From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

Doctoral Dissertations, 2011-2026 Philosophy

2025

Repugnance, Fanaticism, and Cluelessness: Problems for Value Maximization in Ethics and Rational Choice

Baker, Calvin

Princeton, NJ : Princeton University

http://arks.princeton.edu/ark:/88435/dsp01f7623h05h

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

REPUGNANCE, FANATICISM, AND CLUELESSNESS

PROBLEMS FOR VALUE MAXIMIZATION IN ETHICS AND RATIONAL CHOICE

Calvin Baker

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF

PHILOSOPHY

Advisers: Lara Buchak and Jacob Nebel

September 2025

© Copyright by Calvin Collins Baker, 2025. All rights reserved.

Abstract

Some of the most well-known and influential approaches to ethics and rational choice hold that we should be *value maximizers*. According to *total welfarist consequentialism*, for instance, we morally ought to maximize the sum-total of well-being in the world. And according to *expected value theory*, when we want to promote some value—such as well-being—but are uncertain what the consequences of our actions will be, we rationally ought to act in a way that maximizes the expected quantity of our target value. This dissertation explores three fundamental problems for value-maximizing strategies: that maximizing total value turns out to be axiologically *repugnant*; that maximizing expected value requires our decision-making to be dominated by low-probability, high-payoff options—so-called *fanatical* options; and that we are irredeemably *clueless* about what would maximize value in the actual world.

Chapter 1, 'Non-Archimedean population axiologies' (*Economics and Philosophy*, 2025), studies non-Archimedean, or lexical, population axiologies: theories of value which deny that all goods (and bads) can be straightforwardly traded off against each other. Such axiologies can avoid the infamous *Repugnant Conclusion* and boast a number of desirable formal properties to boot. I show that, given some plausible background assumptions, these axiologies have implications that are about as counterintuitive as the Repugnant Conclusion itself. Chapter 2, 'Expected choiceworthiness and fanaticism' (*Philosophical Studies*, 2024), responds critically to Maximize Expected Choiceworthiness (MEC), the dominant approach to choice under moral uncertainty. MEC says that, when faced with moral uncertainty, we should choose an option that maximizes expected moral value. I argue that a number of approaches to avoiding the problem of fanaticism

in cases of empirical uncertainty fail to carry over to decision-making under moral uncertainty. Chapter 3, 'New work for cluelessness,' addresses the problem of cluelessness, which many have taken to be the problem of benevolent decision-making under deep empirical uncertainty. I argue that the phenomenon of cluelessness is not limited to empirical uncertainty, however deep. It also involves higher-order uncertainty, which is a form of normative uncertainty. I propose a way to deal with choice under higher-order uncertainty that ties it to the more general problem of choice under normative uncertainty.

Table of contents

Abstract 3

Chapter one: Non-Archimedean population axiologies 6

Chapter two: Expected choiceworthiness and fanaticism 50

Chapter three: New work for cluelessness 89

Chapter one: Non-Archimedean population axiologies (*Economics and Philosophy*, 2025)

1. Introduction

By a 'heavenly life' I mean a life at a very high positive welfare level. By a 'barely good life' I mean a life at a very low positive welfare level. According to the

> *Repugnant Conclusion*, for any number of heavenly lives, there is a number of barely good lives that would be better.[1]

Similarly, by a 'hellish life' I mean a life at a very low negative welfare level. And by a 'barely bad life' I mean a life at a very high negative welfare level. According to the

> *Negative Repugnant Conclusion*, for any number of hellish lives, there is a number of barely bad lives that would be worse.

Pace Zuber *et. al.* (2021), many continue to believe that we must avoid the (Negative) Repugnant Conclusion. *Non-Archimedean* population axiologies—also called *lexical* views—can do so. For they accept at least one of the following two claims:

[1] I assume throughout that we are dealing with finite populations that solely contain human persons. See Parfit (1984: §131) for the classic introduction of the Repugnant Conclusion and an argument in its favor. See Ng (1989), Carlson (1998), and Huemer (2008) for other arguments for the Repugnant Conclusion.

The Weak Superiority of Heavenly Lives, which says that a sufficient number of heavenly lives would be better than any number of barely good lives,[2] and

The Weak Inferiority of Hellish Lives, which says that a sufficient number of hellish lives would be worse than any number of barely bad lives.[3]

The Weak Superiority of Heavenly Lives is inconsistent with the Repugnant Conclusion; and the Weak Inferiority of Hellish Lives is inconsistent with the Negative Repugnant Conclusion. The former captures the intuition that sufficiently large losses in *quality* (welfare) cannot be compensated for by gains in *quantity* (number), no matter how large; and the same goes for the latter, *mutatis mutandis*. Thus, non-Archimedean axiologies are *prima facie* compelling.

Nevertheless, the purpose of this paper is to offer a novel argument against non-Archimedean axiologies. Here's the plan. Section 2 introduces non-Archimedean axiologies in greater detail, showing why they have great initial appeal. Section 3 offers a novel argument against them. To

[2] The Weak Superiority of Heavenly Lives features a *value superiority* relation. x is *strongly superior* to y iff any quantity of x is better than any quantity of y. x is *weakly superior* to y iff some quantity of x is better than any quantity of y. The Weak Superiority of Heavenly Lives says that heavenly lives are weakly superior to barely good lives. For value superiority proposals, see Parfit (1986) and (2016), Griffin (1988: 340 n.27), Lemos (1993), Portmore (1999), Kitcher (2000), Rachels (2004), Temkin (2012), Chang (2016), Klocksiem (2016), Thomas (2018), Andersson (2021), Carlson (2022), and Nebel (2022). (Not all of these authors *endorse* these proposals.)

[3] The Weak Inferiority of Hellish Lives features a *value inferiority* relation. x is *strongly inferior* to y iff any quantity of x is worse than any quantity of y. x is *weakly inferior* to y iff some quantity of x is worse than any quantity of y. The Weak Inferiority of Hellish Lives says that hellish lives are weakly inferior to barely bad lives.

preview, the argument will be that the non-Archimedean solution to the Negative Repugnant Conclusion is untenable, and that without it, the remaining non-Archimedean view entails a different unacceptable claim, which I call the *Repugnant Elitist Conclusion*. Sections 4-5 explore the possibilities that are left open if one insists on retaining the non-Archimedean solution to the Negative Repugnant Conclusion. One option, which I call the *Somber View*, turns out to be deeply pessimistic, in that it pushes us towards favoring extinction. The remaining option, which I call *Heavy Tails*, entails a weaker version of the Repugnant Elitist Conclusion. Section 6 considers a final escape route from the argument in section 3. Section 7 concludes.

2. The Simple Additive Picture

This section introduces non-Archimedean population axiologies in greater detail. A *population* is a set of lives, each of which has a lifetime welfare level. A *population axiology* is an 'is all-things-considered better than' relation over the set of possible populations. As stated in the introduction, non-Archimedean axiologies accept at least one of the Weak Superiority of Heavenly Lives and the Weak Inferiority of Hellish Lives. Moreover, as I'll understand them, non-Archimedean axiologies endorse an attractive understanding of the structure of value that I call the *Simple Additive Picture*. The Simple Additive Picture is the conjunction of three principles—*Addition*, *Separability*, and *Transitivity*—which we introduce presently.

We begin by building to the principle of Addition. It's plausible that adding a *good life*—a life at a positive welfare level—to a population makes the population better. Why? Michael Huemer provides a simple rationale: "Worthwhile lives are good. More of a good thing is better. Therefore,

increasing the number of worthwhile lives makes the world better."[4] Similarly, it's plausible that adding a *bad life*—a life at a negative welfare level—to a population makes the population worse. We can capture these thoughts precisely by introducing some notation. Following Nebel (2023), for all populations A and B, let '$A \cup B$' refer to the union of disjoint populations A^* and B^*, where A^* is identical to A in size and welfare distribution, and likewise for B^* and B. We can now state the principle of

> *Addition*: for any population X and good (bad) life l, $X \cup l$ is better (worse) than X.

Here's another plausible claim: when we're evaluating a change to a population, people who are entirely unaffected by that change should not factor into our evaluation. Imagine, for example, that we're thinking about adding a new person to Earth's present population. Intuitively, how good or bad it would be to add this person doesn't depend on how many people lived in the Achaemenid Empire (ca. 550 - 330 BCE), or on how well-off these ancient Persians were. This thought can be captured more precisely in the principle of

> *Separability*: for all populations X, Y, and Z, X is better than Y iff $X \cup Z$ is better than $Y \cup Z$.[5]

[4] Huemer (2008: 923). Cf. Parfit (2016: 110) on the Simple View, which says that "Anyone's existence is in itself good, and makes the world in one way better, if this person's life is good to live, or worth living." See also Broome (2004) for a classic case against the view that it's axiologically neutral to add good lives to the world.

[5] See Thomas (2022b) for an explication and defense of Separability; Goodsell (2021) for an objection to a principle to which Thomas (2022b) appeals in arguing for Separability; and the Appendix of this paper for an objection to Variable Value population axiologies, which paradigmatically violate Separability.

Finally, it's plausible that the 'is all-things-considered better than' relation is transitive. Call this claim *Transitivity*.

Addition, Separability, and Transitivity comprise the Simple Additive Picture. Now, the most well-known population axiology that endorses the Simple Additive Picture is *Standard Totalism*. According to Standard Totalism, (i) welfare is a scalar quantity (which means that every welfare level can be represented by a single real number) and (ii) for every population A and B, A is better than B just in case total welfare in A is greater than total welfare in B. Standard Totalism infamously entails both the Repugnant Conclusion and the Negative Repugnant Conclusion. Non-Archimedean views consequently enjoy great initial appeal, for, like Standard Totalism, they endorse the Simple Additive Picture, but, unlike Standard Totalism, they are able to avoid the (Negative) Repugnant Conclusion.[6] What's more, non-Archimedean views are able to avoid Gustaf Arrhenius's famous impossibility results in population axiology—at least in cases of choice under certainty.[7] For these reasons, I used to believe that non-Archimedeanism was the best game in town. I have recently come to believe that this was a mistake. The next several sections explain why.

[6] They can do so by understanding welfare as a vector quantity. See Nebel (2022) for discussion.

[7] See Thomas (2018) and Carlson (2022) for discussion of non-Archimedeanism *vis-à-vis* Arrhenius's (2011) impossibility theorem and Thornley (2021b) for a new impossibility theorem for population *prospect* axiology, which non-Archimedeanism cannot avoid.

3. Main argument

This section lays out my main argument against non-Archimedean axiologies. The argument proceeds in three stages. Firstly, I argue that hellish lives and barely bad lives are *exchangeable*. Loosely, this means that we can trade off these types of lives against one another. (A precise definition is given below.) I argue for this exchangeability claim primarily by arguing against the Weak Inferiority of Hellish Lives. Secondly, I introduce a relation that I call *outweighing* and suggest that non-Archimedeans should accept that barely good lives and barely bad lives can outweigh each other. Finally, given the claims reached thus far, I derive the Repugnant Elitist Conclusion, which I regard as a *reductio*.

3.1 Preliminaries

Two preliminaries before we begin. First, I assume that a population is good (bad) iff it is better (worse) than the *empty population*, which contains zero lives. Secondly, it will be useful to prove up front that the Simple Additive Picture entails a claim that I call *Scaling*. To state Scaling, we'll have to introduce some further notation. For any population X, let 'n(X)' refer to a population composed of the union of X and (n-1) *copies* of X, where a copy of a population is a disjoint population identical to the original in size and welfare distribution. Thus, for example, '2(X)' refers to $X \cup X^*$, where X^* is identical to X in size and welfare distribution. We can now state

Scaling: for all populations A and B, A is better than B iff for every natural number q, $q(A)$ is better than $q(B)$.[8]

We'll start by showing that for all populations A and B, if A is better than B, then for every natural number q, $q(A)$ is better than $q(B)$.[9] Take an arbitrary pair of populations A and B and assume that A is better than B. We want to show that it follows that for every natural number q, $q(A)$ is better than $q(B)$. We'll do so by induction. We already have the base case, where $q = 1$: we've assumed that A is better than B. Here's the inductive step: for every natural number q, if $q(A)$ is better than $q(B)$, then $(q + 1)A$ is better than $(q + 1)B$. To establish the inductive step, assume that $q(A)$ is better than $q(B)$. It follows from Separability that $q(A) \cup B$ is better than $(q + 1)B$. And, given that A is better than B, it also follows from Separability that $(q + 1)A$ is better than $q(A) \cup B$. Since $(q + 1)A$ is better than $q(A) \cup B$ and $q(A) \cup B$ is better than $(q + 1)B$, by Transitivity, $(q + 1)A$ is better than $(q + 1)B$. This establishes the inductive step, thereby completing the left-to-right half of the proof. Going from right to left is straightforward. Assume that for every natural number q, $q(A)$ is better than $q(B)$. Then in particular A is better than B (this is the special case where $q = 1$).

[8] Following Thornley (2021a), I assume that lives are individuated by the persons leading them. I assume that we are not dealing with fractions of persons; therefore, every non-empty population contains a natural number of lives. Sometimes, for brevity, I will omit 'natural' in the main text when referring to numbers of lives.

[9] For proof strategy I follow Nebel (2022: 212-13), who proves a different result about heavenly lives and barely good lives.

3.2 Hellish lives and barely bad lives are exchangeable

We now turn to stage one of the argument. The argument will be that hellish lives and barely bad lives are *exchangeable*.

> *Exchangeability*: one bad (good) is exchangeable with another iff for any quantity of the former, there is a quantity of the latter that would be worse (better), and *vice versa*.

So, if hellish lives and barely bad lives are exchangeable, then for any number of hellish lives, there is a number of barely bad lives that would be worse, and *vice versa*.

3.2.1 Against the Weak Inferiority of Hellish Lives

To argue that hellish lives and barely bad lives are exchangeable, I first argue against the Weak Inferiority of Hellish Lives. Here's the argument in a nutshell: in conjunction with the Simple Additive Picture, the Weak Inferiority of Hellish Lives entails

> *Strong Nonsuperiority Across Adjacent Levels*: for any finite, ascending sequence of negative welfare levels beginning with a level that corresponds to a hellish life and ending with a level that corresponds to a barely bad life, there are two welfare levels that are (i) adjacent to one another in the sequence and (ii) such that *no* number of lives at the higher level is worse than *any* number of lives at the lower level.

Strong Nonsuperiority Across Adjacent Levels is false, so, since we're holding the Simple Additive Picture fixed, we must reject the Weak Inferiority of Hellish Lives.

I'll begin by deriving Strong Nonsuperiority Across Adjacent Levels; then I'll argue against it.[10] It will be useful to have the following two definitions at hand for easy reference:

> *Weak nonsuperiority*: x is weakly nonsuperior to y iff some quantity of x is not better than any quantity of y.

> *Strong nonsuperiority*: x is strongly nonsuperior to y iff any quantity of x is not better than any quantity of y.

To save space, I'll refer to a finite, ascending sequence of negative welfare levels beginning with a level that corresponds to a hellish life and ending with a level that corresponds to a barely bad life as a 'negative sequence'. Consider an arbitrary negative sequence S. Call the first level in S 'l_1' and the final level 'l_n'.

[10] Notice that the proof of Strong Nonsuperiority Across Adjacent Levels does *not* assume *Small Steps*, which says that for any two welfare levels, there is a finite sequence of intuitively small differences between them. See Arrhenius (2016), Thomas (2018), and Thornley (2021a) for discussion of Small Steps (alternatively called 'Finite Fine-Grainedness').

We begin by showing that there are two welfare levels in S that are (i) adjacent to one another in S and (ii) such that weak nonsuperiority holds across lives at these levels.[11] Assume for contradiction that there are no such welfare levels. Then for every pair of welfare levels that are adjacent in S, each number of lives at the lower (i.e., worse) welfare level is better than some number of lives at the higher (i.e., better) welfare level. It follows by Transitivity that each number of lives at l_1 is better than some number of lives at l_n. But this contradicts the Weak Inferiority of Hellish Lives, which says that some number of hellish lives (i.e., lives at l_1) is worse than any number of barely bad lives (i.e., lives at l_n). So there are (at least) two welfare levels that are (i) adjacent to one another in S and (ii) such that weak nonsuperiority holds across lives at these levels. Call these welfare levels 'l_j' and 'l_k'.

We'll now show that lives at l_j are strongly nonsuperior to lives at l_k.[12] Assume for contradiction that lives at l_j are not strongly nonsuperior to lives at l_k. This means that for some natural numbers m and n, n lives at l_k would be worse than m lives at l_j. Now take an arbitrary natural number x and consider x lives at l_j. There must be a multiple of m—say, qm—such that qm is greater than x. By Addition, qm lives at l_j would be worse than x lives at l_j. However, since n lives at l_k would be worse than m lives at l_j, by Scaling, qn lives at l_k would be worse than qm lives at l_j. Since qn lives at l_k would be worse than qm lives at l_j and qm lives at l_j would be worse than x lives at l_j, by Transitivity, qn lives at l_k would be worse than x lives at l_j. Since x was arbitrary, we have that for any natural number of lives at l_j, there is a natural number of lives at l_k that would be worse. But

[11] The sub-proof in this paragraph follows Thornley (2021a: 10), who shows the parallel claim in the context of good lives.

[12] The sub-proof in this paragraph is due to Jake Nebel (pc.).

that contradicts the fact that lives at l_j are weakly nonsuperior to lives at l_k. So lives at l_j are strongly nonsuperior to lives at l_k. And since S was an arbitrary negative sequence, we have Strong Nonsuperiority Across Adjacent Levels.[13]

I'll now give a counterexample to Strong Nonsuperiority Across Adjacent Levels. Here's the setup. We construct a finite sequence of lives that contain nothing but pain. The pain in each life is constant and qualitatively uniform. The lives differ with respect to two variables: duration and intensity of pain. The first life is relatively long and the pain in it is agonizing. It's a hellish life. The final life is one second long and the pain in it is barely noticeable. It's a barely bad life. We can get from the duration of the first life to the duration of the final life via a finite number of small decreases in length, e.g. in increments of 0.5 seconds. Plausibly, we can also get from the intensity of pain in the first life to that in the final life via a finite number of small decreases in pain intensity. We can therefore construct a finite sequence of lives beginning with the first life and ending with the final life such that the sole descriptive difference between lives that are adjacent to one another in the sequence is either a small decrease in duration or a small decrease in pain intensity (but not both). Thus, as I imagine the sequence, the second life contains pain of equal intensity to the first

[13] According to *Completeness*, for all populations X and Y, exactly one of the following is true: X is better than Y; Y is better than X; or X and Y are equally good. Given Completeness, strong nonsuperiority collapses into *strong inferiority*, where x is strongly inferior to y iff any quantity of x is worse than any quantity of y. So, if we assume Completeness, Strong Nonsuperiority Across Adjacent Levels will collapse into Strong Inferiority Across Adjacent Levels. Since Strong Inferiority Across Adjacent Levels is even more counterintuitive than Strong Nonsuperiority Across Adjacent Levels, the most plausible non-Archimedean views will allow for incompleteness (so that for some populations A and B, each of the following is false: A is better than B; B is better than A; A and B are equally good). See Thornley (2021a), Nebel (2022), and subsection 6.1 of this paper for further discussion.

life, but is 0.5 seconds shorter; the third life is equal to the second life in duration, but contains pain that is slightly less intense; and so on.

Now consider the sequence of welfare levels that map one-to-one onto the lives in this sequence. This sequence of welfare levels—*S'*—is a negative sequence. So, according to Strong Nonsuperiority Across Adjacent Levels, there are two welfare levels adjacent to one another in *S'* such that strong nonsuperiority holds across the lives at these levels. I submit that this is false. For (1) if strong nonsuperiority holds across lives at two different negative welfare levels, then the welfare difference between the lives is, in an intuitive sense, very large. But (2) for every pair of lives that are adjacent to one another in our sequence, it's not the case that the welfare difference between them is very large. So strong nonsuperiority does not hold across any pair of lives that are adjacent in our sequence. Therefore, Strong Nonsuperiority Across Adjacent Levels is false.

I'll now defend (1) and (2), beginning with the former. Consider an arbitrary pair of negative welfare levels, which we'll call 'level 1' and 'level 2'. If lives at level 1 are strongly nonsuperior to lives at level 2, then no number of lives at level 2 would be worse than any number of lives at level 1. So, in particular, ten billion lives at level 2 would not be worse than one life at level 1. For it to be plausible that ten billion lives at one negative welfare level would not be worse than one life at another negative welfare level, the difference between the welfare levels must be very large, in a straightforward intuitive sense. To illustrate, it is plausible that ten billion barely bad lives would not be worse than one hellish life. In contrast, it is not plausible that ten billion terrible-but-not-quite-hellish lives would not be worse than one hellish life. They would be worse.

Turning now to (2): by hypothesis, the descriptive difference between any two lives that are adjacent in our sequence is small. It therefore seems that the welfare difference between any two lives that are adjacent in our sequence is small. However, one might object that small descriptive differences in pain intensity can make a difference to the *type* of pain that is at issue, which itself is welfare-relevant.[14] For instance, one might hold that although the *descriptive* difference between agony and pain that is terrible-but-not-quite-agony is small, the *welfare* difference between agony and pain that is terrible-but-not-quite-agony is not small. There would then be two lives that are adjacent in our sequence—one containing agony and the other containing pain that is terrible-but-not-quite-agony—such that the welfare difference between them is not small. Is it then plausible to claim that strong nonsuperiority holds across these lives? Intuitively, I find that the answer is No. For even if this gambit gives us a plausible rationale for resisting the inference from small descriptive difference to small welfare difference, it does not support the further claim that the welfare difference is *so* large as to ground the holding of strong nonsuperiority across the lives in question. To continue with our example for concreteness: even granting *arguendo* that the welfare difference between the life containing agony and the adjacent life containing pain that is terrible-but-not-quite-agony is not *small*, it does not follow that the welfare difference is *very large* in the sense intended in (1).

Finally, it's worth considering the following partners-in-guilt defense of Strong Nonsuperiority Across Adjacent Levels.[15] Consider a finite, ascending sequence of welfare levels that solely

[14] See Klocksiem (2016: 1324) for statement and defense of this position. More generally, one might think that *intrapersonal spectrum arguments* motivate the existence of an evaluatively significant threshold somewhere along the pain continuum; see e.g. Pummer (2018) for discussion.

[15] Adapted from Thornley (2021a: §4).

contains some negative welfare levels and a *neutral level* (=df a welfare level such that adding a life at this level does not make the population better or worse). Suppose that the final two levels in the sequence are a barely negative level and the lone neutral level in the sequence. Many will agree that strong nonsuperiority holds across lives at these two levels, for intuitively, no number of bad lives would be better than any number of neutral lives. What's more, many will retain this judgment even if it is stipulated that the descriptive difference between life at the barely negative level and life at the neutral level in question is small. By way of illustration, return to the sequence of pain-filled lives that we constructed above. The final life in this sequence is one second long and contains pain that is barely noticeable. Append to this sequence a life that is one second long and hedonically neutral. Plausibly, this is a neutral life; and the descriptive difference between it and its predecessor is small. Thus, the holding of strong nonsuperiority across lives that are adjacent to one another in a sequence and that are descriptively similar is not a unique commitment of non-Archimedeans who endorse the Weak Inferiority of Hellish Lives.

I am not convinced by this partners-in-guilt defense for two reasons. Firstly, the non-Archimedean must claim that strong nonsuperiority holds across different bad lives—e.g., across hellish lives and lives that are terrible-but-not-quite-hellish. In contrast, in claiming that bad lives are strongly nonsuperior to neutral lives (as we just did), our view is that strong nonsuperiority holds across lives that are bad and lives that aren't. This is a particular instance of the more general principle that no amount of a bad thing would be better than any amount of a not bad thing (i.e., that bad things are strongly nonsuperior to not bad things). Secondly, to avoid the Negative Repugnant Conclusion, non-Archimedeans must claim that there are at least *two* instances of strong

nonsuperiority holding across bad lives.[16] As Jensen observes, this has the unpalatable consequence of creating a 'zone' of incommensurability within the welfare levels, which appears to "dissolve the difference in value between the members of the zone."[17] It therefore seems to me false that non-Archimedeans are partners-in-guilt with those of us who simply wish to claim that bad things are strongly nonsuperior to not bad things.

This concludes my case against Strong Nonsuperiority Across Adjacent Levels.[18] Since the Weak Inferiority of Hellish Lives, in conjunction with the Simple Additive Picture, entails Strong Nonsuperiority Across Adjacent Levels, and we are keeping the Simple Additive Picture fixed, I conclude that we must reject the Weak Inferiority of Hellish Lives.[19]

---

[16] Assuming, as I do in footnote 13, that the most plausible non-Archimedean views allow for incompleteness. Rabinowicz (2022: 437 and 2022: 437, fn. 12) shows that non-Archimedean views with this structure require at least two instances of strong noninferiority holding across good lives to avoid the Repugnant Conclusion. His argument applies to the Negative Repugnant Conclusion, strong nonsuperiority, and bad lives *mutatis mutandis*.

[17] Jensen (2020: 308). See Jensen (2020: §6) for further discussion.

[18] I have left unaddressed the claim that we can soften the blow of Strong Nonsuperiority Across Adjacent Levels (and similar implications; see footnote 19 of this paper) by holding that it is *vague* which levels the strong nonsuperiority relation holds across. I am not optimistic about this move but lack the space to address it here. See Nebel (2022: §8.4), Pummer (2022), and Thomas (2022c) for discussion.

[19] By deploying arguments parallel to those in this subsection, we can show that in conjunction with the Simple Additive Picture, the Weak Superiority of Heavenly Lives entails *Strong Noninferiority Across Adjacent Levels*: for any finite, descending sequence of positive welfare levels beginning with a level that corresponds to a heavenly life and ending with a level that corresponds to a barely good life, there are two welfare levels that are (i) adjacent to one other in the sequence and (ii) such that no number of lives at the lower level would be better than any number of lives at the higher level. Insofar as we're happy to reject the Weak Inferiority of Hellish Lives on the ground that it entails Strong Nonsuperiority Across Adjacent Levels, why can't we similarly reject the Weak Superiority of Heavenly Lives on the ground that it entails

3.2.2 In favor of exchangeability

If hellish lives are not weakly inferior to barely bad lives, does it follow that hellish lives are exchangeable with barely bad lives? No. Handfield and Rabinowicz (2018) and Rabinowicz (2022) prove (assuming Transitivity) that *Trilemma* holds for all goods (bads) that are quantifiable and such that more is better (worse).[20]

> *Trilemma:* for all such goods (bads) *x* and *y* such that *x* is better (worse) than *y*, exactly one of the following is true: *x* is weakly superior (inferior) to *y*, *x* and *y* are exchangeable, or *x* is *radically incommensurable* with *y*.

> *Radical incommensurability: x* is radically incommensurable with *y* iff there is a quantity k of *x* such that for every quantity k^+ of *x* at least as great as k, there is some quantity k' of *y* such that k^+x is incommensurable with every quantity of *y* at least as great as k'.

Strong Noninferiority Across Adjacent Levels—and thereby dispense with non-Archimedean views? I have chosen not to pursue this argumentative strategy because, to my mind, Strong Nonsuperiority Across Adjacent Levels is significantly more counterintuitive than Strong Noninferiority Across Adjacent Levels; and I expect that many readers will share this intuition. Thus, I consider it to be an advantage of my argument against non-Archimedean views that it does *not* depend on the intuitive rejection of Strong Noninferiority Across Adjacent Levels.

[20] Strictly speaking, Handfield and Rabinowicz's (2018) proof specifically concerns *harms* and Rabinowicz's (2022) proof specifically concerns *goods*, but their proofs generalize to the Trilemma stated in the main text of this paper.

Could hellish lives be radically incommensurable with barely bad lives? If they were, then there would be some natural number k such that for every natural number k^+ at least as great as k, there is a natural number k' such that k^+ hellish lives would be incommensurable with every natural number of barely bad lives at least as great as k'. To illustrate, suppose that $k = 100$ and consider 101 hellish lives. It would follow that there is a natural number—say, 10,000—such that 101 hellish lives would be incommensurable with every natural number of barely bad lives at least as great as 10,000. Could this be true? I shall argue that the answer is No.

Suppose hellish lives are radically incommensurable with barely bad lives. Now consider an arbitrary negative sequence and suppose for contradiction that lives at every pair of levels that are adjacent in the sequence are exchangeable. It follows by Transitivity that for any number of lives at the first level (which corresponds to a hellish life), there is a number of lives at the final level (which corresponds to a barely bad life) that would be worse. But that contradicts the supposition that hellish lives are radically incommensurable with barely bad lives. So, it's not the case that lives at every pair of adjacent levels are exchangeable. It follows by Trilemma that either weak inferiority or radical incommensurability holds across lives at at least one pair of adjacent levels. I assume the former option is a non-starter, so, since the negative sequence was arbitrary, we have

> *Radical Incommensurability Across Adjacent Levels*: in any negative sequence, radical incommensurability holds across lives at at least one pair of welfare levels that are adjacent in the sequence.

This is implausible for essentially the same reasons that Strong Nonsuperiority Across Adjacent Levels is implausible. Moreover, the claim that hellish lives are radically incommensurable with barely bad lives does not even enjoy the intuitive appeal of the Weak Inferiority of Hellish Lives. I conclude against the radical incommensurability of hellish and barely bad lives. By Trilemma, we are left with exchangeability.

3.3 Outweighing

We now proceed to stage two of the main argument. Here, I introduce the *outweighing* relation and suggest that non-Archimedeans should accept that barely good lives and barely bad lives can outweigh each other.

> *Outweighing*. *x* can outweigh *y* iff for any quantity k of *y* that is good (bad), there is a quantity k' of *x* such that k*y* together with k'*x* is bad (good).

Non-Archimedeans should accept that barely bad lives can outweigh barely good lives, because denying this claim is intuitively intolerable. Here's why. Suppose that barely bad lives could not outweigh barely good lives. Then there would be some natural number *n* such that *n* barely good lives along with *any* number of barely bad lives would not be a bad population. That is unacceptable. *n* barely good lives along with 10^n barely bad lives would be a bad population. So barely bad lives can outweigh barely good lives.

Can barely good lives outweigh barely bad lives? If they could not, then there would be some natural number *m* such that *m* barely bad lives along with any number of barely good lives would not be a good population. To me, this does not register as obviously false in the way that the corresponding claim about good lives does. However, we can show a stronger result: if *m* barely bad lives along with any number of barely good lives would not be a good population, then a *single* barely bad life along with any number of barely good lives would not be a good population. To see how, assume for contradiction that there is some natural number *x* such that *x* barely good lives ∪ one barely bad life would be a good population, i.e., would be better than the empty population. It follows from Scaling that for every natural number *q*, *qx* barely good lives ∪ *q* barely bad lives would be better than the empty population. But that is inconsistent with the claim that barely good lives can't outweigh barely bad lives—i.e., that there is a natural number *m* such that *m* barely bad lives ∪ any number of barely good lives would not be good, i.e., would not be better than the empty population. So, if barely good lives can't outweigh barely bad lives, then a single barely bad life cannot be taken together with any number of barely good lives to form a good population.

However, it seems *internally* ill-motivated for non-Archimedeans to deny that there is some number of barely good lives that can be taken together with a single barely bad life to form a good population. Despite being barely so, barely good lives are good. They are positively valuable for the persons leading them. By Addition, the addition of each barely good life makes the population better. In virtue of what, then, would the goodness of enough such lives—we can make the number as large as we like—fail to overcome the badness of a single barely bad life? For now, I'll assume

that non-Archimedeans lack a compelling answer to this question, but we'll revisit this assumption in section 6.[21]

3.4 The Repugnant Elitist Conclusion

So far, I've argued that hellish lives and barely bad lives are exchangeable (stage 1) and that non-Archimedeans should accept that barely good lives and barely bad lives can outweigh each other (stage 2). We can now derive the

> *Repugnant Elitist Conclusion*: some number of heavenly lives ∪ any number of hellish lives would be a good population.

[21] On the standard conception of the welfare levels, the welfare levels consist exhaustively of positive welfare levels, negative welfare levels, and a single neutral level. In contrast, Gustafsson (2020) explores an alternative picture on which there are positive welfare levels, negative welfare levels, and multiple *undistinguished* welfare levels, where life at an undistinguished welfare level is neither good, nor bad, nor neutral for the person leading it. An anonymous referee suggests that Gustafsson's alternative picture undermines intuitive support for the claim that barely good lives can outweigh barely bad lives, because, on the alternative picture, barely bad lives may be significantly worse than barely good lives. However, even granting Gustafsson's alternative picture *arguendo*, it's not clear to me why the fact that barely bad lives are significantly worse than barely good lives would, in itself, undermine the outweighing thesis. After all, if barely bad lives are significantly worse than barely good lives, then barely good lives are significantly better than barely bad lives. What *would* undermine the outweighing thesis is the claim that barely bad lives are (significantly) more bad than barely good lives are good—i.e., that barely bad lives have a (significantly) greater absolute value than barely good lives. But as far as I am aware, this is a substantive further claim that (i) does not simply fall out of Gustafsson's alternative picture and (ii) is unmotivated, at least as things stand at present.

We begin by proving that for all good populations X and Y and every bad population Z, if X is better than Y and $Y \cup Z$ is good, then $X \cup Z$ is good. To do so, assume that we have two arbitrary good populations X and Y such that X is better than Y and an arbitrary bad population Z such that $Y \cup Z$ is good, i.e., better than the empty population. Since X is better than Y, by Separability, $X \cup Z$ is better than $Y \cup Z$. Since $X \cup Z$ is better than $Y \cup Z$ and $Y \cup Z$ is better than the empty population, by Transitivity, $X \cup Z$ is better than the empty population—i.e., $X \cup Z$ is good. Using the same strategy, we can also show that for all bad populations X and Y and every good population Z, if X is worse than Y and $X \cup Z$ would be good, then $Y \cup Z$ would be good. The proof is omitted for brevity.

Now for the Repugnant Elitist Conclusion:

1. There is a natural number t such that t heavenly lives would be better than any natural number of barely good lives. (Weak Superiority of Heavenly Lives)
2. For any natural number of barely bad lives, there is a natural number of barely good lives such that the barely bad lives ∪ the barely good lives would be a good population. (Barely good lives can outweigh barely bad lives)
3. For all good populations X and Y and every bad population Z, if X is better than Y and $Y \cup Z$ is good, then $X \cup Z$ is good.
4. t heavenly lives ∪ any number of barely bad lives would be a good population. (1-3)
5. For any natural number of hellish lives, there is a natural number of barely bad lives that would be worse. (Exchangeability)

6. For all bad populations X and Y and every good population Z, if X is worse than Y and $X \cup Z$ would be good, then $Y \cup Z$ would be good.
7. t heavenly lives $\cup$ any number of hellish lives would be a good population. (4-6)

(7) is the Repugnant Elitist Conclusion, which I regard as a *reductio*. (Intuitively, for any natural number of heavenly lives, there is *some* natural number of hellish lives such that the heavenly and hellish lives together would not be a good population.) To drive a final nail into the coffin, we can also show that one heavenly life $\cup$ any natural number of hellish lives would not be a bad population. To see how, assume for contradiction that there is a natural number x such that one heavenly life $\cup$ x hellish lives would be bad, i.e., worse than the empty population. Then, by Scaling, for every natural number q, q heavenly lives $\cup$ qx hellish lives would be worse than the empty population. But this contradicts the Repugnant Elitist Conclusion.

3.5 Argument recap

Non-Archimedean population axiologies endorse the Simple Additive Picture and the Weak Superiority of Heavenly Lives and/or the Weak Inferiority of Hellish Lives. We can name and taxonomize the resulting positions as follows:

Table 1: Non-Archimedean population axiologies

	Weak Superiority of Heavenly Lives	Weak Inferiority of Hellish Lives
The Rosy View	✓	✗
The Somber View	✗	✓
Heavy Tails	✓	✓

Here's my main argument against non-Archimedeanism in a nutshell: in conjunction with the Simple Additive Picture, the Weak Inferiority of Hellish Lives entails Strong Nonsuperiority Across Adjacent Levels. That's false, so we should reject the Weak Inferiority of Hellish Lives. That rules out the Somber View and Heavy Tails, but leaves open the Rosy View. However, the Rosy View entails the Repugnant Elitist Conclusion, which is false. So all non-Archimedean positions fail.

The next few sections proceed as follows: in due recognition of the fact that not everyone will find Strong Nonsuperiority Across Adjacent Levels as damning as I do, I consider the Somber View in the next section and Heavy Tails in the one after. Then, in section 6, I consider the option of denying that barely good lives can outweigh barely bad lives.

4. The Somber View

The Somber View is the mirror image of the Rosy View. It accepts the Weak Inferiority of Hellish Lives and the claim that barely good lives and barely bad lives can outweigh one another. And, just as the Rosy View holds that hellish lives and barely bad lives are exchangeable, the Somber View holds that heavenly lives and barely good lives are exchangeable. Thus, just as the Rosy View bites the bullet on the Negative Repugnant Conclusion, the Somber View bites the bullet on the Repugnant Conclusion.

We can run an argument parallel to the one given for the Repugnant Elitist Conclusion to show that the Somber View implies the following: any population with a sufficient number of hellish lives is bad, no matter the number of heavenly lives it also contains (see Appendix for proof).[22] I find this implication *prima facie* plausible, unlike the Repugnant Elitist Conclusion. For it is plausible that the extremes of suffering are so horrific that an outcome is irredeemably marred if

[22] Mogensen (fc.) discusses *Lexical Threshold Negative Utilitarianism* (LTNU), which is stronger than the Somber View. According to LTNU, one hellish life along with any number of good lives would be a bad population. Mogensen derives this conclusion from Separability, Transitivity, the Weak Inferiority of Hellish Lives (which he calls the 'Contrary Reverse Repugnant Conclusion'), and two claims that I do not appeal to: firstly, Completeness, which says that for all populations *A* and *B*, exactly one of the following is true: *A* is better than *B*, *B* is better than *A*, or *A* and *B* are equally good; and secondly, that bad lives can outweigh good lives. (In contrast, I claim—more weakly—that barely bad lives can outweigh barely good lives). Mogensen later shows that if one drops the assumption of Completeness, one can derive *Weak Lexical Threshold Negative Utilitarianism* (WLTNU), which, like the Somber View, implies that once a population has a sufficient number of hellish lives, it's bad. He does not consider the objection that (W)LTNU implies Strong Nonsuperiority Across Adjacent Levels, however.

it contains enough of them.[23] That is the chief reason why the Somber View is worthy of note and discussion. However, in light of the grim inductive evidence from the historical record and the many different ways in which the future could play out, there is, disturbingly, a large expected number of hellish lives in the future.[24] Now, strictly speaking, the Somber View does not say anything about the future, for, as stated, it does not say anything about the evaluation of risky prospects. But a natural synthesis of the Somber View with expected value reasoning will yield the result that the future is bad in expectation.[25] So, on this view, it would be expectedly better if

[23] Crisp (2021) is sympathetic to this perspective.

[24] As Edward Gibbon is famously quoted: "history…is, indeed, little more than the register of the crimes, follies, and misfortunes of mankind" (Gibbon 1996: Vol. 1, ch. 3, pt. 2). Gibbon's quip is undoubtedly hyperbolic, but the fact that it's humorous, rather than simply puzzling, testifies to the element of truth it contains. See, e.g., Glover (2012) on the untold horrors contained within just one century of human history (the 20th); Alexander (2015) on the misery present in the contemporary United States; and Caviola *et al*. (2021, reproduced in MacAskill 2022: 200), who find in a survey administered to 240 people in India and to 240 people in the United States that 13.5% of respondents believe that their lives have contained more suffering than happiness, 7.7% would prefer to have never been born, and 25% would not relive their lives. As far as possible futures go, all of the following—in which hellish lives would very likely be realized—are live epistemic possibilities: dystopia (e.g., a long-lasting global totalitarianism), the unrecoverable collapse of industrial civilization (perhaps due to a nuclear WWIII), and longterm economic stagnation. See Ord (2020) and MacAskill (2022) for discussion of possible future scenarios.

[25] The Somber View implies that any population with at least some number x of hellish lives is bad. So, on the synthesis I'm imagining, if the expected number of hellish lives in a future population is at least as great as x, then that future population is bad in expectation. Given that there is a large expected number of hellish lives in the future of humanity, our future is expectedly bad by the lights of the Somber View—at least in terms of human welfare, and barring the *ad hoc* maneuver of stipulating that x is so high as to be practically unreachable (a move that Mogensen (fc.) also dismisses; cf. Mogensen as well on setting the lifetime welfare level of a hellish life so low as to be practically unrealizable).

we succumbed to an immediate, painless extinction. That is a difficult pill to swallow.[26] Moreover, the Somber View implies that no number of heavenly lives along with a single hellish life would be a good population (see Appendix for proof). The Somber View therefore has two options: either a single hellish life along with any number of heavenly lives would constitute a bad population, or it would constitute a population that is neither good, nor bad, nor neutral—i.e., a population that is incommensurable with the empty population. Neither option seems right. For a population containing one hellish life and ten trillion heavenly lives seems good overall.[27]

5. Heavy Tails

The preceding observations suggest to me that the Somber View is false. It remains for us to examine Heavy Tails. Heavy Tails affirms the Weak Superiority of Heavenly Lives; the Weak Inferiority of Hellish Lives; that barely good lives and barely bad lives can outweigh each other; and that heavenly and hellish lives can outweigh each other.[28] Heavy Tails thus avoids both the

[26] Similarly, Mogensen (fc.) writes, "the greatest cost of accepting [Lexical Threshold Negative Utilitarianism, on which see footnote 22] is surely that it appears to support the desirability of human extinction".

[27] For agreement, see Mogensen's (fc.) assessment of Ursula K. Le Guin's "The Ones Who Walk Away from Omelas."

[28] Why assume that heavenly lives can outweigh hellish lives? Three reasons. First, it gives Heavy Tails a natural, appealing symmetry. Second, assume that they can't. Then there's a natural number n such that n hellish lives along with any number of heavenly lives wouldn't be good. Since there is in expectation a large number of hellish lives in the future (see section 4), we expect that the future will not be good, no matter how many heavenly lives it contains (at least in welfarist terms, and again, barring the *ad hoc* maneuver of setting n so high that it could never be reached in practice). I am unprepared to accept this implication, though I grant that this is a reasonable point of disagreement.

Repugnant Elitist Conclusion and the Somber View's proclivities towards extinction—in addition to both the Repugnant Conclusion and the Negative Repugnant Conclusion. Heavy Tails therefore seems to me the *relatively* most plausible non-Archimedean view, despite the fact that it bites the bullet on Strong Nonsuperiority Across Adjacent Levels.[29] However, Heavy Tails implies that one heavenly life along with any number of barely bad lives would not be a bad population (see Appendix for proof). This implication—call it the *Weak Repugnant Elitist Conclusion*—is rather counterintuitive. For instance, it seems that a population containing one heavenly life and ten trillion barely bad lives would be bad. Heavy Tails must deny this. To be thorough, it is also worth noting that Heavy Tails has the parallel implication that one hellish life along with any number of barely good lives would not be a good population. It is unclear to me whether this implication counts in favor of or against Heavy Tails, though I appreciate that many will find it welcome.

6. Outweighing redux

Might the preceding arguments be taken as a *reductio* of the claim that barely good lives and barely bad lives can outweigh one another (§3.3)? I'll take the claim that barely bad lives can outweigh barely good lives as unassailable, since it's extremely plausible that one barely good life can be conjoined with *some* number of barely bad lives to form a bad population. In contrast, the claim that one barely bad life can be conjoined with some number of barely good lives to form a good population admittedly enjoys less intuitive support. So, the place to push is against the claim that barely good lives can outweigh barely bad lives.

[29] And Strong Noninferiority Across Adjacent Levels—see footnote 19 for discussion.

The non-Archimedean could alternatively claim that one barely bad life along with any number of barely good lives would be neither good, nor bad, nor neutral (i.e., that it would be neither better than, nor worse than, nor equally as good as the empty population, but rather incommensurable with the empty population). This move continues to strike me as *ad hoc*, given that, by non-Archimedean lights, the world just gets better and better as we add (barely) good lives.[30] Still, it does have the welcome effect of blocking the arguments for the Repugnant Elitist Conclusion and the Weak Repugnant Elitist Conclusion; and for this reason, I expect that many non-Archimedeans will go in for it. What, then, is the best view for the non-Archimedean who takes this route?

Denying that barely good lives can outweigh barely bad lives does not help the Somber View, so we'll examine revisions of Heavy Tails and the Rosy View, in that order. *Revised Heavy Tails* is the same as Heavy Tails, except that it denies that barely good lives can outweigh barely bad lives—thereby avoiding the Weak Repugnant Elitist Conclusion. Unfortunately, Revised Heavy Tails—alongside Heavy Tails—implies that there is a natural number n such that n heavenly lives ∪ any natural number of barely bad lives would be a good population (see Appendix for proof).[31]

[30] Though see Hájek and Rabinowicz (2021) for a possible rationale.

[31] Cf. Mogensen (fc.) on *Lexical Total Utilitarianism*, which is very similar, if not identical, to Revised Heavy Tails and which also entails Heavenly Dominance. I have assumed throughout this paper that the populations we consider solely contain human lives. (For simplicity, I wanted to avoid the thorny topic of interspecies welfare comparisons and aggregation.) However, Mogensen has an interesting proposal—which he does not endorse—for why it might be desirable for a more general population axiology to accept Heavenly Dominance: we might feel that a world containing a sufficient number of heavenly *human* lives would be good, even if it contained arbitrarily many more barely bad *non-human animal* lives—e.g., lives of insects and fish that are just barely negative in welfare. I do not have this intuition and, more strongly, worry that whatever intuitive force the example may have rests on an objectionable form of speciesism.

This claim—call it *Heavenly Dominance*—is counterintuitive. For instance, it seems that for every natural number q, q heavenly lives ∪ 10^{100q} barely bad lives would *not* be a good population. As Mogensen (fc.) remarks, "it may well strike us most repugnant of all to assert that there are some lives so good that, for their sake, we should be willing to accept that arbitrarily many individuals may have to have lives bad enough that, for their sake, we should wish that they had never been born." It's also worth noting, though, that (Revised) Heavy Tails has the parallel implication that there is a natural number m such that m hellish lives ∪ any natural number of barely good lives would be a bad population. I lack a clear intuition about this implication, but I expect that many will find it welcome.

A second option is the *Revised Rosy View*. The Revised Rosy View is the same as the Rosy View, except that it denies that barely good lives can outweigh barely bad lives—thereby avoiding the Repugnant Elitist Conclusion. Moreover, unlike Revised Heavy Tails, the Revised Rosy View does not imply Strong Nonsuperiority Across Adjacent Levels or Heavenly Dominance. It is the ability of the Revised Rosy View to avoid these two unattractive implications of Revised Heavy Tails that makes it a live option. (Otherwise, one might wonder about the motivation for a non-Archimedean view that avoids the Repugnant Conclusion but fails to avoid the Negative Repugnant Conclusion.)

7. Conclusion

The main goal of this paper has been to offer a novel argument against non-Archimedean population axiologies, which I did in section 3. A secondary goal has been to assess the relative

plausibility of competing non-Archimedean views, which I attempted to do by identifying their most counterintuitive implications. Results are summarized in the following two tables.

Table 2: Non-Archimedean population axiologies (expanded)

	Weak Superiority of Heavenly Lives	Weak Inferiority of Hellish Lives	Barely good lives can outweigh barely bad lives
Rosy View	✓	✗	✓
Revised Rosy View	✓	✗	✗
Somber View	✗	✓	✓
Heavy Tails	✓	✓	✓
Revised Heavy Tails	✓	✓	✗

Table 3: Itemized billing for non-Archimedean views

View	Key costs
Rosy View	1. Negative Repugnant Conclusion 2. Repugnant Elitist Conclusion
Revised Rosy View	1. Negative Repugnant Conclusion 2. One barely bad life along with any number of barely good lives would not be a good population (problem of internal motivation)
Somber View	1. Repugnant Conclusion 2. Strong Nonsuperiority Across Adjacent Levels 3. No number of heavenly lives along with one hellish life would be a good population 4. Tends to favor extinction—including in the actual world
Heavy Tails	1. Strong Nonsuperiority Across Adjacent Levels 2. Weak Repugnant Elitist Conclusion 3. Heavenly Dominance
Revised Heavy Tails	1. Strong Nonsuperiority Across Adjacent Levels 2. One barely bad life along with any number of barely good lives would not be a good population (problem of internal motivation) 3. Heavenly Dominance

To close, I'd like to make three conjectures that I hope will be productively provocative. Firstly, if there is such a thing as *the correct abductive methodology in ethical theorizing*, and if this methodology gives significant weight to simplicity—as abduction seems to in scientific theorizing—then we ought to prefer Standard Totalism to non-Archimedeanism. Standard Totalism is much simpler than non-Archimedeanism. Still, it has seemed to many that non-Archimedeanism is more extensionally adequate than Standard Totalism,[32] with the chief worry leveled against the former being its implications for choice under risk.[33] I take the results shown in this paper to narrow the gap in extensional adequacy between non-Archimedeanism and Standard Totalism—even in cases of choice under certainty. With a sufficiently narrow gap in extensional adequacy, considerations of simplicity can then carry the all-things-considered judgment of which view to prefer.

Secondly, if Standard Totalism is true, then familiar forms of maximizing welfarist consequentialism are false.[34] According to Standard Totalism, all good lives are exchangeable, all bad lives are exchangeable, and good and bad lives can all outweigh each other. Standard Totalism implies that for any number of hellish lives, there is a number of barely good lives such that it would be better to add the barely good lives to the world than to alleviate the suffering contained

[32] In addition to the Repugnant Conclusion and the Negative Repugnant Conclusion, Standard Totalism implies the *Very Repugnant Conclusion*: for any number of heavenly lives, and for any *greater* number of hellish lives, the hellish lives along with some number of barely good lives would be better than the heavenly lives (Arrhenius 2011: 2).

[33] See Nebel (2022: §8.5) and Thomas (2022a: 495) for worries about non-Archimedeanism under risk. Nebel (2022: 221) calls this "the most serious problem for [non-Archimedean] views."

[34] According to consequentialism, facts about what we morally ought to do are fully determined by axiological facts.

in the hellish lives.[35] Although I am prepared to accept that this is true as a matter of impersonal axiology—i.e., as a matter of what would be better 'from the point of view of the universe'—I am unprepared to accept that we morally ought to bring into existence some large number of people with barely good lives, rather than alleviating the extreme suffering of people who currently exist.

Disturbingly, however, we can pose the paradoxical arguments in population *axiology* that lead us to Standard Totalist repugnance in *deontic* terms.[36] These deontic arguments constitute a major challenge to nonconsequentialism. As I see things, then, one of the most important open questions in ethical theory is whether nonconsequentialists can meet this challenge. If they could not, then I would be inclined to embrace some form of particularism, or to conclude that the impossibility of population ethics supports metaethical error theory.[37] The latter option would admittedly undermine this essay, for it would imply that population ethics has no subject matter. Nevertheless, *reasoning through* the paradoxes of population ethics would still have been worthwhile, insofar as it was instrumental in allowing us to grasp the metaethical truth or to better understand our own (mind-dependent) values.

[35] As dramatized in Temkin (2012: 413).

[36] See Arrhenius (2004).

[37] See Cowie (2022) for just such an argument in favor of error theory. See also McMahan (2013: 34), who writes that "It is these problems [*viz*., the problems of population ethics]...rather than arguments in metaethics about the queerness of objective values, the connections between normativity and motivation, and so on, that seem to me to pose the greatest challenge to realism in ethics."

Appendix

Separability and Variable Value Theories

To my mind, Separability is the least plausible of the three principles that comprise the Simple Additive Picture. Again, according to

> *Separability*, for all populations X, Y, and Z, X is better than Y iff $X \cup Z$ is better than $Y \cup Z$.

Unfortunately, it has proven difficult to formulate a plausible population axiology that violates Separability. Perhaps the most well-known axiologies that do so are *Variable Value theories*.[38] According to Variable Value theories, good lives have diminishing marginal value. However, it's extremely plausible that bad lives have non-diminishing marginal disvalue; and difficulties arise if we believe that there's an asymmetry in the marginal (dis)value of good and bad lives. Firstly, we incur an explanatory burden: what explains the asymmetry? Secondly, we're saddled with the following strongly counterintuitive implication. Imagine that God creates a population consisting of an arbitrarily large number of heavenly lives and one barely bad life. Intuitively, this population is good (i.e., better than the empty population). However, if God iteratively creates new populations with identical welfare distributions (we can imagine that this happens on causally isolated planets), then the universe will eventually become bad and then get ever-increasingly so. This is because the value of the heavenly lives will approach a finite upper bound as their number

[38] For work on Variable Value Theories, see Hurka (1983), Ng (1989), Sider (1991), Asheim and Zuber (2014), and Pivato (2020).

goes to infinity, while the disvalue of the barely bad lives will grow without limit. Eventually, the unbounded disvalue must eclipse the bounded value. So, by successively adding intrinsically good, causally isolated populations to the universe, we can make the universe an arbitrarily bad place. I find this mysterious.[39] It also has dire implications for the future of humanity. For unless we are sanguine about our chances of building a utopia in which there are few bad lives, the longterm survival of humanity is likely to be bad (in expectation, and in welfarist terms) on account of the asymmetry between good and bad lives.

Proofs

Proof: the Somber View implies that some number of hellish lives along with any number of heavenly lives would be a bad population.

1. There is a natural number *t* such that *t* hellish lives would be worse than any natural number of barely bad lives. (Weak Inferiority of Hellish Lives)
2. For any natural number of barely good lives, there is a natural number of barely bad lives such that the barely good lives ∪ the barely bad lives would be a bad population. (Barely bad lives can outweigh barely good lives)

[39] This objection is closely related to, but distinct from, Parfit's Ridiculous and Absurd Conclusions (1984: §138). (The Ridiculous Conclusion takes the unit of aggregation to be discrete experiences, whereas I aggregate over lives (as throughout). The Absurd Conclusion trades on difficulties that arise from capping the quantity of value that can be realized within a given period of time, whereas I adopt a 'timeless' perspective.)

3. For all bad populations X and Y and every good population Z, if X is worse than Y and $Y \cup Z$ would be bad, then $X \cup Z$ would be bad.
4. t hellish lives ∪ any natural number of barely good lives would be a bad population. (1-3)
5. For any natural number of heavenly lives, there is a natural number of barely good lives that would be better. (Exchangeability of good lives)
6. For all good populations X and Y and every bad population Z, if X is better than Y and $X \cup Z$ would be bad, then $Y \cup Z$ would be bad.
7. t hellish lives ∪ any natural number of heavenly lives would be a bad population. (4-6)

Proof: the Somber View implies that no natural number of heavenly lives along with a single hellish life would be a good population.

Assume for contradiction that there is a natural number n such that n heavenly lives ∪ one hellish life would be a good population. By Scaling, for every natural number q, qn heavenly lives ∪ q hellish lives would be a good population. But this contradicts the fact that on the Somber View, there is a natural number t such that t hellish lives ∪ any natural number of heavenly lives would be a bad population.

Proof: Heavy Tails implies that one heavenly life ∪ any number of barely bad lives would not be a bad population (the Weak Repugnant Elitist Conclusion).

We begin by showing that there is a natural number t such that t heavenly lives ∪ any number of barely bad lives would be a good population. This is simply the first half of the proof of the Repugnant Elitist Conclusion:

1. There is a natural number t such that t heavenly lives would be better than any natural number of barely good lives. (Weak Superiority of Heavenly Lives)
2. For any natural number of barely bad lives, there is a natural number of barely good lives such that the barely bad lives ∪ the barely good lives would be a good population. (Barely good lives can outweigh barely bad lives)
3. For all good populations X and Y and every bad population Z, if X is better than Y and $Y \cup Z$ is good, then $X \cup Z$ is good.
4. t heavenly lives ∪ any number of barely bad lives would be a good population. (1-3)

Now assume for contradiction that there is a natural number m such that m barely bad lives ∪ one heavenly life would be a bad population. Then, by Scaling, for every natural number q, qm barely bad lives ∪ q heavenly lives would be a bad population. But this contradicts the fact that there is a natural number t such that t heavenly lives ∪ any natural number of barely bad lives would be a good population. So there is no natural number m such that m barely bad lives ∪ one heavenly life would be a bad population.

Proof: (Revised) Heavy Tails implies that there is a natural number n such that n heavenly lives ∪ any natural number of barely bad lives would be a good population.

The Weak Inferiority of Hellish Lives means that there is a natural number m such that m hellish lives would be worse than any natural number of barely bad lives. And since heavenly lives can outweigh hellish lives, there is a natural number n such that n heavenly lives ∪ m hellish lives would be a good population. Now, for every good population X and all bad populations Y and Z, if $X \cup Y$ would be good and Y is worse than Z, then $X \cup Z$ would be good. (To see this, assume that $X \cup Y$ would be good and that Y is worse than Z. Since Y is worse than Z, by Separability, $X \cup Z$ would be better than $X \cup Y$. And since $X \cup Y$ would be better than the empty population, by Transitivity, $X \cup Z$ would be better than the empty population.) Therefore, n heavenly lives ∪ any number of barely bad lives would be good.

Using a parallel strategy, we can show that (Revised) Heavy Tails similarly implies that there is a natural number m such that m hellish lives ∪ any natural number of barely good lives would be a bad population. The proof is omitted for brevity.

References

Alexander, Scott. 2015. "How Bad Are Things?" *Slate Star Codex* (blog). https://slatestarcodex.com/2015/12/24/how-bad-are-things/.

Andersson, Henrik. 2021. "Spectrum Arguments, Indeterminacy, and Value Superiority." In *Value Incommensurability: Ethics, Risk, and Decision-Making*, ed. H. Andersson and A. Herlitz, 109-25. New York: Routledge.

Arrhenius, Gustaf. 2004. "The Paradoxes of Future Generations and Normative Theory." In *The Repugnant Conclusion: Essays on Population Ethics*, ed. T. Tännsjö and J. Ryberg, 201-18. Dordrecht: Kluwer Academic Publishers.

Arrhenius, Gustaf. 2011. "The Impossibility of a Satisfactory Population Ethics." In *Descriptive and Normative Approaches to Human Behavior Advanced Series on Mathematical Psychology*, ed. E. Dzhafarov and L. Perry, 1-26. Singapore: World Scientific Publishing Company.

Arrhenius, Gustaf. 2016. "Population Ethics and Different-Number-Based Imprecision." *Theoria*, 82 (2): 166–181.

Asheim, Geir B and Stéphane Zuber. 2014. "Escaping the Repugnant Conclusion: Rank Discounted Utilitarianism with Variable Population." *Theoretical Economics*, 9 (3): 629–650.

Broome, John. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Carlson, Erik. 1998. "Mere Addition and Two Trilemmas of Population Ethics." *Economics & Philosophy* 14: 283-306.

Carlson, Erik. 2022. "On Some Impossibility Theorems in Population Ethics." In the *Oxford Handbook of Population Ethics*, ed. G. Arrhenius, K. Bykvist, T. Campbell, and E. Finneron-Burns, 204-25. New York: Oxford University Press.

Caviola, Lucius, A. Hoskin, and J. Lewis. 2021. "Do people think their lives are worth living? A survey in India and the US." Unpublished MS.

Chang Ruth. 2016. "Parity, Imprecise Comparability, and the Repugnant Conclusion." *Theoria* 82: 183–215.

Cowie, Christopher. 2022. "A New Argument for Moral Error Theory." *Noûs* 56: 276-94.

Crisp, Roger. 2021. "Would extinction be so bad?" *New Statesman*. https://www.newstatesman.com/ideas/agora/2021/08/would-extinction-be-so-bad.

Gibbon, Edward. 1996. *The History of the Decline and Fall of the Roman Empire*. Project Gutenberg. https://www.gutenberg.org/files/25717/25717-h/25717-h.htm.

Glover, Jonathan. 2012. *Humanity: A Moral History of the Twentieth Century*. New Haven: Yale University Press.

Goodsell, Zachary. 2021. "A St Petersburg Paradox for risky welfare aggregation." *Analysis* 81 (3): 420–26.

Griffin, James. 1988. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Oxford University Press.

Gustafsson, Johan E. 2020. "Population Axiology and the Possibility of a Fourth Category of Absolute Value." *Economics & Philosophy* 36: 81-110.

Handfield, Toby and Wlodek Rabinowicz. 2018. "Incommensurability and Vagueness in Spectrum Arguments: Options for Saving the Transitivity of Betterness." *Philosophical Studies* 175: 2373–2387.

Hájek, Alan and Wlodek Rabinowicz. 2022. "Degrees of Commensurability and the Repugnant Conclusion." *Noûs* 56: 897-919.

Huemer, Michael. 2008. "In Defence of Repugnance." *Mind* 117 (468): 899-933.

Hurka, Thomas. 1983. "Value and Population Size." *Ethics* 93 (3): 496–507.

Jensen, Karsten Klint. 2020. "Weak Superiority, Imprecise Equality and the Repugnant Conclusion." *Utilitas* 32: 294–315.

Kitcher, Philip. 2000. "Parfit's Puzzle." *Noûs*, 34 (4): 550–577.

Klocksiem, Justin. 2016. “How to Accept the Transitivity of *Better Than*.” *Philosophical Studies* 173: 1309–1334.

Lemos, Noah M. 1993. “Higher Goods and the Myth of Tithonus.” *Journal of Philosophy* 90 (9): 482-96.

MacAskill, William. 2022. *What We Owe the Future*. New York: Basic Books.

McMahan, Jeff. 2013. “Causing People to Exist and Saving People’s Lives.” *Journal of Ethics* 17: 5-35.

Mogensen, Andreas. “The Weight of Suffering.” Forthcoming in the *Journal of Philosophy*.

Nebel, Jacob M. 2022. “Totalism Without Repugnance.” In *Ethics and Existence: The Legacy of Derek Parfit*, ed. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, 200-31. Oxford: Oxford University Press.

Nebel, Jacob M. 2023. “The Sum of Well-Being.” *Mind*. https://doi.org/10.1093/mind/fzad019.

Ng, Yew-Kwang. 1989. “What Should We Do About Future Generations? Impossibility of Parfit’s Theory X.” *Economics & Philosophy*, 5 (2): 235–253.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.

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

Parfit, Derek. 1986. "Overpopulation and the Quality of Life." In *Applied Ethics*, ed. P. Singer, 145–164. Oxford: Oxford University Press.

Parfit, Derek. 2016. "Can We Avoid the Repugnant Conclusion?" *Theoria* 82 (2): 110–27.

Pivato, Marcus. 2020. "Rank-additive Population Ethics." *Economic Theory*, 69 (4): 861–918.

Portmore, Douglas W. 1999. "Does the Total Principle Have Any Repugnant Implications?" *Ratio* 12 (1): 80–98.

Pummer, Theron. 2018. "Spectrum Arguments and Hypersensitivity." *Philosophical Studies* 175: 1729–1744.

Pummer, Theron. "Sorites on What Matters." In *Ethics and Existence: The Legacy of Derek Parfit*, ed. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, 498-523. Oxford: Oxford University Press.

Rabinowicz, Wlodek. 2022. “Can Parfit’s Appeal to Incommensurabilities in Value Block the Continuum Argument for the Repugnant Conclusion?” In *Ethics and Existence: The Legacy of Derek Parfit*, ed. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, 430-60. Oxford: Oxford University Press.

Rachels, Stuart. 2004. “Repugnance or Intransitivity: A Repugnant but Forced Choice.” In *The Repugnant Conclusion: Essays on Population Ethics*, ed. J. Ryberg and T. Tännsjö, 163–186. Dordrecht: Kluwer Academic.

Sider, Theodore R. 1991. “Might Theory X be a Theory of Diminishing Marginal Value?” *Analysis*, 51 (4): 265–271.

Temkin, Larry S. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. New York: Oxford University Press.

Thomas, Teruji. 2018. “Some Possibilities in Population Axiology.” *Mind* 127 (507): 807–832.

Thomas, Teruji. 2022a. “On Evaluative Imprecision.” In *Ethics and Existence: The Legacy of Derek Parfit*, ed. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, 478-97. Oxford: Oxford University Press.

Thomas, Teruji. 2022b. “Separability and Population Ethics.” In the *Oxford Handbook of Population Ethics*, ed. G. Arrhenius, K. Bykvist, T. Campbell, and E. Finneron-Burns, 271-95. New York: Oxford University Press.

Thomas, Teruji. 2022c. “Are Spectrum Arguments Defused by Vagueness?” *Australasian Journal of Philosophy* 100 (4): 743-757.

Thornley, Elliott. 2021a. “A Dilemma for Lexical and Archimedean Views in Population Axiology.” *Economics & Philosophy*, 1-21.

Thornley, Elliott. 2021b. “The Impossibility of a Satisfactory Population Prospect Axiology (Independently of Finite Fine-Grainedness).” *Philosophical Studies* 178: 3671–3695.

Zuber, Stéphane *et al*. 2021. “What Should We Agree on about the Repugnant Conclusion?” *Utilitas* 33 (4): 379-383.

Chapter two: Expected choiceworthiness and fanaticism (*Philosophical Studies*, 2024)

1. Introduction

We face moral uncertainty when the moral theories in which we have nonzero credence issue conflicting verdicts.[40] To illustrate, consider an agent whose credence is split between a nonconsequentialist theory focused on justice and a consequentialist theory focused on welfare. Suppose that in a given decision-situation, the former requires her to rectify injustices in her community, whereas the latter requires her to mitigate existential risks to humanity. Since she is uncertain which moral theory is true, she is uncertain whether she morally ought to rectify injustice or mitigate existential risk (assuming she can't do both).

Maximize Expected Choiceworthiness (MEC) is the most prominent theory of how we ought to make decisions when we face moral uncertainty. I will focus on the version of MEC developed by MacAskill, Bykvist, and Ord (2020) because it is the most recent and systematic formulation of the view.[41] The central thrust of MEC is that, when we face moral uncertainty, we ought to take an option that maximizes expected *choiceworthiness*, where choiceworthiness is a representation

[40] Moral uncertainty is a proper subset of *normative uncertainty*. We can also be normatively uncertain about epistemic rationality, prudence, etc.

[41] See also MacAskill and Ord (2020). For other iterations of MEC, see Lockhart (2000), Ross (2006), Sepielli (2009), and Wedgwood (2013). Carr (2020) proposes a variant of MEC that does not depend on *intertheoretic value comparisons*, which are explained presently in the main text. Riedener (2020) argues that we should handle axiological uncertainty in the way that MEC handles deontic uncertainty.

of the strength of moral reason to choose an option, according to a moral theory.[42] To indicate how MEC works, consider an agent who has one third of her credence in Kantian deontology and two thirds in classical utilitarianism.[43] The agent finds herself in a situation in which it is necessary to lie to avoid some trivial social awkwardness. Suppose that Kantian deontology says to tell the truth, but that classical utilitarianism says to lie. What to do? Notice—the expected choiceworthiness maximizer tells us—that in this case, the moral stakes are intuitively much higher according to Kant than they are according to Bentham. That is, it's much more important to tell the truth—i.e., to respect the categorical imperative—if Kantian deontology is true than it is to lie—i.e., to promote *slightly* more pleasure—if classical utilitarianism is true. Or, in the language of MEC: the choiceworthiness of truth-telling, conditional on Kantian deontology, is significantly greater than the choiceworthiness of lying, conditional on classical utilitarianism (again, in this particular case). So, even though the agent has most credence in classical utilitarianism, she ought in this case to adhere to Kantian deontology, because doing so maximizes expected choiceworthiness.

[42] MacAskill, Bykvist, and Ord remain neutral on whether the ought of moral uncertainty is "moral (second-order), rational, virtue ethical or something else" (2020: 30). In contrast, MacAskill and Ord say that it's a rational ought and suggest that this is the "established view in the literature" (2020: 350, fn. 11). I follow MacAskill, Bykvist, and Ord in remaining neutral, though I flag here that the nature of the ought of moral uncertainty will become relevant in §3.

[43] MacAskill, Bykvist, and Ord remain neutral on whether the relevant probabilities are the agent's actual credences (i.e., their subjective probabilities) or their epistemic credences (i.e., the credences that they should have, given their evidence) (2020: 4). I follow them in remaining neutral, though I note that they later appeal to credences that one "should have," indicating that they have epistemic credences in mind, at least at that stage in their argument (2020: 152).

One question that immediately arises for MEC is how we can compare 'strength of reason *x* to Φ according to Kantian deontology' to 'strength of reason *y* to Ψ according to classical utilitarianism'—which we must if we are to take an expectation of choiceworthiness. This is an instance of the problem of *intertheoretic value comparisons*. To solve the problem, MacAskill, Bykvist, and Ord propose what they call a *universal scale account*, on which many moral theories share a common, *theory-neutral* scale of choiceworthiness.[44] In their words, "intertheoretic [value] comparisons are true in virtue of the fact that there is some independent choice-worthiness scale that is the same across different theories."[45] For now, we will grant this account *arguendo*.

By design, MEC handles moral uncertainty in the way that Expected Utility Theory (EUT) handles descriptive uncertainty.[46] But there are (at least) two distinct views that answer to the name 'Expected Utility Theory', so we need to disambiguate. According to the view that I'll call 'Expected Value Theory', when we're pursuing some quantifiable *target value*, such as *dollars* or *lives saved*, we rationally ought to choose an option that maximizes the expected quantity of our target value.[47] For example, if we're pursuing US dollars, we rationally ought to choose an option that maximizes expected US dollars. So, if offered a choice between $49 for sure and a fair coin toss that pays out $100 on heads and $0 on tails, we rationally ought to take the coin toss. In contrast, according to the view that I'll call 'orthodox EUT', we rationally ought to choose an option that maximizes expected *utility* (see e.g. von Neumann and Morgenstern (1947)). What's utility? Wilkinson (2023) provides a lucid gloss:

[44] MacAskill, Bykvist, and Ord (2020: 133, 145).

[45] MacAskill, Bykvist, and Ord (2020: 133).

[46] MacAskill, Bykvist, and Ord (2020: 47-48).

[47] Here I follow Wilkinson (2023: 626-28) in distinguishing Expected Value Theory from orthodox EUT.

> here I mean *utility* in a decision-theoretic sense, meaning something quite different from (moral) *value*—it also represents the agent's attitude to risk. Depending on how strongly the agent prefers obtaining good outcomes with high probability, the utility of an outcome can be *any* increasing function of its value—one outcome will still have greater value than another if and only if it has greater utility, but it can vary in how much greater its value is.[48]

If an agent is risk-averse, orthodox EUT can permit (indeed, require) her to choose the sure $49 over the coin toss—unlike Expected Value Theory. More on orthodox EUT and risk to come in §3.

As I understand it, MEC in its original formulation is Expected Value Theory applied to the target value of choiceworthiness.[49] Each moral theory with which MEC works comes with its own pre-existing choiceworthiness function. Moreover, the universal scale assumption means that each such moral theory measures choiceworthiness on the same scale. That's what allows us to construct decision tables in which moral theories serve as our states of nature and to take meaningful expectations of choiceworthiness. Thus, the expected choiceworthiness maximizer might represent the 'truth vs. lie' example we just explored as follows in Table 4. (The specific numbers in the table don't matter; they're just for illustration.)

[48] Wilkinson (2023: 627-28). I have omitted a footnote from the quoted text.

[49] However, MacAskill, Bykvist, and Ord (2020: 48, fn. 15) express their openness to decision theories that depart from risk neutrality. I therefore explore in §3 two leading theories that do so.

Table 4: Truth vs. lie

	Kantian deontology	Classical utilitarianism
Tell the truth	10	1
Lie	-100	2

Since the agent has one third of her credence in Kantian deontology and two thirds in classical utilitarianism, the expected choiceworthiness of telling the truth = (1/3)(10) + (2/3)(1) = 4; and the expected choiceworthiness of lying = (1/3)(-100) + (2/3)(2) = -32. That's why, according to MEC, the agent ought to tell the truth, even though she has credence = 2/3 that she ought to lie.

MEC inherits from Expected Value Theory the problem of *finite fanaticism*.[50] Finite fanaticism arises due to the fact that on Expected Value Theory, tiny probabilities of astronomical—though finite—quantities of value correspond to large quantities of *expected* value. Expected Value Theory is thus fanatical in that it can require agents to enter lotteries in which (significant) loss is all but certain, but the gain, were it to obtain, would be enormous. The classic exposition of finite fanaticism is Pascal's Mugging.[51] Here, a mugger tells an expected value maximizer that if she hands over her wallet, he will generate an astronomical quantity of value for her. Since the expected value maximizer assigns nonzero probability to the mugger's veracity, Expected Value

[50] On which see Wilkinson (2022), Beckstead and Thomas (2023), and Russell (2023).

[51] Due to Bostrom (2009).

Theory seems to absurdly demand that she hand over her wallet. The analogue of Pascal's mugger in cases of moral uncertainty is a theory *T* which claims that the moral stakes are extremely high in a given decision-situation—or, perhaps, in most or even all decision-situations. If the moral stakes according to *T* are sufficiently higher than they are according to the other moral theories in which the expected choiceworthiness maximizer has credence, then MEC will require her to act in accordance with *T*, even when her credence in *T* is very low indeed.

Allowing for *infinite payoffs* makes the fanaticism problem even worse. In fact, it threatens a new problem: *infinitarian paralysis*.[52] Here's how: for any option *O*, many agents will have *some* nonzero credence in a moral theory on which *O* has infinitely positive choiceworthiness and *some* nonzero credence in a moral theory on which *O* has infinitely negative choiceworthiness. For such agents, it seems to follow that the expected choiceworthiness of every option is undefined, in which case, MEC has nothing to say about which option to choose.[53] MacAskill, Bykvist, and Ord acknowledge that left unaddressed, this problem "simply *breaks* MEC" (2020: 152). One way to avoid paralysis would be to claim that we ought to choose the option that maximizes (p(infinitely positive payoff) - p(infinitely negative payoff)) (where 'p' denotes probability)—intuitively, the option that 'gives us the best shot at infinity'.[54] But adopting this decision rule would mean

[52] The phrase 'infinitarian paralysis' is due to Bostrom (2011).

[53] Cf. Hájek's (2003) insight, in response to Pascal's Wager, that by Pascal's lights, every option seemingly has infinite expected utility, because every option is associated with some nonzero probability of resulting in theistic belief.

[54] For discussion and more precise formulation of approaches in this vein, see Slesinger (1994), Hájek and Nover (2004), and Bostrom (2011: 35-36). Although this patch succeeds at blocking paralysis, it remains implausible. Imagine that you must choose exactly one of three prospects. The first and second offer some tiny probability of ∞, an even tinier probability of -∞, and a near certainty of an enormous, though finite,

allowing our moral decision-making to be dominated by whichever infinite-stakes moral theory happens to afford us such options. Even if this rule somehow generated reasonable practical verdicts—which I doubt it would—it would do so for the intuitively wrong reason.

Does the paralysis problem arise simply due to an inapposite representation of infinite value? Chen and Rubio (2020) propose that to deal with infinite payoffs, we ditch cardinal arithmetic—on which, for any finite real number x, $\infty \pm x = \infty$; $(x)\infty = \infty$; and $\infty - \infty$ is undefined—in favor of surreal arithmetic. In surreal arithmetic, we can precisify an infinite payoff using the ordinal ω, which does not have the three properties of ∞ just listed: where x is again a positive finite real, $(\omega + x) > \omega$; where $0 < y < 1$ and y is non-infinitesimal, $(y)\omega < \omega$; and $\omega - \omega = 0$. Although fascinating, this move will not be of much help to MEC. For it remains that $(y)\omega > x$—i.e., that the product of any non-infinitesimal real number between zero and one and ω is greater than any finite positive real number. Infinitarian *paralysis* is blocked, but the door is left wide open for swamping by infinite-stakes moral theories. We are right back to fanaticism (Beckstead and Thomas (2023) call this form of fanaticism "infinity obsession"). Similar remarks apply to alternative representations of infinite payoffs, including lexicographic vector utilities (see Hájek (2003: §4) for discussion).

The lesson is that the viability of MEC depends on a satisfactory solution to the problems of finite fanaticism and infinity. MacAskill, Bykvist, and Ord offer the following response:

---

quantity of extreme suffering. The third is a certainty of an astronomical, though finite, quantity of the *summum bonum* (whatever it is). The decision rule we're considering will require you to choose one of the first two prospects—whichever has the greater $(p(\infty) - p(-\infty))$—because it doesn't care about finite prospects when infinite prospects are on the table. This is intuitively intolerable.

> whatever is the best solution to the fanaticism problem under empirical uncertainty is likely to be the best solution to the fanaticism problem under moral uncertainty. This means that this issue is not a distinctive problem for moral uncertainty (2020: 153).

(Here 'the fanaticism problem' seems to refer to infinitarian paralysis, but it is clear that the authors mean to make the same move with respect to finite fanaticism as well.) A distinct response can be located in their *statistical normalization* proposal, which is introduced below in section 5. So far, both responses have gone unscrutinized. The central argument of this paper is that the first response—import solutions to finite fanaticism and the problems of infinity from decision theory—is open to serious doubt and that the second response—statistical normalization—is unsatisfactory. Sections 2 - 4 address the first response. They respectively consider bounded and finite value functions, departures from risk neutrality, and Knowledge-First Decision Theory, each of which can be leveraged to address finite fanaticism and the problems of infinity in the familiar context of decision-making under *descriptive* uncertainty. The argument of these sections will be that each strategy fails as a (straightforward) solution to these problems in the context of decision-making under *moral* uncertainty. Section 5 argues that statistical normalization avoids fanaticism and the problems of infinity at the unacceptable cost of rendering MEC *stakes-insensitive*, in a sense to be explained. Section 6 concludes by considering the implications of the foregoing for the wider debate about decision-making under moral uncertainty.

2. Bounded and finite value functions

Some decision theorists argue that our individual utility functions should be *bounded*.[55] A utility function U is bounded just in case there are natural numbers m and n such that for every option O in the domain of U, $m \leq U(O) \leq n$. An expected utility maximizer whose utility function is bounded won't struggle with infinitarian paralysis or infinity obsession, because no option will have infinite (dis)utility for her. Moreover, if the upper bound on utility is sufficiently low, she will avoid exploitation in finite fanaticism cases like Pascal's Mugging, for her utility function implies that the stakes can only ever be so high—for her, astronomical stakes do not exist.

Can the expected *choiceworthiness* maximizer similarly avoid finite fanaticism and the problems of infinity by claiming that the choiceworthiness function of every live moral theory is bounded?[56] No, because it's not the case that the choiceworthiness function of every live moral theory is bounded. Consider *total welfarist decision-theoretic consequentialism*, for instance. According to this moral theory, the choiceworthiness of an option equals its expected agent-neutral moral value, where agent-neutral moral value is understood in total welfarist terms. The total welfarist value function is unbounded: the world gets better and better—in a non-diminishing manner—as we add moral patients with positive welfare to it. Likewise, the world gets worse and worse in a non-diminishing manner as we add moral patients with negative welfare to it. Indeed, the total welfarist value function isn't even *finite*. A finite value function assigns everything in its domain a finite

[55] See McGee (1999) and Russell and Isaacs (2021).

[56] By a 'live' moral theory I just mean one that gets included in the expected choiceworthiness calculation. I consider the proposal that we can exclude certain moral theories from the calculation on the basis of knowledge in §4.

value.[57] But the total welfarist value function says that a world with infinitely many moral patients at a positive welfare level (and no other moral patients) is infinitely valuable (assuming that the sum of welfare doesn't converge to a finite value). Since total welfarist decision-theoretic consequentialism says that the choiceworthiness of an option equals its expected agent-neutral moral value, and the moral value function is neither bounded nor finite, the choiceworthiness function is neither bounded nor finite.

In case total welfarist decision-theoretic consequentialism strikes you as so implausible as to fail to be a live option, consider Bradley Monton's point that to insist that moral choiceworthiness functions must be bounded would be

> ethically problematic because one can always add more agents into the [choiceworthiness] calculation—one can always consider the possibility of more sweet little orphans…who can benefit from one's decisions…the [well-being] of those added sweet little orphans matter[s] (2019: 5) just as much as that of the already-considered sweet little orphans (2019: 5, fn. 8).[58]

Monton's position strikes me as both coherent and correct: the moral importance of a person's well-being is independent of the existence and well-being of other people.[59]

[57] Note that a value function can be finite but unbounded.

[58] Cf. Beckstead and Thomas (2023: §3.2).

[59] See Thomas (2022) for arguments in favor of this view, called *Separability*, and for explanation of the close connection between Separability and total welfarist consequentialism. However, see Goodsell (2021) for an objection to a related principle (Anteriority) that draws on the St. Petersburg paradox.

Of course, we could consider the weaker consequentialist thesis that the choiceworthiness of an option is greater than that of another iff and because it has greater expected agent-neutral moral value. This would leave open the exact formulation of the choiceworthiness function, which could (e.g.) be increasing with respect to expected agent-neutral moral value but asymptotically approach an upper bound. But considering this alternative means that we simply have rival consequentialist theories on our hands—one being the theory with the infinite choiceworthiness function stated above, and another being the theory with the bounded choiceworthiness function just stated. That's consistent with the key claim on the table, namely that it's not the case that the choiceworthiness function of every live moral theory is bounded. It also highlights the principal issue with the appeal to boundedness (or finitude): it doesn't work if the expected choiceworthiness maximizer has nonzero credence in any moral theory with an unbounded (or infinite) choiceworthiness function.[60]

One might respond by arguing as follows:

(1) The choiceworthiness function of every live moral theory is bounded iff the utility function of every rational individual is bounded.
(2) The utility function of every rational individual is bounded.[61]
(3) So, the choiceworthiness function of every live moral theory is bounded.

[60] Cf. Beckstead and Thomas (2023: §6).

[61] Again, see McGee (1999) and Russell and Isaacs (2021) for arguments to this effect.

Why think that (1) is true? One possibility is that for a moral theory to have a choiceworthiness function *C just is* for the theory to claim that it's fitting for an agent to adopt *C* as her own utility function. If so, and if (2) is true as well, then we have a way to save MEC from both finite fanaticism and the problems of infinity. For the upshot will be that there aren't any live infinite-stakes moral theories and, if the upper bound on the individual utility function is sufficiently low, that there aren't any live astronomical-finite-stakes moral theories either.

I am uncertain whether (1) is true but grant it *arguendo*. Here are two responses to the argument. First, as I've argued in this section, an agent with *bona fide* moral uncertainty will be unable—by her own lights—to rule out *every* moral theory with an unbounded choiceworthiness function, such as total welfarist decision-theoretic consequentialism. So, if she's committed to (1), she won't be in a position to assert (2). Second, it seems rationally permissible to value certain goods, such as happy days of life, linearly (i.e., at constant marginal utility) and to pursue them in a risk-neutral manner. But an agent who values a good linearly and who is risk-neutral in her pursuit of that good has an unbounded utility function. So, it seems rationally permissible to have an unbounded utility function.[62] We therefore have reason to deny (2) that is independent of the considerations in favor of unbounded moral choiceworthiness functions. Of course, this doesn't settle the dispute about the rationality of unbounded individual utility functions. Instead, it shows that we cannot be certain of (2). And if we are uncertain about (2) but committed to (1), we will be unable to rule out unbounded moral choiceworthiness functions. That's bad news if we're expected

[62] Hájek (2012: 422-23), Buchak (2013: 73), Smith (2014: 496-97), and Cibinel (2023) each make a version of this point. See also Wilkinson (2022: 460, fn. 45) for a distinct but related objection to bounded utility functions.

choiceworthiness maximizers, because it means that we'll be vulnerable to swamping by fanatical moral theories.

3. Departures from risk neutrality

Can we avoid finite fanaticism and the problems of infinity by pursuing choiceworthiness in a manner that departs from risk neutrality? In this section, I'll consider two leading models of risk aversion and risk seekingness: that found in Lara Buchak's (2013) Risk-Weighted Expected Utility theory (REU) and that found in orthodox EUT.[63] My treatment of REU will be relatively brief, for REU has serious difficulty with finite fanaticism and the problems of infinity even in the familiar context of decision-making under descriptive uncertainty. I'll spend more time with the orthodox EUT treatment of risk, since—as I'll argue—appealing to it encounters a unique difficulty in the context of decision-making under moral uncertainty.

3.1 Risk-Weighted Expected Utility theory

According to REU, it's not just our credence and utility functions that determine what we rationally ought to do (as orthodox EUT would have it). Instead, a third function is also relevant: our *risk function*, which reflects our risk attitude—i.e., whether we are risk-neutral or, if not, the extent to which we are risk-averse or risk-seeking. For risk-averse agents, the risk function dampens the

[63] In this section, for simplicity, I will focus on risk aversion in the pursuit of positive choiceworthiness. However, see footnote 73 for discussion of the role of risk seekingness *vis-à-vis* negative choiceworthiness.

probability function.[64] However, the risk function either does or does not discount sufficiently small probabilities to zero.[65] If it does not, then 'Maximize Risk-Weighted Expected Choiceworthiness' remains stuck with finite fanaticism. For no matter how low the risk function drives the probability of a high-stakes moral theory *T*, if it doesn't drive p(*T*) to zero, then, as long as the stakes according to *T* are sufficiently high relative to the stakes according to the competing moral theories, *T* will continue to swamp the risk-weighted expected choiceworthiness calculation.[66] Similarly, if the risk function doesn't drive the probability of an infinite-stakes moral theory to zero, then the infinite-stakes theory will still generate infinite risk-weighted expected choiceworthiness. But if the risk function does discount sufficiently small probabilities to zero, then it will encounter a host of objections that have recently been leveled against all such discounting strategies, including: that it treats intuitively terrible lotteries as certainties of small wins (e.g., a lottery offering a tiny probability of doom and a near-certainty of a penny);[67] that it either violates dominance reasoning or becomes vulnerable to money pumps;[68] and that it either fails to avoid fanaticism after all or generates intransitive cycles.[69] Again, to be clear, this dilemma

[64] See Buchak (2013: 49-50). An example Buchak uses to illustrate risk aversion is the risk function $r(p) = p^2$. Consider the gamble *G* in which a fair coin is tossed. If the coin lands heads, you get 10 utils; if it lands tails, you get nothing. The risk function of a risk-neutral agent is $r(p) = p$, so for such an agent, REU(*G*) = 5 utils. In contrast, REU(*G*) for the risk-averse agent is $(0.5)^2(10 \text{ utils}) + 0 = 2.5$ utils.

[65] Cf. Beckstead and Thomas (2023: §2.3) on *tail discounting*. To my knowledge, the objections from Isaacs (2016), Kosonen (2022), and Cibinel (2023) cited presently apply to tail discounting as well.

[66] Buchak (2013: 73-74) acknowledges this *mutatis mutandis* (in discussing individual utility, rather than moral choiceworthiness) in her discussion of the St. Petersburg paradox.

[67] Isaacs (2016).

[68] Kosonen (2022: chapter 4).

[69] Cibinel (2023).

isn't specific to moral uncertainty; it's a general problem for REU. I'll therefore move on to orthodox EUT.

3.2 Orthodox Expected Utility Theory

We can avoid finite fanaticism and the problems of infinity by abandoning expected choiceworthiness maximization in favor of expected utility maximization. Here's one way to implement this strategy. Our utility function, U, takes choiceworthiness as its sole input and outputs utility. We take U to be bounded, strictly convex over negative choiceworthiness, and strictly concave over positive choiceworthiness, so that U is increasing also. Concavity over positive choiceworthiness means that positive choiceworthiness has diminishing marginal utility. U has this shape not because we *care* less about additional positive choiceworthiness, but because we're risk-averse in our pursuit of positive choiceworthiness and we model risk aversion in the way that orthodox EUT says we should (as opposed to the way that REU says we should, for instance).[70] Call this proposal *Maximize Expected Utility* (MEU). Although MacAskill, Bykvist, and Ord appear to reject proposals in this vein, MEU is in my view a plausible response to finite fanaticism and the problems of infinity, so I'll consider it in some detail.[71]

[70] See Buchak (2013: chapter 1) for a critique of the way in which orthodox EUT models risk aversion. Notice, though, that although Buchak herself worries that "bounding the utility function seems *ad hoc*" (2013: 73), bounded utility is compatible with REU. One might therefore consider a risk-weighted iteration of Maximize Expected Utility (MEU), introduced presently in the main text; but the two objections to MEU given below will apply to any risk-weighted iteration as well.

[71] MacAskill, Bykvist, and Ord (2020: 153, fn. 7) express skepticism about bounded value functions in general, citing problems highlighted in Beckstead and Thomas (2023)—on which more below in the main text.

To see MEU in action, consider Table 5.

Table 5: Barebones fanaticism (choiceworthiness)

	Moderate moral theory	Fanatical moral theory
Φ	10	5
Ψ	5	1,000,000

As above, the numbers in the decision table represent choiceworthiness (on MacAskill, Bykvist, and Ord's universal scale). Suppose that an agent has credence = 0.99999 in the moderate moral theory and credence = 0.00001 in the fanatical moral theory. Then the expected choiceworthiness of Φ-ing ≈ 10 and the expected choiceworthiness of Ψ-ing ≈ 15. So, MEC says that the agent ought to Ψ—even though she has credence = 0.99999 that she ought to Φ. That's a fanatical verdict.

In contrast to MEC, MEU says that we should first feed the choiceworthiness scores into a utility function U and then maximize expected utility. Consider, for example, the function $U(x) = 100 \cdot \tanh(x/100)$ (where 'tanh' refers to the hyperbolic tangent function). $U(x)$ approaches -100 as x approaches $-\infty$ and 100 as x approaches ∞. Feeding the choiceworthiness scores from Table 5 into U and rounding to two decimals, we get Table 6.

Table 6: Barebones fanaticism (utility)

	Moderate moral theory	Fanatical moral theory
Φ	9.97	5
Ψ	5	100

Using the same credences as before, EU(Φ-ing) ≈ 9.97; EU(Ψ-ing) ≈ 5. So, MEU says that the agent ought to Φ, which is the desired (i.e., non-fanatical) verdict. This is a neat solution, and it works regardless of whether the moral theories that threaten to 'mug' us are astronomical-finite-stakes theories or infinite-stakes theories. In the remainder of this section, I'll give two objections to this proposal. The first objection applies (*mutatis mutandis*) to bounded utility functions in general. The second is specific to decision-making under moral uncertainty.

First objection: the boundedness of the utility function generates counterintuitive implications of its own.[72] Consider Table 7.

[72] Beckstead and Thomas (2023: §3) make this point in the context of decision-making under descriptive uncertainty.

Table 7: High moral stakes (choiceworthiness)

	Moral theory 1	Moral theory 2
Φ	100,000	0
Ψ	0	10,000

The decision-situation depicted in Table 7 is high stakes according to each moral theory, but it's significantly higher stakes according to moral theory 1. Suppose that an agent's credence in moral theory 1 = 0.49 and that her credence in moral theory 2 = 0.51. Intuitively, she ought to Φ. However, if we feed the choiceworthiness scores into our utility function, we get Table 8 (rounding to two decimals).

Table 8: High moral stakes (utility)

	Moral theory 1	Moral theory 2
Φ	100	0
Ψ	0	100

$EU(\Phi) \approx 49$; $EU(\Psi) \approx 51$. Thus, MEU gives the intuitively incorrect verdict that the agent ought to Ψ. Of course, we could reverse this particular verdict by choosing a different utility function. But doing so would not solve the general problem that this case serves to highlight, namely that

the boundedness of the utility function can effectively obscure massive differences in moral stakes when we are close to the upper bound on decision-theoretic utility.

The second objection stems from the observation that an agent who takes moral uncertainty seriously should—by her own lights—be uncertain whether (i) risk aversion is appropriate in the pursuit of choiceworthiness to begin with and (ii) if it is, whether a degree of risk aversion sufficient to avoid finite fanaticism and the problems of infinity is similarly appropriate.[73] If an agent is uncertain about (i), she'll be uncertain whether MEC or MEU is true. And if she's uncertain about (ii), she'll be uncertain about the value of the upper bound on decision-theoretic utility.[74] In what follows, I'll focus on uncertainty about (i) for presentational simplicity; but a full discussion would incorporate uncertainty about (ii) as well. To be clear before proceeding: what is at issue here is not uncertainty about further first-order moral hypotheses. It is higher-order

[73] What sort of uncertainty is in play here will depend on what sort of ought the agent takes the ought of moral uncertainty to be. If she takes it to be a second-order moral ought, then her uncertainty will be second-order moral uncertainty. If she takes it to be an ought of instrumental rationality—i.e., the sort of rationality with which decision theory is concerned—then her uncertainty will concern the rationality of risk aversion. And the very same considerations that MacAskill, Bykvist, and Ord adduce in favor of taking first-order moral uncertainty seriously are equally strong considerations in favor of taking second-order moral uncertainty and decision-theoretic uncertainty seriously; see MacAskill, Bykvist, and Ord (2020: 11-14). Note also that whereas orthodox-EUT-style risk *aversion* allows us to avoid fanaticism in the pursuit of *positive* choiceworthiness, it's risk *seekingness* that allows us to avoid the corresponding problem in the context of *negative* choiceworthiness. I gloss over this complication in the main text for presentational simplicity; see Beckstead and Thomas (2023: §2.2 and §3.3) for discussion.

[74] MacAskill, Bykvist, and Ord acknowledge that we can be uncertain which theory of moral uncertainty is true (2020: 30-33). Moreover, they "do not want to deny that there might be a need for a theory that can deal with *higher-order* uncertainty" (2020: 31). One might get off the boat here—I introduce a hard externalist response below.

uncertainty; specifically, it is uncertainty about what is the criterion of appropriateness for choice under first-order moral uncertainty.

To appreciate what uncertainty whether risk aversion is appropriate in the pursuit of choiceworthiness involves, return to the barebones fanaticism case from Table 5, reproduced here for ease of reading.

Table 5: Barebones fanaticism (choiceworthiness)

	Moderate moral theory	Fanatical moral theory
Φ	10	5
Ψ	5	1,000,000

As we saw, MEC and MEU disagree about which option the agent ought to choose. If the agent is uncertain whether MEC or MEU is true, she will consequently be uncertain which option she ought to choose. We can represent her uncertainty in a new decision table. To do so, we'll need to introduce a new term. Let the *meta-choiceworthiness* of an option be a representation of the strength of normative reason to choose the option, according to a theory of decision-making under moral uncertainty. Since MEC and MEU disagree about the choice between Φ and Ψ, they assign Φ and Ψ different meta-choiceworthiness scores. Considering the natural interpretations of MEC and MEU according to which the meta-choiceworthiness of an option equals its expected choiceworthiness and its expected utility (respectively), we get Table 9.

Table 9: barebones fanaticism (meta-choiceworthiness)[75]

	MEC	MEU
Φ	10	9.97
Ψ	15	5

What should the agent do? We don't know, because we don't (yet) have a theory for adjudicating this type of normative uncertainty.[76] However, notice that the MEU proposal we are considering tacitly endorses the following answer: ignore MEC and every iteration of MEU that fails to avoid fanaticism due to having an insufficiently low upper bound on utility; choose the option that maximizes expected utility relative to some sufficiently risk averse utility function. *Perhaps* this is the way to go; but at the very least, it needs to be argued for.

Here is a worry about any such argument. Notice that as the fanatical first-order moral theory becomes increasingly fanatical—i.e., as the choiceworthiness the theory assigns to its favored option grows—the meta-choiceworthiness of this option according to MEC will grow correspondingly, whereas it will barely grow at all according to the iteration of MEU to which we must appeal (due to the boundedness of the utility function). At the limit, an infinite-stakes first-

[75] Of course, this isn't the only possible precisification of MEC. But see the discussion of distinct precisifications of total welfarist consequentialism in §2 of the main text.

[76] Again, what type of uncertainty this is will depend on what type of ought the agent takes the ought of moral uncertainty to be. See footnote 73 for further detail.

order moral theory will generate infinite meta-choiceworthiness according to MEC. To illustrate, consider Table 10.

Table 10: Infinite-stakes fanaticism (choiceworthiness)

	Moderate moral theory	Infinite stakes moral theory
Φ	10	5
Ψ	5	∞

Suppose that the agent's credences are the same as in Table 5: she has credence = 0.99999 in the moderate moral theory and credence = 0.00001 in the infinite-stakes moral theory. Then, as in Table 5, the expected choiceworthiness of Φ-ing ≈ 10, but the expected choiceworthiness of Ψ-ing = ∞. In contrast, since $U(x)$ approaches 100 as x approaches infinity, the expected utility of Ψ-ing remains practically unchanged at ≈ 5 (rounding to two decimals).[77] We then get Table 11.

[77] Strictly speaking, to take this expectation, we must extend U to include ±∞ in its domain via completion. To do so, we define $U(\pm\infty)$ to be the limiting value of $U(x)$ as x approaches ±∞, namely ±100.

Table 11: Infinite-stakes fanaticism (meta-choiceworthiness)

	MEC	MEU
Φ	10	9.97
Ψ	∞	5

Again, since we don't (yet) have a theory of how to adjudicate this type of normative uncertainty, nothing *practical* follows from Table 11. What does follow is the observation that if we want to conclude in favor of MEU (i.e., to conclude that we ought to Φ), we must do so despite the fact that the decision-situation appears to be significantly—indeed, infinitely—higher stakes according to MEC.[78] Whether going this route would be *internally* well-motivated for the proponent of MEU is an open question that I will mostly leave to future research, but to which we will return briefly in the conclusion. (To preview: one possibility is to adopt a *hard externalist* view of decision theory, on which there are no higher-order norms governing choice under decision-theoretic uncertainty.)

[78] Cf. MacAskill *et al*. (2021), who argue that in certain cases where (i) causal decision theory (CDT) and evidential decision theory (EDT) issue conflicting verdicts and (ii) the stakes are intuitively much higher according to EDT than they are according to CDT, one ought to act in accordance with EDT, even if one's credence in CDT is significantly greater. Taking a similar approach to the decision problem in Table 11 will naturally militate against choosing in accordance with MEU, modulo worries about intertheoretic comparisons between MEC and MEU.

4. Knowledge-First Decision Theory

In the preceding two sections, I've relied on the thought that things go badly for MEC as long as the expected choiceworthiness maximizer has any nonzero credence in a sufficiently high-stakes moral theory. Perhaps this has been a mistake. According to *Knowledge-First Decision Theory* (KFDT), we can exclude from our decision matrices states of nature that we know do not obtain—even if we have some nonzero credence in them.[79] Plausibly—the KFDT thought goes—we know that Pascal's mugger does not possess godlike powers, so we should keep our wallet. For this knowledge-first maneuver to work for MEC, though, we would need to know that every high-stakes moral theory that threatens to mug us is false. Yet even if we know that Pascal's mugger is a fraud—and more generally, that outlandish descriptive hypotheses are false—we don't know that every high-stakes moral theory is false.

Here are three examples of high-stakes moral theories that are true, for all we know. Firstly, consider *absolutist nonconsequentialist* moral theories. According to such theories, certain action-types, such as murder, are categorically forbidden—full stop. As MacAskill, Bykvist, and Ord acknowledge, it's natural enough to interpret absolutist theories as assigning infinitely negative choiceworthiness to the categorically forbidden action-types.[80] *A fortiori*, it's natural enough to interpret them as assigning extremely low finite choiceworthiness to the categorically forbidden options. I concur with MacAskill, Bykvist, and Ord that it would be epistemically overconfident to have zero or infinitesimal credence in absolutist theories, "despite the testimony of, for example,

[79] See e.g. Hawthorne and Stanley (2008), Weatherson (2012), Liu (2022), and Hong (fc.).

[80] MacAskill, Bykvist, and Ord (2020: 150-151). Cf. Jackson and Smith (2006).

Kant and Anscombe"; and I believe the same assessment of overconfidence applies to those who claim to know that every absolutist theory is false.

Secondly, consider *longtermism*. According to longtermism (roughly), many of our present-day choices are extremely high stakes because their consequences "ripple down the millennia," affecting countless people in the future—all of whom matter morally.[81] Proponents of longtermism argue that, precisely in light of these high stakes, longtermist considerations ought to govern our moral decision-making in a variety of real-life cases (at least when there are no important nonconsequentialist side-constraints in play).[82] Arguably, longtermism follows from total welfarist decision-theoretic consequentialism.[83] I assume that we don't know that total welfarist decision-theoretic consequentialism is false (or that longtermism *doesn't* follow from it); so, we don't know that longtermism is false.

Finally, consider *amplifications* of moral theories. Roughly, an amplification of a moral theory *T* is a distinct moral theory that matches *T*'s extensional verdicts and shares *T*'s explanatory commitments, but which claims that the moral stakes are higher than they are on *T*.[84] Here's how MacAskill, Bykvist, and Ord explain theory amplification:

[81] The quotation is from Greaves (2016: 313).

[82] See Greaves and MacAskill (2021: §9).

[83] On longtermism, see Bostrom (2003), Beckstead (2013), Ord (2020), MacAskill (2022), and especially Greaves and MacAskill (2021). For skepticism that longtermism follows from (total welfarist decision-theoretic) consequentialism, see Mogensen (2021) and Thorstad (2023).

[84] This glosses over some subtleties for the sake of brevity. See MacAskill, Bykvist, and Ord (2020: 125-31 and 147-48) for further detail on theory amplification.

> Because we endorse a universal scale account, we believe that, for any theory T_1 and for any real number k, we can make sense of another theory T_2 whose choice-worthiness function is k times that of theory T_1. That is: every possible amplification of T_1 is itself a distinct theory.[85]

Let a *fanatical amplification* of a moral theory be an amplification that is sufficiently large to qualify as fanatical in the sense intended throughout. We don't know that every fanatical amplification of every moral theory is false. Why not? Firstly, when we attempt to adjudicate between competing moral theories, our most important method is to assess the relative plausibility of the theories' extensional verdicts and explanatory commitments. But a theory and its amplifications are identical in both of these respects. Moreover, a theory and its amplifications will fare about as well as one another with respect to the main abductive virtues, such as parsimony/simplicity, explanatory and predictive power, generality, and coherence with our background knowledge. (A theory's amplifications don't posit additional theoretical entities. And because their extensional verdicts and explanatory commitments are identical to those of the non-amplified theory, there's no difference in explanatory power, predictive power, or generality. Finally, it's difficult to see how a moral theory could cohere better with our background knowledge than its amplifications—unless we take it as part of our background knowledge that moral theories just aren't high stakes, which strikes me as *ad hoc* (more on this presently).)[86] Since we can't

[85] MacAskill, Bykvist, and Ord (2020: 147-48).

[86] The similarity in abductive status between a moral theory and its amplifications marks an important disanalogy with outlandish descriptive hypotheses, such as the hypothesis that Pascal's mugger is telling the truth. Very often (if not always), outlandish descriptive hypotheses fare significantly worse on

adjudicate between a moral theory and its amplifications on extensional, explanatory, or abductive grounds, the chief way to adjudicate between them is on the basis of raw intuition. But we can't rule out every fanatical amplification of every moral theory on the basis of raw intuition. Perhaps (for example) if some form of contractualism is true, it's just *incredibly* morally important to act in a way that is justifiable to each, for reasons of respect.

Here's a second argument that we don't know that every fanatical amplification of every moral theory is false.[87] Consider a non-fanatical moral theory T and suppose that an amplification of T is fanatical just in case it is produced by scaling the choiceworthiness function of T by a real number at least as great as k. Do we know the falsity of the fanatical amplification of T whose choiceworthiness function is exactly k times that of T—call this theory T_k? No. For knowledge requires safety, which is typically cashed out in terms of

> *margin for error principles* of the form: if one knows in a given case, one does not falsely believe in sufficiently close cases. For instance, normally if [a] hall contains just fifty people and you judge at a glance that it contains at least fifty people, you might very easily have made that judgement even if the hall had contained one person [fewer], in which case your belief would have been false. By a margin for error principle, you know at a glance that the hall contains at least fifty people only if in fact it contains more than fifty. Without a margin for error, your judgement is too unreliably based to constitute knowledge.[88]

---

abductive grounds than their run-of-the-mill competitors, such as the hypothesis that Pascal's mugger is lying.

[87] I owe this argument to Sebastian Liu (pc.).

[88] Williamson (2000: 76). I have omitted a footnote from the quoted text.

Similarly, if you judge that T_k is false on the ground that it's fanatical, you may very well have made the same judgment about the non-fanatical theory whose choiceworthiness function is (k - 0.0000001) times that of T, in which case, your belief would have been false. Thus, your belief that T_k is false because it's fanatical does not satisfy the margin for error principle; so, it isn't safe and thereby fails to constitute knowledge. Therefore, we don't know the falsity of every fanatical amplification of every moral theory. I conclude that a cross between MEC and Knowledge First Decision Theory on which we (i) disregard all moral theories we know to be false—even if we have some nonzero credence in them—and then (ii) maximize expected choiceworthiness, given the surviving theories, won't block the problems of finite fanaticism or infinity.

5. Normalization

So far, we've been assuming MacAskill, Bykvist, and Ord's universal scale account *arguendo*. In this section, we drop this assumption to consider an entirely different response to the problem of intertheoretic value comparisons, which we can requisition to serve as a solution to the problems of finite fanaticism and infinity. The alternative response is that the expected choiceworthiness maximizer should employ a *statistical normalization method* to normalize competing moral theories against one another.[89] The goal of normalization is to reflect the *principle of equal say*,

[89] See MacAskill, Bykvist, and Ord (2020: chapter 4 and 153-55) and MacAskill, Cotton-Barratt, and Ord (2020). MacAskill, Bykvist, and Ord propose normalization to deal with moral theories that are interval-scale measurable but *incomparable* with one another. However, they later discuss normalization in the context of fanaticism (2020: 154-55); and at any rate, the thought that we should normalize competing moral theories against each other when we're morally uncertain is *prima facie* plausible and sufficiently

which says that every moral theory that is assigned the same credence should exert the same degree of influence over the deontic verdict given by the theory of moral uncertainty.[90] Thus, if we have equal credence in just two moral theories, these theories should have equal influence over what our theory of moral uncertainty tells us to do.

Unfortunately, incorporating the principle of equal say via normalization renders our theory of moral uncertainty entirely insensitive to intuitive differences in moral stakes. To see this, consider an agent with equal credence in (i) the extremely demanding moral outlook outlined in Peter Singer's (1972) "Famine, Affluence, and Morality" and (ii) *Dudeism*, the self-styled "slowest-growing religion in the world" whose principal teaching is that "Life is short and complicated and nobody knows what to do about it. So don't do anything about it. Just take it easy, man."[91] Intuitively, such an agent should allocate many more of her resources (time, effort, money, etc.)

---

common to warrant assessment. Here are the technical details of normalization: MacAskill, Bykvist, and Ord (2020: 86-94) and MacAskill, Cotton-Barratt, and Ord (2020: 74-86) defend *variance voting*. This normalization procedure "corresponds to linearly rescaling all of the theory's choiceworthiness values so that their variance is equal to 1, while keeping their means unchanged. This doesn't change the ordering of the options by that theory's lights, it just compresses it or stretches it so that it has the same variance as the others. One can then apply MEC to these normalized choiceworthiness functions" (MacAskill, Bykvist, and Ord (2020: 93)). Two further technical details: firstly, if a theory says that all options are equally choiceworthy, then the normalization procedure does not change its choiceworthiness function (MacAskill, Bykvist, and Ord (2020: 87)). Secondly, MacAskill, Bykvist, and Ord note that we can employ variance voting only when there are finitely many options on the table (2020: 94, n.20) and tentatively defend the view that the relevant options are those available to the agent in a given decision-situation (2020: 101-05).

[90] MacAskill, Bykvist, and Ord (2020: 90-91); MacAskill, Cotton-Barratt, and Ord (2020: 72).

[91] Dudeism is modeled on Jeff Bridges' character, The Dude, from the Coen brothers' 1998 film *The Big Lebowski*. The interested reader is invited to read more at https://dudeism.com/whatisdudeism/.

to impartial altruism, which is choiceworthy according to Singer, than to bowling and drinking white Russians, which are choiceworthy according to Dudeism.[92] Yet if we normalize these theories against one another, Dudeism will have equal say over what the agent ought to do.

The preceding reflections leave us with a general problem for MEC. We often want our theory of moral uncertainty to be stakes-sensitive, as in the case of Singer vs. Dudeism. But, to quote MacAskill, Cotton-Barratt, and Ord, we also want it "to avoid 'fanatical' conclusions, where the expected choiceworthiness of [our] options is almost entirely determined by the choiceworthiness function of a theory in which one has vanishingly small credence but which claims that most decision-situations are enormously high stakes."[93] It is difficult to see how MEC, with its commitment to the machinery of an expectation-maximizing decision theory, can pull this off. Perhaps a theory with an entirely different structure can do better.

6. Conclusion

Finite fanaticism and the problems of infinity pose a significant challenge to MEC. It is far from clear that MEC can simply borrow solutions to these problems from decision theory, and statistical normalization fails. It is left to the expected choiceworthiness maximizer to show us a way forward.

[92] If you don't have this intuition, abstract away from the details of the case and imagine that you have credence = 0.5 that it's *extremely* important for you to Φ and credence = 0.5 that you should Ψ, but only in some very weak sense of 'should'. Assuming that you can't both Φ and Ψ, intuitively, you should Φ.

[93] MacAskill, Cotton-Barratt, and Ord (2020: 73-74).

The dialectic of §3 implicitly raised one way forward, which is to argue for the conjunction of four theses: firstly, that the ought of moral uncertainty is a rational ought—specifically, the ought of practical rationality, as investigated in decision theory; secondly, that the same decision theory governs choice under descriptive and moral uncertainty; thirdly, that this one true decision theory is not fanatical; and fourthly, that *hard externalism* is true about decision theory, although it is false about morality. By 'hard externalism is true about decision theory' I mean that all we can say about an agent who is uncertain which decision theory is true is that they ought to act in accordance with the true decision theory.[94] There is no meta decision theory; and if you don't know what the true decision theory is, or what it recommends that you do in your particular decision-situation, so much the worse for you. Going this route would open the possibility of endorsing MEU *and* blocking the objection from normative uncertainty and stakes-sensitivity that I raised to it in §3. I leave exploration of this possibility to future research.

To close, I would like to mention one response I have not considered, which is to bite the bullet. Might the expected choiceworthiness maximizer simply accept that her theory is fanatical? Here are two consequences of this concession, one practical and the other theoretical. The practical consequence of accepting that MEC is both true and fanatical is that research into the relative plausibility of high-stakes moral theories would become a top priority.[95] In particular, we would be led on a quest to discover and assess theories reaching ever higher into the hierarchy of infinities

[94] Here I paraphrase Tarsney's (2020: 1019) gloss on externalism about morality. For discussion see Russell (forthcoming) and Tarsney (forthcoming).

[95] Or, if the relevant probabilities for decision-making under moral uncertainty are simply your actual credences, the top priority will be introspection to discover your own credence distribution over various fanatical moral theories.

and, in doing so, on an extended foray out of philosophy and into comparative religion and mysticism.[96] Worse still, we would then have to live in accordance with whichever infinite-stakes theory won out—or else fall into perpetual *akrasia*.

This leads us to our second and final point, which is an abductive one about theories of decision-making under moral uncertainty. Finite fanaticism and the problems of infinity do not plague the competing accounts of how we ought to handle moral uncertainty, for these accounts do not propose to enlist the machinery of decision theory to do the job. For instance, on certain externalist views, we should simply do whatever the true moral theory says we ought to do (i.e., we should not engage in moral hedging at all).[97] On My Favorite Theory, we should act in accordance with the moral theory in which we have highest credence.[98] And on Moral Parliamentarianism (roughly), we should act in accordance with the 'legislation' passed by an internal moral parliament that represents our credence distribution over moral theories with a proportional number of delegates, who then debate and bargain with each other over what is to be done.[99] Of course, that a theory of choice under uncertainty is fanatical does not entail that it is false. But the implications of fanaticism are so absurd that anyone wishing to argue abductively for a fanatical theory over non-fanatical competitors will be left fighting an uphill battle.[100]

---

[96] Cf. Chen and Rubio (2020: §4.2-4.4).

[97] See Harman (2015) and Weatherson (2019).

[98] Gustafsson and Torpman (2014), though see Gustafsson (2022).

[99] Newberry and Ord (2021); cf. Greaves and Cotton-Barratt (2023).

[100] I am indebted to Lara Buchak, Krister Bykvist, Pietro Cibinel, Adam Elga, Sam Fullhart, Elizabeth Harman, Harvey Lederman, Sebastian Liu, Jake Nebel, Teruji Thomas, Henry Wilson, and several anonymous referees for valuable comments on earlier drafts of this article and to Hezekiah Grayer II for valuable discussion.

References

Beckstead, Nick. 2013. "On the Overwhelming Importance of Shaping the Far Future." Ph.D. thesis, Rutgers University. https://rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/.

Beckstead, Nick and Teruji Thomas. 2023. "A Paradox for Tiny Probabilities and Enormous Values." *Noûs*. DOI: 10.1111/nous.12462.

Bostrom, Nick. 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308-314.

Bostrom, Nick. 2009. "Pascal's Mugging." *Analysis* 69 (3): 443–445.

Bostrom, Nick. 2011. "Infinite Ethics." *Analysis and Metaphysics* 10: 9-59.

Buchak, Lara. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

Carr, Jennifer Rose. 2020. "Normative Uncertainty without Theories." *Australasian Journal of Philosophy* 98 (4): 747-62.

Chen, Eddy Keming and Daniel Rubio. 2020. “Surreal Decisions.” *Philosophy and Phenomenological Research* 100 (1): 54–74.

Cibinel, Pietro. 2023. “A Dilemma for Nicolausian Discounting.” *Analysis* 83 (4): 662-672.

Goodsell, Zachary. 2021. “A St Petersburg Paradox for Risky Welfare Aggregation.” *Analysis* 81 (3): 420–426.

Greaves, Hilary. 2016. “Cluelessness.” *Proceedings of the Aristotelian Society* 116 (3): 311–339.

Greaves, Hilary and Owen Cotton-Barratt. 2023. “A Bargaining-Theoretic Approach to Moral Uncertainty.” *Journal of Moral Philosophy*. DOI: 10.1163/17455243-20233810.

Greaves, Hilary and William MacAskill. 2021. “The Case for Strong Longtermism.” *Global Priorities Institute* Working Paper No.5-2021. https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/.

Gustafsson, Johan E. 2022. “Second Thoughts about My Favorite Theory.” *Pacific Philosophical Quarterly* 103 (3): 448-470.

Gustafsson, Johan E., and Olle Torpman. 2014. “In Defence of My Favourite Theory.” *Pacific Philosophical Quarterly* 95 (2): 159–174.

Hájek, Alan. 2003. “Waging War on Pascal’s Wager.” *The Philosophical Review* 112 (1): 27–56.

Hájek, Alan. 2012. “Is Strict Coherence Coherent?” *Dialectica* 66 (3): 411-424.

Harman, Elizabeth. 2015. “The Irrelevance of Moral Uncertainty.” In *Oxford Studies in Metaethics* vol. 10, ed. Russ Safer-Landau, 53–79. New York: Oxford University Press.

Hawthorne, John and Jason Stanley. 2008. “Knowledge and Action.” *The Journal of Philosophy* 105 (10): 571–590.

Hong, Frank. “Know Your Way Out of St. Petersburg: An Exploration of ‘Knowledge-First’ Decision Theory.” Forthcoming in *Erkenntnis*.

Isaacs, Yoaav. 2016. “Probabilities Cannot Be Rationally Neglected.” *Mind* 125 (499): 759-762.

Jackson, Frank and Michael Smith. 2006. “Absolutist Moral Theories and Uncertainty.” *The Journal of Philosophy* 103 (6): 267-283.

Kosonen, Petra. 2022. “Tiny Probabilities of Vast Value.” Ph.D. thesis. University of Oxford.

Liu, Sebastian. 2022. “Don’t Bet the Farm: Decision Theory, Inductive Knowledge, and the St. Petersburg Paradox.” Unpublished Manuscript.

Lockhart, Ted. 2000. *Moral Uncertainty and its Consequences*. New York: Oxford University Press.

MacAskill, William. 2022. *What We Owe the Future*. New York: Basic Books.

MacAskill, William, Krister Bykvist, and Toby Ord. 2020. *Moral Uncertainty*. New York: Oxford University Press.

MacAskill, William, Owen Cotton-Barratt, and Toby Ord. 2020. "Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons." *The Journal of Philosophy* 117 (2): 61-95.

MacAskill, William and Toby Ord. 2020. "Why Maximize Expected Choiceworthiness?" *Noûs* 54 (2): 327-353.

MacAskill, William, Aron Vallinder, Caspar Oesterheld, Carl Shulman, and Johannes Treutlein. 2021. "The Evidentialist's Wager." *The Journal of Philosophy* 118 (6): 320-342.

McGee, Vann. 1999. "An Airtight Dutch Book." *Analysis* 59 (4): 257–265.

Mogensen, Andreas L. 2021. "Maximal Cluelessness." *The Philosophical Quarterly* 71 (1): 141-162.

Monton, Bradley. 2019. “How to Avoid Maximizing Expected Utility.” *Philosophers’ Imprint* 19 (18): 1-25.

von Neumann, John and Oskar Morgenstern. 1947. *Theory of Games and Economic Behavior*, 2nd ed. Princeton: Princeton University Press.

Newberry, Toby and Toby Ord. 2021. “The Parliamentary Approach to Moral Uncertainty.” *Future of Humanity Institute* Technical Report 2021-2. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/06/Parliamentary-Approach-to-Moral-Uncertainty.pdf.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.

Nover, Harris, and Alan Hájek. 2004. “Vexing Expectations.” *Mind* 113 (450): 237–249.

Riedener, Stefan. 2020. “An Axiomatic Approach to Axiological Uncertainty.” *Philosophical Studies* 177 (2): 483–504.

Ross, Jacob. 2006. “Rejecting Ethical Deflationism.” *Ethics* 116 (4): 742–768.

Russell, Jeffrey Sanford. 2023. “On Two Arguments for Fanaticism.” *Noûs*. DOI: 10.1111/nous.12461.

Russell, Jeffrey Sanford. “The Value of Normative Information.” Forthcoming in *Australasian Journal of Philosophy*.

Russell, Jeffrey Sanford and Yoaav Isaacs. 2021. “Infinite Prospects.” *Philosophy and Phenomenological Research* 103 (1): 178-198.

Sepielli, Andrew. 2009. “What to Do When You Don’t Know What to Do.” In *Oxford Studies in Metaethics* vol. 4, ed. Russ Shafer-Landau, 5–28. New York: Oxford University Press.

Singer, Peter. 1972. “Famine, Affluence, and Morality.” *Philosophy & Public Affairs* 1 (3): 229-243.

Slesinger, George. 1994. “A Central Theistic Argument.” In *Gambling on God: Essays on Pascal’s Wager*, ed. Jeff Jordan. Lanham: Rowman & Littlefield.

Smith, Nicholas J. J. 2014. “Is Evaluative Compositionality a Requirement of Rationality?” *Mind* 123 (490): 457–502.

Tarsney, Christian. 2020. “*Normative Externalism*, by Brian Weatherson.” *Mind* 130 (519): 1018-1028.

Tarsney, Christian. “Metanormative Regress: An Escape Plan.” Forthcoming in *Philosophical Studies*.

Thomas, Teruji. 2022. “Separability and Population Ethics.” In the *Oxford Handbook of Population Ethics*, ed. G. Arrhenius, K. Bykvist, T. Campbell, and E. Finneron-Burns, 271-295. New York: Oxford University Press.

Thorstad, David. 2023. “High Risk, Low Reward: A Challenge to the Astronomical Value of Existential Risk Mitigation.” *Philosophy & Public Affairs* 51 (4): 373-412.

Weatherson, Brian. 2012. “Knowledge, Bets, and Interests.” In *Knowledge Ascriptions*, ed. Jessica Brown and Mikkel Gerken, 75-103. New York: Oxford University Press.

Weatherson, Brian. 2019. *Normative Externalism*. New York: Oxford University Press.

Wedgwood, Ralph. 2013. “Akrasia and Uncertainty.” *Organon F* 20 (4): 484–506.

Wilkinson, Hayden. 2022. “In Defense of Fanaticism.” *Ethics* 132 (2): 445-477.

Wilkinson, Hayden. 2023. “Can Risk Aversion Survive the Long Run?” *The Philosophical Quarterly* 73 (2): 625–647.

Williamson, Timothy. 2000. “Margins for Error: A Reply.” *The Philosophical Quarterly* 50 (198): 76–81.

Chapter three: New work for cluelessness

"If any person had told the Parliament which met in perplexity and terror after the crash in 1720 that in 1830 the wealth of England would surpass all their wildest dreams, that the annual revenue would equal the principal of that debt which they considered as an intolerable burden…that London would be twice as large and twice as populous, and that nevertheless the rate of mortality would have diminished to one half of what it then was…our ancestors would have given as much credit to the prediction as they gave to Gulliver's Travels. Yet the prediction would have been true."

– Macaulay (1830: 563–64)

1. Cluelessness: approaching the problem

The consequences of our actions matter morally. Moreover, it is plausible that the temporal location of a consequence is not *in itself* morally relevant. We should not care less about things like suffering, (in)justice, and well-being solely because they are far away from us in time. As Parfit observes, "Remoteness in time has, in itself, no more significance than remoteness in space" (1984: §121).[101]

Many—if not all—of our actions have consequences that stretch into the far future. Why think so? The consensus in the literature is that the identities of most—if not all—of the people in the (far)

[101] For defense, see Cowen and Parfit (1982); for pushback, see Mogensen (2022).

future are counterfactually sensitive to many of our actions.[102] The argument for this claim has been repeated on numerous occasions, so rather than rehearsing it once more, I shall simply quote Hilary Greaves, whose account is particularly lucid:

> Suppose, for example, that I pause on my way home from work, in order to help an old lady cross the road. As a result, both she and I are in any given place…at different times, at least for the remainder of that day. As a result, we advance or delay the journeys of countless others, if only by a few seconds, relative to the situation in which I had not helped her cross the road; both we and they affect which further parties enjoy chance meetings with whom; and so forth. At least some of these others were destined to conceive a child on the day in question, and if so, even our trivial influences on their day will affect, if not *whether* they conceive, then at least *which particular child* they conceive (since a delay in sexual intercourse of even a few seconds is overwhelmingly likely to affect which particular sperm fertilizes the egg). But once my trivial decision has affected *that*, it equally counts as causally responsible for *everything the child in question does during his or her life* (that is, for the differences between the actions and effects of this child and those that the alternative, in fact unconceived, child would have performed and had)—and of all the causal consequences of all *those* things, stretching down as they do through the millennia (Greaves 2016: 314-15; see also Lenman 2000: 345-48, Dorsey 2012: 51, Greaves and

[102] One might also think that history is a chaotic system, meaning that future states of the world are counterfactually sensitive to minute changes in the present state of the world (Lenman 2000: 351, echoed in Dorsey 2012: 51). So, every time we make even a minute change to the present state, we can expect nontrivial differences in future states. However, as Hare (2023) notes, our best natural science has neither established nor refuted that the world *in toto* is a deterministic system that gives rise to chaotic behavior.

MacAskill 2021, Mogensen and MacAskill 2021: §2, and Hare 2023, all building on Parfit 1984).

If (apparently) trivial things like the timings of road crossings have consequences that stretch into the far future, then it is extremely plausible that our (apparently) more important choices, such as the projects to which we devote our lives, whom we marry, and how many children we have, do as well.

Can we accurately forecast the far future consequences of our actions? It is natural to believe that we cannot. Here is a classic motivating case due to Lenman (2000: 344-45): in the first century BCE, Richard, a bandit, elects not to kill Angie, a villager. Angie turns out to be a distant ancestor of Hitler; thus, Hitler's existence and crimes are counterfactually sensitive to Richard's choice to spare Angie. Can we conclude that Richard's mercy was for the worse, then? No, according to Lenman, for it is possible that if Richard had killed Angie, instead of Hitler there would have been Malcolm the Truly Appalling, whose crimes would have been even more unspeakable than those of Hitler. It is impossible for Richard to know whether killing or sparing Angie would be for the best in light of total history. Lenman's suggestion is that we are in a similar situation: no matter how epistemically virtuous and well-informed we are, it is impossible for us to know which of our available options would be best in terms of total consequences.

We can now state the cluelessness worry.[103] Many of our actions have consequences that stretch into the far future (and perhaps even past the heat death of the universe—see Schwitzgebel 2023 for discussion). These consequences matter and are not to be subjected to pure time discounting. But we are in the dark about the vast majority of them.[104] We therefore appear to be almost entirely clueless about the part of morality that is concerned with promoting good outcomes and preventing bad outcomes.

It is important to flag that not everyone will accept what I have said so far; or at least, not everyone will accept that it is of any great moral concern. In particular, certain nonconsequentialists believe that, although unforeseeable consequences may have *axiological* significance, they do not have *deontic* significance. They believe, that is, that we are only on the hook for the foreseeable, or 'visible,' consequences of our actions (see e.g. Lenman 2000). One way to motivate this thought is to consider a case in which someone buys a meal for a stranger in need. It is very plausible, at least *prima facie*, that buying the meal counts as an act of impartial beneficence, even if the buyer is ignorant—and knows that she is ignorant—of the *total* consequences associated with buying the meal, stretching as they do into the far future. Thus, it might seem that cluelessness is a problem only for consequentialists.

[103] Or at any rate, the worry that I take to be most gripping for ethics in the actual world and that will be our subject in this essay. *N.B.*, however, that several distinct decision-theoretic issues may be classed under the heading of 'cluelessness'—see Greaves and MacAskill (2021: §7) for discussion.

[104] This is how I am understanding cluelessness: an agent is *clueless* in a given decision-situation just in case she cares about all the consequences of her choice but is by her own lights in a position to foresee only a small fraction of them. So, large-scale changes to the identities of future persons are not *necessary* for cluelessness; but many in the literature take them to be *sufficient* in the actual world.

Unfortunately, it is not so straightforward to escape the problem of cluelessness by adopting a nonconsequentialist view. First, Mogensen and MacAskill (2021) argue that standard nonconsequentialist asymmetries between doing and allowing and between harming and benefitting imply, in light of the fact that our actions have consequences that stretch into the far future, that we should do as little as possible, which they regard as a *reductio*. Second, de Kenessey (2024) argues that standard formulations of contractualism, which tether the truth of moral principles to facts about what would happen if they were generally adopted, leave us, in light of the problem of cluelessness, bereft of moral knowledge. Finally, certain nonconsequentialist moral outlooks assign deontic significance to all the consequences of our actions, rather than only the immediate, foreseeable consequences. (Witness classical Indian Mahāyāna Buddhism, which urges us to promote welfare in an utterly impartial manner (see Crosby and Skilton 2008), despite quite plausibly *not* being committed to consequentialism, but rather to a form of moral phenomenology or virtue ethics (see respectively Garfield 2022 and Harris 2024).) In making these points, I do not wish to suggest that the problem of cluelessness is a problem for *everyone*. I wish, more conservatively, to deny that cluelessness is a problem only for consequentialists. The intended audience of this paper, therefore, is the set of theorists who are committed to accounting for all the consequences of their actions—whether they are, like many consequentialists and Mahāyāna Buddhists, so committed by their own professed lights, or whether they are, like the dialectical opponents of Mogensen and MacAskill and de Kenessey, so committed by non-obvious implications of their view.

With the motivation for the problem of cluelessness in place, I shall now outline the contribution I hope to make to our study of the problem. I will examine the hypothesis that we should adhere

to expected utility theory (EUT) in response to cluelessness. Most authors assume that it is misguided to do so in the face of our apparently vast ignorance about the far future. As a result, the case for adherence to EUT has received little systematic assessment in the literature. That is surprising, if for no other reason that EUT is orthodoxy in decision theory. It deserves a fair hearing.

I conduct my examination of the case for EUT in two stages. The first stage is critical: I argue that extant responses to cluelessness that depart from EUT face serious challenges. In particular, I argue against the view that we should treat cluelessness as a case of *opaque sweetening* (§2), that we should have *imprecise credences* and defer to the *maximality rule* (§3), and that we should ignore the far future consequences of our actions altogether (§4). Collectively, these sections motivate the thought that we should at least consider adherence to EUT. With this thought in mind, the next stage of the paper turns to EUT itself. I begin by defending EUT from the objection that it cannot adequately address the problem of *conscious unawareness* (§5). My discussion here is in line with recent work on Bayesian theories of awareness expansion; it will not be surprising to those familiar with that literature. The major innovation with respect to EUT comes with the introduction of *higher-order uncertainty* (§6). I suggest that for many agents, grappling with the phenomenon of cluelessness will involve higher-order uncertainty. Higher-order uncertainty is a form of normative uncertainty, rather than empirical uncertainty; and it poses a problem for EUT that appears especially foreboding when viewed against the backdrop of the far future. I close by sketching a route to a solution that ties choice under higher-order uncertainty to choice under moral uncertainty. Certain implementations of this solution require minimal revision to EUT; others require significant revision. If I am right, then—surprisingly—the extent to which cluelessness

prompts departure from orthodoxy depends not on the depth of our empirical uncertainty, but on what we ought to do when we are uncertain what morality requires of us.

2. Parity and opaque sweetening

Roughly, two value-bearers *x* and *y* are *on a par* if it is neither the case that *x* is better than *y*, nor that *y* is better than *x*, nor that *x* and *y* are equally good, nor that *x* and *y* are utterly incomparable (Chang 2002). More precisely, *x* and *y* are on a par iff *x* is at least as good as *x*, *y* is at least as good as *y*, it's not the case that *x* is at least as good as *y*, and it's not the case that *y* is at least as good as *x* (Dorr, Nebel, and Zuehl 2023: 418). The parity relation is *in*sensitive to mild sweetening but *is* sensitive to major sweetening: if two value-bearers are on a par, then a small improvement to one will not make it better than the other, but a major improvement to one *will* make it better than the other. If two postdoctoral positions are on a par, then increasing the salary of one position by $1 will not make it better than the other, but increasing its salary by $10,000 will. Hare (2023) believes that the *total consequences*—the foreseeable consequences together with the unforeseeable consequences, stretching to the end of time—associated with almost all of our options are on a par. He reasons as follows: due to the fragility of personal identity, the identity of (nearly) everyone in the future is counterfactually sensitive to our choices. And Hare finds it convincing that histories populated by entirely different "casts" of persons will be on a par. The thought here seems to be that the different people in the different histories will get up to different things, realizing different values (e.g., creating different great works of art) along the way.

If the total consequences associated with our options are on a par, and if it is permissible to choose any option when presented with a menu of options whose outcomes are all on a par, then, in Hare's words, the part of morality that is concerned with promoting good consequences collapses into "a kind of moral nihilism—we always know that everything we can do is morally permissible, in both objective and subjective senses of the term" (Hare 2023). Hare, however, resists this conclusion. On his view (2023, building on Hare 2010), we morally ought to choose options with foreseeably good consequences, despite our knowledge of the parity of the *total* consequences of our options. His argument, in brief, is that our knowledge of universal parity does not supply us with a reason in any direction, whereas our knowledge of foreseeably good consequences does supply us with reason to choose the option(s) associated with those good consequences. We therefore have *most* reason to choose the options associated with foreseeably good consequences, and we ought to do what we have most reason to do. Thus, nihilism is avoided.

However, *pace* Hare, we do not know that the total consequences associated with each of our options are on a par. Consider the actual historical case of

> *Stanislav Petrov*. On 26 September 1983, a Soviet missile detection system erroneously reported that the United States had launched a nuclear missile strike against the USSR. In contravention of standing military orders, Petrov elected not to authorize a retaliatory nuclear strike.

We don't know how history would have played out had Petrov authorized a nuclear strike against the U.S. In particular, we don't know (nor was Petrov in a position to know) that history as it will

actually unfold is on a par with history as it would have unfolded, had Petrov chosen to launch. That is because in some epistemically possible histories in which Petrov launches, a nuclear WWIII ensues, leading to human extinction in some and to the unrecoverable collapse of industrial civilization in others. And, plausibly, histories in which humanity goes extinct or suffers the unrecoverable collapse of industrial civilization near the end of the 20th century are *worse than* (and so, not on a par with) histories in which human civilization flourishes well past the 20th century.

Petrov's case demonstrates the epistemic possibility that it is not just the identities of future generations that are counterfactually sensitive to our decisions. It is epistemically possible—even if improbable—that a myriad of future contingencies—including whether humanity succumbs to a premature extinction, endures the (unrecoverable) collapse of industrial civilization, or blossoms into an interstellar spacefaring civilization—are *also* counterfactually sensitive to our choices (see Bostrom 2003, Ord 2020, MacAskill 2022, and Tarsney 2023 for defense and discussion). Some of these events would be extremely consequential, both descriptively and evaluatively. It is therefore not the case that "we can almost always be almost perfectly confident that we are making the world neither better nor worse" due to the widespread parity of possible future histories (Hare 2023). Some future histories are better than others, and it is epistemically possible that whether they obtain is counterfactually sensitive to our decisions.

Now, one might grant this point but object that *Petrov scenarios*—scenarios in which an individual's decision makes a parity-breaking difference to the future—are so rare that Hare's thesis is all but established. Notice, however, that the *identity* of the agent in a Petrov scenario can

be decisive. For instance, in a possible world in which Petrov's counterpart—Petrov*—is more disposed than Petrov to follow hierarchical authority and less disposed to follow his own judgment, Petrov* may comply with his standing orders and launch the retaliatory strike. And we have already seen that the identities of most or all future people are counterfactually sensitive even to our mundane decisions. So, although it may be rare for us to find *ourselves* in a Petrov scenario, it may not be rare for the *outcome* of a Petrov scenario to be counterfactually sensitive to our decisions.

3. Imprecise credences and the maximality rule

According to Greaves (2016: 323), cases of *complex cluelessness* feature a choice between a pair of options A and B such that (i) we have evidence E1 that the consequences of A would be systematically and substantially better than those of B, (ii) we have evidence E2 that the consequences of B would be systematically and substantially better than those of A, and (iii) it is unclear how to weigh E1 against E2. Thus, although we have strong reason to suspect that the consequences of our choice will be significant, we are deeply uncertain about the *sign* of their significance, which is to say, about whether they will be good or bad. Greaves offers the example of donating to the Against Malaria Foundation (AMF), which reliably safeguards children in developing countries from contracting malaria by distributing long-lasting insecticide-treated bed nets. The short-run consequences of the work done by AMF are very good. But what of the total consequences? Greaves (2016: 325) notes that reducing childhood mortality plausibly increases population size and argues that "it is radically unclear" whether population increases in developing countries tend to be good or bad for them in the medium run, let alone over the very long run.

Continuing Greaves's discussion, Mogensen (2021: 152-53) focuses on further possible effects of population increases on the environment, rate of technological innovation, long-run economic growth, and probability of existential catastrophe. He concludes that "[w]e simply do not know enough about the long-run impact of acts whose short-term effects involve improvements in the health and well-being of people living in poverty in developing countries" to say much with confidence about their total consequences (Mogensen 2021: 156).

Greaves and Mogensen consider the view that agents confronting the type of uncertainty at issue in cases of complex cluelessness should have *imprecise credence functions*. Rather than having a single precise credence function, as in orthodox Bayesianism, an agent with an imprecise credence function has a *representor*, a set of probability functions containing at least two elements. As Greaves puts it, "the idea here is that when the evidence fails conclusively to recommend any particular credence function above certain others, agents are rationally required to remain neutral between the credence functions…[and] to include all such equally recommended credence functions in their representor" (2016: 328). We can then model cases of complex cluelessness as those in which different options are expectedly best relative to different credence functions in the representor. In the AMF case, this will mean that according to at least one credence function in the representor, donating to AMF is best in expectation, whereas according to at least one other credence function in the representor, doing something other than donating to AMF is best in expectation. This raises the question of what is rationally permissible to do when there is 'disagreement' among the credence functions in one's representor.

Mogensen (2021) considers the *maximality rule*, which says that for any pair of options O_1 and O_2, O_1 is rationally preferred to O_2 just in case the expected utility of O_1 is greater than that of O_2 according to *every* credence function in the agent's representor.[105] Thus, the choice of an option is rationally permissible just in case there is no other option that has greater expected utility according to every credence function in the representor. Mogensen's contention is that "we cannot rule out the maximality rule. As a result, we ought to avoid drawing any conclusions that are inconsistent with it" (2021: 150-51). To showcase the maximality rule in action, Mogensen compares AMF to the Make-A-Wish Foundation (MAW), which grants 'wish experiences' to American children with critical illnesses. MAW is significantly less cost-effective than AMF at promoting welfare in the short run. However, given the state of our evidence bearing on the long-run consequences of donating to AMF and to MAW—"incomplete, imprecise, and equivocal"—a rational agent will have at least one probability function in her representor according to which donating to the former maximizes expected utility and at least one according to which donating to the latter maximizes expected utility (Mogensen 2021: 154; cf. Tarsney 2023: §6.1). So, per the maximality rule, either choice is rationally permissible—even for the impartially beneficent agent.

My main worry about this picture concerns overgeneralization. Suppose that whenever we confront a body of evidence for which there is "no obviously canonical way of weighing up [the] reasons or arguments against one another" (Greaves 2016: 334-35), we are rationally required to have imprecise credences. Suppose further that we avoid drawing any conclusions that are

[105] See Mogensen (2021: §3.2 and §5) for introduction and criticism of competitors to the maximality rule.

inconsistent with the maximality rule because we cannot rule it out (as per Mogensen 2021: 150-51). It follows that *many* options will be rationally permissible.[106] To illustrate, consider

> *Social planning*: you are an impartially benevolent social planner who must implement a public policy to address some social problem. You are deciding between policy 1 and policy 2. Most of your evidence points strongly in favor of policy 1, but a bit of evidence you can't rationally ignore weakly favors policy 2. Social science being what it is, it's not entirely clear how to weigh the evidence for policy 1 against the evidence for policy 2. In your representor, therefore, are five credence functions. According to four of the five, policy 1 has significantly greater expected social welfare than policy 2. According to the fifth, however, policy 2 has *slightly* greater expected social welfare than policy 1.

The maximality rule implies that neither policy is rationally preferred—both are permissible. That does not seem right. Intuitively, a benevolent social planner should prefer policy 1. After all, policy 1 'wins' by a significant margin according to four out of five credence functions and loses only narrowly according to the fifth.

Of course, *Social planning* is a stylized case. But it illustrates an acute problem from which the maximality rule would suffer if deployed in earnest in public policy. Research in social science is subtly complex and often fails to yield unequivocal answers. Macroeconomists, for instance, do not agree on the effect of raising the minimum wage on employment (see e.g. Neumark and

[106] This problem is all the worse if, as Yim (2019) suggests, the distinction between cases of *complex* cluelessness and those of *simple* cluelessness (on which see Greaves 2016: 322-25) cannot be upheld.

Wascher 2006, Doucouliagos and Stanley 2009, Boockmann 2010, Dube *et al*. 2010, and Clemens and Wither 2014). To take another example, after a review of the literature on the relationship between political and economic institutions and economic growth, Koyama and Rubin conclude,

> This is what makes the problem of economic development in the poorest parts of the world so vexing. There is no silver bullet. We know what has worked in various historical contexts. But merely transplanting what worked elsewhere to poverty-stricken societies isn't the solution. Context matters. Culture and the historical past matter. So do demography and geography (2022: chapter 17).

As these examples suggest, if one is antecedently committed to a picture on which we are rationally required to have imprecise credences when confronted with sufficiently opaque evidence, given the murkiness of research in social science, it will be the case ubiquitously that no single policy maximizes expected utility relative to *every* credence function in our representor. It will then follow from the maximality rule that much of what we could conceivably do is rationally permissible. This will strike many, myself included, as too lax a standard. Surely there is *something* more action-guiding to be said in the context of setting public policy, even if it is not entirely clear what that is.[107] Generalizing from public policy, the lesson is this: the maximality rule gives us too

[107] What Koyama and Rubin go on to say immediately after the seemingly pessimistic conclusion quoted in the main text is suggestive: "Of course, this does not mean that bringing wealth to the world's poorest is a lost cause. Knowing what did work and why it worked matters precisely because it allows us to build a cumulative body of knowledge. This, in turn, helps provide a framework for understanding what economic policies are likely to succeed. It will take substantial local knowledge to know how to apply this framework to a given society. Which parts can be used and which can be discarded? While the answer is always 'it depends,' this knowledge is still vital to even know where to begin" (2022: chapter 17).

much wiggle room.[108] To show that an option is not rationally required, we need only find *one* credence function that is consistent with our—by hypothesis—ambiguous body of evidence and relative to which the option fails to maximize expected utility. It seems that it will be possible to do this across a myriad of interesting decision-situations. If so, it will follow that little is rationally required of us.

4. Keep calm and carry on

Schwitzgebel (2023) argues that "our decisions should be *not at all* influenced by our expectations about their effects more than a billion years in the future." Among Schwitzgebel's grounds for this position is skepticism about the entire quantitative, decision-theoretic approach to thinking about the far future exemplified in, e.g., Bostrom (2003). Here is Schwitzgebel:

> It's easy to concoct stories about how what we do now might affect the billion-year future, and then to attach decision-theoretic numbers to those stories. We lack good means for evaluating these stories...it is *no more plausible* that the best thing we can do for the billion-year future is...fund studies of existential risk...than that the best thing we can do...is suffer near-term extinction or global catastrophe...I am making a stronger claim of radical uncertainty...maybe the best approach to radical uncertainty is to say that...mathematical decision theory fails—more likely to mislead than enlighten.

[108] Mogensen and Thorstad (2022: §4.2) briefly raise this worry as well.

Similarly, Dorsey (2012) submits that "we have *no way* to assess the probability" that the positive far future consequences of our actions will cancel out the negative far future consequences (2012: 55-56, italics added). A wholly different approach is needed (Dorsey suggests that the consequentialist can respond to the problem of cluelessness in whatever way is best for the metaphysical realist to respond to the problem of external world skepticism).[109]

Against Schwitzgebel and Dorsey, I concur with Cowen (2006) that this level of skepticism "relies heavily on a *complete* lack of information…[which] is not a plausible general assumption" (2006: 392-93). To see this, consider the credence you should have that humanity will be extinct in one billion years. Would it be reasonable to have a credence of 1? No. It would be absurd to have absolute confidence that humanity will be extinct in one billion years. Would it be reasonable to have a credence of 0? No, for the same reason. It would be absurd to have absolute confidence that humanity will *not* be extinct in one billion years. So, it is not the case that 'any credence goes' regarding the far future. Some credences are rationally impermissible (i.e., not warranted by our evidence). We can therefore ask which credences regarding the far future are rationally permissible and what is rational to do in light of them.

A response on behalf of Schwitzgebel, Dorsey, and like-minded skeptics is to concede that *some* credences about the far future are unreasonable—specifically, credences that reflect very high confidence that particular events will or will not occur—but to hold that such unreasonable credences are exceptional. On this view, our evidence about the far future is so scant and/or equivocal that it is consistent with a staggeringly wide array of credence functions. And in

[109] Mogensen (2021: 159) also suggests that cluelessness can be seen as a sort of skeptical challenge.

evidential environments like *this*, one might think, attempting to deploy orthodox decision theory is farcical. From this perspective, decision theory has a limited domain of applicability. We can use it in the right sort of evidential environments—specifically, those in which we have a decent grasp of the probabilities (e.g. in well-defined games of chance). But when our evidence is such that the rationally permissible probabilities are sufficiently opaque or underdetermined, the thing to do is to suspend judgment and focus on possible consequences to which we have better epistemic access (if any). This proposal secures Schwitzgebel's contention that "our decisions should be *not at all* influenced by our expectations about their effects more than a billion years in the future." Call this proposal *limited applicability*.

I have two concerns about limited applicability, one theoretical and the other practical. Beginning with the former: limited applicability is, in slogan form, 'use decision theory, but only sometimes.' So, to make limited applicability precise, we will need to demarcate the boundaries of decision theory. Specifically, we will need to distinguish the evidential environments in which decision theory is applicable from those in which it is not. To accomplish this, it seems that we will need to (i) introduce a measure on the extent to which a set of evidence underdetermines or makes opaque the rational probabilities and then (ii) introduce a rule to the effect that decision theory is applicable when our evidence scores within a specific range and inapplicable when it scores outside of this range. (To illustrate with a crude toy model: we might imagine that sets of evidence receive scores on a scale of one to ten, with one representing maximally opaque evidence and ten representing maximally clear evidence; and that we should use decision theory just in case our evidence scores at least a five.) A worry about (i) is that there is no unique such measure—there might be reasonable disagreement, or no fact of the matter, about how opaque a set of evidence is. More

importantly, a worry about (ii) is that the range will be arbitrary—why does the rational permissibility of using decision theory drop out at exactly *that* level of evidential opacity, as opposed to a level slightly above or below it? These worries show that limited applicability will quickly saddle our theory of rational choice with undesirable complexity and explanatory commitments.

The practical worry is that limited applicability will overgeneralize when deployed in the actual world, similarly to the maximality rule. To appreciate the worry, let us compare our evidence regarding (i) the consequences of public policy for fertility rates and (ii) the consequences of biosecurity and pandemic preparedness for the future of humanity. The question will be whether our evidence regarding the latter is significantly more opaque than our evidence regarding the former.

Beginning with policy and fertility: consider the dwindling global trendline in the total fertility rate (here I rely on Spears *et al.* 2024 and Geruso and Spears forthcoming; see also United Nations 2022 and Thorstad forthcoming: §5.2). Roughly two-thirds of the world's population lives in countries in which average fertility is below replacement. Many national governments view below-replacement fertility as undesirable for a variety of reasons, including the costs of caring for an aging population and the retarding effect of depopulation on economic growth and consequently on average welfare.[110] Unfortunately, however, no country has enacted a public policy that has had

[110] Though more controversial, one might also worry that depopulation translates into a diminished capacity to project military force and, consequently, less bargaining power—and so, ultimately, less security—in international politics. See Mearsheimer (2001) for discussion of population size in relation to power and security in the international system.

a nontrivial, sustained, and positive impact on total fertility and that is compatible with basic human freedom, despite a myriad of attempts. (This is true even of quite liberal policies in countries such as Sweden.) Consequently, many are left with the sense that depopulation is a significant problem, that the solution to the problem may necessitate significant revisions to culture and economic organization, but that *which* such revisions are at present unclear. The hunt is on for a solution that is both politically tractable and efficacious.

Next consider biosecurity and pandemic preparedness. Many who have studied existential and catastrophic risks to humanity have concluded that pandemics caused by engineered pathogens (as opposed to naturally-occurring pathogens) are among the most significant such risks (see e.g. Ord 2020).[111] *Pace* Schwitzgebel, it is plausible that the future of humanity is worse, in expectation, in worlds in which humanity suffers a global catastrophic biological event that causes the death of over half the world's population in the 21st century than in worlds in which it does not. (For such an event could precipitate the breakdown of civil society and democratic institutions, long-term economic stagnation by way of demographic disruption, the collapse of industrial civilization, or outright extinction by way of driving the number of humans below the minimum viable population.) Moreover, there are a variety of interventions, such as screening for novel pathogens in public water systems, that would promote biosecurity and pandemic preparedness (see Millett and Snyder-Beattie 2017 for a cost effectiveness model, though see Thorstad forthcoming for critique). By deploying these interventions, humanity could decrease the probability of a global catastrophic biological event this century.

[111] Engineered pathogens might be designed by benevolent actors undertaking gain-of-function research or by malevolent actors such as cults (like Aum Shinrikyo) or terrorist organizations.

Suppose our evidence for the claim that promoting biosecurity and pandemic preparedness will be good for the future of humanity is so murky that we should suspend judgment and not allow it to influence our decision-making. Then—and here is the challenge for limited applicability—it seems that our evidence for the claim that policy X is politically tractable, will result in a nontrivial and sustained rise in total fertility, and will not have any unfortunate knock-on effects on the economy, gender norms, etc. (on which see Connelly 2010) is *also* so murky that we should suspend judgment on policy X. Now, you might not be persuaded by this particular example. Perhaps you think there is an obvious solution to declining fertility that every government has missed, or that biosecurity is not the thing to focus on if we want to promote the welfare of future generations. But, abstracting away from these particular examples, the thought behind the challenge is that while it is easy to be impressed by the long time horizons at issue when thinking about the far future, it is also easy to underestimate the complexity of the causal models of impact that we must bring to bear on thorny issues in politics and social science. The challenge to the proponent of limited applicability is to show us why the unique, non-arbitrary demarcation of decision theory (supposing there is one) requires us to suspend judgment about the far future (how far, exactly?) but does *not* require us to become skeptics about contemporary political issues. Said differently, the challenge is to show that our measure of evidential opacity (see the theoretical challenge above) will require suspension of judgment on the far future but will *not* require suspension of judgment on present public policy. Perhaps these theoretical and practical challenges can be met, but until they are, I am hesitant to throw out decision theory.

4.1 Gandalf's perspective

A final skeptical reaction, which I do not attribute to Schwitzgebel himself, is nevertheless suggested in his claim that when applied to the far future, "mathematical decision theory fails—*more likely* to mislead than enlighten" (Schwitzgebel 2023, italics added). From this perspective—which anecdotal evidence suggests is common amongst those who have encountered the writings of authors in the *longtermism* movement (see e.g. Bostrom 2003, Beckstead 2019, Ord 2020, and MacAskill 2022)—choosing what to prioritize on the basis of far future expected value calculations is likely to be *worse* overall than sticking with common sense and helping those who are in need in the present.[112] This perspective is consonant with Gandalf's sentiment in Tolkien's *Return of the King* that

> it is not our part to master all the tides of the world, but to do what is in us for the succour of those years wherein we are set, uprooting the evil in the fields that we know, so that those who live after may have clean earth to till. What weather they shall have is not ours to rule (Tolkien 1955: book 5, chapter 9).[113]

I won't take a stand on the truth of Gandalf's perspective here. Instead, I will simply note that it is inconsistent with skepticism about our ability to foresee the far future consequences of our choices. If someone is skeptical about our ability to see into the future, they cannot at the same time insist

[112] Cf. Temkin (2022: Appendix A), who argues that we should sometimes prioritize the present-day global poor over future generations *even if* we would do more good in expectation by mitigating existential risk.

[113] Cf. the 'crowd control example' in Jackson (1991).

that prioritizing present-day welfare will be for the best overall. They must limit themselves to the claim it will be best (in expectation) in the *present* while admitting that, by their own lights, it is entirely inscrutable what the consequences of doing good in the present will be for moral patients in the future. This is internally consistent, but it amounts to admitting that we are indeed clueless about the part of morality concerned with promoting good outcomes. (For, to restate a point from §1, the expected ratio of consequences that obtain in the far future to those that obtain in the near present is extremely large.) Anyone wishing to argue contrariwise must explain why we should rationally expect that 'taking care of our own' in the present will be better *over the long run* than, e.g., attempting to mitigate global catastrophic biological risks by investing in pandemic preparedness.

5. Conscious unawareness

I've argued over the preceding three sections that unorthodox responses to cluelessness face significant challenges. These challenges motivate the thought that we should at least consider sticking with the familiar combination of subjective Bayesianism in epistemology and orthodox expected utility theory (EUT) in decision theory. On this picture, there is no parity (or other form of incommensurability in value), at least as far as the rational agent's preferences are concerned; subjective probabilities should be sharp (i.e., one should have a single precise credence function); and one should simply choose an option that maximizes expected utility—uncertainty about the (far future) consequences of one's actions be damned. Mogensen calls this the 'Naive Response' and criticizes it on the ground that it "fails to take seriously the depth of our uncertainty" (2021:

145). Greaves, in contrast, allows that "it *may* in the end be the correct" response to cluelessness, despite her amenability to imprecise credences (2016: 327).

Before we proceed, an important clarification is in order. Properly understood, EUT is a criterion of rationality. It says that the choice of an option is rationally permissible iff and because it maximizes expected utility. As such, EUT says absolutely nothing to the effect that the rational agent must explicitly consider probabilities or crunch numbers when making decisions. However, an important subset of participants in the cluelessness debate, namely the longtermists mentioned in the previous section, frequently assume (if only tacitly) that, at least sometimes, explicit number-crunching is appropriate.[114] That is, they assume that, sometimes, the best way to realize the criterion of rationality with which EUT presents us is to consciously deploy the theory as a decision procedure. Therefore, to keep my discussion pertinent to the current debate, I will not limit my discussion to EUT *qua* criterion of rationality.

[114] For instance, Tarsney (2023) suggests that an impartially benevolent agent who embraces expected value maximization and is happy to choose prospects associated with very low probabilities and very high payoffs will tend to favor interventions aimed at reducing existential risk (see in particular Tarsney 2023: §7, though see Tarsney 2025 for an interesting proposal for avoiding dominance by such 'fanatical' or 'Pascalian' prospects). Unfortunately—but interestingly—it is precisely interventions aimed at existential risk reduction—such as attempts to reduce the probability of a catastrophic outcome brought about by a misaligned artificial intelligence—that resist analysis based on well-understood *systematic effects* (Burch-Brown 2014). According to Burch-Brown, although we may be clueless about the effects of any *token* action, certain action *types*, such as murder, are systematically associated with bad effects (and others with good effects). We can therefore recommend (against) these action types on expected value grounds. Burch-Brown acknowledges, however, that this response to cluelessness is less compelling when applied to "actions that are fairly unique, not to be repeated many times" and without much of an observable track record (2014: 117), which aptly describes actions taken to mitigate existential risks during a presumed 'time of perils' (on which see Ord 2020 and Thorstad 2023).

Greaves and MacAskill (2021: §7.2) acknowledge that *conscious unawareness* is a problem for EUT. (They do not specify whether conscious unawareness, and the introduction of the catchall state (introduced presently), are respectively a problem and a solution for EUT *qua* criterion of rationality or *qua* decision procedure. But it is natural enough to interpret them as focusing on the latter: conscious unawareness appears to be a problem for bounded agents like ourselves who are explicitly attempting to guide their decisions by the lights of EUT.) What is conscious unawareness? Roughly, it is the knowledge that one is quite likely to be unaware of considerations that are pertinent to one's decision. Greaves and MacAskill provide the illuminating example of an impartially beneficent agent living in Europe in the Middle Ages:

> the considerations most relevant to their decision—such as the benefits of science, and therefore the enormous value of efforts to help make the scientific and industrial revolutions happen sooner—would not have been on their radar. Rather, they might instead have backed attempts to spread Christianity, perhaps by violence (2021: 20-21).

This style of pessimistic historical induction—alongside reflection both on the incomplete state of our empirical knowledge (e.g. about fundamental physics) and on our limited ability to see into the future—suggests that we, like our medieval counterparts, are unaware of important facts that bear on temporally impartial beneficence.

One particular upshot of conscious unawareness is that we should believe that "we are likely to be omitting some important possible states of nature from our model altogether" (Greaves and

MacAskill 2021: 21), just as someone deliberating at the dawn of the 19th century would have neglected to include the invention of atomic weaponry in their own model.[115] Greaves and MacAskill do not believe that this poses a significant issue for EUT, however. For on their view, we can simply incorporate into our decision matrix a 'catchall' state that corresponds to "'all the other possibilities I haven't thought of'" (Greaves and MacAskill 2021: 21).[116] The main worry I see with this proposal is that the assignment of a utility score to every option-catchall-state pair will be wholly arbitrary.[117] By hypothesis, we can't say anything about the state of the world represented by the catchall state, since every possibility we *have* thought of will already be taken into account in its own state of nature. The catchall state is thus, in essence, a giant question mark. And it is difficult to see how we could meaningfully assign a utility score to the outcome 'I take action Φ and the state of the world turns out to be <question mark>.'

A response to this objection is that the utility—and also the probability—that we assign to the catchall state should be the same for each available option. The rationale for this response is that because we have no idea what the world is like in the catchall state, for all options A and B, we have no reason to believe that p(catchall state | A) > p(catchall state | B) or that, if the catchall state obtains, u(A) > u(B). And—crucially—if the probability and the utility of the catchall state is the same for each available option, then including it in our decision matrix will make no difference to the ordering of our options in terms of expected utility. Call this the *uniformity response*.

[115] Cf. Wells (2001), included in full in the Appendix.

[116] For recent, more detailed work on the catchall state proposal in the context of Bayesianism, see Steele and Stefánsson (2021: chapter 6), Steele and Stefánsson (2022, particularly §3), and de Canson (2024).

[117] Cf. Henderson *et al*. (2010: 190), Wenmackers and Romeijn (2016: 1232-34), and Bradley (2017: 255), who worry that the assignment of any particular probability to the catchall state will be arbitrary.

Here are two worries about the uniformity response. The first is that the uniformity response tacitly relies on the Principle of Indifference, which says that evidentially symmetric propositions should receive equal credence (e.g., if your evidence no more supports proposition *p* than it supports proposition *q* and vice versa, then your credence that *p* should equal your credence that *q*). For in our case, again, the thought is that for all options A and B, we have no more reason to believe that the catchall state will obtain if we do A than it will if we do B, and if it does obtain, that it will have been better to have done A than it will have to have done B. The Principle of Indifference is controversial—for the standard objection appealing to the *problem of multiple partitions*, see van Fraassen (1989) and White (2010). However, both Mason (2004) and Greaves (2016) argue that appealing to (a restricted version of) the Principle of Indifference is rationally permissible in the context of cluelessness, and Liu (forthcoming) offers a plausible general defense of the principle. I am therefore not inclined to take the reliance of the uniformity response on indifference reasoning as strong evidence against it.

The second worry is that the uniformity response amounts to saying that we can just ignore conscious unawareness, which is unsatisfying. As Lenman alleges in reply to a distinct, though related, indifference-based response to cluelessness, it

> is just a rather fancy technical way of saying we have no clue what will happen. It is surely a sophistry to treat a zero expected value that reflects our knowledge that an act will lack significant consequences as parallel in significance to one that reflects our total ignorance of what such consequences…will be (2000: 357).

In just the same way—one might object—assigning the same probability and utility to the catchall state for every available option is for practical purposes equivalent to eschewing its inclusion in the first place, i.e., to assigning it zero probability (for again, a uniform assignment guarantees that no difference is made to the ordering of options in terms of EU).

Although I appreciate the intuitive force of this objection, I believe that we should reject it. It will be helpful to reemphasize that our guiding question is what is rational to *do*, given the goal of impartial beneficence. I assume that what is rational to do is determined by the balance of our reasons. So the question we need to ask is whether conscious unawareness changes the balance of our reasons. It seems to me that it does not. Consider an arbitrary pair of options A and B. Does conscious unawareness give us a reason to choose A over B? No. A reason to choose A over B is, roughly, something that can be said in favor of A that cannot be said in favor of B.[118] But possibilities of which we are irredeemably unaware at the time of decision cannot supply us with any such consideration. So, on reflection, it strikes me as appropriate that the inclusion of the catchall state in our decision matrix should have no effect on the ranking of our options in terms of expected utility. Therefore, although I accept that *which* utility we assign the catchall state is arbitrary, it is *not* arbitrary to adopt a uniform such assignment. Rather, it is principled.

[118] Here I follow Hare (2023) in assuming that practical reasons are *contrastive*.

6. Higher-order uncertainty

This defense notwithstanding, I anticipate that many will continue to feel that adherence to EUT "fails to take seriously the depth of our uncertainty" (Mogensen 2021: 145). I share this concern. In this section, I will introduce the hypothesis that adherence to EUT in the context of cluelessness fails to account for *higher-order uncertainty*. To introduce higher-order uncertainty, we will begin with a concrete example; then we will give an abstract characterization. Here is Adam Elga:

> Will it rain tomorrow? I'm not sure. My credence (subjective probability) that it will rain is 40%.
>
> Is 40% the rational credence for me to have that it will rain? I'm not sure about that either. Properly taking all of one's evidence into account can be tricky. I'm not sure that I've done it exactly right.
>
> So: I am uncertain whether it will rain. And I am uncertain about the rational degree of belief for me to have that it will rain (2013: 127).[119]

As Elga's example illustrates, higher-order uncertainty is uncertainty about which credence function is rational.[120] Here is the general characterization: let 'C' denote your actual credence

[119] I have omitted a footnote from the quotation.

[120] Thus characterized, it is not uncertainty about which credence function one in fact has. For reasons of space, I must bracket the issue of introspective access to one's own credences.

function—your 'subjective probabilities'—and 'R' denote the credence function it would be rational for you to have, given your evidence. You're higher-order uncertain about a proposition *p* iff C(*p*) = n, but C(R(*p*) = n) < 1 (where $0 \leq n \leq 1$).[121] In words: you're higher-order uncertain just in case you're some degree confident in a proposition, but you're less than fully certain that your level of confidence in the proposition is rational. Thus, higher-order uncertainty is one way of precisifying the attractive thought that we should sometimes be *modest* about our credences (see Dorst 2019 for discussion).

I suggest that most of us will face higher-order uncertainty when thinking about the long-run consequences of our actions, at least upon reflection. Indeed, consideration of concrete cases motivates the thought that *not* having higher-order uncertainty would amount to a form of epistemic hubris. To illustrate, return to the example of global catastrophic biological risk introduced in §4. Suppose you want to decrease the probability that a global catastrophic biological event occurs this century. Consequently, you pursue a career in biosecurity policy rather than one in banking. How much lower is the risk of catastrophe given your choice of career? That is, what is the difference between R(biological catastrophe | banking career) and R(biological catastrophe | biosecurity career)? As a rough first pass, the answer to this question appears to be 'a very small, but nonzero, number'—for you drive the probability of catastrophe down a little bit by working

[121] Strictly speaking this is not quite right. Since higher-order uncertainty in the most general sense is simply uncertainty about which credence function is rational, someone can be higher-order uncertain if she is uncertain which credence function is rational, given her evidence, while failing to have any precise credence function at all. In that case, the first conjunct of the biconditional stated in the main text, namely that C(*p*) = n, won't hold, but we will still have a case of higher-order uncertainty on our hands. In this section, I ignore this complication—both for presentational simplicity and analytical tractability—and assume that the decision-maker does have a single precise credence function.

on the problem, but since you're only one person, it seems difficult for you to drive it down significantly. Now for the key question in relation to higher-order uncertainty: will you be *fully* confident in any precise numerical answer to the question 'what is R(biological catastrophe | banking career) – R(biological catastrophe | biosecurity career)?'? If your answer is 'No, I wouldn't,' then you would face higher-order uncertainty in such a case.

What is rational for us to do when we face higher-order uncertainty? The question I am asking here is not, 'how should we resolve our higher-order uncertainty into a single precise credence function?'. The question is rather, 'given that we face higher-order uncertainty, what is rational for us to do in light of our goals and values?'. Thus, the question is not about what to believe; it is about what to do, given uncertainty about what to believe.

EUT contains an implicit answer to this question. It says that we should maximize expected utility, where the expectation of utility is defined relative to our *first-order* credences—i.e., to our credence distribution over pertinent empirical hypotheses. As such, it implicitly says that higher-order uncertainty—uncertainty about which first-order credences we *should have*, given our evidence—is wholly irrelevant to what we ought to do. That is a way in which, to my mind, EUT "fails to take seriously the depth of our uncertainty" (again borrowing Mogensen's phrase): it tells us to *disregard* a proper subset of our uncertainty, namely our uncertainty about what we should believe in the first place. And although, as I freely admit now, I lack a quick and decisive argument to the conclusion that we *should* somehow take our higher-order uncertainty into account, I anticipate I will not be alone in feeling hesitant to accept the view that says, 'don't worry even if you are *significantly* uncertain whether your credences are rational; just forge on ahead and keep

maximizing expected utility.'[122] Assuming we do want to say something about higher-order uncertainty as it relates to rational choice, then (other than 'just ignore it'), what should we say? A well-developed answer to this question would require a paper of its own. In the remainder of this paper, therefore, I set myself the more modest goal of proposing a route forward.

When we ask about the implications of higher-order uncertainty for rational choice, multiple proposals spring to mind. Perhaps we should attempt to repurpose one of the candidate criteria governing choice under imprecise credences (see Buchak 2022: §4 for an overview). Alternatively, perhaps we should repurpose one of the proposals from the literature on interpersonal credence aggregation (see e.g. Russell *et al*. 2015). Another option would be to borrow tools from voting theory (see e.g. Weatherson 1998: 6, Mogensen and Thorstad 2022: §5.2). Given the breadth of theoretical options, it would be helpful to have a principled method of search. Fortunately, I believe there is one. The key thing to notice is that choice under higher-order uncertainty is a particular instance of the more general phenomenon of choice under normative uncertainty. How so? On the picture I have in mind, the clueless agent is uncertain in the first instance about *instrumental rationality*: she is uncertain what is rational for her to do, relative to her goal of impartial beneficence. And—importantly—her practical uncertainty itself arises in part due to uncertainty about *epistemic rationality*: she is uncertain which credence function she ought to have, given her evidence. Thus, she is uncertain which option she ought to choose at least in part because she is uncertain which credence function she ought to have. Such an agent isn't just grappling with

[122] Cf. Mogensen and Thorstad (2022: §4.1): "a policy that is optimal relative to a subjective best-guess probability assignment may in principle be suboptimal, or even catastrophic in expectation, relative to the other probability assignments that were not excluded by our evidence. Intuitively, when this is the case, it is unwise to optimize relative to one's subjective best-guess probability estimate."

empirical uncertainty regarding the (far future) consequences of her options, therefore; she is also grappling with normative uncertainty.[123]

Our default expectation should be that there is a uniform solution to the problem of choice under normative uncertainty. It would be surprising if the criterion for choice under *prudential* uncertainty were, for instance, entirely different from the criterion for choice under *moral* uncertainty. Rather than expecting the correct account of choice under normative uncertainty to be disjoint and fractured, we should seek a theory that is simple and unified. These considerations support the following

> *Linking Claim*: whatever is the correct account of choice under moral uncertainty is also (*mutatis mutandis*) the correct account of choice under higher-order uncertainty.[124]

[123] A complication: it is standardly thought that higher-order uncertainty can arise either from uncertainty about what your evidence *is* or from uncertainty about what your evidence *supports* (Horowitz 2014: 737-38, Worsnip 2018: 19, Greco 2019: 85-86). And whereas uncertainty about what your evidence supports is an instance of normative uncertainty, uncertainty about what your evidence *is* plausibly is not. So, what I go on to say in the main text is relevant only to higher-order uncertainty that arises from uncertainty about evidential support. I am not particularly troubled by this limitation in scope, however, because canonical cases from the cluelessness literature, such as Greaves's (2016) and Mogensen's (2021) discussions of the Against Malaria Foundation, seem to involve precisely this sort of uncertainty.

[124] Following MacAskill and Ord (2020: 350, fn.11), I assume that the ought of moral uncertainty is a rational ought as opposed to a second-order moral ought. (This is the dominant view in the moral uncertainty literature, though it is not entirely uncontroversial (see MacAskill, Bykvist, and Ord 2020: chapter 1, section III, particularly pp. 29-30, for further discussion).) To my mind, various proposals in the moral uncertainty literature, such as Maximize Expected Choiceworthiness (MacAskill and Ord 2020) and moral bargaining theory (see presently in the main text), make the most sense on this assumption; and the Linking Claim is also more plausible if we make it, since the question of choice under moral

I use moral uncertainty in the Linking Claim because it has received the most attention among the varieties of normative uncertainty, but we could replace 'moral' with 'prudential' or 'axiological.' The point is to situate the question of choice under higher-order uncertainty within the wider phenomenon of normative uncertainty.[125]

To illustrate what I have in mind, I will *briefly* introduce three proposals from the moral uncertainty literature and indicate how each can be applied to choice under higher-order uncertainty. First, consider My Favorite Theory. According to My Favorite Theory, the morally-uncertain agent should act in accordance with whichever moral theory she finds most plausible (Gustafsson and Torpman 2014). Applied to choice under higher-order uncertainty, the thought would be that the agent should choose the option that maximizes expected utility relative to the credence function she thinks is most likely to be rational. Second, consider *externalism*. According to externalism, although the decision-maker's *empirical* uncertainty is decision relevant, her moral uncertainty is not. All we can say about the morally-uncertain agent is that she should act in accordance with the true moral theory, whatever that is (for defense, see Weatherson 2019; for criticism, see Tarsney

uncertainty and that of choice under higher-order uncertainty are then both properly understood as questions of rationality.

[125] A clarification: my proposal is addressed to impartially beneficent agents who face higher-order uncertainty. More precisely, it is addressed to agents (i) whose preferences can be represented by a cardinal utility function that is unique up to positive affine transformations, (ii) who believe that rationality requires them to have a single precise credence function, but (iii) are uncertain which credence function is rational given their evidence. It is *not* addressed to agents who are uncertain whether they should have precise credences, imprecise credences, or perhaps no numerical credences at all in cases of deep uncertainty. This more thoroughgoing form of epistemic uncertainty is certainly interesting, but it is beyond the scope of this paper to address.

2020). Applied to choice under higher-order uncertainty, externalism would say that the decision-maker's uncertainty about which credence function is rational is inconsequential; all we can say is that the decision-maker should choose in accordance with the rational credence function, whatever that is. Finally, consider *moral bargaining theory*. According to moral bargaining theory, we should model the moral theories in which we have credence as agents in a bargaining process; and we should choose the option these agents would settle on via bargaining (see Greaves and Cotton-Barratt 2023 and Kaczmarek *et al*. forthcoming). Applied to choice under higher-order uncertainty, the idea would be that the decision-maker should model the credence functions she thinks may be rational as distinct agents, each of whom shares her utility function, and that she should choose the option these agents would settle on via bargaining.

Where does this leave EUT as a response to cluelessness? The answer depends on what the correct account of choice under moral uncertainty is. If—just for example—My Favorite Theory is correct, no significant revision to EUT will be necessary. If, on the other hand, moral bargaining theory is correct, EUT—both *qua* criterion of rationality and *qua* decision procedure—will be in for significant revision, at least when deployed in the context of temporally-impartial beneficence. Sadly, I don't know the correct account of choice under moral uncertainty (or any other type of normative uncertainty). But if I am right that higher-order uncertainty will be at issue in paradigm cases of cluelessness, then we have made two important steps of progress. First, we have seen that cluelessness does not just involve (deep) empirical uncertainty—as many have assumed. It also involves normative uncertainty. Second, however, in making this observation we have to an extent subsumed the problem of cluelessness under the problem of choice under normative uncertainty.

A solution to the latter will push us some distance towards a solution to the former. I leave this as new work for cluelessness.

References

Beckstead, Nick. 2019. "A brief argument for the overwhelming importance of shaping the far future." In *Effective Altruism: Philosophical Issues*, ed. H. Greaves and T. Pummer, 80–98. Oxford: Oxford University Press.

Boockmann, Bernhard. 2010. "The combined employment effects of minimum wages and labor market regulation—a meta-analysis." *Applied Economics Quarterly* 61: 167-188.

Bostrom, Nick. 2003. "Astronomical waste: The opportunity cost of delayed technological development." *Utilitas* 15 (3): 308-314.

Bradley, Richard. 2017. *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Buchak, Lara. 2022. "Normative theories of rational choice: rivals to expected utility." In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. https://plato.stanford.edu/entries/rationality-normative-nonutility/.

Burch-Brown, Joanna M. 2014. "Clues for consequentialists." *Utilitas* 26 (1): 105-119.

de Canson, Chloé. 2024. “The nature of awareness growth.” *The Philosophical Review* 133 (1): 1-32.

Clemens, Jeffrey and Michael Wither. 2014. “The minimum wage and the great recession: Evidence of effects on the employment and income trajectories of low-skilled workers.” *National Bureau of Economic Research* working paper 20724. doi:10.3386/w20724.

Connelly, Matthew. 2010. *Fatal Misconception: The Struggle to Control World Population*. Cambridge: Harvard University Press.

Cowen, Tyler. 2006. “The epistemic problem does not refute consequentialism.” *Utilitas* 18 (4): 383-399.

Cowen, Tyler and Derek Parfit. 1992. “Against the social discount rate.” In *Justice Between Age Groups and Generations*, ed. P. Laslett and J. S. Fishkin, 144–161. New Haven: Yale University Press.

Chang, Ruth. 2002. “The possibility of parity.” *Ethics* 112 (4): 659–88.

Crosby, Kate and Andrew Skilton (trans.). 2008. *A Guide to the Bodhisattva Way of Life* (*Bodhicaryāvatāra*), by Śāntideva. New York: Oxford University Press.

Dorr, Cian, Jacob M. Nebel, and Jake Zuehl. 2023. “The case for comparability.” *Noûs* 57: 414-453.

Dorsey, Dale. 2012. “Consequentialism, metaphysical realism and the argument from cluelessness.” *Philosophical Quarterly* 62: 48-70.

Dorst, Kevin. “Higher-order uncertainty.” In *Higher-Order Evidence: New Essays*, ed. M. Skipper and A. Steglich-Petersen, 35-61. New York: Oxford University Press.

Doucouliagos, Hristos and T. D. Stanley. 2009. “Publication selection bias in minimum-wage research? A meta-regression analysis.” *British Journal of Industrial Relations* 47 (2): 406 -428.

Dube, Arindrajit, T. William Lester, and Michael Reich. 2010. “Minimum wage effects across state borders: Estimates using contiguous counties.” *The Review of Economics and Statistics* 92 (4): 945–964.

Elga, Adam. 2013. “The puzzle of the unmarked clock and the new rational reflection principle.” *Philosophical Studies* 164: 127-139.

van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.

Garfield, Jay L. 2022. *Buddhist Ethics: A Philosophical Exploration*. New York: Oxford University Press.

Geruso, Mike and Dean Spears. “With a whimper: Depopulation and longtermism.” Forthcoming in *Essays on Longtermism*, ed. J. Barrett, H. Greaves, and D. Thorstad. Oxford University Press.

Greaves, Hilary. 2016. “Cluelessness.” *Proceedings of the Aristotelian Society* 116 (3): 311–339.

Greaves, Hilary and Owen Cotton-Barratt. 2023. “A bargaining-theoretic approach to moral uncertainty.” *Journal of Moral Philosophy*. DOI: 10.1163/17455243-20233810.

Greaves, Hilary and William MacAskill. 2021. “The case for strong longtermism.” *Global Priorities Institute* Working Paper No.5-2021. https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/.

Greco, Daniel. 2019. “Fragmentation and higher-order evidence.” In *Higher-Order Evidence: New Essays*, ed. M. Skipper and A. Steglich-Petersen, 84-104. Oxford: Oxford University Press.

Gustafsson, Johan E., and Olle Torpman. 2014. “In defence of My Favourite Theory.” *Pacific Philosophical Quarterly* 95 (2): 159–174.

Hare, Caspar. 2010. “Take the sugar.” *Analysis* 70 (2): 237-247.

Hare, Caspar. 2023. *Living in a Strange World*. Unpublished manuscript. Under contract with Oxford University Press.

Harris, Stephen E. 2024. *Buddhist Ethics and the Bodhisattva Path: Śāntideva on Virtue and Well-Being*. London: Bloomsbury.

Henderson, Leah, Noah D. Goodman, Joshua B. Tenenbaum, and James F. Woodward. 2010. "The structure and dynamics of scientific theories: A hierarchical Bayesian perspective." *Philosophy of Science* 77 (2): 172–200.

Horowitz, Sophie. 2014. "Epistemic akrasia." *Noûs* 48 (4): 718-744.

Jackson, Frank. 1991. "Decision-theoretic consequentialism and the nearest and dearest objection." *Ethics* 101 (3): 461-482.

Kaczmarek, Patrick, Harry R. Lloyd, and Michael Plant. "Moral uncertainty, proportionality, and bargaining." Forthcoming in *Ergo*. https://philpapers.org/archive/KACMUP.pdf.

de Kenessey, Brendan. 2024. "Ethics and the limits of armchair sociology." *Journal of Philosophy*. https://doi.org/10.5840/jphil202410176.

Koyama, Mark, and Jared T. Rubin. 2022. *How the World Became Rich: The Historical Origins of Economic Growth*. Cambridge: Polity.

Lenman, James. 2000. "Consequentialism and cluelessness." *Philosophy & Public Affairs* 29 (4): 342–370.

Liu, Sebastian. "How to be indifferent." *Noûs*, forthcoming.

Mason, Elinor. 2004. "Consequentialism and the principle of indifference." *Utilitas* 16 (3): 316-321.

MacAskill, William. 2022. *What We Owe the Future*. New York: Basic Books.

MacAskill, William, Krister Bykvist, and Toby Ord. 2020. *Moral Uncertainty*. New York: Oxford University Press.

MacAskill, William and Toby Ord. 2020. "Why Maximize Expected Choiceworthiness?" *Noûs* 54 (2): 327-353.

Macaulay, Thomas Babbington. 1830. "Sir Thomas More; or, Colloquies on the Progress and Prospects of Society by Robert Southey." *Edinburgh Review* 50: 528-565.

Mearsheimer, John J. 2001. *The Tragedy of Great Power Politics*. New York: W. W. Norton.

Millett, Piers and Andrew Snyder-Beattie. 2017. "Existential risk and cost-effective biosecurity." *Health Security* 15: 373–384.

Mogensen, Andreas L. 2021. "Maximal cluelessness." *Philosophical Quarterly* 71 (1): 141-162.

Mogensen, Andreas L. 2022. "The only ethical argument for positive δ? Partiality and pure time preference." *Philosophical Studies* 179: 2731-2750.

Mogensen, Andreas L. and William MacAskill. 2021. "The paralysis argument." *Philosophers' Imprint* 21 (15): 1-17.

Mogensen, Andreas L. and David Thorstad. 2022. "Tough enough? Robust satisficing as a decision norm for long-term policy analysis." *Synthese* 200: 36. https://doi.org/10.1007/s11229-022-03566-5.

Neumark, David and William Wascher. 2006. “Minimum wages and employment: a review of evidence from the new minimum wage research.” *National Bureau of Economic Research* Working Paper 12663. https://www.nber.org/system/files/working_papers/w12663/w12663.pdf.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.

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

Russell, Jeffrey Sanford, John Hawthorne, and Lara Buchak. 2015. “Groupthink.” *Philosophical Studies* 172 (5): 1287-1309.

Schwitzgebel, Eric. 2023. “The washout argument against longtermism.” Unpublished manuscript.

Spears, Dean, Sangita Vyas, Gage Weston, and Michael Geruso. 2024. “Long-term population projections: Scenarios of low or rebounding fertility.” *PloS One* 19 (4):e0298190. doi:10.1371/journal.pone.0298190.

Steele, Katie and H. Orri Stefánsson. 2021. *Beyond Uncertainty: Reasoning with Unknown Possibilities*. Cambridge: Cambridge University Press.

Steele, Katie and H. Orri Stefánsson. 2022. “Transformative experience, awareness growth, and the limits of rational planning.” *Philosophy of Science* 89 (5): 939–48.

Tarsney, Christian. 2020. "*Normative Externalism*, by Brian Weatherson." *Mind* 130 (519): 1018-1028.

Tarsney, Christian. 2023. "The epistemic challenge to longtermism." *Synthese* 201: 195. https://doi.org/10.1007/s11229-023-04153-y.

Tarsney, Christian. 2025. "Expected value, to a point: Moral decision-making under background uncertainty." *Noûs*. https://doi.org/10.1111/nous.12544.

Thorstad, David. 2023. "High risk, low reward: A challenge to the astronomical value of existential risk mitigation." *Philosophy & Public Affairs* 51 (4): 373 -412.

Thorstad, David. "Three mistakes in the moral mathematics of existential risk." *Ethics*, forthcoming.

United Nations. 2022. *World Population Prospects 2022*. United Nations Department of Economics and Social Affairs. https://population.un.org/wpp/.

Temkin, Larry. 2022. *Being Good in a World of Need*. New York: Oxford University Press.

Tolkien, J. R. R. 1955. *The Return of the King*. London: George Allen & Unwin.

Weatherson, Brian. 1998. Decision-making with imprecise probabilities. Unpublished manuscript. http://brian.weatherson.org/vdt.pdf.

Weatherson, Brian. 2019. *Normative Externalism*. New York: Oxford University Press.

Wells, Lin. 2001. "Thoughts for the 2001 quadrennial defense review." *The Rumsfeld Archive*. https://library.rumsfeld.com/doclib/sp/2382/2001-04-

12%20To%20George%20W%20Bush%20et%20al%20re%20Predicting%20the%20Future.pdf.

Wenmackers, Sylvia and Jan-Willem Romeijn. 2016. “New theory about old evidence.” *Synthese* 193 (4): 1225–50.

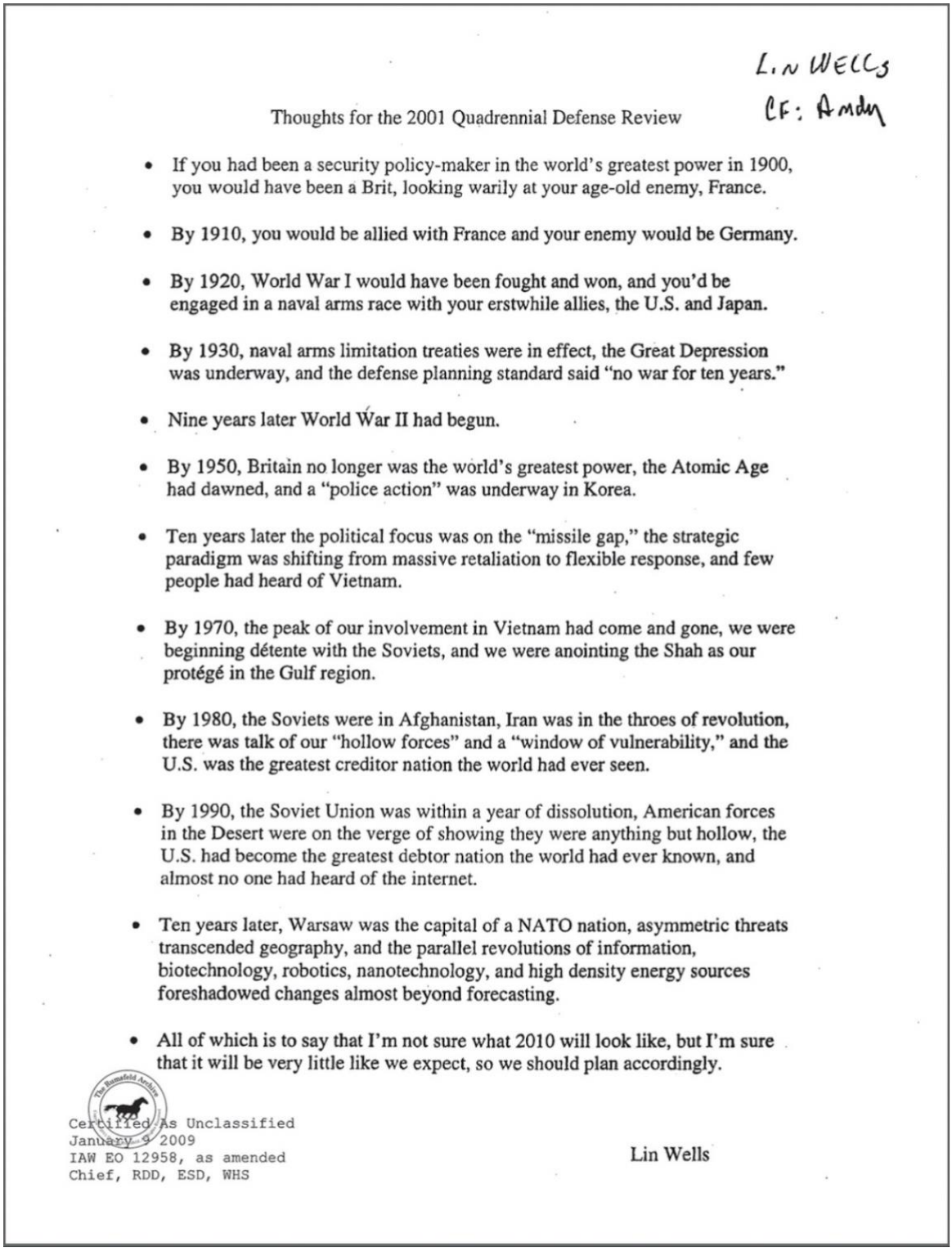
White, Roger. 2010. “Evidential symmetry and mushy credence.” In *Oxford Studies in Epistemology* vol. 3, ed. T. S. Gendler and J. Hawthorne. Oxford: Oxford University Press.

Worsnip, Alex. 2018. “The conflict of evidence and coherence.” *Philosophy and Phenomenological Research* 96 (1): 3–44.

Yim, Lok Lam. 2019. “The cluelessness objection revisited.” *Proceedings of the Aristotelian Society* 199: 321-324.

Appendix

Figure 1: Wells on the difficulty of forecasting in international politics



L. N WELLS
CF: Andy

Thoughts for the 2001 Quadrennial Defense Review

- If you had been a security policy-maker in the world's greatest power in 1900, you would have been a Brit, looking warily at your age-old enemy, France.
- By 1910, you would be allied with France and your enemy would be Germany.
- By 1920, World War I would have been fought and won, and you'd be engaged in a naval arms race with your erstwhile allies, the U.S. and Japan.
- By 1930, naval arms limitation treaties were in effect, the Great Depression was underway, and the defense planning standard said "no war for ten years."
- Nine years later World War II had begun.
- By 1950, Britain no longer was the world's greatest power, the Atomic Age had dawned, and a "police action" was underway in Korea.
- Ten years later the political focus was on the "missile gap," the strategic paradigm was shifting from massive retaliation to flexible response, and few people had heard of Vietnam.
- By 1970, the peak of our involvement in Vietnam had come and gone, we were beginning détente with the Soviets, and we were anointing the Shah as our protégé in the Gulf region.
- By 1980, the Soviets were in Afghanistan, Iran was in the throes of revolution, there was talk of our "hollow forces" and a "window of vulnerability," and the U.S. was the greatest creditor nation the world had ever seen.
- By 1990, the Soviet Union was within a year of dissolution, American forces in the Desert were on the verge of showing they were anything but hollow, the U.S. had become the greatest debtor nation the world had ever known, and almost no one had heard of the internet.
- Ten years later, Warsaw was the capital of a NATO nation, asymmetric threats transcended geography, and the parallel revolutions of information, biotechnology, robotics, nanotechnology, and high density energy sources foreshadowed changes almost beyond forecasting.
- All of which is to say that I'm not sure what 2010 will look like, but I'm sure that it will be very little like we expect, so we should plan accordingly.

The Rumsfeld Archive

Certified As Unclassified
January 9 2009
IAW EO 12958, as amended
Chief, RDD, ESD, WHS

Lin Wells