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 $a > 0$.)

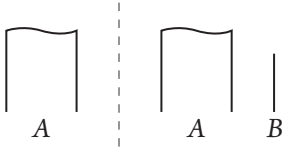
We have listed the Critical Level theory here for completeness, but for ease of exposition, we set it aside until section 5. In the present section, we consider the case in which credences are split between the other theories on the above list.

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

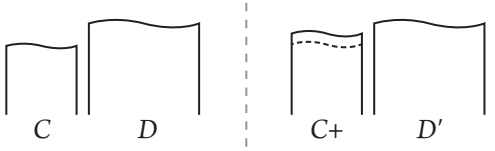
extinction as the nonexistence of any generation after the present one. There are then three possibilities:

(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.

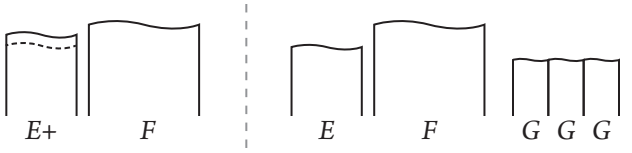
of extinction risk for simplicity of exposition; including it would complicate the detailed expression of our analysis, but would not affect its basic points.

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).

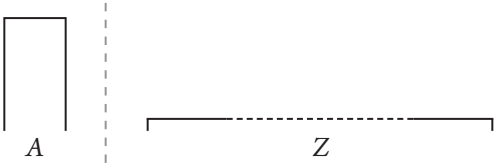
6.3. Structuralist Intertheoretic Comparisons

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.