Philosophical Studies
https://doi.org/10.1007/s11098-026-02579-x

In defense of moderation

Jacob Barrett[1]

Received: 9 October 2025 / Accepted: 12 June 2026
© The Author(s) 2026

Abstract
A decision theory is fanatical if it says that, for any sure thing of getting some finite amount of value, it would be better to almost certainly get nothing while having some tiny probability (no matter how small) of getting sufficiently more finite value. Fanaticism is extremely counterintuitive; common sense requires a more moderate view. However, a recent slew of arguments purport to vindicate it, claiming that moderate alternatives to fanaticism are sometimes similarly counterintuitive, face a powerful continuum argument, and violate widely accepted synchronic and diachronic consistency conditions. In this paper, I defend moderation. I show that certain arguments for fanaticism raise trouble for some versions of moderation—but not for more plausible moderate approaches. Other arguments raise more general difficulties for moderates—but fanatics face these problems too. There is therefore little reason to doubt our commonsensical commitment to moderation, and we can rest easy not worrying too much about tiny probabilities of enormous value.

Keywords Decision theory · Fanaticism · Moderation · Expected value theory · Expect utility theory · Bounded utility · Probability discounting · Tiny probabilities

1 Introduction

Certain theories that apply in ordinary contexts break down at the extremes. For example, while classical mechanics works well when dealing with moderately sized objects traveling at moderate speeds, it falters with extremely small masses or extremely high speeds.

Many otherwise plausible approaches to decision theory face a similar difficulty with extremely low probabilities and extremely high values. Suppose you wish to maximize the number of happy lives in the world, and you value such lives linearly

✉ Jacob Barrett
jacob.a.barrett@vanderbilt.edu

[1] Philosophy Department, Vanderbilt University, Nashville, TN, USA

Published online: 06 July 2026

(viewing each additional life as just as valuable as the last). You then encounter the following choice (compare Wilkinson, 2022: 445–446):

> *Long-Shot Utopia*: You can donate a large sum of money to one of two organizations. The first is an anti-malaria charity that would use your donation to certainly save a thousand happy lives. The second would use it to conduct speculative research on a mythical substance that might solve all our problems and produce a long and glorious future full of an astronomical number of happy lives. This research would almost certainly fail, with a tiny probability of succeeding. Specifically, there is a one-in-a-billion-in-a-billion (10^{-18}) probability that funding the research would produce a billion billion billion (10^{27}) happy lives. No other effects are relevant.

According to one popular decision theory, we should rank risky prospects by their "expected value": the probability-weighted sum of the values of their outcomes. Funding the speculative research has a one-in-a-billion-in-a-billion chance of producing a billion billion billion happy lives, and so the same expected value as producing a billion happy lives with certainty. This is greater than the expected value of donating to the charity, which saves only a thousand happy lives with certainty. So expected value theory says that funding the speculative research is better—even though the charity would certainly save a moderate number of lives, and the research would almost certainly do nothing.

This result is extremely counterintuitive. A decision theory provides an account of instrumental rationality. When it ranks one risky prospect as "better" than another, this means we should choose the former over the latter—at least insofar as we aim to promote value. But it seems problematically *fanatical* to pursue tiny probabilities of enormous value, taking actions that almost certainly won't work, over sure things of moderate value. We thus appear to have an obvious counterexample to expected value theory, or to any approach that similarly treats tiny probabilities of enormous value in this way:

> *Fanaticism*: for any sure thing of getting some finite amount of value, it would be better to almost certainly get nothing while having some tiny probability (no matter how small) of getting sufficiently more finite value (Wilkinson, 2022: 449; Beckstead and Thomas, 2024: 434).

Instead, we should endorse a *moderate* decision theory that rejects fanaticism, and so denies that funding the speculative research is better.

Some fanatics warn against trusting our "initial intuitions" here, since we are prone to various cognitive biases (or to misapplying otherwise reliable heuristics) when thinking about small probabilities (Wilkinson, 2022: 452–453; compare Russell and Isaacs, 2021: 186). Fair enough. But it's not just our initial intuitions that should trouble us. The more we reflect on these issues, the stronger and more considered our anti-fanatical convictions grow.

For one thing, fanaticism raises a deep theoretical problem. Suppose again that you value happy lives linearly, but must now choose between donating to the charity

and creating a St. Petersburg coin. The coin, if created, will be flipped until it lands heads—as many times as it takes. If it lands heads on the first flip, it produces 2 happy lives. Landing heads on the second flip produces 4 happy lives, landing heads on the third flip produces 8 happy lives… and so on. So, for any n, the prospect of creating the coin has a $1/2^n$ probability of yielding 2^n happy lives. How valuable is this prospect? Well, it involves infinitely many outcomes—there's *some* probability it takes *any* number of flips before landing heads—each with the probability-weighted value of 1 happy life (a ½ probability of 2 lives, a ¼ probability of 4 lives,…). The prospect's expected value therefore sums to *infinity*, such that expected value theory implies that creating the coin is better than donating to the charity—for *any* finite number of lives the charity might save (Bernoulli, 1954; Hájek and Nover, 2006: 4). This is troubling, not only because it's implausible to value the St. Petersburg prospect so highly, but also because when the coin lands heads this only ever yields finitely many lives, meaning that we must value the prospect of creating the coin *higher than any outcome it might realize*, which seems downright paradoxical. And fanaticism, it turns out, is the real culprit here. Any fanatical decision theory implies the existence of prospects with precisely this paradoxical feature, given two minimal and widely accepted further assumptions: statewise dominance (if prospect X yields a better outcome than prospect Y in every possible state of the world, then X is better than Y) and transitivity (if X is better than Y, and Y is better than Z, then X is better than Z) (Beckstead and Thomas, 2024: 446–447).

For another, fanaticism provides absurd real-world guidance. This is most obvious if we consider the possibility of outcomes that—unlike those yielded by the St. Petersburg coin—are themselves infinitely valuable (involving, say, infinitely many happy lives). For, given the same minimal assumptions, fanaticism implies that *any* probability of achieving an infinitely valuable outcome is better than any sure thing of finite value (Beckstead and Thomas, 2024: 448–449). And this suggests that we should, in practice, be entirely focused on chasing tiny probabilities of infinite value. After all, even if we are nearly certain that we can't bring infinitely valuable outcomes about, we can't be 100% sure of this. And that's all fanaticism requires to get a grip. For it would be suspiciously convenient if we could avoid this problem by claiming that *exact* symmetries between the probabilities of our actions bringing about infinite value always cancel out, as in the "many gods" reply to Pascal's wager (Askell, 2019: 515–516). In practice, then, fanaticism compels us to constantly seek out, and to pursue, tiny asymmetries in the probability of infinite value—at least insofar as we pursue value at all.

So fanaticism is extremely counterintuitive, and it would take extremely strong arguments to convince me to accept it. I imagine most readers agree. Perhaps they'd endorse fanaticism if they could be convinced that rejecting it committed them to even worse results. But few cures could be worse than this disease.

As it turns out, a recent slew of arguments purport to show precisely this: that rejecting fanaticism leads to even less tenable results than accepting it (Beckstead, 2013; Beckstead and Thomas, 2024; Cibinel, 2023; Kosonen, 2022, 2024; Wilkinson, 2022, 2024). We must embrace fanaticism, these arguments suggest, on pain of endorsing even less plausible results in other simple comparisons (Sect. 2) or in more complex "continuum" cases (Sect. 3), or of violating plausible synchronic (Sect. 4)

or diachronic (Sect. 5) consistency conditions. Moderation may be commonsensical and presumptively correct, but it commits us to violating other claims that are even more plausible and incontrovertible. We seem to be stuck with fanaticism after all.

My aim in this paper is to defend moderation. My argument proceeds as a defense in the traditional sense: I canvass the various arguments on offer for fanaticism, showing that none give us much reason to doubt, let alone abandon, our presumptive commitment to moderation. That said, these arguments do suggest important lessons. In some cases, they bear on what sort of moderates we should be, as they raise serious problems for only some forms of moderation. In others, moderates cannot plausibly avoid running into these problems. But fanatics face nearly identical problems, so that they are problems for everyone, not merely moderates.

I should be clear that while I defend moderation against fanaticism, I don't defend any specific moderate decision theory. Admittedly, as the paper progresses I'll be transparent about where my own sympathies lie, explaining why I am attracted to a view I call "difference-making discounting" (see Kosonen, 2022: ch. 4; Cibinel, 2023: 667). But it's not my aim to defend this theory in particular, and my argument doesn't stand or fall with it. Instead, my primary goal is simply to show that we lack good reasons to embrace fanaticism, such that remaining debates should turn not on whether to be a moderate but on what type of moderate to be (Sect. 6).

My secondary goal is to make this debate more accessible. As the literature stands, it's easy to get the sense that fanaticism, despite being so counterintuitive, has the balance of impenetrably technical arguments on its side. To show that this isn't so, I have tried to gather all the major issues in one place and to explain them in as close to plain English as the subject matter allows. As we clear our way through this technical thicket, we'll discover that moderation is not only more intuitive than fanaticism, but just as theoretically respectable.

2 Boundedness and probability discounting

Fanatical decision theories imply implausible verdicts about tiny probabilities of enormous value. There are two natural approaches to avoiding this. First, we might avoid massive values by appeal to "boundedness": perhaps values or "utilities" can't get large enough to make trouble. Second, we might tame tiny probabilities by appeal to "probability discounting": perhaps when probabilities get low enough, we should treat them as "negligible" and (roughly) ignore them.

In this section, I explain why boundedness, despite being the orthodox reply to fanaticism (Arrow, 1971: 64; Joyce, 1999: 37), cannot capture our anti-fanatical intuitions. However, I also show that probability discounting, and especially a version of it that discounts tiny probabilities of *making an enormously valuable difference*, is more intuitive. This sets the stage for the rest of our discussion, which mainly focuses on comparing fanaticism to probability discounting.

2.1 Boundedness

On its face, appealing to bounded value seems straightforwardly to solve the problem of fanaticism. If there's a limit to how good outcomes can be, then we don't need to worry about tiny probabilities of enormous value: such enormous values can't exist.

Unfortunately, this approach proves implausible, at least in *moral* decision-making contexts where we are aiming to promote moral value (Monton, 2019: 5; Kosonen, 2022: 32–33). For there are certain goods that are widely believed to increase moral value not only without bound, but linearly. To stick with our above example: if happy lives are valuable, then, morally speaking, each additional happy life must be just as valuable as the last, on pain of moral arbitrariness. And even if one denies this, it's even less controversial that certain *bads* decrease moral value linearly: for example, each additional *suffering* life is just as bad as the last (Beckstead and Thomas, 2024: 436–437). If our decision theory is to handle moral decision-making, then, bounded value won't do. And if it's rationally permissible to value outcomes by their moral value—as, presumably, it is—then theories of rational decision-making must permit unbounded value too.

A more promising move is to admit value lacks bounds, while embracing a view like expected utility theory, which evaluates prospects by the probability-weighted sum of the *utility* of their outcomes. The utility of an outcome is an increasing function of its value, but since this function needn't be linear, utility can be bounded even when value isn't. For example, the utility of an outcome might be its value after adjusting for risk attitudes. If so, our utility function could take this shape: increases in value or disvalue always correspond, respectively, to increases or decreases in utility, but increases in value or disvalue have steeply diminishing effects on utility, such that utility asymptotes to upper and lower bounds.

Bounded utility allows us to maintain that, strictly speaking, each additional happy (or suffering) life is just as (dis)valuable as the last. Instead, we avoid fanaticism by claiming that each additional happy (or suffering) life has less *(dis)utility* than the last, where utility rather than value is what matters to decision-making (Beckstead and Thomas, 2024: 437–439; Bottomley and Williamson, 2025). Here, we needn't adjudicate whether this solves the moral arbitrariness problem or merely pushes the bump in the rug. Regardless, there's a more definitive problem: neither version of boundedness adequately captures our anti-fanatical intuitions.

Consider a variation of our opening case, assuming, as always, that more happy lives are better (going forward I always have happy lives in mind and drop the "happy"):

> *Likely Utopia*: You can donate a large sum of money to one of two organizations. The first is an anti-malaria charity that would use your donation to certainly save a thousand lives. The second would use it to conduct research on a mythical substance that might solve all our problems and produce a long and glorious future full of an astronomical number of lives. This research would likely succeed, with some chance of failing. Specifically, there is a 0.9 probability that funding the research would produce a billion billion billion (10^{27}) lives. No other effects are relevant.

Likely Utopia is like Long-Shot Utopia, except funding the research has a large (0.9) rather than tiny (10^{-18}) probability of producing enormous value. In this case, funding the research is intuitively better. However, boundedness can't necessarily vindicate this verdict, since it makes this depend on how close we are to the upper bound of value or utility. Specifically, if saving 1,000 lives will already bring us arbitrarily close to this upper bound (because there's already lots of other value in the world), then donating to the charity will turn out better: once we are already close *enough* to the bound, any additional lives the research might produce make only an arbitrarily small difference to value or utility. Clearly, this isn't the result we are after. And the problem, too, is clear: boundedness can't deal with tiny probabilities of enormous value per se, as its resistance to fanatical verdicts overgeneralizes to cases involving larger probabilities of enormous value (Kosonen, 2022: 33).

Boundedness can also produce even worse results. Suppose you are unsure how many lives exist in the world: you are almost certainly in a big world with 10^{10} lives, but there's some tiny probability (10^{-18}) that you are in a small world containing only yourself (compare Tarsney, 2023). You must now choose between two prospects: "Big if Big" provides 10^{60} additional lives if you are in the big world; "Small if Small" provides 1 additional life if you are in the small world (compare Kosonen, 2022: 33–34; Beckstead and Thomas, 2024: 444) (Table 1):

Table 1 Big vs. Small

	Big World	Small World
Probability	$1–10^{-18}$	10^{-18}
Big if Big	10^{10} lives $+10^{60}$ lives	1 life
Small if Small	10^{10} lives	1 life +1 life

Suppose now that 10^{10} lives is arbitrarily close to the upper bound. Then, boundedness implies that Small if Small is better than Big if Big, since, if 10^{10} lives is close *enough* to the bound, even a nearly certain chance of adding 10^{60} lives to a world of 10^{10} lives provides less value or utility than a 10^{-18} probability of adding 1 life to a world with 1 life. This verdict is worse than fanatical: it favors a tiny probability of a tiny gain over a large probability of an enormous gain.

Boundedness therefore fails to capture our intuitive resistance to fanaticism and sometimes implies worse-than-fanatical verdicts. If our reason for rejecting fanaticism is that it is too unintuitive, then, boundedness doesn't seem to fit the bill. Can probability discounting do better?

2.2 Probability discounting

The rough idea behind probability discounting is also straightforward: to deal with tiny probabilities of enormous value, simply ignore such probabilities. Though less common than boundedness, probability discounting also has a long pedigree (see Monton, 2019). However, when we try to formulate it carefully we run into a question: we should discount tiny probabilities of *what*?

An initial observation is that, to avoid fanatical verdicts, it isn't enough to discount tiny probabilities of particular goods obtaining, such as some individual receiving an enormous benefit (compare Kowalczyk, 2025: 1015). After all, even if each individual has a reasonably large probability of receiving the benefit, the probability of everyone benefiting might still be so tiny as to be negligible. A natural approach, then, is to discount *outcomes* with tiny probabilities of occurring (Smith, 2014). However, this renders our evaluation of prospects extremely sensitive to how we individuate outcomes—if we individuate finely enough, every outcome is discounted. Most state-of-the-art views, then, avoid this problem by discounting not tiny probability outcomes themselves, but tiny probabilities of achieving outcomes with very high or low values (Beckstead and Thomas, 2024: 439−40; Kosonen, 2022: 164–178; though see Smith, 2024). These views also promise to better isolate our anti-fanatical intuitions, given their targeted concern with tiny probabilities *of enormous value.*

To explain, let the discounting threshold be *e*. A value is in the "left tail" of a prospect if its probability of yielding that value or less is under *e*, in the "right tail" of a prospect if its probability of yielding that value or more is under *e*, and "moderate" if in neither tail. According to *tail discounting* we should, in the first instance, evaluate prospects conditional on the assumption that they yield a moderate value (Beckstead and Thomas, 2024: 439–440). For example, if we otherwise endorse expected value theory and modify it only in this way (as I assume for simplicity), *the tail-discounted value* of prospect *X* is the expected value of *X* conditional on it yielding a moderate value. From here, we say that one prospect is better than another if its tail-discounted value is greater. Finally, when two prospects tie in tail-discounted value, we let tiny probabilities of enormous value serve as tiebreakers, claiming that the prospect with greater undiscounted expected value is better.

Tail discounting avoids fanatical results in cases like Long-Shot Utopia, assuming the discounting threshold is greater than 10^{-18}. And it avoids the implausible results of boundedness, making it more intuitive still. However, tail discounting can't capture *all* our anti-fanatical intuitions. Suppose that, as in Long-Shot Utopia, funding the speculative research has a 10^{-18} probability of producing 10^{27} lives. However, now add that, independently of your choice, there's a 0.1 baseline probability of there being 10^{27} lives anyway. Then, since 0.1 is above any plausible discounting threshold, the 10^{-18} probability that funding the research brings about this amazing outcome is no longer discounted, and we should fund the research after all (Table 2):

Table 2 Charity vs. Research

Probability	$0.9–10^{-18}$	10^{-18}	0.1
Charity	1,000 lives	1,000 lives	1,000 lives
Research	0	10^{27} lives	0
Baseline	0	0	10^{27} lives
Charity + Baseline	1,000 lives	1,000 lives	1,000 lives $+10^{27}$ lives
Research + Baseline	0	10^{27} lives	10^{27} lives

Tail discounting says that Charity is better than Research because it discounts the tiny (10^{-18}) probability of Research producing 10^{27} lives. However, it also says that Research + Baseline is better than Charity + Baseline, since it no longer discounts the

probability of the former producing 10^{27} lives: the probability of achieving an outcome that good or better is now $0.1 + 10^{-18}$. So, tail discounting can't capture the anti-fanatical intuition that we shouldn't donate to speculative research that will almost certainly *make no difference*, since it makes our evaluation sensitive to the baseline probability that at least as good an outcome is achieved regardless (Kosonen, 2022: 167–173; Cibinel, 2023: 665–666).

Despite this, tail discounting remains more intuitive than fanaticism. It avoids rendering our decision-making hostage to tiny probabilities of large finite or infinite value in our initial cases, where the probability of achieving at least as good outcomes for other reasons is also negligible. This deals with the practical problem of having to exclusively chase tiny probabilities of infinite value too, so long as we estimate that bringing about such value is almost certainly impossible. Under the same conditions, tail discounting handles the theoretical problem arising from the St. Petersburg prospect, since it lets us simply "cut off" its tails once its probabilities get low enough (more on which shortly).

Still, some may want a view that gets intuitive results in cases like Charity + Baseline and Research + Baseline, too. To articulate such a view, define the "baseline" as the status quo prospect that would obtain regardless of your choice, including every feature of the world unaffected by it. Define the "difference" a prospect makes as the outcomes yielded by that prospect minus the outcomes yielded by the baseline in every state of the world, such that, above, if "Baseline" is the baseline, then Research is the difference Research + Baseline makes. Finally, when evaluating prospects, discount tiny probabilities of making enormously positive or negative *differences*, rather than tiny probabilities of yielding enormously valuable or disvaluable outcomes, in the same way as tail discounting: one prospect is better than another if the tail-discounted value of the *difference* the former makes is greater than the tail-discounted value of the *difference* the latter makes (compare Kosonen, 2022: ch. 4; Cibinel, 2023: 667). For example, because the difference Research + Baseline makes is simply Research, and because the probability Research produces an outcome at least as good as 10^{27} lives is 10^{-18}, we discount the tiny probability Research + Baseline makes this difference, reaching the verdict that Charity + Baseline is better after all.

Call this latter approach to probability discounting "difference-making discounting." As it captures more of our anti-fanatical intuitions, it seems intuitively superior. However, we'll see that difference-making discounting also has certain features some might resist, leading them to prefer tail discounting. For this reason, I keep both views in play, sometimes treating them separately. Other times, since the two views operate identically when choosing from the "null" baseline (providing nothing with probability 1), it will be convenient to assume this null baseline and treat the two views together.

A remaining question for either view is *where* to set the threshold. Probability discounters typically answer by appeal to intuitions about cases. For example, one calibration exercise asks: how much should we value the St. Petersburg prospect (from the null baseline)? Monton (2019: 17) suggests a discounting threshold of 1 in two quadrillion. This implies discounting any values that might arise after the 46th flip, so that the prospect (which produces 2^n lives with probability $1/2^n$) has equivalent tail-discounted value to certainly saving (roughly) 46 lives. Other suggested

discounting thresholds are as high as 1 in 144,768 (Condorcet, 1785) or 1 in 10,000 (Buffon, 1777), based on calibrations involving probabilities we routinely neglect (see Monton, 2019). For our purposes, though, we needn't settle on a discounting threshold, since it's enough to note that we can avoid problematic results of fanaticism by holding that certain probabilities are clearly low enough to neglect, even if others are borderline.

3 Continuums and vagueness

This brings us to a related difficulty for moderates generally, which for probability discounters takes this form: perhaps there's *no* plausible way to set the discounting threshold. I have in mind a recent "continuum" argument for fanaticism. In this section, I first present the argument, and then show how it can be avoided if we embrace a *vague* discounting threshold—which, I suggest, is independently more plausible than a sharp threshold anyway. For purposes of this section I assume we are always adding different prospects to the null baseline, so that we can treat tail and difference-making discounting together. The distinction between them reemerges in later sections, where I consider synchronic and diachronic consistency arguments for fanaticism.

3.1 The continuum argument

Informally, the continuum argument goes like this (Beckstead, 2013: ch. 6; Beckstead and Thomas, 2024: 431–435; Wilkinson, 2022: 458–460). Intuitively, a slightly lower probability of significantly more value is better than a slightly higher probability of significantly less value. For example, a 0.99 probability of 10 lives is better than a sure thing of 1 life. Denying this seems problematically "timid" (Beckstead and Thomas, 2024: 434). But then, we should also accept that a 0.98 probability of 100 lives is better than a 0.99 probability of 10 lives, on pain of similar timidity. And we should accept that a 0.97 probability of 1,000 lives is even better, and that a 0.96 probability of 10,000 lives is better still. And, well, you see where this is going. We construct a continuum, at each step slightly lowering the probability of bringing about a good outcome while simultaneously increasing that outcome's value significantly (in this case, we are multiplying the probability by 0.99 and the value by 10). Yet if we keep doing this we eventually reach fanatical verdicts. For example, if we continue the above sequence 6,000 steps, we end up saying that a 10^{-27} probability of $10^{6{,}000}$ lives is better than every previous prospect: that a one-in-a-million-in-a-billion-in-a-trillion shot at ginormous value is better than a significant probability of moderate value.

Formally, the argument has two premises. First:

> *Anti-timidity*: for any non-zero probability p of getting finite value v, and for any fixed standard of when one probability counts as "slightly lower" than another, there is always a sufficiently large value V such that getting V with

> slightly lower (but still non-zero) probability q is better than getting v with p (Russell, 2024: 569; compare Beckstead and Thomas, 2024: 434).

Anti-timidity says that we can always compensate for a small loss in probability with a large enough gain in value. This doesn't imply that we can construct exactly the above continuum, but it does imply that we can construct a structurally similar one.

Second:

> Transitivity: for any prospects X, Y, and Z, if X is better than Y, and Y is better than Z, then X is better than Z.

Transitivity allows us to infer that each step in the continuum is better not only than the step immediately preceding it, but than all previous steps. So, while anti-timidity says that we can always make a prospect better by slightly decreasing the probability it yields a good outcome while significantly increasing that outcome's value, transitivity lets us iterate this operation until we arrive at the fanatical verdict that a tiny probability of enormous value is better than the sure thing we began with.

Transitivity is widely accepted. Rejecting it comes at serious theoretical cost, and many may prefer a transitive fanatical theory over an intransitive moderate one. Furthermore, not all arguments for fanaticism require transitivity, making the rejection of transitivity itself an insufficient reply to fanatics. For these reasons, I assume transitivity and focus on the possibility of rejecting anti-timidity.

At first glance, rejecting anti-timidity doesn't look good. Choosing a slightly higher probability of way less value over a slightly lower probability of way more value doesn't seem rational and moderate, but irrational and timid. This seems no better than fanaticism, and plausibly worse.

Specifically, if we reject anti-timidity, there must be some point where the continuum breaks. At each step we slightly decrease the probability of receiving a good outcome and significantly increase that outcome's value—until we come to a point where, if we decrease the probability further, no increase in value can compensate for this. For example, perhaps we again start with 1 life for sure, and find (by repeatedly multiplying the probability by 0.99 and the value by 10) that 10 lives with probability 0.99 is better, that 100 lives with probability 0.98 is better still, and so on, until the point where probability 10^{-16} of 10^{3665} lives is better than all previous prospects. However, here we draw a line. We say that if we reduce this probability any further, *no* increase in lives can compensate for this. Probability 10^{-16} of 10^{3665} lives isn't worse than probability 10^{-17} of 10^{3666} lives, nor, for that matter, than probability 10^{-17} of $10^{999,999,999}$ lives. Of course, rejecting anti-timidity doesn't commit us to saying that *this* is the exact point the continuum breaks, but it does say that it must break somewhere.

This is troubling for two reasons. First, it's *implausible* that there's some point where a reduction in probability cannot be compensated for by any increase in value, no matter how large. Indeed, this seems especially implausible when we realize that the probability losses in question might themselves be arbitrarily small, rendering moderates open to the charge that they—and not just fanatics—are "absurdly sensitive to tiny changes in probability" (Wilkinson, 2022: 464). Second, any specific

location of such a point seems *arbitrary*. Why should the threshold be exactly probability 10^{-17}, rather than something slightly higher or lower?

3.2 Dissolving the continuum argument

Thankfully, first appearances can be deceiving. While it might seem implausible and arbitrary that there's some point where the continuum argument breaks, I now explain that this is exactly what probability discounting predicts, if we hold that the discounting threshold at which probabilities become "negligible" is vague.

Such a claim is independently plausible: "negligible" carries all the hallmarks of vagueness (compare Peterson, 2002). First, as just noted, any precise threshold seems arbitrary. Recall how ridiculous it struck us to learn that Condorcet suggested a discounting threshold of exactly 1 in 144,768. Second, negligibility is subject to borderline cases. Some probabilities seem clearly tiny enough to be negligible; others are obviously too large to discount; and others, well, it's hard to say. Finally, and most tellingly, negligibility is susceptible to a sorites paradox. For any point where we might propose a discounting threshold, it will seem problematically fanatical not to discount probabilities just barely larger than that. And this is exactly what we should expect if the boundary between negligible and non-negligible probabilities is vague. In much the same way, it seems implausible and arbitrary that if we remove any particular grain of sand, we no longer have a heap, or that if we remove any given hair, we make someone bald. Heaps and baldness involve vagueness, and vagueness exhibits resistance to small changes: although there are clear cases falling on either side of a boundary, small changes never seem to bring us across it. Similarly, small decreases in probability never seem to cross the threshold from non-negligible to negligible. But just like in these other cases, we cannot infer from the fact that small changes never seem to cross a boundary that no such boundary exists; instead, we can only infer that if there is a boundary, it must be vague.

What exactly goes wrong in sorites arguments is a difficult question about which different theories of vagueness disagree. For example, epistemicists claim that there really is a sharp boundary which no small change ever *seems* to cross because its location is unknowable, whereas supervaluationists deny the existence of a sharp boundary altogether: a vague term like "negligible" has multiple admissible precisifications, and while certain probabilities are determinately negligible or non-negligible when this holds on all admissible precisifications, borderline cases are indeterminate (see Keefe, 2000). Here, I take no stand on such matters but simply assume the mainstream position that sorites arguments, despite their intuitive force, must go wrong somewhere: a grain of sand isn't a heap. And this suggests an intriguing possibility. If negligibility is vague and so susceptible to a sorites paradox, then perhaps we can show that the continuum argument itself relies on this sorites argument, thereby blocking the argument while simultaneously explaining why it seems so intuitive (compare Thomas, 2022 on other continuum arguments).

Now, to be clear, embracing a vague threshold doesn't yet reduce the above continuum argument to a sorites argument, since neither premise contains the vague term "negligible." But this very absence should give us pause. The core idea of probability discounting is that not all probabilities should be treated the same. Tiny probabilities

of enormous value should *not* count the same as more significant probabilities, since some really are so small as to be negligible. And now note that anti-timidity isn't as innocent as it seems. It doesn't distinguish negligible from non-negligible probabilities, so it's no wonder that it produces fanatical verdicts.

This suggests that probability discounters have independent reason to reject anti-timidity in unqualified form, and instead to endorse the following, restricted version:

> *Moderate timidity*: for any *non-negligible* probability p of getting finite value v, and for any fixed standard of when one probability counts as "slightly lower" than another, there is always a sufficiently large value V such that getting V with slightly lower (but still *non-negligible*) probability q is better than getting v with p.

Moderate timidity is just like anti-timidity, but it refers to non-negligible instead of non-zero probabilities. It agrees with anti-timidity that we can usually compensate for a tiny loss in probability with a large gain in value. But it disagrees when a tiny loss switches a probability from non-negligible to negligible.

Once we have swapped out anti-timidity for moderate timidity we can still construct a continuum argument for fanaticism through repeated application of moderate timidity, but the argument now requires an additional premise. At each step in the continuum we must now hold both:

(a) The slightly lower probability of greater value is better than the slightly higher probability of less value, *conditional on both probabilities being non-negligible;* and
(b) The slight decrease in probability doesn't switch it from non-negligible to negligible.

Like anti-timidity, moderate timidity entails that it's possible to construct a continuum that satisfies (a) at every step. But unlike anti-timidity, it makes the continuum argument rest on there being no step where (b) fails. The implausibility and arbitrariness of there being some point where the continuum breaks therefore comes to turn on the implausibility and arbitrariness of there being some point where slightly decreasing a probability crosses a threshold from non-negligible to negligible. But this, as we have seen, is precisely what we should expect if the boundary between negligible and non-negligible probabilities is vague—indeed, the claim that (b) never fails is *exactly* the soritical claim discussed above, namely, that no small decrease in probability ever crosses the boundary from non-negligible to negligible. Thus, whereas the original version of the continuum argument relying on anti-timidity begs the question against probability discounters by assuming away any distinction between negligible and non-negligible probabilities, the version that proceeds from moderate timidity reduces to a sorites argument—and this explains both why it fails and why it seems so intuitive.

The upshot is that the continuum argument can be blocked by endorsing a vague discounting threshold and replacing anti-timidity with moderate timidity, where both of these moves are independently motivated rather than *ad hoc* responses to

an objection. Still, some may see some theoretical cost to this approach. Probability discounting was always going to be less simple than fanatical theories given its extra parameter for the discounting threshold—and now probability discounters need to allow vagueness into their theory, too? But beginning with simplicity generally: while we should certainly prefer a simpler theory all else equal, we should also prefer additional complexity that's needed to adequately explain a phenomenon and get the cases right. Classical mechanics is simpler than quantum mechanics, but this is no reason to prefer it when dealing with the extreme cases where it falters—and the situation is parallel, I have suggested, when it comes to the comparison between fanaticism and moderation. To be sure, simplicity would favor fanaticism if moderation's additional complexity were explanatorily idle or accommodated unreliable intuitions. But as we saw at the outset, our anti-fanatical convictions grow stronger rather than weaker upon reflection, suggesting that fanaticism underfits the data rather than moderation overfitting it.

As for vagueness specifically: vagueness is strange, and perhaps, all else equal, we should prefer to do without it. I'm actually not so sure. While *normative* vagueness has been less widely discussed than the descriptive vagueness arising for baldness or heaps, I suspect that it's endemic and unavoidable. Just consider, for example, the fuzzy line between being innocently inattentive and culpably negligent: could there be some *precise* number of seconds one can permissibly look away from one's child before one crosses the line (Schoenfield, 2016: 262–263)? It seems more plausible that negligence, like negligibility, is vague—and arguably the same is true of many other normative distinctions, such as between "unreasonable" and "reasonable" or "excessive" and "proportional." Furthermore, if one is uncomfortable with vagueness being an intrinsic feature of our normative landscape, the epistemicist account—on which vagueness is only apparent, arising due to the unknowability of the locations of sharp thresholds—is always available. Regardless, even if we grant that vagueness carries some theoretical cost, fanaticism is extremely counterintuitive and theoretically costly itself, such that if the only cost to avoiding fanaticism is embracing vagueness, this is a cost we should bear. Probability discounting therefore remains far more plausible than fanaticism on intuitive grounds, with theoretical considerations not yet tipping the balance.

4 Synchronic consistency and separability

So far, we've seen that probability discounting, and especially difference-making discounting, best explains our anti-fanatical intuitions in simple cases, and that a vague discounting threshold extends this explanation to continuum cases. We now turn to another argument for fanaticism, which claims that moderates violate a synchronic consistency condition known as *separability*. This says, roughly, that our evaluation of prospects should depend only on features of them we affect—and not, for example, on what happens in distant regions of space and time.

Separability is well-known from population ethics, where it's often employed in an argument against average utilitarianism (McMahan, 1981: 115; Parfit, 1984: 420). According to average utilitarianism, we should maximize the average welfare of all

sentient beings. Now, suppose we can bring someone into existence with a neutral life. Average utilitarianism implies that whether we should do so depends on the average welfare level: if the average is negative, we should; if the average is positive, we shouldn't. This implausibly violates separability, because average welfare can depend on the welfare of beings in the distant past or in distant galaxies beyond our causal reach—and these aren't plausibly relevant to our choice. It's implausible, for example, that whether we should bring someone into existence now might depend on how well off ancient Egyptians were, or on the welfare of aliens in faraway galaxies.

In debates over fanaticism, separability is often formalized like this:

> *Prospect Separability*: for any prospects X, Y, and Z, X is at least as good as Y if and only if X+Z is at least as good as Y+Z (Wilkinson, 2022: 466–467; Russell, 2024: 570).

Here, "X+Z" is the prospect derived from adding X's outcomes to Z's outcomes in every state of the world. Suppose a coin will be flipped. If X gives \$5 on heads and \$10 on tails, and Z gives \$1 on heads and \$2 on tails, then X+Z gives \$6 on heads and \$12 on tails.

I now explain a recent argument that separability is inconsistent with moderation, and then a recent rebuttal suggesting that separability is inconsistent with the other premises leading to this result. From there, I consider two moves fanatics might make to revive the separability argument. I argue that the first move, which has been more widely discussed, leaves fanatics with few grounds on which to prefer their view to *tail* discounting. But I also argue that existing discussions have overlooked a second move available to fanatics—only it leaves fanatics with no grounds on which to prefer their view to *difference-making* discounting.

4.1 Separability against moderation

The argument is that moderates violate prospect separability. In fact, we've already seen that tail discounting does so. Recall the problem leading us to consider difference-making discounting: tail discounting says that Charity is better than Research, but that Research + Baseline is better than Charity + Baseline (Table 2). This violates prospect separability.

The problem generalizes: if value is unbounded (as I here assume), *all* moderate views violate prospect separability, as long as they satisfy a principle of *stochastic dominance* (Russell, 2024: 572–573). We can explain this principle as combining two ideas. The first, which we encountered a version of earlier, is that if X yields at least as good an outcome as Y in every possible state of the world, and a better outcome in some, then X "statewise dominates" Y. Second, if, for any value, X and Y have the same probability of yielding that value, then X and Y are "stochastically equivalent." We can now say that X "stochastically dominates" Y if X statewise dominates some stochastically equivalent prospect to Y, and define the principle of stochastic dominance as follows:

Stochastic Dominance: for all prospects X, Y, X is better than Y if X stochastically dominates Y, and X is equally good as Y if X is stochastically equivalent to Y.

Suppose a fair coin will be flipped, yielding the following payoffs (Table 3):

Table 3 Stochastic Dominance

	Heads	Tails
Probability	0.5	0.5
X	10 lives	5 lives
Y	9 lives	4 lives
Z	4 lives	9 lives

Here, X statewise dominates Y, since it produces more value in both states (or columns). Likewise, Y is stochastically equivalent to Z, since both provide the same probabilities of any given value. Finally, X stochastically dominates Z, since X statewise dominates Y, and Y is stochastically equivalent to Z.

Stochastic dominance is a plausible principle satisfied by most decision theories (see Tarsney, 2025), including expected value theory, expected utility theory, risk-weighted expected utility theory (Buchak, 2013), and tail discounting—indeed, one reason tail discounters treat undiscounted expected value as a tiebreaker is to satisfy this principle. It therefore seems bad news for moderates that stochastic dominance, prospect separability, and moderation are inconsistent. We can generate this inconsistency with the following recipe (Russell, 2024: 590; my presentation follows Kosonen, 2022: 17–20; see also Beckstead and Thomas, 2024: 441–442; Wilkinson, 2022: 466–473). First, if we deny fanaticism, then there's some case where a Fanatical prospect providing a small probability p of enormous value V isn't better than a Moderate prospect certainly providing moderate value v. Next, take a prospect "Increasing" that yields nothing if Fanatical pays off, but otherwise has p probability of yielding each of several linearly increasing values (none as great as V). We can now show that Fanatical + Increasing stochastically dominates Moderate + Increasing.

This is easiest to see with an example. For convenience, suppose Moderate yields a sure thing of 1 life, and Fanatical a 0.2 probability of 1,000 lives. (Of course, 0.2 isn't really small enough to raise fanatical concerns, but any probability will do, and a large probability like 0.2 lets us illustrate this with fewer columns). Applying the above recipe (Table 4):

Table 4 Stochastic Dominance Against Moderation

Probability	0.2	0.2	0.2	0.2	0.2
Fanatical	1,000 lives	0	0	0	0
Moderate	1 life	1 life	1 life	1 life	1 life
Increasing	0	1 life	2 lives	3 lives	4 lives
Fanatical + Increasing	1,000 lives	1 life	2 lives	3 lives	4 lives
Moderate + Increasing	1 life	2 lives	3 lives	4 lives	5 lives

Here, a glance at the two bottom rows reveals that Fanatical + Increasing stochastically dominates Moderate + Increasing. However, moderation implies that Fanatical isn't better than Moderate, from which it follows by prospect separability

that Fanatical + Increasing isn't better than Moderate + Increasing. So we have our inconsistency: moderate views cannot satisfy both stochastic dominance and prospect separability. This appears a powerful argument against moderation.

4.2 Separability against stochastic dominance

Once again, however, appearances are deceiving. For it has been shown that stochastic dominance and prospect separability are quite generally inconsistent (as usual, assuming unbounded value), and not just in the presence of moderation (Russell, 2024: 571). So the conflict we find here is a problem for everyone, and it provides no argument against moderation.

The conflict involves the St. Petersburg prospect, which, recall, yields 2^n lives for any probability $1/2^n$. Now suppose a new mechanism generates these payoffs. First, a pointer is spun with a 50/50 chance of landing "left" or "right." If it lands "left," it certainly yields 2 lives. If it lands "right," a coin is flipped and yields St. Petersburg payoffs (skipping the first payoff of 2 lives) depending on how long it takes to land heads. Call this prospect "Lefty." However, the twist is that another prospect, "Righty," is isomorphic to Lefty, except it certainly yields two lives if the pointer lands "right" and the coin's payoffs if it lands "left" (this example is adapted from Russell, 2024: 571–572; see also Goodsell, 2021: 425).

Our prospects therefore look like this (I omit the label "(happy) lives") (Table 5):

Table 5 Lefty vs. Righty

	Left					Right				
Flips	1	2	3	4	…	1	2	3	4	…
Probability	0.25	0.125	0.0625	0.03125	…	0.25	0.125	0.0625	0.03125	…
Lefty	2	2	2	2	…	4	8	16	32	…
Righty	4	8	16	32	…	2	2	2	2	…

By design, Lefty and Righty each have the same payoffs as the St. Petersburg prospect. They are therefore stochastically equivalent and so, by stochastic dominance, equally good.

Now, by prospect separability, since Lefty is equally good as Righty, Lefty + Lefty should also be equally good as Righty + Lefty. This is intuitive. But is it so? Let's check (Table 6):

Table 6 Lefty + Lefty vs. Righty + Lefty

	Left					Right				
Flips	1	2	3	4	…	1	2	3	4	…
Probability	0.25	0.125	0.0625	0.03125	…	0.25	0.125	0.0625	0.03125	…
Lefty	2	2	2	2	…	4	8	16	32	…
Righty	4	8	16	32	…	2	2	2	2	…
Lefty + Lefty	4	4	4	4	…	8	16	32	64	…
Righty + Lefty	6	10	18	34	…	6	10	18	34	…

Bizarrely, it turns out that Righty + Lefty stochastically dominates Lefty + Lefty. We can see this by reorganizing their payoffs to respectively yield the following stochastically equivalent prospects (Table 7):

Table 7 Lefty + Lefty* vs. Righty + Lefty*

Probability	0.5	0.25	0.125	0.0625	0.03125	…
Lefty + Lefty*	4	8	16	32	64	…
Righty + Lefty*	6	10	18	34	66	…

It follows that, according to stochastic dominance, Lefty is just as good as Righty, but Righty + Lefty is better than Lefty + Lefty. Prospect separability says this cannot be. So, stochastic dominance and prospect separability are jointly inconsistent (assuming value is unbounded), regardless of whether we are moderates or fanatics.

This is a startling result in itself. And it also defeats the separability argument for fanaticism. That separability, stochastic dominance, and moderation are inconsistent can't be an argument against moderation, when separability and stochastic dominance are jointly inconsistent (Russell, 2024: 576).

4.3 Reviving the separability argument: take one

The dialectic doesn't end here, however: fanatics might patch their argument by refining separability or stochastic dominance, arguing that while no one can satisfy these principles in full generality, only fanatics can satisfy suitably restricted versions of them. Here, I consider two ways this might go, depending on whether fanatics are more strongly committed to stochastic dominance or to (the motivation behind) separability. I argue that if they go the first way, they cannot claim a significant advantage over *tail* discounting. And if they go the second, they cannot claim any advantage over *difference-making* discounting.

The first move notes that stochastic dominance only commits fanatics to violating prospect separability in cases resembling the St. Petersburg paradox, where prospects can yield infinitely many outcomes. Thus, fanatics can at least satisfy stochastic dominance in combination with:

> *Finite Separability*: for all prospects X, Y, and Z, such that X, Y, and Z can each yield only finitely many outcomes, X is at least as good as Y if and only if X + Z is at least as good as Y + Z (Russell, 2024: 574).

Since moderates who satisfy stochastic dominance are committed to violating even this restricted version of separability, this might seem to count in favor of fanaticism.

This move doesn't, however, leave fanatics with a strong case that their view is preferable to tail discounting. (I come back to difference-making discounting shortly.) To be sure, tail discounters do violate separability in cases involving only finitely many outcomes, as we have seen. But is this a great theoretical cost? It is not. For recall that the basic idea behind separability is that our ranking of prospects should depend only on features of prospects we can affect, and not on spatiotemporal regions unaffected by our choice. Regardless of whether we think of this as a foundational constraint or as grounded in some deeper thesis—for example, about value being decomposable across causally isolated regions—any minimally adequate version of separability must be able to capture this idea by delivering the right verdicts in cases standardly used to motivate it, namely, those involving ancient Egyptians or

distant aliens. Finite separability cannot do this, since it allows that our evaluation of prospects may depend on what happens in causally isolated spatiotemporal regions whenever we assign some positive probability to there being infinitely many outcomes in them. This will occur, for example, if we assign any probability to ancient Egyptians or distant aliens having at some point flipped a St. Petersburg coin, or (less extravagantly) to their welfare levels depending on continuous variables that can take infinitely many values. And since, in practice, we can never altogether rule such possibilities out—we cannot assign probability *zero* to them, even if they are extremely unlikely—finite separability is therefore not only unmotivated but toothless (Wilkinson, 2024: 1912). It cannot deliver the verdicts motivating prospect separability in the first place, and so undercuts its own motivation, leaving no argument from finite separability to fanaticism (compare Russell, 2024: 574).

The same problem arises for another recent attempt at weakening prospect separability to prevent its conflict with stochastic dominance. In particular, a distinctive feature of the Lefty and Righty examples (Tables 5, 6 and 7) generating this conflict isn't only that the prospects in question can yield infinitely many outcomes, but also that they are probabilistically dependent on each other: whenever Lefty yields a small finite payoff Righty yields a St. Petersburg prospect, and vice versa. So perhaps we could avoid this problem by weakening prospect separability to:

> *Independent Separability*: For any prospects X, Y, Z, such that the values of Z's outcomes are probabilistically independent of the values of both X's and Y's outcomes, X is at least as good as Y if and only if X+Z is at least as good as Y+Z (Wilkinson, 2024: 1912).

However, this principle is once again unable to capture the motivating cases behind prospect separability, since it renders our evaluation of prospects sensitive to what happens in causally isolated spatiotemporal regions *whenever they contain some probabilistic dependence with the parts of the world we can affect*. For example, if the welfare levels of ancient Egyptians or distant aliens depend on certain variables about which we are uncertain, which may also affect how things will go in the future, then independent separability will again fail to rule out their welfare levels being relevant to our choices today. And this again is plausibly the case, quite generally, in practice. For example, it may be that the probabilistic dependency arises due to certain facts about fundamental physics, human nature, or the supernatural (say, whether karma or a deity who rewards good behavior and punishes bad exists) about which we cannot be 100% certain, and which introduce a correlation between the welfare levels of ancient Egyptians, distant aliens, and those we affect through our choices.

Appealing to finite or independent separability therefore fails to support fanaticism over tail discounting because neither captures the motivation behind separability—each fails to get our initial cases involving ancient Egyptians and distant aliens right, and more generally fails to apply in nearly any realistic case. Indeed, on reflection, the failure of both restrictions is unsurprising. The motivating cases are compelling precisely because the welfare of ancient Egyptians and distant aliens lies beyond our causal reach. Finite separability cannot capture these cases because its proposed restriction cross-cuts causation: whether prospect cans yield finitely or infinitely

many outcomes has nothing to do with whether they involve causally isolated spatiotemporal regions. Independent separability cannot capture these cases because it substitutes the wrong kind of independence: as is familiar from debates over causal and evidential decision theory, probabilistic independence can come apart from causal independence, and this is why independent separability predictably fails to capture separability intuitions whenever there's any correlation between causally isolated regions. Any adequate restriction of prospect separability must therefore track causal isolation more directly. Can fanatics find such a restriction that preserves the separability argument for fanaticism?

4.4 Reviving the separability argument: take two

This brings us to fanatics' second possible move: instead of restricting prospect separability in ways that no longer capture its motivation, they might stick with this motivation and see where it leads. Indeed, fanatics might note that this motivation never really supported prospect separability in unrestricted form. To capture the cases that motivate separability, we have seen, our ranking of prospects must depend only on the parts of the world we affect, so that we can ignore features unaffected by our choice—for example, those far away in space and time. But this doesn't support the stronger claim that adding two prospects X and Y to *any* third prospect Z shouldn't change our ranking of them—it only implies this in cases where Z is a causally unaffected *baseline*.

A weaker principle encoding this motivation behind separability (which I haven't seen discussed elsewhere) is:

> *Baseline Separability*: for any prospects X, Y, B, C, B + X is at least as good as B + Y when B is the baseline if and only if C + X is at least as good as C + Y when C is the baseline.

Baseline separability says that our ranking of the prospects obtained from adding X and Y to the baseline must hold constant regardless of what the baseline is. This follows from but is weaker than prospect separability, since it doesn't also imply that we can add different combinations of prospects to the baseline in a separable way. For example, if the baseline is B, baseline separability doesn't imply that B + X is better than B + Y if and only if B + X + Z is better than B + Y + Z.

An equivalent formulation of baseline separability appeals to the null baseline (probability 1 of nothing). When the baseline is null, any difference made to the baseline is simply the resulting prospect. So we can also articulate baseline separability as the idea that we can evaluate prospects obtained from making differences to a baseline *as if* those differences are complete prospects:

> *Baseline Separability*: for any prospects X, Y, B, B + X is at least as good as B + Y when B is the baseline if and only if X is at least as good as Y.

This, again, directly captures the motivation behind separability: what matters to the ranking of prospects is the parts of them we affect, not what happens at a caus-

ally unaffected baseline. And, indeed, we can now formalize what went wrong with other attempts to rescue separability in terms of their failure to accommodate baseline separability: restricting separability to prospects that have finitely many outcomes or are probabilistically independent doesn't prevent violations of baseline separability in cases where the baseline prospect can yield infinitely many outcomes or contains some probabilistic dependence with other prospects under consideration.

Baseline separability therefore seems uniquely suited to capturing the driving intuitions behind separability. And with it in mind, we can now diagnose the conflict between prospect separability and stochastic dominance—and begin to see our way toward a possible resolution. Recall that stochastic dominance says that X is better than Y if X stochastically dominates Y, and that X is equally good as Y if X is stochastically equivalent to Y. It says this generally, including in the special case where we are adding X or Y to the null baseline. So stochastic dominance, combined with baseline separability (in its second formulation), immediately entails:

> *Difference-Making Dominance*: for any prospects X, Y, B, where B is the baseline, B + X is better than B + Y if X stochastically dominates Y, and B + X is equally good as B + Y if X is stochastically equivalent to Y.

Difference-making dominance applies stochastic dominance reasoning to *differences* rather than prospects. It says that if we can make difference X or Y, and X stochastically dominates Y, then the prospect resulting from adding X to the baseline is better than the prospect resulting from adding Y.

Now recall the earlier conflict involving two St. Petersburg coins: Lefty and Righty are stochastically equivalent, but Righty + Lefty stochastically dominates Lefty + Lefty (Tables 5, 6 and 7). This yields a conflict between stochastic dominance and prospect separability, but it can only even potentially violate *baseline* separability if Lefty is the baseline (otherwise, baseline separability doesn't apply). And if Lefty is the baseline, we find something interesting. Righty + Lefty still stochastically dominates Lefty + Lefty, so baseline separability and stochastic dominance remain inconsistent. However, Righty + Lefty doesn't *difference-making* dominate Lefty + Lefty, since these now involve adding Righty or Lefty to a baseline of Lefty, and Righty doesn't stochastically dominate Lefty (but is rather stochastically equivalent to it). *Baseline* separability and *difference-making* dominance are thus mutually consistent, even though prospect separability and stochastic dominance are not. Indeed, we can now see that the conflict between prospect separability and stochastic dominance arises because the two principles together imply difference-making dominance, but stochastic dominance and difference-making dominance can come apart. Adding two stochastically equivalent prospects to the same baseline can sometimes (bizarrely!) yield one prospect that stochastically dominates the other.

This suggests that fanatics can avoid inconsistency by abandoning prospect separability and stochastic dominance in favor of baseline separability and difference-making dominance. And this seems promising. Although stochastic dominance is *prima facie* plausible, once we see it implies that adding two stochastically equivalent prospects to the same baseline sometimes yields one prospect better than the other, it becomes far from obvious that we should stick with stochastic dominance

rather than its difference-making cousin. And, anyway, we must side with difference-making dominance to maintain the motivation behind separability, as encoded in baseline separability.

At this point, fanatics might claim victory over tail discounting, which (in the example of Charity + Baseline vs. Research + Baseline) violates not only prospect but baseline separability. However, moderates have a new rejoinder available: while this is a problem for *tail* discounting, *difference-making* discounting emerges unscathed. (For that matter, so do some approaches to bounding value (Russell, 2024), which I set aside here.) For difference-making discounting says that one prospect is better than another when the tail-discounted value of the *difference* it makes to the baseline is greater. By design, then, difference-making discounting satisfies baseline separability: it always yields the same verdicts when evaluating the same differences made to any baseline.

Of course, something has to give, since we've just seen that baseline separability and stochastic dominance are inconsistent. So now's a good time to mention the dirty secret of difference-making discounting, namely, that it indeed sometimes violates stochastic dominance—siding with difference-making dominance whenever the two conflict. This follows from the fact that tail discounting satisfies stochastic dominance, and difference-making discounting ranks the prospects obtained by adding X and Y to a baseline the same way tail discounting ranks X and Y. But to see it in action, consider again the conflict between moderation, stochastic dominance, and prospect separability: moderation entails that Moderate isn't worse than Fanatical, but Fanatical + Increasing stochastically dominates Moderate + Increasing (Table 4). Now, there are two possibilities: either Increasing isn't the baseline, or it is. If Increasing isn't the baseline, then baseline separability doesn't apply and can't be violated. But if Increasing is the baseline, then difference-making discounting does *not* deem Fanatical + Increasing better than Moderate + Increasing (in violation of baseline separability), even though the former stochastically dominates the latter. For, in this case, difference-making discounting evaluates Moderate + Increasing and Fanatical + Increasing the same way that tail discounting evaluates Moderate and Fanatical. And, *ex hypothesi*, tail discounting says that Fanatical isn't better.

The upshot is that fanatics aren't the only ones who can capture the motivation behind separability by endorsing baseline separability and difference-making dominance. A version of probability discounting does the same: difference-making discounting. This second move therefore provides fanatics no advantage over moderates generally; it only works against tail discounters.

Given my strong anti-fanatical intuitions in difference-making cases, and my strong separability intuitions captured by baseline separability, I incline toward this latter route: toward the package of baseline separability, difference-making dominance, and difference-making discounting, and toward a decision theory effectively concerned with evaluating the (tail-discounted) differences we make to baselines rather than complete prospects themselves. However, I don't mean to suggest that this choice is easy, and some may find the cost of giving up ordinary stochastic dominance too high to bear (see Greaves et al., 2024). In that case, though, we are back to the earlier comparison between tail discounting and fanaticism, where fanaticism lacked an advantage. So, either way, considerations of separability don't support

fanaticism over moderation. If fanatics resolve the conflict between prospect separability and stochastic dominance by siding with stochastic dominance, they must abandon the motivation behind separability, and so the hopes of providing a compelling separability argument that fanaticism is preferable to tail discounting. And if they instead stick with this motivation and modify stochastic dominance accordingly, they can no longer mount a separability argument providing any reason to accept fanaticism over difference-making discounting.

5 Diachronic consistency and reflection

We come now to the final argument against moderation, which appeals to considerations of diachronic rather than synchronic consistency (Kosonen, 2024; compare Wilkinson, 2022: 473–476). It's well known that alternatives to expected utility theory yield inconsistent behavior in diachronic cases (Machina, 1989; Gustafsson, 2022). Specifically, agents who always choose prospects that such theories deem better are sometimes liable to "money pumps": making sequences of choices that incur certain losses relative to other possible sequences. This sometimes occurs because such theories violate a "sure thing" or a related (but stronger) "reflection" principle:

> *Reflection*: for all prospects X, Y, if X is at least as good as Y conditional on any possible answer to some question, then X is at least as good as Y unconditionally (Russell, 2024: 580).

Suppose the answer to some question is "true" or "false." If X is better than Y conditional on either answer, then reflection says that X is better than Y unconditionally. This is intuitive: if given either "true" or "false" I'll judge X better than Y, then it seems I should judge X better than Y already. However, I don't insist that we accept reflection—I only use it to help explicate certain money pumps that arise when it's violated.

There literature contains some discussion of the minimal premises needed to show that moderates violate reflection or related principles (Wilkinson, 2022: 473–476; Russell, 2024: 579–584). But for present purposes we can bypass these issues and admit that probability discounting violates them. (Bounded value or utility approaches, by contrast, can avoid this problem (Russell and Isaacs, 2021)—but as usual I set them aside.) In this section, I first explain how these violations render probability discounters open to money pumps, and how difference-making discounters are subject to another money pump. Then I explain how, much like in the case of separability, fanatics face similar difficulties. I also suggest that, this time, the ensuing dialectic resolves more simply, as fanatics' violations seem, if anything, more concerning than moderates'.

5.1 Diachronic inconsistency against probability discounting

Suppose we begin at the null baseline (so that tail and difference-making discounting agree). There are two boxes, "Risky" and "Safe," which are both "winners" with

probability 0.1 and "losers" with probability 0.9. If the boxes are winners, Risky has a 50/50 chance of containing either $1,000,000 or nothing, and Safe certainly contains $1,000. If they are losers, Risky contains a $20 consolation prize and Safe contains $10 (Table 8):

Table 8 Risky vs. Safe

	Winners		Losers
Probability	0.05	0.05	0.9
Risky	$1,000,000	$0	$20
Safe	$1,000	$1,000	$10

Assume (implausibly) that the discounting threshold is 0.05 and that we value money linearly. (I switch from lives to money for congruence with "money pump" language, and appeal to a high discounting threshold of 0.05 for simplicity.) Then, we don't need to discount Safe, and it has a tail-discounted value of $109. But we do need to discount the probability Risky yields $1,000,000, giving it a tail-discounted value of $20. So Safe is better than Risky.

However, suppose we ask: "are the boxes winners or losers?" If "losers," Risky contains $20 and Safe contains $10. So Risky is better than Safe. If "winners," Risky has a 50/50 chance of yielding $1,000,000 or nothing, and Safe contains $1,000. Since 0.5 is well above the discounting threshold, the tail-discounted value of Risky conditional on "winners" is $500,000, so Risky is again better than Safe. This violates reflection, since Safe is better than Risky unconditionally, even though Risky is better than Safe conditional on either answer to the question: "are the boxes winners or losers?"

Violations of this sort produce diachronic inconsistencies (Kosonen, 2024). Suppose I don't yet know whether the boxes are winners or losers, and I am offered Risky for free or Safe for a cost. Since I value Safe higher than Risky, there's some cost I pay for Safe. I then find out whether the boxes are winners or losers, and am offered a chance to trade boxes. Since, conditional on *either* "winners" or "losers" I value Risky more, I always accept this trade at some cost. So I am subject to a money pump: I first take Safe for some cost and then trade for Risky at some cost, when I could have initially taken Risky for free. And, crucially, this isn't because I got unlucky. I am guaranteed to pay both costs, since my reflection violation means I value Safe higher than Risky before finding out whether the boxes are winners, but Risky higher than Safe afterward—no matter what I find out.

In this example, I have chosen "myopically"—without factoring in predictions about my future choices. If I factor these in and choose in a "sophisticated" way (via "backward induction"), I avoid this problem: anticipating that I'll trade for Risky once I discover whether the boxes are winners, I take Risky for free from the start. However, a tweak traps sophisticated choosers in a money pump too (compare Briggs, 2015). This time, I can either pay some cost for Safe now, or else wait to find out whether the boxes are winners before taking either for free. Since I am a sophisticated chooser, I forecast ahead, realizing that if I wait to find out, my future self will choose Risky (since I'll value it higher regardless of what I discover). From my current perspective, this means that waiting to find out amounts to getting Risky, which I value lower than Safe. So rather than waiting, I pay some cost to take Safe from

the outset. But this means that like my myopic counterpart, I am money pumped: I pay some cost to get Safe when I could have had it for free, had I simply waited to discover whether the boxes were winners before taking it.

Both tail and difference-making discounters face the above problem. But only difference-making discounters face another sort of money pump, which doesn't rely on reflection violations but instead on the idea that changing the baseline sometimes alters how difference-making discounters evaluate prospects (Cibinel, 2023: 667–668). Consider three prospects. Cold certainly provides $1,000,000. Medium has a 0.95 probability of providing $1,000,000 but comes with a certain bonus of $50. Hot has a 0.9 probability of providing $1,000,000 but comes with a certain bonus of $100 (Table 9):

Table 9 Cold vs. Medium vs. Hot

Probability	0.9	0.05	0.05
Cold	$1,000,000	$1,000,000	$1,000,000
Medium	$1,000,000 + $50	$1,000,000 + $50	$50
Hot	$1,000,000 + $100	$100	$100

Suppose (again implausibly) that the probability discounting threshold is 0.05 and that I value money linearly. From a baseline of Cold, I am offered the choice to keep Cold for free or trade for Medium at some cost. Relative to the baseline, Medium guarantees me $50 more but reduces my chance of getting $1,000,000 by 0.05. Difference-making discounting says to discount this small reduction in probability and treat Medium like a pure gain of $50. So I take the trade at some cost. But now that Medium is the baseline, I am offered a trade for Hot. Relative to this new baseline, Hot provides a sure gain of $50 and only reduces the chance of getting $1,000,000 by 0.05. Again, difference-making discounting says to discount the small reduction in probability and accept the trade at some cost. Finally, now that Hot is the baseline, I am offered a trade for Cold. Cold provides $100 less for sure, but also gives me a 0.1 greater probability of getting $1,000,000. Since 0.1 is now above the discounting threshold, I don't discount this and value Cold significantly more than Hot. So I accept the trade for Cold at some cost, and end up right back with Cold—minus the three costs I paid. I am thus money pumped, as I could have simply kept Cold for free.

Interestingly, difference-making discounters face the same type of pump in a diachronic version of the continuum argument (which, this time, even vagueness can't prevent). Starting from a baseline providing a sure thing of some finite amount of value, difference-making discounters accept a costly trade for a prospect providing a slightly lower probability of significantly more value. Then from this new baseline, they again accept a costly trade for a prospect providing a slightly lower probability of significantly more value, and so on, until they eventually end up with a prospect providing a tiny probability of enormous value—which they value less than the sure thing prospect they started with. So they trade back for their initial prospect at some cost.

In these two examples, I again assume myopic choice. But as before, simple tweaks render sophisticated choosers subject to money pumps too. For reasons of space, I leave the development of sophisticated pumps as an exercise to the reader. Going forward, I similarly focus on myopic pumps, since those familiar with money pumps can easily transform myopic pumps into sophisticated pumps, whereas myopic variants tend to be more intuitive for those lacking familiarity.

There is a large literature on money pumps, their normative significance, and how to avoid them (Gustafsson, 2022). Basically, there are three reactions to money pumps such as the above. One can accept that they count against the theory producing them. One can deny this. Or one can deny that the pumps really arise, for example, by abandoning myopic and sophisticated choice for *resolute* choice: at the outset, one commits to the best sequence of choices, and then follows through on that commitment when facing subsequent offers (McClennen, 1990). Thankfully, there's no need to wade into these debates here, because, as I now explain, fanatics face no less troubling problems than probability discounters, regardless of how troubling they ultimately are.

5.2 Diachronic inconsistency against fanaticism

The problem, again, relates to the St. Petersburg prospect. Recall that fanatics are committed to the existence of prospects with precisely its paradoxical feature: despite only ever yielding finitely valued outcomes, it's better than any finite outcome, and so better than any outcome it might yield (Beckstead and Thomas, 2024: 446–447). This, we can now add, also commits fanatics to violating reflection, as is most easily seen if we assume (for simplicity) that they accept this of the original St. Petersburg prospect, and compare it to a prospect with slightly better payoffs—one that, for any n, provides 2^n+1 lives with probability $1/2^n$ (Russell and Isaacs, 2021: 179–182). Call the former prospect Minus and the latter prospect Plus (reinterpreting these prospects to involve money, which we value linearly). Then we have (Tables 10 and 11):

Table 10 Minus

# Flips	1	2	3	4	5	…
Probability	0.5	0.25	0.125	0.0625	0.03125	…
Minus	$2	$4	$8	$16	$32	…

Table 11 Plus

# Flips	1	2	3	4	5	…
Probability	0.5	0.25	0.125	0.0625	0.03125	…
Plus	$3	$5	$9	$17	$33	…

Here, the payoffs of Minus and Plus are uncorrelated: each is a different coin.

Now, since Minus is the original St. Petersburg prospect, fanatics evaluate it as better than each outcome Plus might realize: Minus is better than any finite outcome and Plus can only realize finite outcomes. So, given any answer to the question, "what outcome will Plus realize?", Minus is better than Plus. By reflection, Minus should therefore be better than Plus unconditionally. But it's not. By stochastic (or difference-making) dominance the reverse holds: Plus is better than Minus.

This generates an especially troubling money pump for fanatics (Russell and Isaacs, 2021: 179–182). Suppose I begin with Minus, which I can either keep for free or trade for Plus at some cost. Since Plus is slightly better than Minus, there's some small cost I pay for this trade. However, it's then revealed what outcome Plus realizes, and I am offered a swap back for Minus. Since Minus is better than any outcome Plus might yield, it's always better to swap back. In fact, since Minus is better than *any* finite outcome, I should pay *any* arbitrarily large finite cost to swap back for Minus. So I agree to pay any cost, and end up back where I started with Minus, having paid first a small cost to swap for Plus, and then any arbitrarily large cost to swap back. I am thus money pumped, since initially I could have kept Minus for free. (As usual, sophisticated choosers face a similar pump.)

This leaves us in a similar dialectical situation as with separability. Probability discounters face diachronic inconsistencies or money pumps, which at first seems like a major strike against them. But fanatics turn out to face such violations too, such that diachronic inconsistencies don't favor fanaticism after all.

We might therefore wonder whether, as in the case of separability, fanatics might try to repair their argument by noting that, since their money pumps arise only for prospects with infinitely many possible outcomes, their diachronic inconsistencies are less worrying than moderates'. But this move fares no better here, as we can see by comparing the different sorts of money pump each side faces. For, first, with respect to *severity*, the pumps fanatics face seem worse than those arising for probability discounters, as they involve being pumped for arbitrarily large amounts. Probability discounters avoid these especially bad pumps, since discounting the small probabilities in the tails of Plus and Minus eliminates the paradoxical feature generating the pumps, namely, that such prospects are better than any finite outcomes despite only ever yielding them. And second, if we ask *why* the money pumps arise, the source of fanatics' pumps seems no less troubling. Admittedly, difference-making discounters face an additional sort of pump generated by a shifting baseline, which might somewhat count against the view. But both tail discounters and fanatics are subject to money pumps arising for the exact same reason: they violate reflection. And while it's true that fanatics' reflection violations only arise in cases where prospects involve infinitely many possible outcomes, fanatics can't dismiss these violations on such grounds, since reflection is no less plausible in these cases. For example, if X is better than Y given any answer to the question, "how tall is so-and-so?," the conclusion that X is better than Y unconditionally is just as compelling regardless of whether height is continuous—regardless, that is, of whether the question has finitely or infinitely many possible answers (corresponding to finitely or infinitely many outcomes). Furthermore, like the restricted versions of separability above, restricting reflection to cases involving finitely many outcomes would render the principle toothless in practice, since the existence of infinitely many outcomes follows from us assigning *any* probability to there being something generating them, such as a St. Petersburg-like process or a continuous variable. And we can never altogether rule this out.

Overall, then, considerations of diachronic consistency provide little if any support for fanaticism—and may even count in favor of moderation. Fanatics face more severe pumps than tail discounters', of an equally troubling source. And while difference-making discounters face an additional sort of pump, even here the severity

of fanatics' pumps remains a significant strike against fanaticism, suggesting that difference-making discounters at least do no worse.

6 Conclusion

We began by observing that moderation is far more intuitive than fanaticism, and that fanaticism raises deep theoretical and practical problems. Indeed, our starting idea was that fanaticism is so implausible that it would take very strong arguments to persuade us to endorse it. Since then, we've sought out such arguments and come up empty.

The first argument concerned boundedness. Here, we saw that while boundedness yields even worse verdicts than fanaticism, probability discounting does better. Specifically, tail discounting is more intuitive than fanaticism, and difference-making discounting is more intuitive still.

We next considered a continuum argument for fanaticism, which we saw can be blocked by a vague discounting threshold. Some may think vagueness theoretically costly, but if the best we can say in favor of fanaticism is that it avoids vagueness, this is hardly enough to defeat moderation.

From here, we indeed found little further for fanatics to say. While moderates can't satisfy prospect separability without violating stochastic dominance, neither can fanatics. If we hold onto stochastic dominance and weaken separability accordingly, we can no longer accommodate the motivating cases that support separability, leaving us with no argument for fanaticism over *tail* discounting. If we instead stick with the motivation behind separability, we can maintain consistency by swapping out prospect separability and stochastic dominance for baseline separability and difference-making dominance. This rules out tail discounting, but leaves fanaticism with no advantage over *difference-making* discounting.

Finally, we saw that while probability discounters are diachronically inconsistent or face money pumps, fanatics are subject to money pumps that are, at least in one respect, worse. This leaves us with little new reason to prefer fanaticism, and perhaps some reason to prefer moderation.

The upshot is that there's little if any reason to resist our commonsensical rejection of fanaticism. Probability discounting is far more intuitive and no less theoretically respectable. Although it would be nice to have a neat theory satisfying unbounded value and utility, prospect separability, stochastic dominance, and diachronic consistency, such a theory is simply not in the cards. No theory in this vicinity is wholly satisfactory. But theory choice is always a matter of comparative evaluation. And moderation fares far better in this comparison than fanaticism.

This doesn't settle what decision theory to accept. For one thing, despite raising the initial worry that boundedness cannot capture our anti-fanatical intuitions, I haven't carefully evaluated its merits or recent attempts to motivate it as a solution to fanaticism (Bottomley and Williamson, 2025), and some may think it proves best on theoretical grounds—indeed, I've noted parenthetically throughout that certain versions of boundedness can avoid separability violations (Russell, 2024) and diachronic inconsistencies (Russell and Isaacs, 2021). For another, while I incline

toward difference-making discounting, I haven't systematically compared it to tail discounting, which some might prefer since it satisfies stochastic dominance and faces fewer diachronic inconsistencies. For that matter, I haven't explored the full range of moderate approaches, nor how such approaches might be combined with other deviations from expected value theory.

So there's room for more work on what sort of moderate to be. Nevertheless, I hope to have shown that this is where the debate should go: since at least *some* versions of moderation are preferable to fanaticism, we can eliminate fanaticism from contention. Thus, my defense of moderation is complete, and we can rest easy not worrying too much about tiny probabilities of enormous value.

Acknowledgements For helpful comments and discussion, I would like to thank Adam Bales, Kenny Easwaran, Andreas Mogensen, Sarah Raskoff, Jeffrey Russell, Bradford Saad, Andreas Schmidt, Rhys Southan, Christian Tarsney, David Thorstad, John Weymark, Hayden Wilkinson, Timothy Luke Williamson, and especially Teruji Thomas. Thanks also to two anonymous referees and to an audience at the 12th Oxford Workshop on Global Priorities Research at The University of Oxford.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Arrow, K, J. (1971). *Essays in the theory of risk-bearing*. Markham.

Askell, A. (2019). Prudential objections to atheism. In G. Oppy (Ed.), *A companion to atheism and philosophy* (pp. 506-520). Wiley-Blackwell. https://doi.org/10.1002/9781119119302.ch33

Beckstead, N. (2013). *On the overwhelming importance of shaping the far future* [Doctoral Dissertation, Rutgers University]. https://doi.org/10.7282/T35M649

Beckstead, N., & Thomas, T. (2024). A paradox for tiny probabilities and enormous values. *Noûs*, *58*, 431–455. https://doi.org/10.1111/nous.12462

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*, 23–36. https://doi.org/10.2307/1909829

Bottomley, C., & Williamson, T. L. (2025). On the offence against fanaticism. *Ethics*, *135*, 320–332. https://doi.org/10.1086/732617

Briggs, R. (2015). Costs of abandoning the sure-thing principle. *Canadian Journal of Philosophy*, *45*, 827–840. https://doi.org/10.1080/00455091.2015.1122387

Buchak, L. (2013). *Risk and rationality*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672165.001.0001

Buffon, G-L. L. de. (1777). Essai d'arithmétique morale. In *Histoire naturelle, générale et particulière, Supplément, Tome IV* (pp. 46–148). Imprimerie Royale.

Cibinel, P. (2023). A dilemma for Nicolausian discounting. *Analysis*, *83*, 662–672. https://doi.org/10.1093/analys/anac095

Condorcet, M. de. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale.

Goodsell, Z. (2021). A St Petersburg paradox for risky welfare aggregation. *Analysis*, *81*, 420-426. https://doi.org/10.1093/analys/anaa079
Greaves, H., Thomas, T., Mogensen, A., & MacAskill, W. (2024). On the desire to make a difference. *Philosophical Studies*, *181*, 1599–1626. https://doi.org/10.1007/s11098-024-02102-0
Gustafsson, J. E. (2022). *Money-pump arguments*. Cambridge University Press. https://doi.org/10.1017/9781108754750
Hájek, A., Nover, H. (2006). Perplexing expectations. *Mind*, *115*, 703–720. https://doi.org/10.1093/mind/fzl703
Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511498497
Keefe, R. (2000). *Theories of vagueness.* Cambridge University Press.
Kosonen, P. (2022). *Tiny probabilities of vast value* [Doctoral Dissertation, University of Oxford].
Kosonen, P. (2024). Probability discounting and money pumps. *Philosophy and Phenomenological Research*, *109*, 593–611. https://doi.org/10.1111/phpr.13053
Kowalczyk, K. (2025). Saving fanaticism. *Australasian Journal of Philosophy*, *103*, 1011–1023. https://doi.org/10.1080/00048402.2025.2489120
Machina, M. J. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, *27*, 1622–1668.
McClennen, E. F. (1990). *Rationality and dynamic choice: Foundational explorations.* Cambridge University Press.
McMahan, J. (1981). Problems of population theory. *Ethics*, *92*, 96–127. https://doi.org/10.1086/292301
Monton, B. (2019). How to avoid maximizing expected utility. *Philosophers' Imprint*, *19*, 1–25.
Parfit, D. (1984). *Reasons and persons.* Oxford University Press. https://doi.org/10.1093/019824908X.001.0001
Peterson, M. (2002). What is a *de minimis* risk? *Risk Management*, *4*, 47–55. https://doi.org/10.1057/palgrave.rm.8240118
Russell, J. S. (2024). On two arguments for fanaticism. *Noûs*, *58*, 565–595. https://doi.org/10.1111/nous.12461
Russell, J. S., & Isaacs, Y. (2021). Infinite prospects. *Philosophy and Phenomenological Research*, *103*, 178–198. https://doi.org/10.1111/phpr.12704
Schoenfield, M. (2016). Moral vagueness is ontic vagueness. *Ethics*, *126*, 257–282. https://doi.org/10.1086/683541
Smith, N. J. J. (2014). Is evaluative compositionality a requirement of rationality? *Mind*, *123*, 457–502. https://doi.org/10.1093/mind/fzu072
Smith, M. (2024). Decision theory and *de minimis* risk. *Erkenntnis*, *89*, 2169–2192. https://doi.org/10.1007/s10670-022-00624-9
Tarsney, C. (2023). Average utilitarianism implies solipsistic egoism. *Australasian Journal of Philosophy*, *101*, 140–151. https://doi.org/10.1080/00048402.2021.1962375
Tarsney, C. (2025). Expected value, to a point: Moral decision-making under background uncertainty. *Noûs*, *59*, 1093–1125. https://doi.org/10.1111/nous.12544
Thomas, T. (2022). Are spectrum arguments defused by Vagueness? *Australasian Journal of Philosophy*, *100*, 743–757. https://doi.org/10.1080/00048402.2021.1920622
Wilkinson, H. (2022). In defense of fanaticism. *Ethics*, *132*, 445–477. https://doi.org/10.1086/716869
Wilkinson, H. (2024). Egyptology and fanaticism. *Philosophical Studies*, *181*, 1903–1923. https://doi.org/10.1007/s11098-024-02180-0

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.