man in *Big Man* would be permissible if the big man himself were also a worker on his lunch break. Pushing the big man in that case would not be permissible, however. Thomson's Second Solution thus fails.[12]

What is more, her reasons for abandoning her First Solution were bad ones. Thomson abandoned her First Solution because she came to endorse:

> *The Means Principle for Rights* (MPR): If (i) *X* has a right against *Y* that *Y* not do *β*, and (ii) if *Y* does *α* then *Y* will thereby do *β*, then *X* has a right against *Y* that *Y* not do *α*, that right being at least as stringent as *X*'s right against *Y* that *Y* not do *β*.

12 Kamm raises a similar kind of objection (*Morality, Mortality*, 2:167).

I grant that if MPR is true, then First Solution fails—if MPR is true, the one *does* have a right that the bystander not turn the trolley (because the one has a right that the bystander not kill him, and if he turns the trolley the bystander will thereby kill him). But we should not accept MPR.

Why should we think MPR is true? Thomson offers two reasons. Here is the first:

> Does anybody have a [right] against me that I not press my doorbell? Nicholas, for example? What on earth could be reason to think he does? Then we learn that if I press the doorbell, I will thereby kill him. (He and a battery are wired to the doorbell.) … [MPR] tells us that he does in fact have a [right] that I not press the doorbell. Admittedly, I would be doing him no wrong if it were not the case that by pressing the doorbell I would do him a harm. But that *is* the case. And don't we think he therefore has a [right]—a very stringent [right]—that I not press the doorbell?[13]

This is unpersuasive. What is true is that if I press my doorbell I will thereby infringe Nicholas's right against me that I not kill him. We need not add to this that Nicholas has *another* right against me that I not press my doorbell.

Nothing is gained, and something may well be lost, by supposing Nicholas has a right that I not press my doorbell. If MPR is correct, Nicholas has *two* distinct rights against me, each as stringent as a right not to be killed, both of which would be infringed if I press my doorbell. This is double counting. Often, if *X* has two rights of stringency *S* against *Y* that *Y* would infringe were *Y* to ϕ, the weights of those rights combine, increasing the moral pressure against *Y*'s ϕ-ing above what it would be if *X* had only one such right. But it would be quite odd, as MPR seems to entail, if faced with a choice between causing *X* a harm directly, i.e., not by doing something else, and causing *Y* a similar harm by ϕ-ing, I ought to harm *X*, for in doing so I only infringe one right of *X*'s against me—that I not harm *X*—whereas if I cause *Y* the harm, I infringe two rights of *Y*'s against me—that I not harm *Y* and that I not ϕ—each equal in stringency to that of *X*'s right against me. Now, a proponent of MPR might say that rights connected via MPR do not combine in the way distinct rights do. However, if they do not, then MPR-generated rights, if such exist, are a very special class of rights, different in kind from all others. But, given that, why think there are any such rights? Surely we need not lean on Nicholas's having a right against me that I not press my doorbell to explain why I ought not to press it. We can explain that by appeal to his having a right that I not kill him and that if I press my doorbell I will violate that right.

But maybe there are other reasons for thinking that Nicholas has a right

13 Thomson, *The Realm of Rights*, 157.

against me that I not press my doorbell. Perhaps Nicholas's having that right is necessary to explain why forcible action may be taken, by Nicholas and others, to prevent me from pressing my doorbell. After all, normally, others may not forcibly act to prevent me from pressing my doorbell. But, again, we need not invest in Nicholas a distinct right against me that I not press my doorbell to explain this. All we need appeal to is the fact that forcibly preventing me from pressing my doorbell is both a necessary and proportionate means of preventing me from violating Nicholas's right against me that I not kill him. Furthermore, suppose that I am whistling as I walk toward the doorbell. If forcibly preventing me from whistling were both a necessary and proportionate means of preventing me from violating Nicholas's right against me that I not kill him—because, say, doing so would distract me long enough to allow Nicholas to detach himself from the battery—then doing so would be permissible. But, note, Nicholas certainly does not have a right against me that I not whistle (nor could MPR deliver the result that he has any such right, for were I to kill him I would not do so *by* whistling). So the supposition is misguided that Nicholas must have a right that I not do whatever preventing me from doing would be both a necessary and proportionate means to preventing me from violating his right that I not kill him. Rather, the explanation is far simpler: *X* may ϕ if ϕ-ing is a necessary and proportionate means to preventing *Y* from violating *X*'s right that *Y* not kill *X*.

Thomson's other reason for thinking MPR is true is that it can explain the truth of:

> *The Means Principle for Permissibility* (MPP): If (i) if *Y* does α then *Y* will do β, and (ii) it is permissible for *Y* to do α, then it is permissible for *Y* to do β.

Again, this is unpersuasive. Even if MPP is true, MPR might well be false. MPR also seems superfluous to the explanation Thomson offers. If a right is infringed in achieving an end, we need not appeal to some distinct right infringed in implementing the means to it to explain the permissibility facts. We can simply appeal to the right infringed in achieving the end those means are means to to do that.

Furthermore, the postulation that MPR is that which explains the truth of MPP is dubious, for MPP clearly applies even in cases in which the permissibility facts are not a function of rights. Suppose *X* can press a blue button, thereby saving *A* from drowning, or she can press a red button, thereby saving *A*, *B*, and *C* from drowning. In this case, neither *A*, *B*, nor *C* has a right against *X* that *X* rescue him.[14] Nevertheless, MPP clearly applies here: it is morally impermissible for *X* to press the blue button in this case because it is not morally permissible for

14 That there is no right to be rescued is something for which Thomson argues explicitly and persuasively in *The Realm of Rights*, 160–63.

him to only save *A*, and if he presses the blue button he will thereby only save *A*. As no rights are in play and yet MPP applies in this case, what explains the truth of MPP must be something far more general than any principle like MPR, which only concerns rights, specifically.

What's more, double-counting worries may well threaten the very compatibility of MPR and MPP. Suppose bringing about a good, *G*, justifies infringing a right, *R*. *G* might be sufficient to justify infringing *one* right of the stringency of *R*, but insufficient to justify infringing *two* rights of *R*'s stringency. If so, we get the very odd result that if I can infringe *R* directly to bring about G, then infringing it would be permissible, but if I could do so only by doing something else, then infringing it would not be permissible. MPR and MPP make uncomfortable bedfellows. MPR does no work that could not be done without it, and it creates unnecessary problems.

I cannot see, nor has Thomson offered, any other reason for thinking that there are the special kinds of rights MPR postulates. There is no explanatory work that needs doing that cannot be done simply and efficiently without them, and were they to exist they would be unlike all other rights in many diverse ways. I conclude both that we have no reason to accept MPR and that parsimony considerations, in addition to the fact that the rejection of MPR offers the prospect of a rights-based solution to the Trolley Problem, give us good reason to reject it.

6. *THREE OPTIONS* AND THE *PRO TANTO* OBLIGATION TO MINIMIZE RIGHT INFRINGEMENTS

Before offering my own preferred solution to the Trolley Problem, I want first to briefly revisit Thomson's discussion of *Three Options*. Though it is inessential to her ultimately failed argument in "Turning the Trolley," I think it is illuminating nonetheless.

Three Options supposedly motivates (*) and *Third Principle*. It does not. I agree that only (i) and (iii) are permissible in *Three Options*: the bystander may not turn the trolley onto the one in that case. However, neither (*) nor *Third Principle* is part of the explanation why. Consider:

> *Shepherd (Three Options)*: Things are as they are in *Shepherd* except that the shepherd has a third option, (iii): he can jump onto the track in front of the five. If he does, though he will be killed, he will save the five because his body will bring the trolley to an abrupt halt.

All of the shepherd's options—(i) do nothing, (ii′) pull the five off the track, and (iii) jump in front of the trolley—are permissible. It is not the case, then, that if

doing a good deed has a cost and a person can pay it himself, then, if he does the good deed, he must pay the cost himself.

So (*) is false. *Third Principle* is, too. Consider:

> *Three Options (Consent)*: Things are as they are in *Three Options* except that if the bystander flips the switch to the right (option (ii)), the one on whom he turns the trolley has freely and rationally consented to having it turned on him to save the five.[15]

In this case, the bystander may take any of his three options. Sometimes it *is* permissible to kill one to save five even if one can instead kill oneself to save them.

Thomson might cry foul here. *Third Principle* was only ever offered as a *ceteris paribus* principle, and in *Three Options (Consent) ceteris* are not *paribus*. In that case, the one has consented to the trolley being turned onto him and perhaps that kind of complication was meant to be excluded by *Third Principle*'s *ceteris paribus* clause.

Now that may be, but I think it obscures a deeper truth. For though *Third Principle*'s *ceteris paribus* clause may catch *Three Options (Consent)* in its net, that case has something in common with *Three Options* that explains *why* the data are as they are in those cases. In both, the one whom it is permissible to kill has no right against the bystander that he not kill him to save the five. In *Three Options (Consent)* the one has waived that right by consenting to being killed, and in *Three Options* there is no such right because the bystander does not have such a right *against himself*.

When I chop off your finger, I infringe your right against me that I not harm you. What follows from this? Many things, including:

> (1) to be permissible (a) the good that I achieve in harming you must be substantially greater than the loss you suffer, (b) perhaps certain particular causal relations between my harming you and the good thereby achieved must obtain, (c) …
>
> (2) prior to harming you I must, other things equal, seek a release from you to harm you, and
>
> (3) after harming you I must, other things equal, compensate you for your loss.[16]

15 All of the points I go on to make in what follows could equally well be made concerning a version of *Three Options* in which option (ii) involves the bystander turning the trolley onto the one who villainously launched it toward the five in the first place.

16 An in-depth discussion of these and many other aspects of rights may be found in Thomson's *The Realm of Rights*.

These are (some of) the hallmarks (defeasible, to be sure) of the right not to be harmed. They all apply whether I chop off your finger or kill you by turning a trolley onto you. This is not so when it comes to harming myself, however. To harm myself permissibly the good I bring about in doing so need not be greater than the harm I suffer. Also, I need not seek a release from myself prior to harming myself, nor do I owe myself compensation afterward. In the case of self-harm, all of the hallmarks of a right not to be harmed are absent. That is as it should be; we do not have any such rights against ourselves. What is more, I submit, this is an important part of the special moral authority we each have, and no one else has, over our own bodies.

Instead of *Third Principle*, then, to explain *Three Options* we can instead appeal to

> *Rights Principle*: *A* must not kill *B* to save five if (a) doing so would infringe a right of *B*'s against *A* that *A* not kill *B*, and (b) *A* can instead kill someone who does not have a right against *A* that *A* not kill him to save the five.

Rights Principle explains what is going on in *Three Options* in a way that is consistent with the data in *Three Options (Consent)* and is more unifying than *Third Principle*. The bystander may do nothing (option (i)) for one may let five die if every other option would lead to the death of someone else who otherwise would not. He may also turn the trolley onto himself (option (iii)), for one may sacrifice oneself to save five others. And he may not take option (ii) because, in accordance with *Rights Principle*, he has at least one option available to him (option (iii)) whereby he saves the five by killing someone (himself) who does not have a right against him that he not kill him.

Rights Principle is not only more unifying than is *Third Principle*, it is also more explanatorily powerful. Take the following variant of *Three Options (Consent)*:

> *Three Options (Consent)'*: Things are as they are in *Three Options* except that the bystander's options are: (i) do nothing; (ii) flip the switch to the right, thereby killing the one who has not consented to having the trolley turned onto him; and (iii) flip the switch to the left, thereby killing someone else who has consented to having the trolley turned onto him.

Rights Principle, unlike *Third Principle*, can explain the moral impermissibility of option (ii) in this case, and it can do so in a manner exactly parallel to the way it does in *Three Options*. Option (i) is permissible for the same reason as it is permissible in *Three Options*. Option (iii) is permissible because, as seems intuitive,

it is permissible to kill one to save five if the one who is killed has no right against the bystander that he not kill him. And given the permissibility of option (iii), the impermissibility of option (ii) follows via *Rights Principle* just as it does in *Three Options*.[17]

Rights Principle is itself an instance of the more general *ceteris paribus* principle:

> *Minimize Right Infringements*: *A* ought not to ϕ in achieving result, *R*, if (a) doing so would infringe a right of stringency, *S*, and (b) *A* can instead achieve *R* without infringing a right of stringency greater than or equal to *S*.[18]

Minimize Rights Infringements is consonant with the picture of rights as means-blocking barriers I mentioned above and this further recommends it as a component of the full story about the ethics of harming in these cases. *Minimize Rights Infringements* embodies the thought that rights constrain the means by which we may harm people to minimize harm overall not only in virtue of those means' causal properties, but also in virtue of their relational properties. It is an affront to one's moral integrity to be harmed to minimize harm overall via certain causal routes; it is also an affront to be harmed to minimize harm overall when harm overall could be minimized without treading upon one's rights. *Minimize Rights Infringements*, then, is a natural elaboration of the very conception of rights to which First Solution makes appeal.

Three Options is explicable, but not by way of any principles that ground an argument for the impermissibility of the bystander's taking option (ii) in *Bystander*. The impermissibility of the bystander's taking option (ii) in *Three Options* is explained by *Rights Principle*, which is just an instance of *Minimize Right Infringements*. What is more, these principles are consistent with the intuitive verdict

17 Does *Rights Principle* not entail in *Shepherd (Three Options)* that (ii) is impermissible given that (iii) is permissible? No. It would do so only if in pulling the five off the track, the shepherd would thereby infringe the right of the one standing behind them that he not kill him. But it is not at all clear that in pulling the five off the track the shepherd does kill the one. And even if it does count as a killing, such a killing is not the kind of killing we, in general, have rights against others that they not perpetrate against us. If I duck a bullet that kills you but would not have killed you had I not ducked to save my own life, then, whether we say I have killed you or not, I have not infringed any right of yours in ducking. The morality of ducking is discussed in Boorse and Sorensen, "Ducking Harm."

18 Interestingly, this is a principle Thomson not only might endorse, but has endorsed explicitly: "However weak a [right], a large increment of good to be got by infringing it does not make infringing the [right] permissible if one has a non-[right]-infringing way of producing the same increment of good, or even a comparably large increment of good" (Thomson, *The Realm of Rights*, 164).

that the bystander may take option (ii) in *Bystander*, for in that case the bystander has no non-right-infringing way of preventing the greater harm in question.

7. TOWARD A SOLUTION TO THE TROLLEY PROBLEM

Thomson's reversal on her own Trolley Problem is ill motivated. The principles to which she appeals in her argument that option (ii) in *Bystander* is impermissible are false. What is more, the moral data in *Three Options*—data that I do not question—can be accommodated by *Rights Principle* and *Minimize Right Infringements*, principles consistent with the permissibility of the bystander's taking option (ii) in *Bystander*. The Trolley Problem, then, is alive and well and in need of a solution. Thomson's Second Solution fails and her grounds for ditching her First Solution in favor of it are bad ones. I think it is correct to place rights infringements at the heart of the solution, as First Solution does, but First Solution does not employ rights infringements in precisely the right way. Here I will sketch what seems to be a more promising way of incorporating rights infringements into a solution to the Trolley Problem.

Recall First Solution: the bystander may take option (ii) in *Bystander* because in doing so he (1) makes what was threatening five come to threaten only one, and (2) does so not by any means which, in themselves, constitute an infringement of any right of the one's. The problem with First Solution is that there are cases in which the moral data are not as First Solution would have them. Here are two cases in which it is permissible to harm the few in order to save the many that do not involve making what was threatening the many threaten the few:

> *Trolley (Avalanche)*: Everything is as it is in *Bystander* except that if the bystander takes option (ii) the trolley will be redirected onto an empty spur of track and collide with a cliff wall thereby causing an avalanche that will crush and kill one person trapped below.
>
> *Trolley (Landslide)*: Everything is as it is in *Bystander* except that the bystander cannot redirect the trolley onto the side spur of track. But he can activate a device (option (ii)) that will launch the five safely onto a bluff high above the tracks. Their landing there, however, would cause a landslide that would crush the one trapped below.[19]

In both cases the bystander may cause the lesser harm, but in neither does he make what was threatening the five threaten the one. Rather, both cases involve

19 These are variants of cases presented in Kamm, *Intricate Ethics*.

the creation of a new threat to the one.[20] True, First Solution only purports to offer a sufficient condition for harming some to save others, and so these cases are not, strictly speaking, counterexamples to it. They do demonstrate, however, that First Solution is not the complete story of the permissibility of harming some to save others and they can help point the way toward a more complete one.

Here is a case in which redirecting a threat from the many onto the few by means which themselves are not an infringement of anyone's rights is impermissible:

> *Trolley (Two Effects)*: Everything is as it is in *Bystander* except that the one is not trapped on the side spur of track but is instead standing alongside it. If the bystander presses a button (option (ii)), his doing so will have two distinct effects: (1) it will cause the trolley to be redirected onto the side spur of track and (2) it will activate a machine that will push the one onto the side spur of track right in front of the oncoming trolley (in such a way that he will be unable to avoid being run over and killed by it).

The bystander may not take option (ii) even though he would be redirecting a threat away from the many onto the few by means that would not infringe anyone's rights—the one has no right that the bystander not press the button nor that he not redirect the trolley. *Trolley (Two Effects)* is thus a counterexample to First Solution.

A slightly different solution is called for. I tentatively propose:

> *Harm Prevention Principle* (HPP): One may cause a right-infringing harm, *h*, that otherwise would not have occurred only if
>
> (i) the total harm caused is less than would occur were *h* not caused,
>
> (ii) *h* is caused by some event, *P*, that is the prevention of some greater harm, *H*, and

20 *Trolley (Avalanche)* and *Trolley (Landslide)* also make trouble for the solution to the Trolley Problem offered in Haslett ("Boulders and Trolleys"). That it is permissible to harm others by creating new threats in these versions of the trolley scenario casts doubt upon solutions to the Trolley Problem, like Haslett's, which permit only *shifting* dangers away from the many onto the fewer, and not *originating* dangers to the fewer in order to save the many. That these cases undermine such solutions is significant because Haslett's solution, like the solution I go on to offer, is an attempt to develop and strengthen Thomson's First Solution to the Trolley Problem. Thanks to an anonymous referee for impressing upon me the significance of *Trolley (Avalanche)* and *Trolley (Landslide)* for Haslett's solution to the Trolley Problem.

(iii) any infringement of the serious rights of those who suffer harm in the course of the causing of *h* neither causes *P* nor causes *h*.[21]

According to HPP, causing a lesser harm can only be justified if that harm is a causal consequence of the prevention of some other greater harm, i.e., the prevention event, and also—and here is where HPP borrows from Thomson's First Solution—any infringements of the serious rights of those who suffer the lesser harm are neither a cause of the prevention event nor a cause of the lesser harm. According to HPP, then, neither the prevention event itself, nor any of the events that either cause it or cause the lesser harm, may be something the one harmed has a serious right against the one causing the harm that he not do. Causing a lesser harm is morally permissible, then, only if the lesser harm lies causally downstream from the prevention of the greater harm, and no serious right infringement of the one who suffers the lesser harm lies causally upstream from either the lesser harm or the prevention of the greater harm.

21 HPP is inspired in large part by considering many of the cases Frances Kamm introduces in *Intricate Ethics*, and reflecting upon her discussion of them. Though HPP is similar to Kamm's theory in many ways, the theories are distinct. For Kamm, what is crucial is that the lesser harm not be causally upstream from a "greater good." HPP requires that all lesser harms be causally downstream from the greater harm prevented. (This makes for a significant difference with Kamm's theory, for a lesser harm can fail to be causally upstream from the greater harm prevented without being causally downstream from it.) HPP also does not allow harming in order to bring about goods, however great they may be, that are not the prevention of greater harms. It may be that Kamm's theory (which is too complex to state in detail here) and HPP are fastening onto the very same phenomena. (Whether they are may well depend on whether by the "greater good" Kamm just means what I have identified as the "prevention event," though Kamm's use of the "greater good" does not settle the matter.) I do not, therefore, offer HPP here as a rival to Kamm's theory. However, in structure, it is quite a bit simpler than is hers. HPP, though it does crucially appeal to the notion of a prevention event, does not make recourse to any of the machinery that figures centrally in Kamm's theory, machinery that includes such notions as: "means to the greater good," "non-causal flipside (of the greater good or means to the greater good)," "the structural equivalent of the greater good," and "means which overlap the involvement-of-the-person part of evil.*" Though precise necessary and sufficient conditions for an event being a prevention event are hard to come by, I think it is not so difficult to pick out in various cases which event is the prevention event. By contrast, Kamm's notions, listed above, are not so easy to identify across cases. Also, like Thomson's theory and unlike Kamm's, rights figure centrally in HPP. This is also a merit of HPP, for the infringement of rights, as Thomson highlighted when she presented her First Solution, does seem to play an important role in the correct explanation of which harmings are and are not permissible. Last, HPP delivers different, and I believe correct, verdicts from those delivered by Kamm's theory in a number of Kamm's own cases: Lazy Susan (With Man Trapped Under It), Vibrations, Component Case II, Worse Than Dead, and Sending Back. (See Kamm, *Intricate Ethics*.) For these reasons, I prefer HPP to Kamm's theory. (HPP is also meant to be supplemented by *Minimize Right Infringements*, of course.)

HPP delivers the intuitively correct verdicts in the cases discussed so far. First, here are the prevention events in each of the cases in which it is permissible to harm the one:

CASES		PREVENTION EVENT
Driver, *Bystander*, and *Trolley (Avalanche)*	→	the turning of the trolley
Shepherd	→	the removal of the five from the track
Trolley (Landslide)	→	the launching of the five off the track

In each of these cases all lesser harms caused are caused by the prevention event and no serious right infringements cause either the prevention event or a lesser harm. Thus, HPP declares harming permissible in each of these cases. Next, here are the prevention events in each of the cases in which it is impermissible to harm the one:

CASES		PREVENTION EVENT
Big Man	→	the trolley colliding with the big man
Trolley (Two Effects)	→	the turning of the trolley

In each of these cases there is an infringement of the serious rights of the one harmed that either causes the prevention event or causes the lesser harm. Thus, HPP declares harming impermissible in each of these cases.

Absolutely crucial to the application of HPP is there being some event in the course of the permissible causing of a lesser harm that is the prevention of a greater harm. In each of the cases above I identified which event is the prevention of the greater harm in that case and noted that the moral data in the case follow from HPP and the event so identified. Though it seems clear in many cases what the prevention event is, it is no easy task to give necessary and sufficient conditions for an event's being the prevention of some harm. That said, however, it is not that there is nothing that can be said.

First, it is certainly not sufficient for an event's being the prevention event that it be such that, had it not occurred, the greater harm would have occurred. That counterfactual is true for many events that are not the prevention event. For instance, the bystander's pressing the button in *Trolley (Two Effects)* (or the thin man's pushing the big man in *Big Man*) is an event that is such that had it

not occurred the greater harm would have occurred, and yet it is not the prevention event, the turning of the trolley is (the big man's halting the trolley is). But, though not sufficient, the holding of that counterfactual does seem to be a necessary condition for an event's being the prevention event. How could an event be that in virtue of which a harm is prevented if it was not such that, had it not occurred, the harm would have occurred? Another thing that seems necessary is that the event involve either the process by which the harm would occur or the victims of that harm were the harm not prevented. Whenever a harm occurs there is some identifiable process by which the harm occurs and some victims who suffer it. For an event to be the prevention event, it seems, it must be an event involving either that process or its potential victims.[22] For instance, the reason why the bystander's pressing the button in *Trolley (Two Effects)* (or the thin man's pushing the big man in *Big Man*) is not the prevention event is precisely because it is not an event involving the process by which the harm would occur were it not prevented; in *Trolley (Two Effects)* (or in *Big Man*), were the harm to the five not prevented, the harm would be caused by the trolley barreling into the five, and so the process by which that harm would occur if it were not prevented is the process of the trolley's barreling toward them.

The distinction I am drawing here—that between an event by which one prevents a harm and the prevention event—can be seen even in cases in which agency plays no role:

> *Trolley (Breeze)*: An out-of-control trolley is barreling toward five track workers who are trapped on the track ahead of it. If nothing stops the trolley, the five will be run over and killed. A gentle breeze dislodges a pebble resting along the cliff face. The pebble bounces down the side of the cliff and strikes a boulder stuck in the cliffside, causing it to roll down the cliff. The boulder collides with the trolley and knocks it off course. The five are saved.

In this case, we can say that the gentle breeze prevented the harm to the five by dislodging the pebble (and that the pebble prevented it by dislodging the boulder), but the breeze's dislodging of the pebble is not itself (nor is the dislodging of the boulder) the prevention event. As in many of the other trolley cases, the

22 It is a further merit of HPP, in contrast with other solutions to the Trolley Problem (including Kamm's), that it focuses on the prevention event. Once it is recognized that whenever a harm is prevented there is a process whereby the harm prevented would have occurred had it not been prevented it naturally falls out, as a matter of course, that all permissible harmings fall either into one class—interferences with an independently identifiable harm-causing process—or another—interferences with the potential victims of the harm-causing process.

prevention event is an event involving the process by which the harm would have been caused were it not prevented, viz., the event of the trolley's being knocked off course by the boulder.

A prevention event can sometimes be an event not involving the harm-causing process; it can also sometimes be an event involving the potential victims of the harm whereby they are rendered no longer susceptible to the harm that the harmful process would cause were it not prevented. So, in *Trolley (Landslide)* the prevention event is the event of the victims being launched out of the way of the oncoming trolley. Here is another example:

> *Trolley (Tractor)*: Everything is as it is in *Bystander* except the bystander cannot redirect the trolley away from the five. Instead, he can (option (ii)) launch a tractor that will gently push the five out of the way of the oncoming trolley. Unfortunately, an innocent person is trapped between the tractor and the five, and so if the bystander launches the tractor, it will run over and crush the one en route to pushing the five out of the way of the trolley.

In this case and in *Trolley (Landslide)* if the bystander takes option (ii), the prevention event is the event involving the potential victims whereby they are rendered insusceptible to the harm. In *Trolley (Landslide)* the prevention event is their being launched to safety, and in Trolley (Tractor) it is their being pushed out of the way by the tractor. But whereas the bystander's taking option (ii) in *Trolley (Landslide)* is permissible, his taking option (ii) in *Trolley (Tractor)* is not, and HPP explains why: in the former, the lesser harm the bystander causes to the one is caused by the prevention event, whereas in the latter, it is not (rather, it is caused by an event that is a cause of the prevention event).[23]

23 It is not clear whether, according to HPP, turning the trolley is permissible in Thomson's Loop Variant of *Bystander* ("The Trolley Problem"). In this variant (call it *Loop*), the side track, on which the one is trapped and onto which the bystander can redirect the trolley, loops around and reconnects with the main track headed toward the five. If the bystander redirects the trolley onto the one, the one's being killed will halt the trolley, thereby preventing it from looping back around and running over and killing the five. Though I am inclined to think that it is permissible to redirect in *Loop*, I grant that one might reasonably disagree. In fact there are two simple and natural alternative additions to the partial characterization of what it is for an event to be the harm prevention event I have offered, each of which delivers one of the two possible verdicts about redirecting the trolley in *Loop*:

> Addition #1: An event is the harm-prevention event only if it is the *earliest* event involving either the harm-causing process or the potential victims of the harm such that had it not happened the harm would have occurred.
>
> Addition #2: An event is the harm-prevention event only if it is the *latest* event in-

What more can be said about what it is for an event to be the event in virtue of which a harm is prevented? It is not clear. The notion of a prevention event is no more precise than that either of a threat or of the means by which a threat is redirected, and so HPP has no advantage over Thomson's First Solution in terms of precision. Though in many cases it is intuitively clear which event is the prevention event, in many other cases this is not so clear. With respect to these latter cases, so long as it is also unclear whether it is permissible for the agent in question to proceed, they constitute no grounds for rejecting HPP.

I have argued that *HPP* can explain the moral data in cases in which agents do harm in order to prevent greater harm to others. But even if HPP is correct, one might wonder: *why* is it correct? One reply—the easy reply—is that it is correct because it best accounts for the data. To an extent this is a good reply; what better test of an explanation is there, after all, than whether it best accounts for the data? On the other hand, this reply might leave one cold. One might want something more unifying, that which "lies behind" HPP.

I think that which lies behind HPP is a thought that might plausibly be thought to be at the bedrock of non-consequentialist moral theory. The thought is that which is encapsulated in the foremost of the prescriptions of the Hippocratic oath, viz., "do no harm." Now HPP does indeed allow harming in certain cases, but it does so only when the harms it allows are ones that are the causal upshots of the prevention of yet greater harms. And this is because, in a way, when the lesser harm is caused by the prevention of the greater harm, what one does in jointly preventing the greater harm and causing the lesser harm is to *transform*

---

volving either the harm-causing process or the potential victims of the harm such that had it not happened the harm would have occurred.

With Addition #1, redirecting the trolley in *Loop* is morally permissible; the earliest event involving the harm-causing process such that had it not happened the harm would have occurred is the turning of the trolley, and if that is the harm-prevention event, then HPP declares turning the trolley permissible. With Addition #2, redirecting the trolley in *Loop* is morally impermissible; the latest event involving the harm-causing process such that had it not happened the harm would have occurred is the trolley's hitting the one, and if that is the harm-prevention event, then HPP declares turning the trolley morally impermissible. That the framework provided by HPP can explain why it is that the moral data in *Loop* are much disputed may well be another mark in its favor over other rival solutions to the Trolley Problem—each of Addition #1 and Addition #2 is a natural and plausible filling in of the account of what it is for an event to be the prevention event, and so, perhaps, what is at the root of the dispute between proponents and opponents of the permissibility of turning the trolley in *Loop* is their respective implicit acceptance of one or the other of Additions #1 and #2. And because Addition #1 and Addition #2 are both easily incorporatable into an account of what it is to be the prevention event, HPP is more flexible than other solutions, like Kamm's, for instance, which take the permissibility of turning the trolley in *Loop* as a datum.

the greater harm into the lesser harm. On this way of thinking, when one causes the lesser harm, one is not so much introducing a new harm into the world as transforming a greater harm into a lesser one. And if this is right, then HPP can be seen as a slight modification of the very intuitive thought that we are morally required to avoid harming others. The modification that this interpretation of HPP suggests is that what morality requires is not, strictly speaking, that we not harm others, but that we avoid introducing *new harms* into the world, where a new harm is one that has not been transformed from some other harm. This, if right, can explain the particular requirement that the lesser harm be *caused by* the prevention of the greater harm. This is because, you might think, a greater harm is not transformed *into* a lesser one if the prevention of the greater harm is not causally prior to the lesser harm into which it is transformed. For if the lesser harm does not come out of, or emerge from, the prevention of the greater harm, then its existence is antecedent to, and thus, in a sense, independent of, the prevention of the greater harm. And if the lesser harm has an existence antecedent to, and independent of, the prevention of the greater harm, then the greater harm that has been prevented has not been *transformed into* the lesser one. Another way of putting this is to say that morality requires that we not introduce any harms into the world that are not already "paid for" by the overall reduction of harm suffered in the world. (Here, again, the "already" is crucial; it is what accounts for the requirement that whatever harms one causes lie causally downstream from the prevention of a greater harm.)

But the thought that the transformation of a greater harm into a lesser harm can be seen as not really bringing another new harm into the world is only part of that which lies behind HPP. For not only does HPP require that the harm prevented be greater than the harm caused and that the prevention of the greater harm be that which causes the harm caused, but it also requires that any right infringement that occurs in the transformation of a greater harm into a lesser harm not be the cause either of the prevention of the greater harm or of the lesser harm. Here the thought connects up once again with the conception of rights as means-blocking barriers. If in minimizing harm overall one is transforming a greater harm into a lesser harm, then rights will act as barriers to that transformation. And insofar as the transformation of the greater harm into the lesser harm is constituted jointly by the prevention of the greater harm and the occurrence of the lesser harm, the means to the transformation of the greater harm into the lesser harm are the causes of those two events. So rights, being means-blocking barriers to the transformation of greater harms into lesser ones, are moral barriers to causing the two events that constitute the transformation of the greater harm into the lesser harm. And so, for a transformation of a greater

harm into a lesser harm to be permissible, none of the causes of the two events that constitute it, i.e., the prevention event and the lesser harm, may be a serious right infringement of those suffering harm.

Now all of this is, to be sure, sketchy and somewhat inchoate. I have appealed to the fact that a greater harm causing a lesser harm is not "the bringing into the world of a new harm," but merely "the transformation of a greater harm into a lesser one." And I have also claimed that rights being means-blocking barriers to the transformation of greater harms into lesser harms requires that the means to that transformation, i.e., the causes of the two events that constitute the transformation (the prevention event and the lesser harm), not involve any serious rights infringements of those suffering harm. I admit that all of this is, at best, metaphorical. But in trying to offer the thoughts that lie behind whatever fundamental principles we think constitute the core of the morality of harming, we really are striking bedrock. When we reach this depth of explanation, it is hard to provide something more concrete than HPP, and so if we do try to dig deeper, we are bound to drift somewhat toward metaphor.

HPP, then, is my tentative proposal for the principle governing permissible right-infringing harms. It (or some suitably modified version of it) can, I am optimistic, account for the moral data in these and other cases in which harm is done to prevent yet other harms. It also seems to capture an intuitive thought that we are allowed to cause harm only when our doing so transforms a greater harm into a lesser one without steamrolling others' inviolability, as embodied in the rights they have, on the way to doing it.

8. CONCLUSION

Thomson's twist on Foot's original Trolley Problem introduced a particularly difficult puzzle for non-consequentialist moral theory. Though Thomson has of late come to think she was wrong to reject Foot's solution, I have argued that she should not have changed her tune. I have explained where Thomson's argument goes wrong and how the data in the cases she employs to motivate her dismissal of the Trolley Problem as a nonproblem can be accounted for without licensing that conclusion. I have also gestured toward a solution to the problem that seems more promising.[24]

University of Massachusetts
pgraham@philos.umass.edu

24 I would like to thank Michael Otsuka and two anonymous referees for comments on earlier drafts of this paper.

REFERENCES

Boorse, Christopher, and Roy A. Sorensen. "Ducking Harm." *Journal of Philosophy* 85, no. 3 (March 1988): 115–34.

Dworkin, Ronald. "Rights as Trumps." In *Theories of Rights*, edited by Jeremy Waldron, 153–67. Oxford: Oxford University Press, 1984.

Foot, Philippa. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5 (1967): 5–15.

Haslett, D. W. "Boulders and Trolleys." *Utilitas* 23, no. 3 (September 2011): 268–87.

Kamm, Francis M. *Intricate Ethics*. Oxford: Oxford University Press, 2007.

———. *Morality, Mortality*. Vol. 2, *Rights, Duties, and Status*. New York: Oxford University Press, 1996.

Smart, J. J. C., and Bernard Williams. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press, 1973.

Thomson, Judith Jarvis. "Killing, Letting Die, and the Trolley Problem." *The Monist* 59, no. 2 (April 1976): 204–17.

———. *The Realm of Rights*. Cambridge, MA: Harvard University Press, 1990.

———. "The Trolley Problem." *Yale Law Journal* 94, no. 6 (May 1985): 1395–1415.

———. "Turning the Trolley." *Philosophy and Public Affairs* 36, no. 4 (Fall 2008): 359–74.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 2 · NOVEMBER 2017
https://doi.org/10.26556/jesp.v12i2.224

HYPOCRISY AND MORAL AUTHORITY

Jessica Isserow and Colin Klein

HYPOCRITES INVITE MORAL OPPROBRIUM, and charges of hypocrisy are a significant and widespread feature of our moral lives. Yet it remains unclear what hypocrites have in common, or what is distinctively bad about them. We propose that hypocrites are persons who have *undermined their claim to moral authority*. Since this self-undermining can occur in a number of ways, our account construes hypocrisy as multiply realizable. As we explain, a person's moral authority refers to a kind of standing that they occupy within a particular moral community. This status is both socially important and normatively precarious. Hence, moral agents are right to be vigilant when it comes to hypocrisy, and are often justified in their outrage when they detect it. We further argue that our view can preserve what is attractive in rival accounts, while avoiding their associated problems.

1. INTRODUCTION

Everyone has been outraged by a hypocrite. Perhaps it was a moralistic vegan friend who managed to sneak the occasional steak. Maybe it was a coworker whose proselytizing piety did not keep them from sleeping in on Sundays. Or perhaps it was a friendly neighbor, a self-advertised keen and green recycler who (you couldn't help but notice) was consistently too lazy to separate plastic from cardboard. Each looks like a hypocrite, especially to the uncharitable eye. Yet what (if anything) do they all have in common?

One obvious and distinctive feature that hypocrites seem to share is a kind of mismatch between their pronouncements and their actions.[1] The hypocrite, we tend to think, is someone who says one thing but does another. Yet this cannot be the whole story. Mismatches between our words and our actions are common enough; most of us are occasionally inconsistent. We often change our

1 Szabados and Soifer, "Hypocrisy, Change of Mind, and Weakness of Will," 61; McKinnon, "Hypocrisy and the Good of Character Possession," 716; Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 308.

perspectives on moral issues over time. And sometimes we simply succumb to weakness of will, shamefully scoffing down that chocolate mud cake after weeks of touting the health benefits of the Atkins Diet.[2] So mere mismatches cannot suffice to single out the phenomenon of hypocrisy. The hypocrite's failure seems to differ in kind from these other forms of inconsistency—not just in degree.

What exactly *are* hypocrites, then? The question is worth asking. Judgments of hypocrisy play an important and widespread role in our moral lives. Public figures are the most obvious cases: lapsed politicians and Tartuffian priests are common targets of hypocrisy judgments. But private individuals are equally eligible. The vegetarian roommate caught sneaking bacon or the philandering friend who condemns adultery are no less hypocritical than those in the public eye. Nor are judgments of hypocrisy restricted to individuals. The United States is frequently accused of hypocrisy with respect to foreign policy or domestic surveillance. Given the importance and breadth of the phenomenon, it deserves an adequate account.

Getting clear on hypocrisy is also a morally important project. The ascription of hypocrisy is a serious charge. Hypocrites tend to invite moral opprobrium—we condemn them, and usually quite harshly.[3] One would hope that this moral censure is warranted. It would be disquieting if, on reflection, hypocrites turned out to be guilty of some relatively minor wrong. If so, we ourselves may have subjected them to disproportionate moral sanction.[4]

Yet it is far from obvious just what the hypocrite's distinctive failure *is*. Some hypocrites deceive, and some manipulate. However, it would be surprising if they were *merely* guilty of those vices. If hypocrites were just liars, we would need no special category to condemn them—we *already* condemn liars. Similar-

2 As these examples suggest, hypocrisy is structurally similar to change of mind and weakness of will. Some might even suspect that our failed Atkins dieter *is* a hypocrite. (Though this is likely to depend on how they fill in the details of the case.) Teasing apart hypocrisy from other, closely related phenomena is a worthwhile project; indeed, it seems that weakness of will is often invoked as an alternative defense against a charge of hypocrisy. Sorting out that relation outstrips our project here, but for promising attempts, see McKinnon, "Hypocrisy and the Good of Character Possession," and Szabados and Soifer, "Hypocrisy, Change of Mind, and Weakness of Will."

3 Following Szabados, "hypocrite" seems to be "a well-established term of moral condemnation" ("Hypocrisy," 202). Hypocrisy is at the very least something typically "viewed with repugnance" (Statman, "Hypocrisy and Self-Deception," 57). Indeed, some go so far as to describe it as "the only unforgivable sin" (Shklar, "Let Us Not Be Hypocritical," 1). That hypocrites are traditionally met with disdain comes as no surprise. Dante (literally) had a special place in hell for them.

4 On this issue, see Statman ("Hypocrisy and Self-Deception"), who suggests that hypocrisy does not always warrant the extreme moral censure that it invites.

ly, we find various kinds of manipulation bad quite on their own. An adequate account of hypocrisy can therefore shed some much-needed light on the apparent normative significance of this category: what exactly is the moral failure of which hypocrites are guilty, and why is this something that we should want to keep track of?

In this paper, we develop an account of hypocrisy that vindicates the idea that hypocrites form a distinct moral category, supports the intuition that they typically deserve moral censure, and illuminates the important social roles that hypocrisy attributions play in our moral lives. Specifically, we argue that hypocrites are *persons who have, by mismatch between judgments and actions, undermined their claim to moral authority*, where (very roughly), a person's moral authority is understood as a kind of standing that they occupy within a particular moral community—a status that is intimately tied up with their capacity to (1) warrant esteem, and (2) bestow (dis)esteem on others. Since an agent can undermine their moral authority in many ways, our account construes hypocrisy as multiply realizable.

Two features of our discussion are worth noting from the outset. First, our explanatory target is the *hypocrite*—that is, one who is guilty of hypocrisy. However, we do not assume here that hypocrisy amounts to a full-fledged character trait, or even to a particularly strong disposition. Perhaps it might in fiction, but Tartuffe is, we take it, something of a limiting case. As far as our day-to-day moral evaluations go, we seem perfectly able (and indeed, perfectly willing) to condemn someone as a hypocrite following *just one* instance of hypocrisy on their part. If charges of hypocrisy were ascriptions of some more robust character trait, then it is puzzling why we see fit to levy these charges without evidence of a more consistent pattern of behavior.[5] Many accounts have, we suggest, gone astray by assuming that hypocrisy amounts to a kind of settled disposition. Not all hypocrites exhibit the robust scheming dispositions of Tartuffe or Uriah Heep. To assume as much is to risk offering a caricature of hypocrisy—not something that can shed much light on the normative significance of the phenomenon as it operates in everyday life.

Second, our explanatory project is paradigm-based. Our strategy will be to first identify (what we take to be) exemplars of hypocrisy—pious priests who are secretly corrupt, homophobic senators whose private dalliances conflict

5 It is possible that "hypocrite," like "liar," is ambiguous between (1) an individual who is guilty of hypocrisy (or lying) and (2) an individual with a strong disposition to be hypocritical (or to lie). As we argue, the former understanding seems closer to the term "hypocrite" as it functions within ordinary usage. Since our discussion centers on everyday attributions of hypocrisy, we take (1) to be a more fitting candidate for our *explanandum*.

with their explicit disavowals, and the like. After having pinned down the explanatorily basic features of these paradigm cases, we develop an account of hypocrisy that can accommodate them. We then argue that more common or garden instances of hypocrisy can fruitfully be understood in similar terms.[6] The core features that we find in paradigmatic instances of hypocrisy are, we suggest, present in more banal cases as well, though they are present to varying degrees, and in less obvious and pronounced ways.

The road ahead is as follows. We begin by introducing the basic idea of a moral authority, explaining the important roles that moral authorities play in the broader community (section 2). We then turn our attention to paradigmatic instances of hypocrisy (section 3). Our considered hypothesis will be that paradigmatic hypocrites are best understood as individuals who have, by their actions, undermined their claim to acting as a moral authority. We then explain how other instances of hypocrisy—those which are less clear-cut, or non-paradigmatic—can fruitfully be understood in these terms as well. Following that, we argue that our account fares better than its rivals along a number of dimensions (section 4). Our proposal is particularly helpful in supplying a general framework that incorporates the insights of these other views while also diagnosing where they go astray. Finally, and as we explain, our account sheds some much-needed light on the important social roles that judgments of hypocrisy play within and across contemporary moral communities (section 5). In particular, we suggest that charges of hypocrisy offer morally homogeneous communities some degree of protection, and moderate tensions within heterogeneous ones.

2. MORAL AUTHORITY

Central to our account of hypocrisy is the notion of a *moral authority*. An individual's moral authority refers to a certain kind of social status that they enjoy within a particular moral community. As we now explain, moral authorities are prevalent within society, and they serve a range of important functions in moral life.

6 Our investigative strategy is similar to that of Miranda Fricker, who has recently developed a paradigm-based account of *blame* ("What's the Point of Blame? A Paradigm Based Explanation"). Following Fricker, we think that there is much to be gained from pursuing a paradigm-based explanation for multifaceted and diverse phenomena, which tend to elude simple and straightforward analysis. Our approach also bears some similarities to Aristotle's discussion of "being." On G. E. L. Owen's well-known interpretation, Aristotle takes "being" to have a "focal meaning," which serves as the explanatory basis for its many different senses ("Logic and Metaphysics in Some Earlier Works of Aristotle"). We thank an anonymous referee for pointing this out.

2.1. The Basic Idea

Suppose I am faced with a difficult moral decision, such that I am genuinely uncertain how to act. A natural move would be to turn to others for help. I might ask for advice. Or, I might simply examine how others act in similar situations. Some will be more suited to this role than others. The judgments and actions of my peers, for example, are of mixed utility. It is good to know what my neighbors do, but they might be wrong. (This is especially pressing if, say, my conundrum involves the possibility that I am complicit in a widespread injustice.) What would be far more helpful would be to look to someone whom I considered especially trustworthy in these matters. If I were religious, an obvious choice would be a priest. But I might equally well turn to a wise grandparent, a popular friend, or an experienced colleague. Or I might look further afield, to the judgments and actions of politicians, inspirational speakers, televangelists, or popes.

Call someone who plays this special role a *moral authority*. In paradigmatic cases, a moral authority is someone whose moral pronouncements a community takes especially seriously. We look up to moral authorities, and we turn to them for moral guidance. We try to refrain from the sort of behavior that they condemn, and to become the kind of persons to whom they lend praise. Moral authorities are invested with this special status (in part) because they are thought to be especially good at living a decent moral life. And their moral assessments matter more than most; because others look to them for guidance as well, their judgments usually have more direct and tangible consequences on how *we* are judged within the relevant community.

2.2. The Role of Moral Authority

Moral authorities, we submit, have *practical* rather than just *epistemic* authority. We take the advice or criticism of moral authorities to be *pro tanto* action-guiding—not merely a reliable source of information regarding our moral duties. In this sense, moral authorities serve an analogous function to political authorities. Specifically, they play three important roles, parallel to the three traditional roles of political authorities: making laws, adjudicating whether laws have been broken, and meting out punishment and reward.[7]

The third parallel is the most obvious. Paradigmatic moral authorities (e.g., judges, medieval priests) often have direct access to the traditional tools of coercion provided by the state. Such power gives their moral judgments special bite. Yet this feature is neither necessary nor even especially common. Most moral authorities simply wield the powerful sanctioning tool of moral disapproval.

7 Locke, *Second Treatise of Government*.

Blame and resentment are, we assume, social punishments in their own right. Moral criticism can operate in a similar way as legal sanctions, and have parallel effects on agents' behavior.[8]

The sanctioning power of a moral authority is amplified in several ways by their standing within the relevant community. Because moral authorities are held in high regard, people who look up to them tend to emulate their patterns of disapproval. So their criticisms cannot simply be ignored in the way that we might shrug off the judgment of a less worthy peer. This is not to say that a moral authority need condemn us from the pulpit. Indeed, in many cases they need not know us at all: the brimstone preacher who condemns dancing may not even know me. What is important is that through their sermons my community will come to look askance at all who step into dancehalls, including me.

While enforcement is the most obvious parallel with political authorities, moral authorities also serve the other two functions. We sometimes look to moral authorities to adjudicate in edge cases. I accept that dancing is sinful—but what about a first dance at a wedding? I know fraud is intolerable—but is it wrong to sell my textbook evaluation copies to the shifty guy who makes the rounds every semester? One's peers may well differ on whether these count as transgressions. If so, it is commonplace to seek out a more trusted person for advice and counsel.

Note that the adjudicative function of moral authority really has two roots. For one, moral authorities are taken to be particularly good judges of the right thing to do: one reason they have an esteemed function is that they seem to get things right. It is worth emphasizing, however, that their *authority* is not merely reducible to good judgment. Moral authorities are thought to be especially good at living a morally decent life. Theoretical knowledge of the moral facts clearly is not sufficient for doing so.[9] Moreover, such knowledge would seem to be poor grounds for investing someone with *practical* authority; following Estlund, one ought to be wary of confusing experts with bosses.[10] Thus, moral authorities are not *simply* epistemic authorities. That said, they occupy a trusted position, and so they do have a correspondingly greater responsibility to get things right. If they lack the knowledge or skill, then they ought to have backed down, lest they lead others astray.

8 Dworkin, "Morally Speaking," 187.

9 Indeed, experimental studies suggest that ethics professors (who would seem to be prime candidates for moral experts) are no more likely to act in accordance with their explicit moral views than professors in other fields. Schwitzgebel and Rust, "The Moral Behavior of Ethicists."

10 Estlund, *Democratic Authority*.

For another, we might ask moral authorities to adjudicate precisely because of their sanctioning power. I may ask my department chair about selling my sample textbooks partly because I trust their judgment and partly because they are the one who would decide whether I have overstepped the bounds of propriety. On a smaller scale, we might turn to moral authorities simply because they provide us with *evidence* of the punitive responses that await our moral transgressions. My environmentally conscious friend may not be the CEO of Greenpeace, but they claim far more authority with respect to environmental issues than anyone else within my close circle. So it is useful to turn to them for advice in these matters—doing so renders me less vulnerable to others' judging eyes.

Finally, moral authorities can play something like a law-giving role. This is not to suggest that they literally construct or determine the moral facts (a suggestion at which many would balk). It is only to suggest that they plausibly play an important role in determining what a particular group of agents takes those facts to be. It is wrong to use racial slurs. But *which* speech acts count as slurs is often up in the air, and moral leaders can play an important role in making sure that everyone is on the same page. Similarly so for facts about the severity of a transgression. Everyone agrees that sexual harassment is bad, but respected members of a philosophy department (say) play an important role in determining whether others treat harassment as trivial or grave. In many cases, the adjudicative and the law-giving shade into one another, but both aspects are important.

2.3. The Elevation of Moral Authority

According to our story, a moral authority's considered judgments are taken to be action-guiding because they have come to occupy a special standing within a particular moral community. But we have said very little about how moral authorities get elevated—just how do they earn their moral street credit? This is a complex social question. Perhaps some are born to authority, some achieve it, and some have moral authority thrust upon them.

We suspect that the second is most common; typically, moral authorities earn their status through an active investment in moral issues. For the most part, such persons exhibit their investment by moralizing—through moral criticism, deliverances of praise and blame, and the like. Doing so earns them a special status for a number of reasons.

First, moralizing is itself a morally valenced action, and so worthy of esteem in its own right. The whistle-blower who condemns injustice at great personal cost is worthy of esteem not (just) because they made the right judgment, but because the *action* of moralizing was itself a difficult one. This is even so when moralizing takes the form of praise or blame: both require effort, expend

some degree of social capital, and expose us to criticism and resentment in turn. Punishment thus comes at a cost, and large communities can only form if some are willing to take on that burden.[11] Most of us free ride on the moralizing of others, so we tend to admire those who do the work and take on the associated risks.

Second, an important link is typically taken to hold between practice and preaching. *Ceteris paribus*, we assume that a person's moral pronouncements reflect their dispositions—that they are disposed to perform those acts to which they lend praise, and to refrain from the sorts of behavior that they criticize. This is not merely a feature of paradigmatic moral leaders. For the most part, we presume that there are "certain minimal connections" between people's professed judgments and their "desires, intentions, and actions."[12] We take a religious friend's praising chastity, for example, as evidence that they are likely (or at least more likely than a randomly chosen person) to be someone who is motivated to refrain from a life of sin.[13]

It is worth addressing some potential concerns before proceeding. First, although the idea seems intuitive enough, one might wonder *why* exactly our moral words are typically taken to have implications for our behavior. While fully addressing this question is well beyond the scope of this paper, we suspect that the presumed connection here is, in large part, owing to conversational norms. As Wallace suggests, there seems to be a foundational conversational assumption underlying much of our moral intercourse "that our interlocutors would not put forward criticism of other people if they lacked the standing or entitlement to do so."[14] It is important here to distinguish two senses of "criticism": (1) the expression of a negative moral opinion, and (2) an act of condemnation.[15] On the latter understanding (with which we are concerned here), criticism is a kind of speech act that an agent must be in a position to perform. When we declare that we are "not in a position" to criticize others, we refer to some kind of "disabling fact" about ourselves that undermines the illocutionary force of our utterance. And one prime candidate for such a disabling fact would surely be that we ourselves are guilty of the relevant vice. It is for this reason that an agent who

11 See Boyd et al., "The Evolution of Altruistic Punishment."

12 McKinnon, "Hypocrisy, with a Note on Integrity," 327.

13 This is not to assume the truth of "motivational internalism"—that is, an especially tight conceptual connection between moral judgment and motivation. We only assume that people's moral judgments tend to shape their behavior in important ways. This seems plausible; moral judgments certainly do not appear to be epiphenomenal.

14 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 317.

15 Cohen, "Casting the First Stone?"

criticizes is, absent any evidence to the contrary, typically taken to be free of the relevant vices; for they present themselves as one who is capable of performing the speech act—one who is *in a position* to criticize.

Accordingly, moralizing will not always suffice to lay claim to moral authority. Certain kinds of disowning prefaces can remove the impression that one is actively invested in moral issues, or committed to particular values, among them: "Well, I'm not in a position to criticize him, but . . . ," and "Before I begin, I should specify that you ought to do as I say and not as I do." These prefaces can serve to blunt the force or status of the illocutionary act, perhaps rendering it an act of criticism rather than one of condemnation.[16] The costs of moralizing thus go down; the speaker takes on less risk, and is exposed to lesser reproach. Typically, these prefaces also remove the impression that the speaker is disposed to act on their expressed moral opinions; for they suggest an absence of (or diminished) moral commitment.

A small but important clarification is needed here. We do not deny that those who are in "no position to criticize" can succeed in saying something that is both true and well-supported by reasons.[17] Suppose that my friend criticizes me on account of my penchant for buying fur. She may be perfectly correct in thinking that I am to blame for my latest mink coat purchase. She might even cite the right sorts of reasons in support of her judgment. We should not swiftly infer from the many fur coats that adorn *her* wardrobe that this judgment is wrong or ill-supported.[18] Her inconsistency suggests that we ought to question the strength of her moral commitments. But we should not necessarily question the content of her moral advice. (Though we may sometimes be indirectly justified in doing so—more on this in section 3.1.)

One may also take issue with another aspect of our story thus far: moral authority seems to come far too *cheap*. Perhaps we ought to refrain from investing others with any kind of trusted status prior to seeing them hard at work. A far more straightforward (and often far more costly) way to exhibit investment in moral issues is to *take action.* Surely we can often see the virtuous at work? We can, for instance, observe the religious leader as they shun the seductress, feed

16 One might want to claim that it is only the *perlocutionary* force of the utterance that is compromised; the condemnations are properly counted as condemnations, but the addressee will not feel their sting. (We thank an anonymous referee for suggesting this alternative way of looking at things.) Our arguments do not depend on going one way or the other.

17 We thank an anonymous referee for raising this issue.

18 To do so would be to commit the *ad hominem tu quoque* fallacy; the mistake of inferring from an individual's shortcomings that her judgments must likewise be defective. For an edifying discussion of this tempting fallacy in relation to hypocrisy, see Aikin, "*Tu Quoque* Arguments and the Significance of Hypocrisy."

the poor, and refuse lavish gifts. Perhaps so. But this cannot be the whole story. We cannot, after all, observe everyone all of the time. Constraints on time and resources mean that we often have to take people at their word. Absent any evidence to the contrary, then, we tend to grant them the benefit of the doubt; we assume that their moral pronouncements are more or less accurate reflections of what lies within.

Finally, one might worry that, far from earning an agent esteem, moralizing can often have exactly the opposite effect. The "pursed lipped prigs and professional offence-takers" of the world—those who "travel through life looking for things of which to disapprove"—are hardly liked, let alone turned to for guidance.[19] Such persons seem likely to earn reputations as priggish, sanctimonious, and pompous fools—not *moral authorities*.

This worry mistakes the nature of moral esteem; we need not find someone personally likeable in order to look up to them as a moral agent. That said, the concern is not baseless. An agent is plausibly less likely to earn the status of a moral authority when their moralizing borders on fanaticism (though, no doubt, this does sometimes happen). Like the pursuit of happiness or spontaneity, then, we suspect that the explicit and intentional pursuit of moral authority often proves self-defeating—especially when it is not carried out with finesse and due caution.

3. HYPOCRITES AS SELF-UNDERMINED MORAL AUTHORITIES

We have devoted quite a bit of space to spelling out the notion of a moral authority. Our doing so was not without good reason, for we take this social standing to be central to the phenomenon of hypocrisy. Such thinking forms the basis of the proposal that we shall now proceed to develop. The account is especially well-suited to capture paradigmatic instances of hypocrisy. However, and as we explain, it easily generalizes to the smaller scale as well.

3.1 The Account

We begin by exploring the core features of paradigmatic instances of hypocrisy: the homophobic senator caught with his pants down in the men's room, the vocal PETA advocate who occasionally sneaks in some bacon, the corrupt priest, and the like. We take such cases to be prime candidates for exemplars. The agents involved are uncontroversially hypocrites, and they seem especially likely

19 Lenman, "Ethics without Errors," 396.

to attract opprobrium. To our minds, the features of primary importance in such cases are the following.[20]

First, these cases involve a characteristic sort of mismatch between an agent's behavior and their pronouncements; the hypocrite expresses an unfavorable moral opinion of some action, and is then caught performing that very same action. Second, the pronouncements in question tend to involve *criticism*. Typically, the hypocrite will *blame* others for the very vice of which they themselves are guilty. Third, the hypocrite occupies a trusted position, at least within a particular community, and with respect to a particular moral issue. And more often than not, that trust carries with it a firm expectation that they will *not* behave precisely as they do.

Finally, the hypocrite's inconsistency is taken to carry important implications not only for our estimation of them, but also for their terms of interaction with others. It is not merely that we doubt the hypocrite's integrity, or question their moral compass. We also tend to think that they are no longer warranted in relating to others on particular terms—in condemning certain kinds of behavior, for example. The hypocrite's inconsistency bears upon their status as a moral attester, and (as we explain shortly), it may sometimes undercut our reasons for trusting the soundness of their judgment as well. Though there are perhaps other hallmarks of hypocrisy, we expect that these features will be widely agreed upon, and, to our minds, they are of the most fundamental importance.

In light of these core features, our considered hypothesis is the following: paradigmatic hypocrites are persons who *by mismatch between judgments and actions, have undermined any claim they have to act as a moral authority*. In paradigmatic cases, the relevant individual was considered a moral authority by their community, but their actions are taken to suggest that the esteem and deference extended to them was not deserved.

To demonstrate, consider our PETA advocate. It is plausible that they function as a moral authority within the community of animal rights activists—helping encourage strong disapproval of the animal meat industry, clarifying which products are cruelty-free, and so on. Clearly, however, there is a mismatch between the advocate's pronouncements and their actions. (There is an obvious tension between forcefully campaigning against the meat industry and indulging in bacon.) Once this inconsistency is discovered, moreover, it can be expected to have important implications for their standing within the relevant community. Not only is esteem likely to be retracted, but others are also likely to feel that the

20 Keep in mind that we regard these as *exemplary* features; it is not our contention that each is necessary for hypocrisy more generally.

PETA advocate is no longer warranted in condemning them for the occasional indulgence at the local steak house.

To be clear, none of this is to suggest that the PETA advocate's judgment is *false*. They may very well be correct; I may be vulnerable to legitimate moral criticism on account of supporting the meat industry. But it would seem that *they* are in no position to condemn me, given their own meat-eating habits. Of course, the PETA advocate may take themselves to have good all-things-considered reasons to criticize others, even if they lack the standing or entitlement to do so. (Even meat eaters with a vegetarian streak may think it worthwhile to inform others about the cruelty of the animal meat industry—perhaps even as they sit down to a steak dinner.) But for reasons spelled out above, it is debatable how edifying their message will be. More importantly, we think that they will only have relevant supporting reasons to fall back on as they make their case—not their warrant as a moral authority.[21]

The considerations above notwithstanding, it is not implausible that the PETA advocate's inconsistency *does* give us some reason to question the soundness of their judgment. Since they do not seem to take the cause sufficiently seriously, they may lose a certain measure of trust; perhaps they cannot be fully relied on to distinguish cruelty-free products from others. Or perhaps their failure to live up to their own moral standards should suggest to us that those standards are unacceptably over-demanding.[22]

Understanding cases of paradigmatic hypocrisy through the lens of moral authority can, we propose, illuminate their centrally important features. Our proposal nicely explains why paradigmatic hypocrites tend to be public figures; such persons are especially well-placed to earn our esteem and trust. (One would expect that they have also racked up their fair share of enemies along the way.) So it is likely to be especially infuriating (or especially satisfying, depending on one's perspective), when they fail to live up to their own lofty standards. Moreover, our account captures the sense in which a hypocrite's relations to others are transformed by their actions. These relations are, we propose, precisely their claim to moral authority; it concerns their esteemed status, together with their ability to fulfill the relevant sanctioning and adjudicatory roles.

Though our proposal is well-suited to accommodate such cases, one might worry that paradigmatic moral authorities are far too rare to account for wide-

21 We thank an anonymous referee for pressing us on this point.

22 Aikin, "*Tu Quoque* Arguments and the Significance of Hypocrisy," 166. Notice that neither of these inferences involves committing the *ad hominem tu quoque fallacy*. The inference from an individual's moral defects to the defectiveness of her moral views is indirect; it proceeds via (what we take to be) reasonable supplementing premises.

spread claims about hypocrisy. But the core features that are distinctive of paradigmatic hypocrisy are, we suggest, present in more familiar cases as well—though, as we will now explain, they are present to varying degrees.

First, consider the mismatch that is distinctive of hypocrisy. We have proposed that paradigmatic hypocrisy consists in a mismatch between an *explicit judgment and an action*. This is to be expected. Explicit judgments are, after all, the most straightforward and least costly way to exhibit an investment in moral issues, and actions are the most straightforward way to be caught out. But they are certainly not the only ways; other forms of inconsistency can and do arise. Sometimes hypocrisy seems to involve a mismatch between two or more judgments or two or more actions. If I campaign against factory farms while praising the nobility of dogfights, I am arguably a hypocrite. Similarly so if I donate to Médecins Sans Frontières while buying stock in military contractors, even if I make no statement about either. (The boundary between judgment and action can be fuzzy, of course. We think that this a further reason not to insist on a strict judgment-action mismatch for non-paradigmatic cases.)

Second, paradigmatic hypocrites engage in the very behavior that they *criticize*. Acts of hypocritical criticism are thus closer to the paradigm than acts of hypocritical praise. We suspect that this is owing to the fact that the former generally involve deliverances of *blame*. In comparison with praise, blame tends to be a far more "serious affair," and we are understandably more concerned about undeserved punishment than unmerited reward.[23] However, mismatches involving praise can also attract charges of hypocrisy. The hawkish politician who dodged the draft or the environmentalist who praises fuel-efficient cars while driving a Hummer attract the charge of hypocrisy because of their failure to do the praiseworthy things that they encourage in others. The same may be true of the person who preaches the benefits of clean living and early morning exercise, but frequently skips their morning run to nurse a cruel hangover.[24] Though these cases are perhaps less common, they also seem linked with the expectations to which moralizing gives rise. Moral authorities are assumed to practice what they preach, and preaching can take the form of praise as well as blame. Further, insofar as praise elevates the objects of praise above the masses, there is at least (what might be thought of as) a *relative* element of condemnation—the lazy slugabed may not be positively bad, but they could do better.

Third, paradigmatic hypocrites are paradigmatic *authorities*: community leaders, priests, and the like. But generally speaking, a hypocrite need not occupy any such role. We may very well take a close friend or a colleague to be guilty

23 Watson, "Two Faces of Responsibility," 283.

24 We thank an anonymous referee for the excellent example.

of hypocrisy. On our understanding, however, the notion of a moral authority is sufficiently flexible to account for cases of the latter sort. Everyday hypocrites are likewise failed moral authorities—though their authority is of a more restricted kind.

Let us explain. Moral authority is, on our understanding, a variable and contextual affair. It is a standing that comes in degrees. Although *paradigmatic* moral authorities are community leaders, parents, and priests, any member of a moral community can set themselves up as a moral authority with more limited force or scope. And doing so need not be particularly difficult—moralizing can often suffice. In the film *Mean Girls*, when Gretchen tells Cady that "ex-boyfriends are just off-limits to friends" because "that's just, like, the rules of feminism," she lays claim to moral authority within a group of only four.

Moral authority is therefore partly a relational matter: you might be an authority relative to me given my loose ways, even though most people would outrank you in turn. Likewise, it is community-dependent: the owner of the local BDSM dungeon may be a moral authority about consent and negotiation, even if the community itself is marginalized. Moral authorities may therefore be orthogonal to traditional social hierarchies: the Solomonic bartender can be an authority despite his humble station. Finally, moral authority can be local to a particular issue. I might make for a fantastic moral authority when it comes to relationship advice, but a poor one when it comes to social justice.[25]

We are now in a position to apply our paradigm-based account to the smaller scale. It is our contention that more familiar cases of hypocrisy similarly involve a mismatch that serves to undermine an agent's claim to moral authority. But the moral authority in question need not be authority of the paradigmatic sort. Compare our PETA advocate with a friend who routinely condemns the animal meat industry and often boasts about their vegetarian lifestyle. Suppose that I happen to discover that this friend—who is all too keen to rebuke *me* for my trips to the local butcher—has a habit of indulging in bacon when the opportunity presents itself. Surely my friend is guilty of hypocrisy. And there is, we propose, a clear sense in which their moral authority with respect to animal rights issues has been undermined by their actions. That authority concerns, among other things, their capacity to fulfill the sanctioning role. To do so, they must be in a position to condemn me for consuming meat products. But they are surely in no such position if they are guilty of having done so themselves. The illo-

25 Our remarks here are somewhat reminiscent of an idea that is often raised in discussions of moral testimony. While it is controversial whether individuals can be moral experts *tout court*, it is generally accepted that they may claim expertise with respect to a specific moral issue (Hopkins, "What Is Wrong with Moral Testimony?" 623–26).

cutionary force of their criticism certainly seems to have been blunted by their actions. Likewise, their authority concerns their ability to adjudicate edge cases: is the dairy industry just as cruel as the meat industry? I may in the past have trusted what my friend had to say on the matter. But my confidence in this testimony could very well be shaken by their inconsistency; for I may now suspect that they do not take the cause sufficiently seriously.[26] It would also be understandable if I were moved to retract some of the esteem that I had previously extended to them on account of their vegetarian lifestyle.

Our proposal therefore seems especially well-placed to accommodate and shed light on core features of hypocrisy—among them, the significance of the mismatch between a hypocrite's pronouncements and their actions, and its implications for the terms of their interaction with others. Admittedly, these features are far more obvious and pronounced in paradigmatic cases: the moral authority is more extensive, the pronouncements usually take the form of harsh criticism, and the inconsistency tends to be especially infuriating. But the very same features seem to operate on the smaller-scale as well.

3.2. Caveats and Clarifications

Several features of our account are worth noting in detail. We begin by fleshing out the relationship between hypocrisy and failures of moral authority. Note that our formulation of hypocrisy is phrased in terms of *claims to* moral authority rather than actual moral authority. Thus, the account also covers cases in which someone has never actually played the role. My friend is a notorious philanderer. At the pub one evening, they criticize my infidelity. They are a hypocrite, of course, even if nobody was ever inclined to listen to them in the first place. The important thing is not that the mismatch undermines what authority they have, but what authority they might *reasonably claim*.

On our account, then, all hypocrites are persons who fail in laying claim to moral authority. But it does not follow that all who fail in laying claim to moral authority are hypocrites. An individual might fail in this regard simply by revealing themselves to be too confused or misinformed to count as an authority on a

26 One might worry that the account does not work quite as well when the moral stakes are low. Suppose that my friend chastises me for littering one day. The next day, they unceremoniously toss their soda can into the local river. Our account suggests that have *undermined their authority* on the ethics of littering. But that might strike one as slightly too strong. However, we think that such cases likewise involve an undermining of moral authority. My friend is surely no longer in a position to condemn me for littering if they are guilty of having done so themselves. Another way to assuage this concern (for those still unconvinced) is to note the possibility that when the moral stakes are low a hypocrite might be one whose claim to moral authority has been *diminished* rather than undermined.

particular moral issue. Serious factual errors, for instance, can reasonably reduce our confidence in an individual's suitability for the status without being self-undermining in the way that is distinctive of hypocrisy.

One can also lack the status of a moral authority simply by being (or being considered) a very bad person. The bushranger who rampages across the land is not fit to be a moral authority because they are not trying to be moral in the first place. In order to function as a moral authority within a particular community, one must, at minimum, be considered a good person.[27] And, in order for that status to be deserved, one must actually *be* a good person. The status of moral authority requires not merely goodness but reliable, trustworthy goodness.[28] Hypocrisy represents *a very particular* way of going wrong with regard to this requirement, and it is this requirement at which charges of hypocrisy aim.

Indeed, we suspect that this is precisely why a *single act* is capable of undermining one's claim to moral authority. Moral commitment is not a choose-your-own-adventure type of affair. Moral matters are deeply important to us, and we expect others to treat them with due seriousness. When people exhibit an investment in moral issues, we expect them to follow through on their moral commitments, even (perhaps especially) when it does not suit them.

A further important clarification concerns the truth-conditions of hypocrisy attributions. Although our account makes central reference to the relationship between the hypocrite and the moral community in which they are embedded, this should not be taken to suggest that whether or not someone counts as a hypocrite is entirely a matter of what others happen to think of them. Whether or not someone is a hypocrite is determined by whether the relevant mismatch undermines their claim to moral authority. Certain kinds of inconsistency certainly do have this feature, and so constitute hypocrisy—the militant vegan sneaking bacon, for instance. But others do not. A kind of inconsistency is at play, for example, when an individual changes their moral views over time. Yet the mis-

27 The "considered" proviso is important; for it is surely possible for incredibly *bad people* to function as moral authorities—to garner esteem, earn the status of a trusted moral adviser, and so on. Indeed, we might even think that this is characteristic of the most spectacular cases of hypocrisy (those that typically make the headlines). But it is, at least to our minds, difficult to imagine those who are *taken to be* depraved, vicious, or cruel functioning as *moral* authorities (though they can of course function as authorities of some other kind—heartless tyrants who rule through fear, for instance). A converse, and more complicated, case is that of the "accidental sage" who is taken to be an authority for mistaken reasons (Chance the gardener, say). Again, the important thing is (we think) the way that they are perceived: the accidental sage *is open* to charges of hypocrisy, though in defense they will likely point out that they never took themselves to be giving moral guidance.

28 It is helpful to keep in mind here that (as was suggested earlier) "trustworthy goodness" will often be relative to a particular moral issue.

match here does not necessarily undermine their claim to moral authority; for it might be representative of careful reconsideration of moral issues. Our account therefore makes room for the possibility that moral agents can be mistaken in their attributions of hypocrisy. Insofar as they are mistaken about the details of the case, they might be wrong in thinking that someone has done something to undermine their claim to moral authority.

Finally, one might worry that the hypocrite's undermining of their claim to moral authority is *a consequence* of their hypocrisy rather than constitutive of it.[29] If this is so, then we have failed to deliver on our promise; far from providing an account of hypocrisy, we have merely provided an account of what follows from it.

We have a few things to say in response to this important challenge. To begin with, it is helpful to disambiguate two readings of the objection. On the first, the objector thinks that hypocrisy is constituted by something else (pretense or lying or what have you) and that this something else is what leads to the loss of moral authority. This form of the objection would require providing an alternative account of the nature of hypocrisy. We are obviously skeptical that such an account can be given. On the other hand, the objection could be that hypocrisy is something like a second-order property of actions: it is the property that actions have when they cause the hypocrite to lose moral authority. But the latter comes so close to our account that we confess that we can see little light between it and our own position.

The concern can be further alleviated by distinguishing two senses in which someone can be a moral authority. There is a distinction to be drawn between an agent's (1) being *entitled* to authority, and (2) actually being *considered* an authority by others. As in other realms, the two can come apart. We are concerned with the former: hypocrites undermine their claim to authority, but this may or may not be discovered. Cases in which an action alone is enough to undermine authority are comparatively rare outside of the moral domain, but only because we usually consider the removal of authority a subsequent act carried out by the body who grants it. But comparable cases do exist: a British monarch who converted to Catholicism, or a priest who violates the seal of confession, have *by their act alone* undermined their authority.[30]

29 We thank an anonymous referee for raising this important objection.

30 Note that in these cases the important thing is that the act is what constitutes the undermining of authority, even though the fact that that relation holds is ultimately dependent on external authority (respectively, the Act of Settlement and the canon law relevant to excommunication *latae sententiae*).

3.3. *Advantages of the View*

Having spelled out the finer details of our account, we turn now to its advantages (many of which, we argue shortly, are *distinct* advantages).

First, our view explains why judgments of hypocrisy often have a certain "leaky" quality to them, and why the mismatch between judgment and action need not be direct. Consider a vegan who extols the environmental benefits of their diet, but is condemned as a hypocrite for not riding public transit. At first glance, their judgment and action concern quite different things. Yet the charge of hypocrisy is understandable if it is understood along the lines we have proposed—as a judgment that their transit choices undermine any claim they have to act as a moral authority on environmental matters in general.

Second, our account can accommodate cases in which mismatches involving praise attract charges of hypocrisy. On our view, a moral authority's status arises, in part, as a result of the link that others take to hold between their moral pronouncements and their behavioral dispositions. For this reason, an individual's failure to do the praiseworthy things that they encourage in others can easily undermine their moral authority, and so expose them to allegations of hypocrisy.[31]

Third, our account immediately makes clear why hypocrisy attracts such opprobrium. We place significant trust in moral authorities. But the promotion and consultation of moral authority is something of a fraught matter. Most of us save a certain amount of time and energy because moral authorities exist: our judgments can be accurate and more reliable, we spend less time on moral deliberation, and we take on less risk when we moralize. Yet moralizing is important, and errors are morally weighty. The errors of a moral authority matter more so than others, for they have broader and more wide-ranging consequences. Moreover, moral authorities acquire much of their standing on credit—that is to say, through their words rather than their actions. And this requires an additional level of trust on our part. When they fail, then, we have every right to feel hurt, betrayed, and angry. Indeed, we might dislike the hypocrite even if we never belonged to the community that they address (a point to which we return in section 5).

31 Once again, it is worth noting that not all forms of praise render one vulnerable to allegations of hypocrisy. As we suggested earlier, certain kinds of disowning prefaces can remove the impression that one's criticism is rooted in a deep-seated commitment to moral values. Some ways of lending praise can have parallel effects, and so may not invite charges of hypocrisy—presenting certain actions as laudable but supererogatory, for example. "If only we could be as pure as Saint Agnes!" sets up quite different expectations than "Everyone ought to be praised for chastity."

In our view, however, it is not merely the element of betrayal or the risks to which they expose us that explain our contempt of hypocrites. We are also likely to be moved to anger by the *unmerited esteem* that they garner. Esteem is a good in high demand; we are usually happy to be esteemed by others, and we are more than happy to avoid their disesteem.[32] But esteem, being inherently comparative, is a good in limited supply. We esteem those who score above average along some evaluative dimension, and not everyone can be above average. Since esteem is ultimately a zero-sum game, we invest moral authorities with our esteem at some cost to ourselves: by elevating them, we lose out a bit.

It is therefore unsurprising that hypocrites are traditionally met with disdain; for they have ultimately shown themselves to have been undeserving of this elevated status. Esteem is a sought-after and precious resource—one that we can earn through moralizing. Yet our moral words only buy us esteem for so long. Moralizing is not merely a matter of judgment; when push comes to practical shove, we must be prepared to perform those actions to which we have given praise and refrain from the sorts of behavior that we have criticized—even (or perhaps especially) if doing so would be inconvenient. Our moral pronouncements, then, have something like the status of a "buy now, pay later" scheme. Moral authorities enjoy the good of esteem without paying the costs up front. But eventually those costs must be paid, and they must be paid via action. In effect, hypocrites have purchased esteem, but failed to pay up when the time is right; they have, if you like, earned themselves a bad credit rating in the economy of esteem—something that is bound to attract a strong degree of moral opprobrium.

A fourth advantage of our account is that it does not explain this characteristic opprobrium by attributing any particular kind of deplorable character trait, unworthy motive, or cruel intentions to the hypocrite. On our view, hypocrites are simply persons who have, by mismatch (typically) between word and action, undermined their claim to moral authority. This formulation is permissive; it leaves the source of the mismatch entirely open. It might be explained by an intention to deceive or manipulate. Or it might be owing to an excessive concern with keeping up moral appearances. Each is surely a feature that some hypocrites share, and each may sometimes be helpful in explaining our characteristic dislike of them. But on our view, none is necessary. (We demonstrate the advantages of this permissiveness in section 4).

That said, a mismatch that finds its roots in deplorable motives may sometimes suffice. Our view has the resources to accommodate scheming hypo-

32 Brennan and Pettit, "The Hidden Economy of Esteem," 80–81.

crites.[33] Consider Tartuffe, who pretends to be pious for reasons of self-advancement. His hypocrisy does not consist in a single bout of inconsistency (of the relevant sort), but in a rather extensive pattern of deception. Our account delivers the correct ruling here. Tartuffe is undoubtedly a hypocrite; for the mismatch between his pious pronouncements and his (very) impious motives and behavior surely does undermine his claim to moral authority. Our proposal can also explain why some may be inclined to judge Tartuffe more harshly than those whose hypocrisy is not so widespread; inconsistency as extensive as Tartuffe's would seem to undermine a purported claim to moral authority far more than a single moral lapse.

Finally, our account allows for extension of the concept of hypocrisy to groups. This is not uncommon. China accuses the United States of hypocrisy on human rights; the Labour Party charges Turnbull with hypocrisy in policy changes; activist Naomi Klein charges large corporations with hypocrisy about attitudes toward the environment. These charges are obviously intelligible, and a natural extension of ordinary hypocrisy. It is convenient to have an account that generalizes readily.

4. THE MULTIPLE REALIZABILITY OF HYPOCRISY

We now explore how well our account stacks up against its rivals. Most of these competing views, we suggest, are not so much wrong as incomplete. None succeeds in identifying necessary and sufficient conditions for hypocrisy. But each points toward a feature that can suffice to undermine authority in the right context, and that is common enough to deserve mention. Ultimately, we argue that what is plausible in these accounts can be subsumed under our own view.

4.1. The Pretense Account

It might seem characteristic of the hypocrite that they present themselves as something other than what they truly are. So it is perhaps unsurprising that the most common position regarding hypocrisy takes it to consist in a kind of deception; in pretending "to motives and moods" or to "certain standards" that one does not really have.[34]

Of course, the hypocrite's deception is thought to go in a particular direction. McKinnon emphasizes that the hypocrite "dissembles precisely because she wants people to think better of her than they would were her true motives

33 We thank an anonymous referee for encouraging us to explain this point.

34 Ryle, *The Concept of Mind*, 172; Taylor, "Integrity," 144–45.

revealed."[35] Szabados and Soifer's Aristotelian account similarly characterizes hypocrites as persons who do the right thing for the wrong reasons, all the while pretending that they are responsive to the right reasons.[36] According to what we shall call the "pretense account," then, hypocrites are those who pretend to be *far better* than they really are.

Theorists divide regarding whether the deception here must be intentional. Hare seems to think so, as does McKinnon in early work.[37] Others disagree. Statman, for instance, suspects that hypocrisy is *incredibly likely* to be the product of self-deception; insofar as the hypocrite seeks to cultivate a particular image in the eyes of others, the best means of achieving this might be to believe that image themselves.[38] Some remain on the fence; although Szabados and Soifer take paradigmatic hypocrites to pretend to be virtuous for reasons of self-advancement, they concede that the deceit need not always be intentional.[39] There is also some disagreement as to whether the pretense must be driven by self-interested concerns. Although the matter is often left open, some have answered in the affirmative.[40]

Pretense accounts capture something important about hypocrisy. After all, hypocrites often do portray themselves as exemplary moral agents, and often they are not. That being said, we have a number of worries with pretense accounts—particularly those that attribute deceptive intentions to the hypocrite. An intention to deceive does not appear to be a necessary condition on hypocrisy. It seems perfectly conceivable that someone could live out their days wholly unaware of any hypocrisy on their part. Following Statman, epiphanies of the form, "I suddenly feel that I have been a hypocrite all my life" do not seem to betray any sort of conceptual confusion.[41] Moreover, and as R. Jay Wallace observes, some people are blind to their own shortcomings and do not foresee that they will fall short of their own moral standards in the future.[42] But falling short

35 McKinnon, "Hypocrisy, with a Note on Integrity," 323; see also Kittay, "On Hypocrisy," 281.

36 Szabados and Soifer, "Hypocrisy after Aristotle," 563.

37 Hare, *Freedom and Reason*, 77; McKinnon, "Hypocrisy, with a Note on Integrity," 323. McKinnon has since changed her mind, conceding that "there are many hypocrites who are quite unself-conscious about the extent to which they misrepresent their real reasons for acting" ("Hypocrisy and the Good of Character Possession," 719).

38 Statman, "Hypocrisy and Self-Deception," 68.

39 Szabados and Soifer, "Hypocrisy after Aristotle," 564.

40 See, e.g., McKinnon, "Hypocrisy and the Good of Character Possession," 722; Szabados, "Hypocrisy," 203.

41 Statman, "Hypocrisy and Self-Deception," 68.

42 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 315.

of those professed standards would seem to expose them to the charge of hypocrisy all the same.

Another worry relates to the project of explaining why it is that hypocrisy typically strikes us as a deep and important species of moral failure. According to sponsors of the pretense account, this failure ultimately amounts to a form of deception. In leading others to (falsely) believe that they are moral paragons, hypocrites sever "the act from the intention," and misrepresent what they truly are.[43]

But why think that deception about one's character constitutes a moral failing? The projection of a favorable self-image is arguably a pervasive phenomenon, at least on a small scale. Ordinary people tend to talk more about their charitable donations than their petty thefts, to cite workers' rights rather than aesthetic distaste when they refuse to shop at Walmart, or to pretty up their dating history for new partners. The desire to appear better than you are is *so* pervasive, and *known* to be so pervasive, that it is hard to see what is distinctively bad about it.[44]

Finally, a common theme running through pretense accounts is that hypocrites are motivated by self-interested concerns; they are "out to promote [their] own advantage at the expense of others," their concern being exclusively with their "moral image."[45] Yet it seems that hypocrites can have noble motives. Consider the father who hides his smoking from his children, inveighing against the unhealthy habit out of concern for them. There is no reason to think that his motivation here is self-serving, at least in any obvious sense. Nonetheless, his children would seem right to accuse him of a hypocrisy, should they ever discover his hidden stash.[46]

Pretense theorists attempt to reconcile noble hypocrisy with their accounts in different ways. Szabados suggests that we construe the notion of "self-interest" more broadly, such that having "some personal stake" in the "project of pretence" suffices.[47] We agree with Crisp and Cowton that this strategy confuses motiva-

43 Kittay, "On Hypocrisy," 285; McKinnon, "Hypocrisy and the Good of Character Possession," 725.

44 Wallace notes that one could move from here to the claim that bourgeois life is simply shot through with hypocrisy ("Hypocrisy, Moral Address, and the Equal Standing of Persons," 312). We agree with him that this is a bit hysterical, and that another theory of hypocrisy is probably preferable.

45 Szabados, "Hypocrisy," 203; McKinnon, "Hypocrisy and the Good of Character Possession," 722.

46 For discussions of noble hypocrisy, see Crisp and Cowton, "Hypocrisy and Moral Seriousness"; and Szabados, "Hypocrisy."

47 Szabados, "Hypocrisy," 204–5.

tion and justification.[48] Even if noble hypocrites derive some pleasure or benefit from behaving as they do, they may still act for other-regarding (as opposed to self-interested) reasons.

Unlike Szabados, McKinnon concedes that noble hypocrites act for other-regarding reasons. Yet she argues that we have good reason to distinguish such persons "from the hypocrite who is ashamed of her concealed motives or . . . whose preoccupation is with her reputation rather than with any actual outcomes she could effect."[49] We agree that there is an important distinction between individuals driven by morally laudable motives and those who have morally questionable intentions. (Blame, or harsh judgments about an agent's moral character, for instance, may not strike us as fully appropriate in the former case.) But we do not think that this is a distinction that marks off hypocrisy from the absence thereof. Since hypocrites can sometimes act from noble motives, we see no reason for thinking that it is a constraint on hypocrisy more generally that it must be driven by self-interested concerns.

4.2. The Blame-Centered Account

We turn next to R. Jay Wallace's account of hypocrisy, which emphasizes the role of reactive attitudes—blame in particular.[50] It is worth noting from the outset that Wallace's explanatory ambitions are restricted in two important respects. First, he confines his *explanandum* to instances of *hypocritical moral address*: cases in which an agent is "actively exercised" about a moral issue.[51] It is this particular form of hypocrisy that Wallace finds distinctively objectionable on moral grounds. Second, Wallace limits his investigation to the phenomenon of *hypocritical moral criticism*. He does not propose to offer an account of hypocritical moral advice, whereby an agent fails to follow her own moral recommendations.

According to Wallace, what is objectionable about hypocritical moral address is that it offends against "the commitment to the equality of persons that is constitutive of moral relations."[52] We all share an interest in avoiding the punitive experience of blame.[53] But the hypocrite would take their interest in avoiding blame to be more important than the interests of the person whom they criticize. So their hypocrisy effectively ascribes a moral standing to themselves that they

48 Crisp and Cowton, "Hypocrisy and Moral Seriousness," 348n7.

49 McKinnon, "Hypocrisy, with a Note on Integrity," 325.

50 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons."

51 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 312.

52 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 308.

53 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 328–29.

are unwilling to extend to others. It is because this offends against a central moral precept—that of the equal standing of persons—that Wallace regards hypocritical moral address as distinctively objectionable on moral grounds.

There is yet another way in which hypocritical moral address is thought to offend against the commitment to the equality of persons: the hypocrite also attaches greater importance to the interests of the criticized person's *victims* than the victims of their own moral transgressions.[54] They blame someone for dishonesty while allowing their own dishonesty to remain unscrutinized, effectively demonstrating that they take that person's victims to have a more serious interest in avoiding dishonesty than the victims of their own dishonest conduct. Since a commitment to the equality of persons is central to moral thought, hypocritical moral address is said to offend "against the spirit of morality, subverting it . . . from within."[55]

Wallace's account is insightful. It is one of few that does not specifically depend on attributing hidden agendas or deficiencies of character to hypocrites (e.g., an intention to deceive, or an excessive concern with one's moral reputation). Instead, Wallace focuses on the reactive attitudes that govern our interactions within a moral community, and construes hypocrisy as deriving from particular relations that we enter into with others.

Insightful as it is, the limited nature of Wallace's account is a serious shortcoming. Wallace tackles but one species of hypocrisy—hypocritical moral criticism—and the account seems difficult to extend to cases of hypocritical moral advice. If I praise the chaste while secretly living lasciviously, I am, it seems, just as much of a hypocrite as if I blame you for your sins. Yet there does not appear to be the same harms at play; only in the second instance have I had any sort of objectionable reactive attitude toward you. A substantive theory of hypocrisy ought to account for the wrongfulness of both, but Wallace's account only has the resources to explain the latter.

A further issue with Wallace's account concerns its emphasis on the *victims* of hypocrisy. According to Wallace, the hypocrite values the victims of their own hypocrisy less than the victims of others who engage in the same conduct. But it is surely possible that a hypocrite could be one who cares *considerably more* about their own victims. Consider the person who inveighs against cheating at cards, though does so themselves. Suppose that every time they cheat, they compensate their victims the exact amount extorted from them. Perhaps this person does not care at all about the victims of others' cheating, and merely harbors a

54 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 330.

55 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 335.

special concern for their own victims. But they would seem to be a hypocrite all the same.[56]

Finally, Wallace's blame-centered account would seem to have difficulty accounting for the role that hypocrisy attributions play both within and across communities. Although this account works naturally for moral communities who think that people are fundamentally equal from a moral point of view, charges of hypocrisy arise even in societies with no such commitment. The seventeenth-century Catholic Church, for example, was hardly egalitarian—yet audiences at the time recognized Tartuffe as a hypocrite all the same. Hence, even people who are not committed to something like the equality of persons can clearly recognize and condemn hypocrisy. This is something that Wallace's account seems to have trouble accommodating.

The blame-centered account also struggles with cases that do not involve two *people*. As noted above, countries and organizations can accuse and be accused of hypocrisy. Yet it is not at all obvious that nations, corporations, and political parties *have* reactive attitudes, or that they are the proper targets of reactive attitudes, or even that they are capable of the kinds of propositional attitudes that Wallace's account requires. Some philosophers may think so, of course. But the fact of US hypocrisy does not seem like it should depend on philosophical claims about group attitudes. Claims of intergroup hypocrisy ought to be intelligible regardless of whether groups have the same sorts of attitudes as individuals.

4.3. The Moral Seriousness Account

Roger Crisp and Christopher Cowton understand hypocrisy as a failure to take morality seriously.[57] The proposal certainly has some attraction. Like our own view, the "moral seriousness account" construes hypocrisy as multiply realizable; there are many ways in which someone can fail to take morality seriously—doing so need not necessarily consist in pretense, or misplaced blame.

Although the moral seriousness account has some initial appeal, we are skeptical that a failure to take morality seriously is sufficient for hypocrisy. Avowed egoists openly profess not to take morality very seriously at all, but they surely do not count as hypocrites for that reason.[58] Nor does a failure to take morality seriously seem necessary for hypocrisy. As Szabados and Soifer point out, many hypocrites take morality *far too seriously*. The fanatic who cannot possibly live up

56 We are indebted to Lachlan Umbers for this criticism.

57 Crisp and Cowton, "Hypocrisy and Moral Seriousness."

58 Szabados and Soifer, "Hypocrisy after Aristotle," 562.

to their over-demanding moral prescriptions would seem to take morality very seriously indeed. But they still strike us as a candidate for hypocrisy.

Finally, and as we have suggested already, hypocrisy can sometimes stem from noble motives. The father who hides his smoking from his teenage son does not seem to be playing fast and loose with morality. This is not lost on Crisp and Cowton. Their defense rests on there being a morally significant distinction between hypocritical acts and hypocrites.[59] Their theory is only intended to apply to the latter; it is only the full-blown hypocrite who is flippant about morality.

Crisp and Cowton are certainly entitled to restrict their ambitions. However, we think that this is a notable shortcoming of their proposal. The moral seriousness account is unlikely to shed much light on everyday attributions of hypocrisy; for these certainly do not appear to be restricted to those with a settled disposition to take morality insufficiently seriously. Perhaps such dispositions are to be found in the likes of Tartuffe or Uriah Heep. But the hypocrites of fiction represent something of a limiting case. It seems implausible to us that hypocrisy as it operates in day-to-day life amounts to a full-fledged character trait. Following Shklar, few of us ordinary folk "have the resources to become self-aware, scheming, accomplished hypocrites like Uriah Heep."[60] If one is to capture the broad and varied phenomenon of hypocrisy, restricting the *explanandum* to persons with a particular kind of deplorable character seems ill-advised. We often accuse otherwise perfectly nice people of hypocrisy. An adequate conception ought to be able to account for this more common and banal species of hypocrite with whom we interact.

4.4. Many Ways to Fall

The theories canvassed above represent a number of ways of explaining what is common to all instances of hypocrisy. But they suffer from serious shortcomings. We believe that the moral authority account can preserve what is attractive in these other views while avoiding their associated problems.

Pretense accounts have an obvious appeal. Hypocrites do, after all, tend to portray a favorable self-image that is misleading. So it is natural to think that the relevant vice is that of deceit or manipulation. However, this feature is certainly not present in all cases of hypocrisy and, to our minds, it is not what is of fundamental importance. What is more important, we think, is the *unmerited esteem* that hypocrites often garner. Typically, the hypocrite does not live up to their own lofty standards, and so we feel angry that we invested them with our

59 Crisp and Cowton, "Hypocrisy and Moral Seriousness," 347.

60 Shklar, "Let Us Not Be Hypocritical," 7.

esteem and trust. However—and importantly—a hypocrite need not earn esteem through purposefully engaging in any form of deception. Many hypocrites simply set the bar too high for themselves, and do not foresee that they will fall short of their own standards in the future.

That being said, our view is consistent with the possibility that some (even many) hypocrites harbor deceptive intentions. A hypocrite may very well make lofty moral pronouncements with the goal of portraying a favorable self-image; indeed, we suspect that these cases are likely to attract a special sort of opprobrium. It is one thing to extend esteem where esteem is not due—it is quite another to be conned into doing so.

Moreover, there are other ways to have bad motives, or even to do the right thing for the wrong reasons. The man who rebuffs the advances of a lover only because he wants to ensure that he can take over his father-in-law's business is not concerned with self-image per se, but there is enough of a mismatch between his motives and his behavior that, under the right circumstances, he can rightly be judged a hypocrite. Again, the point is that there is no *particular* bad set of motives necessary to undermine claims to authority.[61]

Our view can also preserve what is right in the blame-centered account. Wallace's primary focus is hypocritical moral criticism, which characteristically involves deliverances of blame. It is understandable that Wallace should want to focus on blame to the exclusion of praise. As we suggested earlier, blame may very well be a hallmark of paradigmatic hypocrisy. Blame is an especially unpleasant experience; none of us wants to be on the receiving end—least of all from those who are guilty of the very same vice. Nonetheless, hypocrisy issuing from praise is a very real phenomenon, and it is one that Wallace's account would seem to have trouble accommodating. We concede that these cases are likely to be less serious, and they are perhaps not quite as common. But they are no less real for that. So we regard it as a virtue of the moral authority account that it can accommodate hypocritical praise as well.

Unlike the blame-centered account, our view can also explain the role that hypocrisy attributions play within and across communities. Wallace's view has difficulty accounting for attributions of hypocrisy across communities (and within non-egalitarian ones). Our account can accommodate cases like state actors who may or may not be the appropriate targets of second-person reactive

61 See Robbie Fulks's song, "Doin' Right (for All the Wrong Reasons)"; at least one of the authors thinks that the narrator is a loathsome person but not a hypocrite. Evaluation of people who act rightly for the wrong sorts of reasons can be complex in more realistic cases; for an extended discussion of the problems of "unprincipled virtue," see Arpaly, *Unprincipled Virtue*.

attitudes. On our view, to call the United States hypocritical for its foreign policy is not (necessarily) to resent the nation; rather, it is first and foremost to claim that the United States's high-handed pronouncements on foreign policy ought not to be taken seriously.

Finally, we think that Crisp and Cowton make progress in allowing that hypocrisy can be multiply realizable.[62] And the common feature that they propose to identify—a failure to take morality seriously—is surely one that many hypocrites share. However, the moral seriousness account lacks the resources to accommodate hypocrites who act from noble motives. Doing so is important; for we often charge otherwise perfectly nice people with hypocrisy. Such allegations certainly do not seem restricted to those with a settled disposition to take morality insufficiently seriously. That some hypocrites might be driven by laudable motives is consistent with the moral authority account. One can judge that someone is not apt to serve as a moral authority with respect to a particular moral issue even if they are, generally speaking, a good person.

In this portion of the discussion, we have been concerned to argue for two claims. First, although no particular motive, lack of motive, or violation of any specific moral principle is necessary for hypocrisy, within context each can suffice when it results in a mismatch of the relevant kind. Second, our account can preserve the benefits of other views while avoiding their associated problems. The moral authority account construes hypocrisy as multiply realizable: a hypocrite is simply someone who has, through a mismatch of the relevant kind, undermined their claim to moral authority. That mismatch might be the product of cultivating an undeserved moral reputation, a lack of moral seriousness, or something else still—perhaps even good intentions, or overreaching moral ambition. Since our proposal is consistent with any of these motives, it preserves what is appealing in these other views. But it is not committed to taking any particular motive to be necessary for hypocrisy, and so it avoids their associated problems.

5. HYPOCRISY AND MORALLY DIVERSE SOCIETIES

We conclude by discussing the role that judgments of hypocrisy play in regulating our social lives. When we introduced the idea of moral authority, we noted that it was something of a double-edged sword. On the one hand, moral authority can be a good thing: skill and time are as unequally distributed in moral reasoning as they are in any other domain, and we are normally better off if we look up to people who are good at what they do. On the other hand, when moral

62 Crisp and Cowton, "Hypocrisy and Moral Seriousness."

authorities are inept or malicious they can cause serious harm. So we have excellent reason to keep a close eye on them

However, it is not only our own moral authorities that we monitor. Many of those whom we charge with hypocrisy function as moral authorities within *different* communities. So it seems that we are equally (and perhaps even more) disapproving of hypocrites whom we have no need to trust.[63] This may seem puzzling at first. If such persons were never moral authorities *for us*—if *we* never held them in high esteem, or followed their advice—then why ought their failures concern us?

We propose that there is another (perhaps more important) role that judgments of hypocrisy play in contemporary societies. One striking feature of modern society is that it is *morally diverse*: multiple communities disagree over pressing moral issues. Yet those with radically different views can *converge on* judgments of hypocrisy. Perhaps we disagree on the matter of homosexuality; I think it is perfectly fine, whereas you condemn it as sin. But we can both *agree* that the homophobic senator's bathroom dalliances are hypocritical, and that they have undermined their claim to moral authority by their actions. We can further agree that the senator's *hypocrisy* is reprehensible—even while disagreeing on whether their expressed opinions or their actions were the right ones. An interesting feature of hypocrisy ascriptions, then, is that they seem capable of *cross-cutting* communities.

It is for this reason that judgments of hypocrisy play an especially important role within morally diverse societies. As Wallace notes, it is often ineffective to try to sway the opposing side by appealing to the very values over which we disagree.[64] A charge of hypocrisy, by contrast, points toward a kind of moral failure that ought to give our opponents pause independently of our disagreement over more substantive issues. This peculiar quality of hypocrisy has not gone unnoticed. Indeed, some have gone so far as to claim that a charge of hypocrisy is the "most effective verbal weapon" in a "world of ideological conflict and moral confusion."[65]

Our account is particularly well placed to accommodate this interesting feature of hypocrisy. On our view, a charge of hypocrisy is, first and foremost, a charge directed at an agent's *standing* rather than a criticism of an isolated action. For this reason, different communities can converge in their judgments of hypocrisy in spite of their substantive disagreements over moral issues.

We suspect that this is one reason why hypocrisy strikes us as something

63 McKinnon, "Hypocrisy, with a Note on Integrity," 326.

64 Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 307.

65 Shklar, "Let Us Not Be Hypocritical," 22–23.

worth tracking. In a world of rampant moral disagreement, we have excellent reason to keep track of moral authorities that we *do not* regard as authorities as well as those we do. All things being equal, we might allow a certain amount of sloppiness in the authorities that we trust. But authorities who command the respect of large numbers of people with whom we disagree can be held to a stricter standard. Even if *we* do not take them seriously, by exposing their hypocrisy we may be able to convince others that *they* should not take them seriously either.[66] A charge of hypocrisy can undermine the status that such persons have come to enjoy—something that can, in turn, remove some of the force and cohesion of a community that we take to be getting the moral facts wrong.

This might sound as if hypocrisy judgments are simply cynical or calculating. They can be. But the purpose of hypocrisy judgments is not *just* to give us another arrow in our quiver against the unrighteous. At the best of times, we suggest, hypocrisy ascriptions can have a softening effect on moral discourse. Fearsome and rigid individuals might serve as moral authorities. Yet they do so at a considerably greater risk: the more stringent one's judgments, the easier it is to fall astray.

A moral authority who tempers their judgments, by contrast, is far less likely to run afoul of their own prescriptions, and more likely to be met with tolerance and forgiveness when they do so. This observation dovetails nicely with our earlier discussion of disowning prefaces, which can blunt the force of one's criticism, and soften one's purported claim to moral authority in turn. In order to *avoid* exposing ourselves to the charge of hypocrisy, we must temper the vehemence of our judgment. Those who are morally inflexible and harsh are more likely to disappear from the moral scene when they inevitably fail to live up to their own standards.

Charges of hypocrisy, then, are not merely a means of regulating our own moral communities. On the whole and in the aggregate, they play an important role in regulating the relationships between communities as well. This suggests a more positive and less cynical diagnosis of the state of play. At the best of times, judgments of hypocrisy can have a moderating effect within morally diverse societies: they tend to weed out the extremes. Insofar as persons disagree over moral issues, and insofar as they continue to turn to moral authorities for guid-

66 This idea is reminiscent of what Aikin calls "*is* (he) or *ea* (she) *quoque* arguments," which appeal to the hypocrisy of some third party rather than that of one's interlocutor ("*Tu Quoque* Arguments and the Significance of Hypocrisy," 161). He proposes (and we agree) that such arguments are not necessarily fallacious. Insofar as an individual's hypocrisy provides us with some reason to question their sincerity or moral competence, it may provide us with reason to question the soundness of their advice as well.

ance, judgments of hypocrisy might help us steer them in the direction of the relatively fair minded and tolerant ones.[67]

Australian National University
u5312691@anu.edu.au

Macquarie University
colin.klein@mq.edu.au

REFERENCES

Aikin, Scott F. "*Tu Quoque* Arguments and the Significance of Hypocrisy." *Informal Logic* 28, no. 2 (2008): 155–69.

Arpaly, Nomy. *Unprincipled Virtue*. New York: Oxford University Press, 2003.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences* 100, no. 6 (March 2003): 3531–35.

Brennan, Geoffrey, and Philip Pettit. "The Hidden Economy of Esteem." *Economics and Philosophy* 16, no. 1 (April 2000): 77–98.

Cohen, G. A. "Casting the First Stone: Who Can, and Who Can't, Condemn the Terrorists?" In *Finding Oneself in the Other*, 134–42. Princeton, NJ: Princeton University Press, 2012.

Crisp, Roger, and Christopher Cowton. "Hypocrisy and Moral Seriousness." *American Philosophical Quarterly* 31, no. 4 (October 1994): 343–49.

Dworkin, Gerald. "Morally Speaking." In *Reasoning Practically*, edited by Edna Ullmann-Margalit, 182–88. New York: Oxford University Press, 2000.

Estlund, David. *Democratic Authority: A Philosophical Framework*. Princeton, NJ: Princeton University Press, 2008.

Fricker, Miranda. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* 50, no. 1 (March 2016): 165–83.

Hare, R. M. *Freedom and Reason*. Oxford: Oxford University Press, 1963.

Hopkins, Robert. "What Is Wrong with Moral Testimony?" *Philosophy and Phenomenological Research* 74, no. 3 (May 2007): 611–34.

67 Thanks to Edward Elliott, Nic Southwood, Lachlan Umbers, two anonymous reviewers, and audiences at the Australian National University and Macquarie University for helpful feedback on previous drafts. Special thanks to Esther Klein for feedback and debate over many years. Work on this paper was partly supported by Australian Research Council grant FT140100422 to Colin Klein.

Kittay, Eva Feder. "On Hypocrisy." *Metaphilosophy* 13, no. 3–4 (July 1982): 277–89.

Lenman, James. "Ethics without Errors." *Ratio* 26, no. 4 (December 2013): 391–409.

Locke, John. *Second Treatise of Government*. 1690. Edited by C. B. Macpherson. Indianapolis: Hackett Publishing Company, 1980.

McKinnon, Christine. "Hypocrisy and the Good of Character Possession." *Dialogue* 41, no. 4 (Fall 2002): 715–39.

———. "Hypocrisy, with a Note on Integrity." *American Philosophical Quarterly* 28, no. 4 (October 1991): 321–30.

Owen, G. E. L. "Logic and Metaphysics in Some Earlier Works of Aristotle." In *Aristotle and Plato in the Mid-Fourth Century*, edited by Ingeman During and G. E. L. Owen, 163–90. Göteborg, Sweden: Almqvist and Wiksell, 1960.

Ryle, Gilbert. *The Concept of Mind*. London: Hutchinson, 1949.

Schwitzgebel, Eric, and Joshua Rust. "The Moral Behavior of Ethicists: Peer Opinion." *Mind* 118, no. 472 (October 2009): 1043–59.

Shklar, Judith. "Let Us Not Be Hypocritical." *Daedalus* 108, no. 3 (Summer 1979): 1–25.

Statman, Daniel. "Hypocrisy and Self-Deception." *Philosophical Psychology* 10, no. 1 (1997): 57–75.

Szabados, Béla. "Hypocrisy." *Canadian Journal of Philosophy* 9, no. 2 (June 1979): 195–210.

———. "Hypocrisy, Change of Mind, and Weakness of Will: How to Do Moral Philosophy with Examples." *Metaphilosophy* 30, nos. 1–2 (January/April 1999): 60–78.

Szabados, Béla, and Eldon Soifer. "Hypocrisy after Aristotle." *Dialogue* 37, no. 3 (Summer 1998): 545–70.

Taylor, Gabriele. "Integrity." *Proceedings of the Aristotelian Society, Supplementary Volume* 55 (1981): 143–59.

Wallace, R. Jay. "Hypocrisy, Moral Address, and the Equal Standing of Persons." *Philosophy and Public Affairs* 38, no. 4 (Fall 2010): 307–41.

Watson, Gary. "Two Faces of Responsibility." In *Agency and Answerability: Selected Essays*, 260–88. Oxford: Oxford University Press, 2004.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 2 · NOVEMBER 2017
https://doi.org/10.26556/jesp.v12i2.192

CONSENT AND DECEPTION

Robert Jubb

According to what Tom Dougherty labels the "lenient view," it "is only a minor wrong to deceive another person into sex by misleading him or her about certain personal features such as natural hair colour, occupation, or romantic intentions."[1] The lenient view sees such lies as sleazy and so morally unattractive, but a long way from deceiving someone into sex by successfully pretending to be his or her spouse, for example. Dougherty forcefully argues that this view, though most likely fairly widely held, is false. If we deceive someone about a feature of a sexual encounter we have with her or him she or he would be "all things considered unwilling to engage in" if she or he knew it had this feature, that sex is nonconsensual.[2] Since having nonconsensual sex with someone is a serious wrong, for example frequently seen as the distinctive wrong of rape, deceiving someone into sex is seriously wrong.[3] According to Dougherty, when you lie to me about one of my "deal-breakers" for sex, you wrong me in much the same way as we often think of rapists wronging their victims.

This conclusion may well be both shocking and disturbing to us. The levels of rape in our societies are already extremely disturbing. A UK Ministry of Justice survey for England and Wales, for example, suggests that nearly one hundred thousand working-age women are raped every year in England and Wales.[4] That is a horrifying statistic. If Dougherty is correct, something at least in a similar moral category to rape is in all probability far more common. It would be hard, I think, not to be shocked and disturbed by that. Given that his argument's conclusion is in that way difficult to bear, it seems to me that Dougherty's argument should itself be pressed upon pretty hard. If we accept it, however honest our own sex-

1 Dougherty, "Sex, Lies, and Consent," 718.

2 Dougherty, "Sex, Lies, and Consent," 719.

3 Dougherty himself does not take a position on whether rape should be defined as nonconsensual sex, although he notes that many do define it in that way ("Sex, Lies, and Consent," 721).

4 See UK Ministry of Justice, Home Office, and the Office for National Statistics, *An Overview of Sexual Offending in England and Wales*, 6.

ual advances have been, we will have to understand our societies as even greater sites of wrongdoing than we already know them to be. Our relations to them and consequently to each other would have to become even more fraught and difficult. That is not a trivial cost, and not one we should bear without good reason.

There is also a broader issue here. Dougherty supports his conclusion about the wrongness of sexual deception with claims about how deception relates to permissions given by others more generally. His more general claim is that where we need another's permission to rightfully do something, we do it wrongly if we get that permission by deceiving her or him. In that sense, Dougherty's view is one that sees insincerity as almost always a deeply corrupting feature of human life. Concealing one's religious beliefs or lack of them from one's family so as to avoid arguments is, for Dougherty, a serious wrong because it involves deceiving people in continuing a relationship that does not have the character they want it to have. They are not obliged to continue inviting us round for family occasions, and if they knew how often we really attend religious services, they would not. We receive invitations from them by deception, and so on Dougherty's view, we are wronging them. That view seems to me much too moralistic about insincerity, condemning what appear to be central mechanisms of smoothing normal human interaction. Our idiosyncrasies and different commitments can quite easily come into conflict and shock each other, and without a way of concealing them, all kinds of valuable relationships would be much more difficult and much rarer. It is not only because of its implications about our sexual behaviour that I want to challenge Dougherty's argument then. Because of how it supports its central claim, it has further implications well beyond sexual ethics, and in many ways, those are even more concerning.

Dougherty attacks the lenient view in two related ways. First, he argues that it lacks the resources to distinguish different kinds of sexual deceptions before moving on to connect that failing with a broader view about the morality of consent and deception. I begin by trying to show that there are resources unrelated to the morality of sex to distinguish between different sexual deceptions and then move to argue that deception does not always undermine consent.

Let us begin with the distinctions between different kinds of sexual deceptions Dougherty claims are problematic for the lenient view. For example, few would deny that when D'Artagnan successfully pretends to be Milady's lover in order to have sex with her in *The Three Musketeers*, he seriously wrongs her.[5] Dougherty then asks what distinguishes D'Artagnan's deception of Milady and deceptions of the sort the lenient view sees as less serious.[6] If we cannot distin-

5 Dougherty, "Sex, Lies, and Consent," 724.

6 Dougherty, "Sex, Lies, and Consent," 728.

guish D'Artagnan's deception from that of someone who laughs at jokes he or she does not find funny to get someone else into bed, then we must admit that the two cases are alike. Since it is difficult to claim that there is nothing seriously wrong with pretending to be someone's lover into order to deceive her or him into sex, not being able to distinguish that case from apparently more trivial ones suggests that they are in fact not so trivial.

Dougherty argues that the lenient view typically relies on an unacknowledged implausible and illiberally moralistic view of sex, and that once we give up that view, it is not clear on what basis the lenient view distinguishes seemingly trivial sexual deceptions from ones it acknowledges are more serious. As he puts it, "the Lenient Thesis can only plausibly be based on [an] account of consent that makes a fundamental distinction between different features of a sexual encounter."[7] On this view, deception about a sexual encounter's core features, like whom it is actually with, makes it nonconsensual whereas deception about its peripheral features, like what the person it is with finds funny, does not. One way to defend the lenient view then is to explain why an example of seriously wrongful sexual deception is seriously wrongful without using such a distinction between consent to different features of a sexual encounter, so showing that it may not need it. The only alternative explanation for the serious wrongfulness of sexual deception like D'Artagnan's of Milady that Dougherty considers is that it is harmful, which he persuasively argues is inadequate.[8]

However, the serious wrongfulness of D'Artagnan's deception of Milady can be explained in other ways. That deceit seems like it may well be wrong in at least three ways that do not depend on the morality of consent to sex or on its harmfulness.[9] First, whether D'Artagnan lacks consent for sex with Milady, he wrongs both her and her lover by pretending to be him. Pretending to be someone with whom anyone is intimate is wrong because of the risk of corrupting or damaging their relationship, especially if the pretence is exploited for some kind of intimacy. It would be wrong for D'Artagnan to pretend to be Milady's father or brother, and even more so if he used that deception to discover secrets from her childhood, for instance. Second, D'Artagnan relies on the exclusivity of

7 Dougherty, "Sex, Lies, and Consent," 728.

8 Dougherty, "Sex, Lies, and Consent," 725–27.

9 Note that this defence of the lenient view against Dougherty's inference to the best explanation of this case is not committed to these features being sufficient to explain either the wrongfulness of D'Artagnan's deception or even the difference between that deception and deceptions the lenient view sees differently. Dougherty needs there to be nothing to explain the difference other than a moralized view of sexual consent. If there is something more than a moralized view of sexual consent, then there is an alternative explanation for the distinction the lenient view draws.

Milady's sexual relationship with her lover to undermine that relationship by destroying its exclusivity. Exploiting the vulnerability created by the central feature of a person's commitment in that way is particularly cruel. Someone who seriously damages an athlete's ability by goading them to push her- or himself too hard wrongs him or her. Third and relatedly, D'Artagnan causes Milady to fail to keep a promise to her lover not to have sex with anyone else. Tempting someone into not going to see the aging and sick relative she or he has promised to visit is wrong, for example. The more weighty the promise, the more serious the wrong, and promises of sexual fidelity are typically fairly weighty.

There are then at least sometimes the resources to explain what is seriously wrong with obviously objectionable sexual deception without claiming that sexual deception itself is automatically seriously wrong. The lenient view can be defended in ways that Dougherty does not discuss. The wrongfulness of deceit does not always relate only to the way it invalidates a particular act of consent it makes possible, as Dougherty's explanation of the wrong of sexual deceit requires. However, Dougherty does not merely assume that the lenient view distinguishes apparently trivial sexual deception from cases like D'Artagnan's by distinguishing between forms of consent. As well as arguing that the lack of consent explains the seriousness of sexual deceptions like D'Artagnan's and that they cannot be distinguished from the deceptions the lenient view treats as trivial, Dougherty also argues that all nonconsensual sex is seriously wrong.[10] If that argument holds, then the distinction the lenient view relies on must be one relating to consent. Otherwise, in virtue of being nonconsensual, sex must be seriously wrong. Either deception does not undermine consent, or the absence of consent to sex does not necessarily make it seriously wrong. Accepting the explanation I have just given of why D'Artagnan's deception of Milady is seriously wrong then requires accepting one of two burdens. Anyone accepting it must either accept that nonconsensual sex is not automatically seriously wrong or explain why deception does not always undermine consent. It is only if one of those two horns is grasped that the explanation can do the work its defence of the lenient view needs it to, and show that D'Artagnan's deception is not like pretending to find someone's jokes funny so that she or he will have sex with you. Otherwise, pretending to find someone's jokes funny so that he or she will have sex with you makes that sex nonconsensual and, in virtue of being nonconsensual, seriously wrong.

This is where Dougherty's view about the morality of sexual deception begins to link with his views about the morality of deception in general. The link is through a dilemma Dougherty's view faces parallel to that which confronts any-

10 See, for example, Dougherty, "Sex, Lies, and Consent," 720.

one defending the lenient view. On the one hand, if Dougherty adopts a comparatively demanding view of consent, then deception will undermine consent, but it will be less clear that full consent is necessary to grant permissions. It will then be harder to maintain that all nonconsensual sex is seriously wrong since it will seem like something less demanding than full consent is enough for permission for sex. On the other hand, if Dougherty adopts a comparatively lax view of consent, then although it will seem like consent is the standard mechanism for giving permissions to act, it will be compatible with deception. Deceptive sex will then sometimes still be consensual sex. Dougherty typically resolves this dilemma by opting for a comparatively demanding view of consent. Insofar as he outlines his view, it requires that agreement is or at least would be given to any eventualities covered by the consent.[11] If deception occurs about what Dougherty calls a deal-breaker, a feature of whatever is consented to whose absence or presence is required for consent, consent has not been given.

Dougherty's view seems overly demanding about the role agreement over details plays in consent. Imagine a landlord who demands that tenants sing in the shower every morning. I do not think that tenants who lie about whether they will sing in order to rent the property are in some way wronging the landlord, even though the landlord could refuse to rent them the property if they did not promise to sing.[12] Equally, imagine a tenant who will only rent properties that have never been lived in by children. I do not see anything seriously wrong with a landlord lying about this. At least if either of these deceptions is wrong, it is not wrong in the way that many nonconsensual bargains are. They are certainly not wrong in the way that deceiving someone about whether you in fact own a property or can pay the rent are, typically independently of whether that deception ends up harming its victim.

That suggests that deception about a deal-breaker is not always as important as Dougherty thinks. One form of resistance to that suggestion would be to deny that the relevant standards for the wrongfulness of deceptions can be transferred from agreements over property to agreements over sexual contact. However, Dougherty is not in a position to distinguish between consent to those two different kinds of agreements. He rejects the moralism about sexual consent associated with the lenient view. Equally, his examples suggest that he sees our

11 See, for example, "Sex, Lies, and Consent," 719.

12 This is important, since it suggests that this is not a case in which the landlord is subject to a nondiscrimination requirement. If she or he were subject to a nondiscrimination requirement, he or she would have no right to require singing and so there would be no need to gain permission. It might also not be analogous to choices over sexual partners, where at least it seems more likely to be permissible to discriminate on grounds of race, religion, gender, and other grounds standardly protected by nondiscrimination requirements.

rights over property as similar in this way to our rights over our bodies.[13] At the level of generality with which Dougherty is concerned, interests in controlling our property often seem similar to interests in the sexual autonomy Dougherty sees as underlying his rejection of the lenient view.[14] In general, people have an interest in being able to use their property as they please, and we tend to abjure moralistic judgments on how they use it just as Dougherty urges that we abjure moralistic judgments about people's sex lives. Yet it seems we are prepared to allow for deception in the rental market, and do not judge all deceptions alike, even though some deceptions are very serious. Why not also allow deception when we are choosing whom to have sex with, or at least distinguish different kinds of deception?

The kinds of deception we allow in the rental market seem to relate to the interests we think people typically have in renting property. This might seem a form of moralism about how we use property of the sort Dougherty condemns about sexual consent, but it does not have to be. When we decide what lies are permissible in the rental market, we consider what kind of impact allowing them would have on interests whose value is established independently of the existence of the rental market. We have interests, for example, in having a private space in which we are able to act as we please without needing permission from others. This explains why we think tenants may lie to their landlord about whether they are singing in the shower and, perhaps, given that interest extends to setting terms on how others use our property, why the landlord may insist that they do sing in the shower. This interest though is not dependent on views about how we use property. It depends instead on a view about the value of privacy. It is compatible with individuals voluntarily giving up that privacy, for example.

Similarly, the deceptions we think are acceptable when we seek sexual partners, I suggest, are deceptions that relate to the interests we typically think are at stake in how and whether we seek sexual partners. Obvious candidates for those interests include satisfying physical desire, securing a sense of esteem, achieving a kind of intimacy and forming lasting emotional bonds. In particular, being transparent will often be in tension with securing a sense of esteem, since others are not likely to value everything that we do or have done. Especially when we may also be vulnerable because of our intimacy, a little dishonesty may well be warranted. Intimacy may be unattainable without insincerity in some circumstances precisely because of our wariness about the vulnerability it involves. The

13 See Dougherty, "Sex, Lies, and Consent," 732–34, 737–39.

14 See for example, Dougherty, "Sex, Lies, and Consent," 730, on which Dougherty insists that "it is up to each individual to determine which features of a sexual encounter are particularly important to her."

value of these interests is not dependent on a particular view about what sex is for, nor will they only be at stake in our seeking of sexual partners if we adopt a particular view about what sex is for. We cannot choose how we or our sexual partners respond to our sexual encounters, and we may reasonably want to secure ourselves against certain kinds of response.

Of course, this is not a full defence of the lenient view. Insofar as it is a standard and widely held view about an area of life that Dougherty rightly points out contained in the recent past a great deal of injustice, it is very likely that the lenient view is not fully defensible.[15] We have good reason to think that we have not begun to identify all the injustices that mar our sexual lives. As I have already pointed out, we are certainly a very long way from eradicating them. However, it does seem like some of what the lenient view excuses ought to be excused. There are grounds for thinking that not all sexual deceptions are similarly seriously wrong. We can give reasons that do not relate to consent why some sexual deception is seriously wrong, and it seems like we might often have legitimate interests in deceiving sexual partners. Indeed, many of the interests at stake in deceiving sexual partners are also threatened by many other forms of human interaction, which often cannot be straightforwardly separated from attempts to seek sexual partners anyway. This is why it is permissible to deceive relatives, colleagues, friends, and people you do business with, as well as total strangers, about your religious affiliation or lack of it, your political views, your job, your sexual preferences, and so on. They do not always need to know the whole truth in order for your relationships with them not to be exploitative and so wrongful and, for many of the same reasons, nor do your sexual partners.[16]

University of Reading
r.s.d.jubb@reading.ac.uk

References

Dougherty, Tom. "Sex, Lies, and Consent." *Ethics* 123, no. 4 (July 2013): 717–44.

UK Ministry of Justice, Home Office, and Office for National Statistics. *An Overview of Sexual Offending in England and Wales*. January 10, 2013. https://www.gov.uk/government/statistics/an-overview-of-sexual-offending-in-england-and-wales.

15 See Dougherty, "Sex, Lies, and Consent," 722.

16 I would like to thank Tom Dougherty, Hugh Lazenby, Patrick Tomlin, and Maria Carla Zizolfi for comments on earlier versions of this piece, as well as an anonymous referee.

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
http://www.jesp.org
ISSN 1559-3061

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.