Calvin Baker

Expected Choiceworthiness and Fanaticism

Penultimate draft as of March 2024 | Forthcoming in *Philosophical Studies*; kindly cite published version

Abstract

Maximize Expected Choiceworthiness (MEC) is a theory of decision-making under moral uncertainty. It says that we ought to handle moral uncertainty in the way that Expected Value Theory (EVT) handles descriptive uncertainty. MEC inherits from EVT the problem of fanaticism. Roughly, a decision theory is fanatical when it requires our decision-making to be dominated by low-probability, high-payoff options. Proponents of MEC have offered two main lines of response. The first is that MEC should simply import whatever are the best solutions to fanaticism on offer in decision theory. The second is to propose statistical normalization as a novel solution on behalf of MEC. This paper argues that the first response is open to serious doubt and that the second response fails. As a result, MEC appears significantly less plausible when compared to competing accounts of decision-making under moral uncertainty, which are not fanatical.

Keywords: Maximize Expected Choiceworthiness (MEC), Expected Value Theory, fanaticism, infinity, normalization

1. Introduction

We face moral uncertainty when the moral theories in which we have nonzero credence issue conflicting verdicts.[1] 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.[2] 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 of the strength of moral reason to choose an option, according to a moral theory.[3] To indicate how MEC works, consider an agent who has one third of her credence in Kantian deontology and two thirds in classical utilitarianism.[4] 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

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

[2] 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.

[3] 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.

[4] 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).

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.

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.[5] 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."[6] 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.[7] 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.[8] 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:

[5] MacAskill, Bykvist, and Ord (2020: 133, 145).
[6] MacAskill, Bykvist, and Ord (2020: 133).
[7] MacAskill, Bykvist, and Ord (2020: 47-48).
[8] 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."[9]

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.[10] 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 1. (The specific numbers in the table don't matter; they're just for illustration.)

Table 1: 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,

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

[10] 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.

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*.[11] 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.[12] 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 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*.[13] 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.[14] 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

[11] On which see Wilkinson (2022), Beckstead and Thomas (2023), and Russell (2023).
[12] Due to Bostrom (2009).
[13] The phrase 'infinitarian paralysis' is due to Bostrom (2011).
[14] 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.

infinity'.[15] But adopting this decision rule would mean 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:

> "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

[15] 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 $-\infty$, and a near certainty of an enormous, though finite, 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.

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*.[16] 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?[17] 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.

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

[17] 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.

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 value.[18] 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).[19]

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.[20]

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

[18] Note that a value function can be finite but unbounded.
[19] Cf. Beckstead and Thomas (2023: §3.2).
[20] 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.

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.[21]

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.[22]
(3) So, the choiceworthiness function of every live moral theory is bounded.

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

[21] Cf. Beckstead and Thomas (2023: §6).
[22] Again, see McGee (1999) and Russell and Isaacs (2021) for arguments to this effect.

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.[23] 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 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.[24] 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

[23] 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.

[24] In this section, for simplicity, I will focus on risk aversion in the pursuit of positive choiceworthiness. However, see footnote 34 for discussion of the role of risk seekingness *vis-à-vis* negative choiceworthiness.

function dampens the probability function.[25] However, the risk function either does or does not discount sufficiently small probabilities to zero.[26] 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.[27] 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);[28] that it either violates dominance reasoning or becomes vulnerable to money pumps;[29] and that it either fails to avoid fanaticism after all or generates intransitive cycles.[30] Again, to be clear, this dilemma 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

[25] 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 utils) + 0 = 2.5 utils.

[26] 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.

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

[28] Isaacs (2016).

[29] Kosonen (2022: chapter 4).

[30] Cibinel (2023).

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).[31] 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.[32]

To see MEU in action, consider Table 2.

Table 2: 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). $U(x)$

[31] 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.

[32] 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.

approaches -100 as x approaches $-\infty$ and 100 as x approaches ∞. Feeding the choiceworthiness scores from Table 2 into U and rounding to two decimals, we get Table 3.

Table 3: 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.[33] Consider Table 4.

Table 4: High moral stakes (choiceworthiness)

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

The decision-situation depicted in Table 4 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

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

Φ. However, if we feed the choiceworthiness scores into our utility function, we get Table 5 (rounding to two decimals).

Table 5: High moral stakes (utility)

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

EU(Φ) ≈ 49; EU(Ψ) ≈ 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.[34] 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.[35] In what follows, I'll focus on uncertainty about (i) for

[34] 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.

[35] 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.

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 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 2, reproduced here for ease of reading.

Table 2: 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 6.

Table 6: barebones fanaticism (meta-choiceworthiness)[36]

	MEC	MEU

[36] 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.

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.[37] 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-order moral theory will generate infinite meta-choiceworthiness according to MEC. To illustrate, consider Table 7.

Table 7: 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 2: she has credence = 0.99999 in the moderate moral theory and credence = 0.00001 in the infinite-stakes moral theory. Then, as in Table 2, 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

[37] 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 34 for further detail.

of Ψ-ing remains practically unchanged at ≈ 5 (rounding to two decimals).[38] We then get Table 8.

Table 8: 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 8. 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.[39] 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.)

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

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

[39] 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 8 will naturally militate against choosing in accordance with MEU, modulo worries about intertheoretic comparisons between MEC and MEU.

know do not obtain—even if we have some nonzero credence in them.[40] 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.[41] *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, 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.[42] 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).[43] Arguably, longtermism follows from total welfarist decision-theoretic consequentialism.[44] 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.

[40] See e.g. Hawthorne and Stanley (2008), Weatherson (2012), Liu (2022), and Hong (fc.).
[41] MacAskill, Bykvist, and Ord (2020: 150-151). Cf. Jackson and Smith (2006).
[42] The quotation is from Greaves (2016: 313).
[43] See Greaves and MacAskill (2021: §9).
[44] 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).

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.[45] Here's how MacAskill, Bykvist, and Ord explain 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."[46]

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).)[47] Since we can't 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

[45] 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.

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

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

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.[48] 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."[49]

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.

[48] I owe this argument to Sebastian Liu (pc.).
[49] Williamson (2000: 76). I have omitted a footnote from the quoted text.

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.[50] The goal of normalization is to reflect the *principle of equal say*, 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.[51] 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."[52] Intuitively, such an agent should allocate many more of her resources (time, effort, money, etc.) to impartial altruism, which is choiceworthy according to Singer, than to

[50] 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 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).

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

[52] 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/.

bowling and drinking white Russians, which are choiceworthy according to Dudeism.[53] 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."[54] 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.

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.[55] There is no meta decision

[53] 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.
[54] MacAskill, Cotton-Barratt, and Ord (2020: 73-74).
[55] Here I paraphrase Tarsney's (2020: 1019) gloss on externalism about morality. For discussion see Russell (forthcoming) and Tarsney (forthcoming).

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.[56] In particular, we would be led on a quest to discover and assess theories reaching ever higher into the hierarchy of infinities and, in doing so, on an extended foray out of philosophy and into comparative religion and mysticism.[57] 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).[58] On My Favorite Theory, we should act in accordance with the moral theory in which we have highest credence.[59] 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.[60] 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

[56] 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.
[57] Cf. Chen and Rubio (2020: §4.2-4.4).
[58] See Harman (2015) and Weatherson (2019).
[59] Gustafsson and Torpman (2014), though see Gustafsson (2022).
[60] Newberry and Ord (2021); cf. Greaves and Cotton-Barratt (2023).

abductively for a fanatical theory over non-fanatical competitors will be left fighting an uphill battle.[61]

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.

[61] 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.

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.