In Defence of Moderation

Jacob Barrett (Vanderbilt University)

Global Priorities Institute | December 2024

GPI Working Paper No . 32-2024

Please cite this working paper as: Barrett, J. In Defence of Moderation. *Global Priorities Institute Working Paper Series, No. 32-2024.* Available at https://globalprioritiesinstitute.org/in-defence-of-moderation-jacob-barrett

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

In Defense of Moderation

Abstract: A decision theory is fanatical if it says that, for any sure thing of getting some finite amount of value, it would always 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.

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. For example, suppose you wish to maximize the number of happy lives in the world, and you value such lives linearly (you view 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 your donation 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 is almost certain to fail, but there is a tiny probability it succeeds. Specifically, there is a one in a billon billion (10^{-18}) probability that funding the research would produce a billion billion billion (10^{27}) happy lives. No other effects of either option are relevant.

According to one popular decision theory, we should rank risky prospects by their "expected value": the probability-weighted sum of the value 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 the following way:

> *Fanaticism*: for any sure thing of getting some finite amount of value, it would always 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 2023: 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 when thinking about probabilities generally, and small probabilities specifically (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 you must now choose between donating to the anti-malaria 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, …). Standard decision-theoretic reasoning therefore implies that the prospect is better than any finitely valued outcome, so that creating the coin is better than donating to the charity—no matter how many lives the charity would 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 this implies valuing the prospect higher *than any outcome it might realize*, which seems downright paradoxical. And fanaticism, it turns out, is the real culprit here: it implies the existence of prospects with precisely this paradoxical feature, given minimal and widely accepted further assumptions (Beckstead and Thomas 2023: 446-447).

For another, taking fanaticism's real-world implications seriously yields absurd practical guidance. This is most obvious if we consider the possibility of not only enormous finite value, but infinitely valuable outcomes (involving, say, infinitely many happy lives). For, given the same minimal assumptions, fanaticism implies that *any* probability of infinite value is better than any sure thing of

finite value (Beckstead and Thomas 2023: 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 highly confident that infinite value is impossible or that we can't bring it 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 cancel out in every case, 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 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 would 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, Kosonen 2024, Wilkinson 2022, Wilkinson 2024). We must embrace fanaticism, this literature suggests, on pain of endorsing even less plausible results in other simple comparisons (section 2) or in more complex "continuum" cases (section 3), or of violating plausible synchronic (section 4) or diachronic (section 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. At the same time, however, we'll see that these arguments suggest important lessons. In some cases, they constrain

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 from the outset that while I defend moderation against fanaticism, I do not defend any *particular* moderate decision theory. Admittedly, as the paper progresses I will be transparent about where my own sympathies lie, and will explain why I myself am most attracted to a view I introduce in the next section as "difference-making discounting" (Kosonen 2022: ch. 4, Cibinel 2023: 667). But it is not my aim to defend this theory in particular, and the success of my argument does not stand or fall with it. Instead, my ambition is only 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 (section 6).

2. Boundedness and Probability Discounting

Fanatical decision theories yield 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) round them down to zero.

In this section, I briefly 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 versions of it that discount tiny probabilities of *making an enormously valuable difference*, is much more intuitive. This sets the stage for the rest of our discussion, which mainly focuses on the comparison between fanaticism and 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 not only without bound, but linearly in moral value. 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 linearly in value: for example, each additional *suffering* life is just as bad as the last (Beckstead and Thomas 2023: 436-437). If our decision theory is to handle moral decision-making, then, appealing to 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 allow for 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 moral 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 marginal utility, such that utility asymptotes to upper and lower bounds (Beckstead and Thomas 2023: 437-439).

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. Here, we needn't adjudicate whether this really solves the moral arbitrariness problem or merely pushes the bump in the rug. For, 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 your donation 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 will likely succeed, but there is some chance it fails. Specifically, there is a 0.9 probability that funding the research would produce a billion billion billion (10^{27}) lives. No other effects of either option 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 implies that all depends on how close we are to the upper bound of value or utility. Specifically, if saving 1000 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 can 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 moderate or large probabilities of enormous value (Kosonen 2022: 33).

Even worse, boundedness sometimes favors a tiny probability of a tiny gain in value over a

large probability of enormous gain. Suppose you are uncertain how many happy 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}) you are actually 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 2023: 444):

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

Table 1: Big World vs. Small World

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 massive probability of a massive gain.

Boundedness therefore fails to capture our intuitive resistance to fanaticism and sometimes yields worse-than-fanatical verdicts. At least 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 straightforward: to deal with tiny probabilities of enormous value, simply ignore such tiny 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*?

A natural approach is to discount *outcomes* with a tiny probability 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, but tiny probabilities of achieving very high or low values (Beckstead and Thomas 2023: 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. Now, let's say that 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 2023: 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, allowing undiscounted expected value to serve as a tiebreaker.

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. For example, 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 0.01 baseline probability of there being 10^{27} lives anyway. Then, since 0.01 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.

This may be clearer in a table:

Probability	$0.99\text{-}10^{-18}$	10^{-18}	0.01
Charity	1000 lives	1000 lives	1000 lives
Research	0	10^{27} lives	0
Baseline	0	0	10^{27} lives
Charity+Baseline	1000 lives	1000 lives	1000 lives $+10^{27}$ lives
Research+Baseline	0	10^{27} lives	10^{27} lives

Table 2: Charity vs. Research

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 as good or better is now $0.01+10^{-18}$. 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 the cases we began with, 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" the tails of this prospect once its probabilities get low enough (more on which in a moment).

Still, some may want a view that gets the intuitive result 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 all the features of the world unaffected by your

decision. Now, 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, we 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 (Kosonen 2022: ch. 4 and Cibinel 2023: 667 consider similar views). 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 wish to resist, leading them to prefer tail discounting. For this reason, I keep both views in play, sometimes treating them separately. At the same time, since the two views operate identically when choosing from a "null" baseline (providing nothing with probability 1), it will also sometimes be convenient to assume a 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 a null baseline)? Monton (2019: 17) suggests a discounting threshold of 1 in two quadrillion. This implies discounting any values that might arise after the 46^{th} 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

include values as high as 1 in 144,768 (Condorcet 1785) or 1 in 10,000 (Buffon 1777), based on calibrations involving mundane 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 harder cases.

3. Continuums and Vagueness

This brings us to a related difficulty for moderates generally, which in the context of probability discounting 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. For purposes of this section I assume we are always adding different prospects to a null baseline, so that we can treat tail discounting 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 2023: 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 2023: 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.

More formally, the argument has two premises. First:

> *Anti-timidity:* for any non-zero probability p of getting finite value v, and for some 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 2023: 569; compare Beckstead and Thomas 2023: 434).

Anti-timidity says 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, we need:

> 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 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 a widely accepted consistency condition. Rejecting it comes at serious theoretical cost, and many may prefer to endorse a transitive fanatical theory than an intransitive moderate one.

Furthermore, not all arguments for fanaticism require transitivity, making the rejection of transitivity itself an insufficient reply to the fanatic. For these reasons—and to keep things manageable—I assume transitivity and focus on the possibility of rejecting anti-timidity.

At first glance, rejecting anti-timidity doesn't look very 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.

In more detail, if we reject anti-timidity, there must be some point where the above 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 some such point must exist.

Intuitively, the existence of such a point 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. This, ironically, renders 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 some slightly higher or lower probability?

3.2. Dissolving the Continuum Argument

Thankfully, first appearances can be deceiving. For 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 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 moderates shouldn't accept anti-timidity in unqualified form, but rather the following, restricted version:

> *Moderate timidity*: for any *non-negligible* probability p of getting finite value v, and for some 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 like anti-timidity, but it refers to non-negligible in place 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.

Does swapping out anti-timidity for moderate timidity block the continuum argument? It at least transforms it. For 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 a point where the continuum breaks 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.

Seen in this light, the continuum argument reveals itself to rely on a sorites paradox (compare Thomas 2022 on other continuum arguments). Although (b) may seem plausible and hard to resist in any pairwise comparison between two close probabilities, 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 a threshold from non-negligible to negligible, and anti-timidity seems to hold. But just like in other cases of vagueness, we cannot infer from the fact that very small changes never *seem* to cross a boundary (each instance of (b) *seems* true) that there is no boundary to cross.

What exactly goes wrong in sorites arguments is a difficult question that different theories of vagueness answer in different ways (see Thomas 2022: 748-749). But for present purposes it's enough to note that, regardless of how exactly to explain their failure, we know sorites arguments don't work, so we know the continuum argument from moderate timidity doesn't work either. Although it seems implausible and arbitrary that there's any point where the continuum breaks, this is exactly the implausibility and arbitrariness a vague threshold predicts. Furthermore, a vague discounting threshold yields two further predictions: first, that when comparing two obviously non-negligible probabilities, a slightly lower probability of way more value is always better; and second, that when comparing an obviously non-negligible to an obviously negligible probability, the negligible probability isn't better. These two predictions track exactly our intuitions, suggesting that a vague threshold not only dissolves the continuum argument but best explains our otherwise puzzling intuitions. Fanatics cannot similarly explain our intuitions as they bite the bullet and deny the latter.

The upshot is that the continuum argument can be blocked by replacing anti-timidity with moderate timidity and endorsing a vague discounting threshold. Now, some may see some theoretical cost to this: vagueness is strange, and, all else equal, we might prefer to do without it. I'm not so sure: normative vagueness may simply be endemic and unavoidable. Consider, for example, the fuzzy line between being innocently inattentive and culpably negligent: could there be a *precise* number of seconds one can permissibly look away from one's child before one crosses the line (Schoenfield 2016: 262-263)? But regardless, even if 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. Furthermore, we've seen that our anti-fanatical intuitions point us toward probability discounting, and if we endorse such a view, a vague threshold is more plausible than a sharp one independently of the continuum argument (Peterson 2022). Just recall how ridiculous it struck us to learn above that Condorcet suggested a discounting threshold of exactly 1 in 144,768. Probability discounting with a vague threshold is therefore far more plausible than either fanaticism or probability discounting with a sharp threshold, at least on intuitive grounds.

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 can affect—and not, for example, on what's going on in distant regions of space and time.

Separability is well-known from population ethics, where it's often wheeled out 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 bring them into existence; 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 the 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 2023: 570).

Here, "X+Z" is the prospect derived by adding X's outcomes to Z's outcomes in every state of the world. For example, 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 of this argument suggesting that separability is inconsistent with the other premises needed for 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 the fanatic with little grounds on which to prefer their view to *tail* discounting. But I also argue that existing discussions have overlooked a second, perhaps more plausible move available to fanatics—only this second move 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 straightforwardly 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 widely endorsed principle of *stochastic dominance* (Russell 2023: 572-573). We can explain this principle as the combination of two ideas. First, if, X is sure to turn out at least as well as Y, and it might turn out better, then X "statewise dominates" Y. Second, if, for any value, X and Y have the same probability of yielding that value, X and Y are "stochastically equivalent." Finally, 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.

For example, suppose a fair coin will be flipped, yielding the following payoffs:

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

Table 3: Stochastic Dominance

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 dominate 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 2020), including expected value theory, expected utility theory, risk-weighted expected utility theory (Buchak 2013), and tail discounting. It therefore seems bad news for the moderate that stochastic dominance, prospect separability, and moderation are inconsistent. We can generate this inconsistency

with the following recipe (Russell 2023: 590; my presentation follows Kosonen 2022: 17-20; see also Beckstead and Thomas 2023: 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 isn't better than a Moderate prospect providing a sure thing of moderate value v. Next, take a prospect "Increasing" that yields nothing if Fanatical pays off, but has p probability of yielding each of several linearly increasing values (none as great as V) when Fanatical doesn't pay off. It can now be shown that Fanatical+Increasing stochastically dominates Moderate+Increasing.

This is easiest to see with a concrete example. For convenience, suppose Moderate yields a sure thing of 1 life, and Fanatical a 0.2 probability of 1000 lives. (Of course, 0.2 isn't really a small enough probability 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:

Probability	0.2	0.2	0.2	0.2	0.2
Fanatical	1000 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	1000 lives	1 life	2 lives	3 lives	4 lives
Moderate+ Increasing	1 life	2 lives	3 lives	4 lives	5 lives

Table 4: Stochastic Dominance Against Moderation

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 can't be 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 with each other (as usual, assuming unbounded value), and not just in the presence of moderation (Russell 2023: 571). So the conflict we find here is a problem for everyone, and fanatics can't wield it in an argument against moderation.

This conflict relates to the St. Petersburg prospect, which, recall, yields 2^n lives for any probability $1/2^n$. Now suppose a complicated 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 "Leftie." However, the twist is that another prospect, "Rightie" is isomorphic to Leftie, except that it yields two lives for sure if the same pointer lands "right," and St. Petersburg payoffs depending on the same coin if it lands "left" (this example is adapted from Russell 2023: 571-572).

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

Pointer	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	…
Leftie	2	2	2	2	…	4	8	16	32	…
Rightie	4	8	16	32	…	2	2	2	2	…

Table 5: Leftie vs. Rightie

By design, Leftie and Rightie 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 Leftie is equally good as Rightie, Leftie+Leftie should also be equally good as Rightie+Leftie. This is also intuitive. But is it so? Let's check:

Pointer	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	…
Leftie	2	2	2	2	…	4	8	16	32	…
Rightie	4	8	16	32	…	2	2	2	2	…
Leftie+Leftie	4	4	4	4	…	8	16	32	64	…
Rightie+Leftie	6	10	18	34	…	6	10	18	34	…

Table 6: Leftie+Leftie vs. Rightie+Leftie

Bizarrely, it turns out that Rightie+Leftie stochastically dominates Leftie+Leftie. We can see this by reorganizing their payoffs to respectively yield the following stochastically equivalent prospects:

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

Table 7: Leftie+Leftie* vs. Rightie+Leftie*

It follows that, according to stochastic dominance, Leftie is just as good as Rightie, but Rightie+Leftie is better than Leftie+Leftie. 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. But in the present context, 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 2023: 576).

4.3. Reviving the Separability Argument: Take One

The dialectic doesn't end here, however. For fanatics might patch their argument by refining separability or stochastic dominance, arguing that while no one can satisfy these principles in full generality, fanatics can better satisfy restricted versions of them. Here, I consider two ways this might go, depending on whether the fanatic is more strongly committed to stochastic dominance or to (the core 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 contain infinitely many

options. 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 each contain only finitely many options, X is at least as good as Y if and only if X+Z is at least as good as Y+Z.

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 does not, however, leave the fanatic with a strong case that their view is preferable to tail discounting. (I come back to difference-making discounting shortly.) Now, tail discounters are committed to violations of prospect separability beyond cases involving infinitely many options, as we have seen. So they cannot even satisfy finite separability. But is this a great theoretical cost? It is not. For recall that the core motivation of 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. This motivation supports the unrestricted principle of prospect separability—there's no way to argue from it to the idea that prospect separability fails but finite separability holds. So absent an argument for finite separability that doesn't also imply prospect separability, there's no argument from the *motivation* behind prospect separability to finite separability, and so to fanaticism (Russell 2023: 574).

To sharpen this problem, note that finite separability makes our evaluation of prospects depend on what happens in faraway regions in space and time—and so violates the core motivation behind separability—whenever we assign some positive probability to there being infinitely many prospects in any such region. 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 (Wilkinson 2024: 1912). And, since in practice we cannot altogether rule such possibilities out—we cannot assign probability *zero* to them, even if we think them very unlikely—finite separability is therefore not only unmotivated but toothless in practice (Wilkinson 2024: 1912).

The same problem arises for other recent attempts to weaken separability to prevent it from conflicting with stochastic dominance. In particular, a distinctive feature of the Leftie+Rightie (Tables 5-7) example generating this conflict is not only that the prospects in question involve infinitely many options, but also that these options are probabilistically dependent in an obvious way: whenever Leftie yields a small finite payoff Rightie 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 value of Z's outcomes are probabilistically independent of the value 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 motivation behind separability, since it still renders our evaluation of prospects sensitive to what happens in causally unaffected regions of space and time *whenever they contain some probabilistic dependence with the parts of the world we can affec*t. For example, if how well off Ancient Egyptians or distant aliens are depends on certain variables about which we are uncertain, which will also affect how things will go in the future as the result of our choice, then independent separability will again be rendered toothless. And this again is plausibly the case in practice. For example, it may be that the probabilistic dependency arises due to certain facts about fundamental physics, human nature, or the supernatural (for example, whether karma exists or there is a deity who rewards good behavior and punishes bad) about which we cannot be 100% certain, and which introduce a correlation between the welfare levels of Ancient Egyptians, distant aliens, and those we are affecting in the immediate future through our choices.

Appealing to finite or independent separability therefore fails to support fanaticism over tail discounting, because neither can capture the core motivation behind separability—either theoretically or in practice. While such principles are consistent with stochastic dominance, they cannot be used in arguments *from* separability *to* fanaticism, since neither captures this core motivation.

4.4. Reviving the Separability Argument: Take Two

But the fanatic isn't out of moves yet. Instead of restricting prospect separability so that it no longer captures its motivation, they might stick with this motivation and see where it leads. Indeed, upon reflection, the fanatic might note that this motivation doesn't really support prospect separability in unrestricted form. That motivation, recall, was that our ranking of prospects should depend 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. 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 case where Z is a causally unaffected *baseline*.

A weaker principle encoding the motivation behind separability (which I have not 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 must be able to 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.

Another way to understand baseline separability appeals to the notion of a null baseline. When there's a null baseline (probability 1 of nothing), the difference made to a baseline is simply the resulting prospect. So we can 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 given a null baseline.

This, again, captures the motivation behind separability: what matters to the ranking of prospects is the parts of them we affect, not what happens at a causally unaffected baseline. Indeed, we can now see that precisely what went wrong with other attempts to rescue separability was their failure to accommodate baseline separability: restricting separability to prospects that have finitely many outcomes or are probabilistically independent does not prevent violations of baseline separability in cases where the baseline contains infinitely many options or contains some probabilistic dependence with the other prospects under considerations.

Baseline separability therefore seems uniquely well suited to capturing the core intuition behind separability. And with it in mind, we can also now better diagnose the conflict between prospect separability and stochastic dominance—and begin to see our way toward a possible resolution. Recall that stochastic dominance says 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 a 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: Leftie and Rightie are stochastically equivalent, but Rightie+Leftie stochastically dominates Leftie+Leftie (Tables 5-7). This yields a conflict between stochastic dominance and prospect separability, but it can only even potentially violate *baseline* separability if Leftie is the baseline (otherwise, baseline separability doesn't

apply). And if Leftie is the baseline, we find something interesting. Of course, Rightie+Leftie still stochastically dominates Leftie+Leftie, so baseline separability and stochastic dominance remain inconsistent. However, Rightie+Leftie doesn't *difference-making* dominate Leftie+Leftie, since these now involve adding Rightie or Leftie to a baseline of Leftie, and Rightie doesn't stochastically dominate Leftie (but is rather stochastically equivalent to it). *Baseline* separability and *difference-making* dominance are thus 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 rather 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 if we are to maintain the motivation behind separability, encoded in baseline separability.

At this point, fanatics might claim victory over tail discounting, which (in the example of Charity+Baseline versus Research+Baseline) violates not only prospect but baseline separability. However, now the moderate has 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 2023), 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 Fanaticism isn't better.

The upshot is that not only fanatics can capture the core motivation behind separability by endorsing baseline separability and difference-making dominance. An off-the-shelf version of probability discounting does the same: difference-making discounting. This second move therefore provides fanatics no advantage over moderates generally: it only gives them an advantage over 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. That is, I

incline to the package of baseline separability, difference-making dominance, and difference-making discounting, and to 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 core 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, Gustaffson 2020). Specifically, agents who always choose prospects that are better, according to such theories, are sometimes liable to "money pumps": making sequences of choices that incur certain losses relative to other possible sequences of choices. 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 2023: 580).

Suppose the answer to some question is "true" or "false." If X is better than Y conditional on either

answer, then reflection says X is better than Y unconditionally. This is intuitive: if given either "true" or "false" I will judge X better than Y, it seems I should judge X better than Y already. However, I do not here insist that decision theories must satisfy reflection—I only use the principle to help explicate the existence of certain money pumps that arise when it is violated.

In the literature there is some discussion of the minimal premises needed to show that moderates violate reflection or related principles (Wilkinson 2022: 473-476, Russell 2023: 579-584). But for present purposes we can bypass these issues and admit that probability discounting commits such violations. (Bounded value or utility approaches, by contrast, can avoid such problems (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 are, if anything, more concerning than moderates'.

5.1. Diachronic Inconsistency Against Probability Discounting

Suppose we begin at a null baseline (so that tail discounting and difference-making discounting agree). There are two boxes, "Risky" and "Safe," who 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 the boxes are losers, Risky contains a $20 consolation prize and Safe contains $10:

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

Table 8: Risky vs. Safe

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: 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 is some cost I pay for Safe. But I then find out whether the boxes or 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 am 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 this in and choose in a "sophisticated" way (via "backward induction"), I avoid this problem: anticipating that I will 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 they are winners before taking either box 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). However, 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 much 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 changes how difference-making discounters evaluate prospects (Cibineli 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:

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

Table 9: Cold, Medium, Hot

Suppose (again implausibly) that the probability discounting threshold is 0.05 and that I value money linearly, and that the baseline is Cold. I am now offered the choice to keep Cold for free or to 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 money pump in a diachronic version of the continuum argument (which, this time, even vagueness can't prevent). Starting from a baseline of getting a sure thing of some finite amount of value (money or lives), 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 prospect they started with, and which they are now willing to trade back for 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 money pumps as an exercise to the reader and, going forward, I similarly focus on myopic money pumps. This is because those familiar with money pumps can relatively easily transform the myopic pumps into sophisticated pumps, whereas myopic variants tend to be more intuitive for those lacking this familiarity.

There is a large literature on money pumps, their normative significance, and how to avoid them (Gustaffson 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 money pumps really arise, for example, by abandoning myopic and sophisticated choice in

favor of *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 is no need to wade into these debates here, because, as I now explain, fanatics face *at least as* troubling problems as probability discounters, regardless of how troubling they ultimately are.

5.2. Diachronic Inconsistency Against Fanaticism

The problem, again, relates to the St. Petersburg prospect. For recall that fanatics are committed to this prospect as having a paradoxical feature: it is better than any finite outcome, including any outcome it might yield (Beckstead and Thomas 2023: 446-447). This violates reflection, as is easily seen if we compare our original St. Petersburg prospect to a prospect with slightly better payoffs, which, 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):

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

Table 10: Minus

# Flips	1	2	3	4	5	…
Probability	0.5	0.25	0.125	0.0625	0.03125	…
Plus	$3	$5	$9	$16	$33	…

Table 11: Plus

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

Per fanaticism, Minus is better than each outcome Plus might realize, since 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. For, by stochastic (or difference-making) dominance the reverse holds: Plus is better than Minus.

This generates an even more troubling money pump (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 is 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 realize, 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 turned into a money pump, 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 discounting faces diachronic inconsistencies or money pumps, which at first seems like a major strike against it. But it turns out that fanatics face such violations, too. Thankfully, in this case, the dialectic resolves more simply, since there isn't even a *prima facie* move fanatics can make to repair the argument. For, first, the violations fanatics face seem *worse* than the violations probability discounters face, as they involve being pumped for an arbitrarily large amount. And, second, unlike with separability, the situation isn't one where moderates are committed to strictly more violations. Rather, it's one where fanatics and probability discounters are committed to different violations. Probability discounters aren't committed to especially bad money pumps in St. Petersburg cases, since discounting the small probabilities in Plus or Minus eliminates the paradoxical feature that such prospects are better than any of their outcomes. And fanatics aren't committed to the same violations as moderates.

The issue of diachronic inconsistency therefore favors moderates—if we take the implausibility of their respective violations at face value—or at least ends in a draw—if, say, we appeal to resolute choice to avoid all such violations, deeming none of them problematic. Probability discounters (and especially difference-making discounters) face strange diachronic inconsistencies, but

fanatics face even worse inconsistencies, so this can't support fanaticism.

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 much 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 avoided if the probability discounting threshold is vague. Some may think vagueness theoretically costly, but if the best we can say in favor of fanaticism is that it doesn't imply vagueness, this is hardly enough to defeat moderation.

From here, we found that there is indeed little further for fanatics to say. While moderates can't satisfy prospect separability without violating stochastic dominance, fanatics can't jointly satisfy these conditions either. If we hold onto stochastic dominance and try to weaken separability accordingly, we must abandon the core motivation of separability, leaving us with little reason to prefer fanaticism over *tail* discounting. If we instead stick with the motivation behind separability, we can maintain a consistent view 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 considered the argument that probability discounters are diachronically inconsistent or run into money pumps. Here, we saw that fanatics are subject to worse money pumps.

This leaves us with no new reason to prefer fanaticism, and perhaps some reason to prefer moderation.

The upshot is that there is 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, it turns out that 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 does not settle what decision theory to accept. For one thing, despite raising the initial worry that boundedness cannot capture our anti-fanatical intuitions, I didn't carefully evaluate its merits, and perhaps some may ultimately think it proves best—indeed, I've noted in parentheticals throughout that there are ways to bound value to avoid separability violations (Russell 2023) and to bound value or utility to avoid diachronic inconsistencies (Russell and Isaacs 2021). For another, while I incline toward difference-making discounting, I haven't systematically compared it to tail discounting, and some might prefer tail discounting since it satisfies stochastic dominance and faces fewer diachronic inconsistencies. For that matter, I haven't tried to explore anything like the full range of moderate approaches, nor how such approaches might be combined with other deviations from expected value theory, for example, involving other forms of sensitivity to risk or ambiguity.

So there is 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.

Works Cited

Arrow, Kenneth. 1971. *Essays in the Theory of Risk-Bearing*. Markham.

Askell, Amanda. 2019. "Prudential Objections to Atheism." In Oppy (ed.) *A Companion to Atheism and Philosophy*. Wiley.

Beckstead, Nicholas. 2013. *On the Overwhelming Importance of Shaping the Far Future*. PhD Thesis, Rutgers.

Beckstead, Nick, Teruji Thomas. 2023. "A Paradox for Tiny Probabilities and Enormous Values." *Noûs* 58: 431-455.

Bernoulli, Daniel. 1954. "Exposition of a New Theory on the Measurement of Risk." *Econometrica* 22: 23-36.

Briggs, Rachel. 2015. "Costs of abandoning the Sure-Thing Principle." *Canadian Journal of Philosophy* 45: 827-840.

Buchak, Lara. 2013. *Risk and Rationality*. Oxford.

Buffon, Georges-Louis Leclerc de. 1777. "Essai d'Arithmétique Morale." In *Supplement a l'Histoire Naturelle, Volume IV*. L'Imprimerie Royale.

Cibinel, Pietro. 2023. "A dilemma for Nicolausian discounting." *Analysis* 83: 662-672.

Condorcet, Marquis de. 1785. *Essai Sur L'application De L'analyse À La Probabilité Des Décisions Rendues À La Pluralité Des Voix*. L'Imprimerie Royale.

Gustafsson, Johan E. 2022. *Money-Pump Arguments*. Cambridge.

Greaves, Hilary, William MacAskill, Andreas Mogensen, Teruji Thomas. 2024. "On the Desire to Make a Difference." *Philosophical Studies* 181: 1599-1626.

Hájek, Alan, Harris Nover. 2006. "Perplexing Expectations." *Mind* 115: 703-720.

Kosonen, Petra. 2022. *Tiny Probabilities of Vast Value*. PhD thesis, Oxford.

Kosonen, Petra. 2024. "Probability discounting and money pumps." *Philosophy and Phenomenological*

Research 109: 593-611.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice.* Cambridge.

McMahan, Jeff. 1981. "Problems of Population Theory." *Ethics* 92: 96–127.

Monton, Bradley. 2019. "How to Avoid Maximizing Expected Utility." *Philosophers' Imprint* 19: 1-25.

Parfit, Derek. 1984. *Reasons and Persons.* Oxford.

Pascal, Blaise. 1670. *Pensées* (translated by W. F. Trotter, 1910). Dent.

Peterson, Martin. 2002. "What is a *De Minimis* Risk?" *Risk Management* 4: 47-55.

Russell, Jeffrey Sanford. 2023. "On Two Arguments for Fanaticism." *Noûs* 58: 565-595.

Russell, Jeffrey Sanford and Yoaav Isaacs. 2021. "Infinite Prospects." *Philosophy and Phenomenological Research* 103: 178-198.

Schoenfield, Miriam. 2016. "Moral Vagueness is Ontic Vagueness." *Ethics* 126: 257-282.

Smith, Nicholas J.J. 2014. "Is Evaluative Compositionality a Requirement of Rationality?" *Mind* 123: 457-502.

Smith, Martin. 2024. "Decision Theory and *De Minimis* Risk." *Erkenntnis* 89: 2169-2192.

Tarsney, Christian. 2020. "Exceeding Expectations: Stochastic Dominance as a General Decision Theory." *GPI Working Paper No. 3-2020*. https://globalprioritiesinstitute.org/wpcontent/uploads/Christian-Tarsney_Exceeding-expectations-stochastic-dominance-as-a-general-decision-theory.pdf

Tarsney, Christian. 2023. "Average utilitarianism implies solipsistic egoism." *Australasian Journal of Philosophy* 101: 140-151.

Thomas, Teruji. 2022. "Are Spectrum Arguments Defused by Vagueness?" *Australasian Journal of Philosophy* 100: 743-757.

Wilkinson, Hayden. 2022. "In Defense of Fanaticism." *Ethics* 132: 445-477.

Wilkinson, Hayden. 2024. "Egyptology and fanaticism." *Philosophical Studies* 181: 1903-1923.