Against Anti-Fanaticism

Christian Tarsney*

Forthcoming in *Philosophy and Phenomenological Research*.

Abstract

Should you be willing to forego any sure good for a tiny probability of a vastly greater good? *Fanatics* say you should, *anti-fanatics* say you should not. Anti-fanaticism has great intuitive appeal. But, I argue, these intuitions are untenable, because satisfying them in their full generality is incompatible with three very plausible principles: acyclicity, a minimal dominance principle, and the principle that any outcome can be made better or worse. This argument against anti-fanaticism can be turned into a positive argument for a weak version of fanaticism, but only from significantly more contentious premises. In combination, these facts suggest that those who find fanaticism counterintuitive should favor not anti-fanaticism, but an intermediate position that permits agents to have incomplete preferences that are neither fanatical nor anti-fanatical.

1 Introduction

How much practical weight should you be willing to give to extremely remote possibilities? For any positive probability p, no matter how small, and any good g, no matter how great, should you be willing to forego a certainty of g in exchange for probability p of a greater good g^*, if the latter is great enough? *Fanaticism* is (roughly for now) the view that this sort of preference is rationally required: for any positive probability p and good g, there must be a good g^* such that you prefer probability p of g^* over certainty of g. *Anti-fanaticism* (again roughly) is the view that the *opposite* sort of preference is rationally required: There must be some positive probability p and good g such that you prefer g for sure over *any* good with probability p or less.

*Population Wellbeing Initiative, UT Austin; christian.tarsney@austin.utexas.edu

When we consult our intuitions about cases, anti-fanaticism holds a clear advantage over fanaticism. Suppose, for instance, that you are given a choice between certainty of a long and happy life (say, 100 years full of all the things that ordinarily make a human life good), or a gamble that gives you a one-in-a-googol (10^{-100}) chance of an even better life and a complementary ($1-10^{-100}$) chance of instant death. Most of us will intuit, I think, that *no matter what* that even better life consists in (no matter how long it lasts or what goods it would involve), it is more prudentially rational to choose the first option. Similarly, in a moral context, suppose you must choose between guaranteeing a very good future for all sentient life on Earth (say, hundreds of millions of years in which large populations will enjoy prosperity, justice, and happiness) or a gamble that gives a one-in-a-googol chance of an even better collective future and a complementary chance of instant collective annihilation. Again, most of us will intuit that no matter what super-utopian future we would get by winning the gamble, it would be better to take the sure thing.

The issue of how to weigh small probabilities isn't merely hypothetical. On the contrary, it is central to some of the most important questions of prioritization in practical ethics: With our limited resources, individually and collectively, should we focus on modest improvements to the world that we can achieve with confidence, like reducing the burdens of infectious disease and the suffering of farmed animals? Or should we instead focus on increasing the probability of a flourishing long-term future (e.g., by reducing the risk of near-term human extinction), even if we can affect the latter probability only very slightly? If we are risk-neutral expected value maximizers (evaluating each risky prospect by the probability-weighted sum of the values of its possible outcomes), then there is a strong case to be made for the latter view (Cowen, 2007; Beckstead, 2019; Greaves and MacAskill, 2021). But this case can have a decidedly fanatical tinge to it. For instance, in arguing for moral importance of reducing risks of human extinction and other permanent global catastrophes, Nick Bostrom estimates that a future interstellar civilization could support the equivalent of at least 10^{52} human lives in digital form, and reasons that '[e]ven if we give this allegedly lower bound...a mere 1 per cent chance of being correct, we find that the expected value of reducing existential risk by a mere *one billionth of one billionth of one percentage point* is worth a hundred billion times as much as a billion human lives' (Bostrom, 2013, p. 19). This suggests that we should pass up opportunities to do enormous amounts of good to maximize the probability of an astronomically good future, even if the difference we can make to that probability is on the order of, say, 10^{-30}. Those with anti-fanatical intuitions,

I think, will find these intuitions triggered as strongly by this reasoning as by the hypothetical cases in the last paragraph.

Largely because of its significance for practical ethics, the question of fanaticism has become the focus of a growing literature in ethics and decision theory. To name just a few recent contributions: Wilkinson (2022) offers an extended, multi-pronged defense of fanaticism. Monton (2019) defends "Nicolausian discounting" (ignoring small probabilities) largely as a way to avoid fanaticism. Balfour (2021) highlights the counterintuitive fanaticism of expected value maximization with respect to existential risks. Russell and Isaacs (2021) describe some of the unwelcome theoretical implications of fanaticism. Finally, Russell (2023) and Beckstead and Thomas (2024) present compelling arguments both for and against fanaticism, without committing themselves to either conclusion.

Here's what this paper will add. First (§2): While the recent literature has focused on the truth or falsity of fanaticism, which requires agents to give unlimited weight to small probabilities, I highlight the opposing thesis of *anti*-fanaticism, which requires agents to give only *limited* weight to small probabilities. Anti-fanaticism is not merely the negation of fanaticism, because there is a middle ground between the two, which does not require agents to be either fanatical or anti-fanatical. And so arguments against fanaticism need not be arguments for anti-fanaticism, and vice versa.

Second (§§3–4): The recent literature has focused on a narrow formulation of fanaticism involving choices between binary gambles and sure outcomes. But this setting is overly restrictive: Decision theories that are fanatical or anti-fanatical in this limited setting may not be so in general. I introduce a more general setting, where we must choose between two ways of altering an uncertain baseline prospect: modestly improving every outcome, or shifting a small amount of probability from a much worse outcome to a much better one. I then present a formulation of anti-fanaticism that captures our anti-fanatical intuitions in this setting.

Third (§5): I argue that fully satisfying our anti-fanatical intuitions comes at an unacceptable cost, by showing that this more general anti-fanatical thesis is incompatible with three very plausible principles: acyclicity, a minimal dominance principle, and the principle that any outcome can be made better or worse. This impossibility result is the central contribution of the paper.

Fourth (§§6–7): I show that, because they satisfy these three principles, two canonically "anti-fanatical" decision theories—bounded expected utility maximization and "tail discounting"—are *not* generally anti-fanatical. In particular, while they are anti-fanatical in the restricted setting of binary

gambles vs. sure outcomes, they are not necessarily any less fanatical than expected value maximization when it comes to small changes in intermediate probabilities of very good or very bad outcomes—and, I emphasize, nearly every case of practical interest is of this latter kind. These arguments serve to unify recent observations about the potentially fanatical character of bounded expected utility maximization (Beckstead and Thomas, 2024) and tail discounting (Kosonen, 2022; Cibinel, 2023), as well as Cibinel's argument that Nicolausian discounting can avoid fanaticism only at the cost of preference cycles.

Fifth (§8): I show that the preceding negative argument against anti-fanaticism can be turned into a positive argument for a weak version of fanaticism, but only by means of significantly more contentious premises (completeness and transitivity in place of acyclicity).

From these facts I conclude (§9) that those who find fanaticism counter-intuitive should favor not anti-fanaticism, but an intermediate position that permits agents to have incomplete preferences that are neither fanatical nor anti-fanatical.

2 Fanaticism and anti-fanaticism

Let's start with some basic setup. Our central question will be what rationality requires of an agent in terms of her preferences over *prospects*. Prospects are understood as probability distributions over *outcomes*, where an outcome is a specification of all evaluatively significant features of the world. Our focus will be on discrete prospects, which can be represented as a set of ordered pairs of an outcome and a probability, with the probabilities summing to 1. For the special case of a binary prospect with two possible outcomes, we will write $\langle o_i, p, o_j \rangle$ to denote the prospect that yields outcome o_i with probability p and outcome o_j otherwise. For the prospect that yields outcome o_i with certainty, we write $\langle o_i \rangle$.

In giving examples, we will sometimes use the idea of states of nature, and understand prospects as mapping states (each with an assigned probability) to outcomes. But states will play only a didactic role; none of the formal principles or arguments below will make any reference to them.

We assume that outcomes can be compared in terms of value (e.g. moral or prudential), and that these comparisons are given independent of and prior to any ranking of prospects. Specifically, where $\mathbb{O}$ denotes the set of all possible outcomes, we assume a preorder (a reflexive, transitive binary relation) $\succsim_{\mathbb{O}}$ on $\mathbb{O}$, where $o_i \succsim_{\mathbb{O}} o_j$ means that o_i is at least as good as (or *weakly better than*) o_j. If $o_i \succsim_{\mathbb{O}} o_j$ but $o_j \not\succsim_{\mathbb{O}} o_i$, we say that o_i is *strictly better*

than o_j, denoted $o_i \succ_{\mathbb{O}} o_j$. If $o_i \succsim_{\mathbb{O}} o_j$ and $o_j \succsim_{\mathbb{O}} o_i$, we say that o_i and o_j are *equally good*, denoted $o_i \sim_{\mathbb{O}} o_j$. If neither relation holds, then we say that o_i and o_j are *incomparable*, denoted $o_i \bowtie_{\mathbb{O}} o_j$.

An agent is assumed to have ranking of prospects, a preorder $\succsim$, which we will describe as a *preference relation* (while remaining neutral about what preferences are, e.g., whether they are choice dispositions, subjective value judgments, or beliefs about some more objective evaluative relation). Thus $P_i \succsim P_j$ means that prospect P_i is *preferred at least equally* (or *weakly preferred*) to prospect P_j. If $P_i \succsim P_j$ but not $P_j \succsim P_i$, we say that P_i is *strictly preferred* to P_j, denoted $P_i \succ P_j$. If $P_i \succsim P_j$ and $P_j \succsim P_i$, we say that they are *equally preferred*, denoted $P_i \sim P_j$. And if neither relation holds, we say that there is a *preference gap* between P_i and P_j, denoted $P_i \bowtie P_j$.

This gives us enough machinery to precisely state one version of fanaticism. *Narrow Fanaticism*, as we will call it, is the conjunction of two theses:

Narrow Positive Fanaticism It is rationally required that, for any outcomes $o^+ \succ_{\mathbb{O}} o^-$ and probability $p > 0$, there is an outcome o^{*+} such that $\langle o^{*+}, p', o^- \rangle \succ \langle o^+ \rangle$ for all $p' \geq p$.

Narrow Negative Fanaticism It is rationally required that, for any outcomes $o^+ \succ_{\mathbb{O}} o^-$ and probability $p > 0$, there is an outcome o^{*-} such that $\langle o^- \rangle \succ \langle o^{*-}, p', o^+ \rangle$ for all $p' \geq p$.[1]

An agent who satisfies Narrow Positive Fanaticism will forego certainty of a very good outcome, o^+, for a risky prospect that is almost certain to yield a very bad outcome o^- but carries some minuscule probability of an outcome o^{*+}—as long as o^{*+} is good enough. An agent who satisfies Narrow Negative Fanaticism will *accept* certainty of the very bad outcome o^-, rather than take a risky prospect that is almost certain to deliver the very good outcome o^+, but carries some minuscule probability of an outcome o^{*-}—as long as o^{*-} is bad enough. Narrow Fanaticism is, roughly, the version of fanaticism that has been the focus of the recent literature.[2]

[1] Here and elsewhere, I will use o^+/o^- to suggest good/bad outcomes, and o^{*+}/o^{*-} to suggest *astronomically* good/bad outcomes. These superscripts do not have formal meaning—they do not restrict quantification, for instance—but merely indicate what sort of outcome is of most interest, e.g., the cases in which a principle is non-trivial. Thus, for instance, a statement like "for any outcomes o^+ and o^-" can be read as "for any outcomes o^+ [no matter how good] and o^- [no matter how bad]".

[2] It differs from other recent statements of fanaticism in four ways: First, I treat fanaticism as a thesis about rational requirement, rather than a thesis about the goodness of prospects or simply a property of an agent's preferences. Second, it is common to restrict statements

While Narrow Fanaticism holds that we are rationally required to have a certain kind of preference between sure outcomes and binary lotteries, Narrow Anti-Fanaticism holds that we are rationally required to have the opposite sort of preference. It likewise consists of two theses:

Narrow Positive Anti-Fanaticism It is rationally required that there are some outcomes $o^+ \succ_{\mathbb{O}} o^-$ and probability $p > 0$ such that, for any outcome o^{*+}, $\langle o^+ \rangle \succ \langle o^{*+}, p', o^- \rangle$ for all $p' \leq p$.

Narrow Negative Anti-Fanaticism It is rationally required that there some outcomes $o^+ \succ_{\mathbb{O}} o^-$ and probability $p > 0$ such that, for any outcome o^{*-}, $\langle o^{*-}, p', o^+ \rangle \succ \langle o^- \rangle$ for all $p' \leq p$.

An agent who satisfies Narrow Positive Anti-Fanaticism will at least sometimes prefer certainty of a very good outcome o^+ to a tiny chance of an astronomically good outcome o^{*+} (and a very bad outcome o^- otherwise), no matter how good o^{*+} may be. On the other hand, an agent who satisfies Narrow Negative Anti-Fanaticism will sometimes prefer to take a tiny risk of an astronomically bad outcome o^{*-} rather than settle for o^-, no matter how bad o^{*-} may be.

Narrow Fanaticism and Narrow Anti-Fanaticism assert contrary rational requirements. These theses are not jointly exhaustive, therefore, since rationality might impose *neither* requirement on us. Let's call the intermediate view, which denies both requirements, *permissivism*. Permissivism can

---

of fanaticism to "finite" outcomes. This ensures, for instance, that an agent who regards some pair of outcomes (e.g., Heaven and Hell) as infinitely better/worse than any others, and will always pay any finite cost to increase the probability of the former or reduce the probability of the latter, counts as fanatical. I omit this restriction merely for simplicity; inserting it would have no effect on my arguments. Third, I treat the "baseline" outcomes that occur with probability $1-p$ in the risky option (o^- for Narrow Positive Fanaticism, o^+ for Narrow Negative Fanaticism) as variables, whereas fanaticism is often characterized relative to a fixed baseline outcome (typically designated "0"). This will allow us, in the next section, to generalize fanaticism and anti-fanaticism to a context where the baseline outcome is uncertain. And since extant arguments for and against fanaticism do not depend on the choice of baseline, this extra generality seems unobjectionable. Fourth, rather than just making a claim about the prospect that gives an extreme outcome with probability p, I treat fanaticism as a thesis about all probabilities p' greater than or equal to p, and will likewise treat anti-fanaticism as a thesis about all probabilities p' less than or equal to p. The latter fact plays some role in the argument of §5—without it we would have to slightly strengthen the dominance principle employed there (see fn. 11 for details). But I think it is unobjectionable: If, for instance, you should prefer probability p of an astronomical gain to certainty of a modest gain, then clearly you should prefer any larger probability of the astronomical gain as well; and if you should prefer the modest gain to probability p of the astronomical gain, then clearly you should also prefer it to any smaller probability.

take multiple forms. For instance, it might permit both fanatical and anti-fanatical preferences. Or it might permit (or even, its name notwithstanding, require) *incomplete* preferences that are neither fanatical nor anti-fanatical. But apart from noting its existence, we will say no more about the permissivist alternative for now, returning to it only in the concluding section.

3 Small probabilities and small differences in probability

The narrow versions of fanaticism and anti-fanaticism are attractively simple. But they do not fully capture the question of whether we should give potentially unlimited weight to arbitrarily small probabilities. That's because they only consider a special case, where an agent is choosing between certainty on the one hand and a binary prospect involving a tiny probability of an astronomically good/bad outcome on the other. The more general (and more realistic) case is a choice between two prospects that *differ* slightly in how they distribute probability between much better and much worse outcomes—in other words, a case where the agent must decide how much she is willing to pay in order to *shift* a small amount of probability from a much worse outcome to a much better outcome.

To see the difference, consider the two choice situations described in tables 1a and 1b, each of which involves a choice between two prospects, with uncertainty between two possible states of nature. The first is a simple choice between certainty of a small gain (1) and a small probability (2ε) of a large gain ($\frac{1}{\varepsilon}$). The second case is slightly more complicated: Here the astronomically large gain $\frac{1}{\varepsilon}$ has a "baseline" probability of $0.5-\varepsilon$, and the choice is between taking a sure gain (improving each outcome by 1) or slightly *increasing* that baseline probability (by 2ε)—in this case, by moving the astronomically good outcome from a less probable to a more probable state. In the second case, small probabilities are nowhere to be seen—every state, and every outcome, has a quite substantial probability. What is small is the *difference* in probabilities between s_1 and s_2, and hence between the probabilities of a very good outcome associated with P_3 and P_4 respectively. It seems to me, though, that anyone who finds expected value maximization counterintuitively fanatical in the first case will have the same intuition about the second case.

Choices like 1b are also much more common and realistic than choices like 1a. Consider, for instance, two important real-world cases: voting and existential risk mitigation. If you are deciding whether it is worth your while to vote in an election, in terms of the difference you might make to the

(A) EXTREMAL PASCALIAN CHOICE

	s_1 $(1-2\varepsilon)$	s_2 (2ε)
P_1	1	1
P_2	0	$\frac{1}{\varepsilon}$

(B) INTERMEDIATE PASCALIAN CHOICE

	s_1 $(0.5-\varepsilon)$	s_2 $(0.5+\varepsilon)$
P_3	$\frac{1}{\varepsilon}+1$	1
P_4	0	$\frac{1}{\varepsilon}$

TABLE 1: Left: A simple choice between certainty of a small gain and a small probability of a large gain. Right: A small probability difference without a small probability. The choice is between a sure gain of 1 and a small (2ε) increase in the probability of a large gain ($\frac{1}{\varepsilon}$).

outcome, it is never the case that your preferred candidate is certain to lose if you don't vote, but has a tiny probability of winning if you do; nor that they are certain to win if you vote, but have a tiny probability of losing if you don't. Rather, they have some intermediate probability of winning, which your vote would slightly increase. Similarly, if you are deciding whether to devote some unit of resources to mitigating existential risks (from engineered pandemics, nuclear war, AI or the like) or to providing some more certain good (like direct cash transfers to the poor), the situation is not that near-term human extinction is certain if you do not act to prevent it, and your intervention represents humanity's only slim hope of survival; nor that humanity is certain to survive if you act to assure its survival, and the only chance of doom comes from your inaction. Rather, there is some non-trivial probability that humanity will succumb to a near-term existential catastrophe, and some non-trivial probability that it will not, and your intervention (if well-chosen) might very slightly reduce the former probability and increase the latter.

Our tendency to think of these cases in terms of "very small probabilities" rather than "very small probability differences" may reflect our tendency to focus on the *difference* we make to the value of outcomes, rather than their absolute value. In the case of voting, for instance, there is indeed a very small probability that your vote makes a very large difference, complemented by a very large probability that it makes no difference.[3] But what we ought to care about, ultimately, is not the difference we make but how well things actually turn out.[4] The more accurate framing of these choices, then, is in terms of small differences in intermediate probabilities, not absolutely-small probabilities.

[3]The case of existential risk mitigation might be more complicated. If the course of history is sufficiently sensitive to very small changes, then perhaps every action you take has substantial—but almost exactly equal—probabilities of causing and of preventing near-term existential catastrophe.

[4]For defense of this claim, see Greaves et al. (2024).

In summary, insofar as we find fanaticism counterintuitive, the counterintuitiveness is not confined to simple choices between risk and certainty, but extends to the more general case involving small changes to an uncertain baseline prospect. And insofar as anti-fanaticism is meant to resist fanatical applications of expected value reasoning in real-world contexts, it must apply to this more general case as well. The thesis must be that it is irrational to trade certainty of a substantial gain for arbitrarily small shifts in probability from one outcome to another.

4 Anti-fanaticism generalized

In this section, we generalize anti-fanaticism to the wider context of small probability shifts. (We will do the same for fanaticism in §8.) To do this, we need to introduce two new concepts. In the simple context of small probabilities, the "safe" option could be characterized as yielding a single outcome with certainty. In the more general context, we assume that astronomically-better and astronomically-worse outcomes will have some non-zero probability whatever option one chooses, and so the safe option can no longer be characterized in that way. We can characterize it instead, however, as offering a sure *improvement* to the outcome of the baseline prospect. To illustrate: In the case of voting, if you choose not to vote, you can instead spend an hour watching television. This doesn't deliver a single outcome with certainty since (among other things) it leaves you uncertain who will win the election—but it does mean that both possible election outcomes will be *improved*, from your perspective, by the addition of one hour of television. Similarly, if instead of spending a fixed sum of money to mitigate existential risks, you spend it on direct cash transfers to the very poor, you are not left with certainty of any particular outcome, but instead with certainty that whatever otherwise would have happened will be improved by certain very poor people being made slightly less poor.[5]

We therefore introduce the following two concepts:

An *improvement* is a function $I : \mathbb{O} \mapsto \mathbb{O}$ that maps every outcome to a strictly better outcome if one exists, or else to a weakly better outcome.

A *worsening* is a function $W : \mathbb{O} \mapsto \mathbb{O}$ that maps every outcome to a strictly worse outcome if one exists, or else to a weakly worse outcome.

[5] Of course, in reality even these improvements are not literally *certain*. But treating them that way is, as far as I can see, a harmless idealization for present purposes, that does not sacrifice any interesting generality.

Paradigmatically, an improvement might be a fixed change that can be applied to any outcome, like giving one individual an additional year of life at a fixed level of positive well-being. But the concept is more flexible than that, and does not presuppose that there is any concrete change that can be applied to every possible outcome and that always makes an outcome (even weakly) better. An improvement (or worsening) *could* correspond to intuitively very different concrete changes to different outcomes. (And of course analogous remarks apply to worsenings.)

In rough terms, the generalized form of anti-fanaticism will say (in the positive case) that a large enough *improvement* to every outcome in a baseline prospect should always be preferred to a small enough shift in probability from one outcome to another, no matter how disparate those outcomes might be. But this gloss requires an important caveat: One extreme way of being an anti-fanatic is to hold that some outcomes and prospects are "good enough", in the sense that a rational agent may be completely indifferent to any further improvements. Formally, let's say that a prospect P is *maximal* if no other prospect is strictly preferred to it, and that an outcome o is maximal if the prospect $\langle o \rangle$ is maximal. (Likewise, P is *minimal* if no other prospect is strictly *dis*preferred to it, and o is minimal if $\langle o \rangle$ is minimal.) In a choice between a sure improvement and a small probability shift, the prospect resulting from the small shift might be maximal, either because the baseline prospect itself consisted entirely of maximal outcomes, or because it involved only a small probability of a non-maximal outcome, which the small probability shift eliminates. In this case, the anti-fanatic need not require that the sure improvement be strictly preferred.

These points in hand, we can now state a more general version of anti-fanaticism:

General Positive Anti-Fanaticism It is rationally required that there is some improvement I and probability $p > 0$ such that, for any outcomes $o^{*+} \succ_{\mathbb{O}} o^{-}$ and any probability $q < 1$, the prospect $\langle I(o^{*+}), q, I(o^{-}) \rangle$ is weakly preferred to $\langle o^{*+}, q + p', o^{-} \rangle$ for any $p' \leq p$ (and $\leq 1 - q$). Moreover, this preference is strict unless o^{*+} and o^{-} are both maximal, or o^{*+} is maximal and $q + p' = 1$.

General Negative Anti-Fanaticism It is rationally required that there is some worsening W and probability $p > 0$ such that, for any outcomes $o^{+} \succ_{\mathbb{O}} o^{*-}$ and any probability $q < 1$, the prospect $\langle o^{*-}, q + p', o^{+} \rangle$ is weakly preferred to $\langle W(o^{*-}), q, W(o^{+}) \rangle$ for any $p' \leq p$ (and $\leq 1 - q$). Moreover, this preference is strict unless o^{+} and o^{*-} are both minimal, or o^{*-} is minimal and $q + p' = 1$.

Intuitively, General Positive Anti-Fanaticism says that there is some improvement I large enough, and (non-zero) probability p small enough, that given a choice between (i) improving every outcome in your "baseline" prospect by I or (ii) increasing the probability of an astronomically good outcome by p, you always at least weakly prefer the former. For example, perhaps you should always prefer adding twenty happy years to someone's life over increasing the probability of any outcome (no matter how good) by 10^{-30} or less. And General Negative Anti-Fanaticism says that there is some worsening W large enough, and (non-zero) probability p small enough, that given a choice between (i) worsening every outcome in your baseline prospect by W or (ii) increasing the probability of an astronomically bad outcome by p, you always at least weakly prefer the latter. For instance, perhaps you should always be willing to increase the risk of any outcome (no matter how bad) by 10^{-30} or less rather than *shorten* a happy life by twenty years. *General Anti-Fanaticism* is the conjunction of these two theses.[6]

General Anti-Fanaticism is stronger than Narrow Anti-Fanaticism, since it universally quantifies over the baseline probability q, where Narrow Anti-Fanaticism only covers the special case of $q = 0$.[7]

[6]An essential feature of General Anti-Fanaticism is that, while what counts as a "small" probability difference (i.e., the value of p) may depend on the improvement I, it does not depend on the baseline probability of the extreme outcome (i.e., on q). A weaker generalization of anti-fanaticism that allowed such dependency would not be susceptible to the argument against General Anti-Fanaticism in the next section. But the independence of p from q is, I think, an essential feature of our anti-fanatical intuitions and their real-world applications. The intuition that one is not required to prefer a 10^{-30} reduction in the risk of human extinction over certainty of saving a single life, for instance, does not depend on the baseline probability of extinction.

[7]More precisely, General Positive Anti-Fanaticism implies Narrow Positive Anti-Fanaticism given the minimal assumption that there are two outcomes $o^{*+} \succ_{\mathbb{O}} o^{-}$ where o^{-} is non-maximal. (An analogous assumption is needed in the negative case.) Spelling this out: Assume two outcomes $o^{*+} \succ_{\mathbb{O}} o^{-}$ where o^{-} is non-maximal. Let I and p be an improvement and a probability that satisfy General Positive Anti-Fanaticism. Without loss of generality, we can assume that $p < 1$, since if General Positive Anti-Fanaticism holds for a given p, it holds for any smaller p as well. Since we're interested in the case where $q = 0$, we then have that $q + p' < 1$. Thus, General Positive Anti-Fanaticism implies that $\langle I(o^{*+}), 0, I(o^{-})\rangle$ is strictly preferred to $\langle o^{*+}, 0 + p', o^{-}\rangle$ for any $p' \leq p$—or in other words, $\langle I(o^{-})\rangle$ is strictly preferred to $\langle o^{*+}, p', o^{-}\rangle$ for any $p' \leq p$. Letting $I(o^{-}) = o^{+}$, this is exactly Narrow Positive Anti-Fanaticism. (We know that $I(o^{-})$ is strictly better than o^{-} since by assumption there is *some* outcome, namely o^{*+}, strictly better than o^{-}.)

5 Against anti-fanaticism

Anyone who wants to do justice to our anti-fanatical intuitions and to resist fanatical real-world applications of expected value reasoning must, I have argued, endorse not just Narrow but General Anti-Fanaticism. But unfortunately, this thesis is subject to a very powerful objection. To state the objection, we need to introduce a few new principles.

No Best Outcome For every outcome, there is a strictly better outcome.

No Worst Outcome For every outcome, there is a strictly worse outcome.

Minimal Dominance If $o_i \succ_{\mathbb{O}} o_j$, then it is rationally required that $\langle o_i \rangle \succ \langle o_j \rangle$.

Acyclicity It is rationally required that, if $P_1 \succ P_2$, $P_2 \succ P_3$, …, $P_{n-1} \succ P_n$, then it's not the case that $P_n \succ P_1$.

These principles are very weak, and hard to deny. No Best Outcome and No Worst Outcome do not imply anything like unboundedness of cardinal value or utility. They only assert that there is always *some* way of making an outcome at least a little better/worse—for instance, by extending a happy/unhappy life, or adding a positive/negative experience to a life without changing its duration.[8] Minimal Dominance is perhaps the weakest possible expression of the idea that the desirability of a prospect depends on the value of its possible outcomes. It is weaker than widely accepted principles like statewise dominance, stochastic dominance, and even "superdominance", the principle that if the worst possible outcome of P_i is better than the best possible outcome of P_j, then P_i should be strictly preferred to P_j.[9] Finally, Acyclicity is the least controversial of the standard

[8] An argument for No Best Outcome is that (i) it is always possible to add an additional good to a life (e.g., an extra happy experience) without harming anyone else, (ii) adding a good to someone's life makes things better for that person, and (iii) making things better for one person without harming anyone else yields a strictly better outcome. No Best Outcome might be denied by an extreme negative utilitarian who holds that there *are* no welfare goods, or that welfare goods make no contribution to the value of outcomes (even as tiebreakers), and therefore that a world with no welfare bads (e.g., an empty world) is the best possible outcome. (Thanks to Andreas Mogensen for pointing this out.) It might also be denied by those like Leibniz who claim that the actual world is the best possible outcome.

[9] Although there are various arguments in the literature for denying statewise or stochastic dominance, I am not aware of anyone who would deny Minimal Dominance, or of any motivation for denying it. It is worth emphasizing that "outcomes" and "betterness" can be construed very broadly to include agent-relative, non-consequentialist considerations, so that Minimal Dominance and other dominance principles do not carry any commitment to consequentialism.

coherence constraints on rational preference associated with expected utility theory. It is a significant weakening of Transitivity, the requirement that if $P_i \succsim P_j$ and $P_j \succsim P_k$, then $P_i \succsim P_k$. And it is supported by a particularly strong instance of the "money-pump" arguments commonly used to justify the expected utility axioms (Gustafsson, 2022, Ch. 2).

But surprisingly, General Anti-Fanaticism is incompatible with these very weak principles. Specifically:

Theorem 1. *Acyclicity, Minimal Dominance, and No Best Outcome rule out General Positive Anti-Fanaticism. Acyclicity, Minimal Dominance, and No Worst Outcome rule out General Negative Anti-Fanaticism.*

Here is the proof (focusing on the positive case—the negative case is exactly parallel): No Best Outcome and Minimal Dominance together imply that no outcome is maximal. (For every outcome o there's a strictly better outcome o', and $\langle o' \rangle$ must be strictly preferred to $\langle o \rangle$.) General Positive Anti-Fanaticism therefore requires without qualification that there is some improvement I and positive probability p such that, for any outcomes $o^{*+} \succsim_{\mathbb{O}} o^-$ and probability $q < 1$, the prospect $\langle I(o^{*+}), q, I(o^-) \rangle$ is strictly preferred to $\langle o^{*+}, q + p', o^- \rangle$ for any $p' \leq p$ and $\leq 1 - q$.[10] Given such an I and p, choose an integer n such that $\frac{1}{n} \leq p$, and an arbitrary "baseline" outcome o. Then consider the case described in Table 2. Here we have n equiprobable states and $n+1$ prospects. The various possible outcomes are generated from the baseline outcome o by applying improvement I one or more times, with I^k representing k iterations of I (that is, the result of applying I to o, k times over). At each step from P_{i-1} to P_i, we make two changes: We "slightly" improve the outcome in every state by adding one iteration of I, while "astronomically" *worsening* the outcome in state s_i by *removing* n iterations of I. (Here "slightly" and "astronomically" do not imply anything about cardinal value—they just mean "by a single iteration of I" and "by many iterations of I" respectively.)

General Positive Anti-Fanaticism implies that each step from P_{i-1} to P_i is a strict improvement: In a choice between improving every outcome by I or shifting probability $\frac{1}{n} \leq p$ to an astronomically better outcome, we always prefer the former.[11] Thus, $P_1 \succ P_0$, $P_2 \succ P_1$, and so on. But at the

[10] No Best Outcome and Minimal Dominance will not be used in the rest of the proof. This shows that a stronger version of General Anti-Fanaticism requiring the sure improvement to be strictly preferred in every case would be intrinsically cyclic.

[11] In more detail: For all $i < n$, $P_i = \langle I^{n+i}(o), \frac{n-i}{n}, I^i(o) \rangle$ and $P_{i+1} = \langle I^{n+i+1}(o), \frac{n-i-1}{n}, I^{i+1}(o) \rangle$. Letting $o^{*+} = I^{n+i}(o)$, $o^- = I^i(o)$, $p' = \frac{1}{n}$, and $q = \frac{n-i-1}{n}$, General Positive Anti-Fanaticism implies that the latter prospect is strictly preferable.

Here we are using the fact that General Anti-Fanaticism tells us to prefer certainty of

	$s_1\,(\frac{1}{n})$	$s_2\,(\frac{1}{n})$	$s_3\,(\frac{1}{n})$	$\cdots$	$s_n\,(\frac{1}{n})$
P_0	$I^n(o)$	$I^n(o)$	$I^n(o)$	$\cdots$	$I^n(o)$
P_1	$I(o)$	$I^{n+1}(o)$	$I^{n+1}(o)$	$\cdots$	$I^{n+1}(o)$
P_2	$I^2(o)$	$I^2(o)$	$I^{n+2}(o))$	$\cdots$	$I^{n+2}(o)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
P_n	$I^n(o)$	$I^n(o)$	$I^n(o)$	$\cdots$	$I^n(o)$

TABLE 2: An illustration of the cyclicity objection to General Anti-Fanaticism

end of this sequence of strict improvements, we end up exactly where we started—P_n is identical to P_0! So we have a cycle: $P_n \succ P_{n-1} \succ \ldots P_1 \succ P_n$.[12]

improvement I to probability shifts of p *or less*. Without the "or less" clause, we would either have to restrict the argument against General Anti-Fanaticism to the case where $p = \frac{1}{n}$ for some integer n, or else strengthen Minimal Dominance. Taking the latter approach, we could use the following (still quite weak and uncontroversial) principle:

Binary Monotonicity If $o_i \succ_{\mathbb{O}} o_j$ and $p > q$, then it is rationally required that (1) $\langle o_i, p, o_j\rangle \succ \langle o_i, q, o_j\rangle$, and moreover that (2) for all P_i, $P_i \succ \langle o_i, p, o_j\rangle$ implies $P_i \succ \langle o_i, q, o_j\rangle$, and likewise $\langle o_i, q, o_j\rangle \succ P_i$ implies $\langle o_i, p, o_j\rangle \succ P_i$.

Replacing Minimal Dominance with Binary Monotonicity would allow us to complete the argument against a version of General Anti-Fanaticism that only applied to p itself and not to all $p' \leq p$ (modulo some further adjustment to handle cases where the baseline probability q is greater than $1-p$).

[12]A slight modification of this argument, that's less elegant but more narratively compelling, subtracts $n+1$ iterations of I (rather than n iterations) from a single state at each step. (Thanks to Jeffrey Russell for suggesting the simpler example used in the main text.) In this modified case (which will be used in the proof of Theorem 2, and is depicted in Table 3 below), following the preferences of the General Anti-Fanatic from P_0 to P_n leaves us not back where we started, but with a worse prospect: certainty of $I^{n-1}(o)$ rather than $I^n(o)$. Minimal Dominance and No Best Outcome guarantee a preference cycle in this case as well.

This presentation of the argument highlights a structural similarity to an argument against non-aggregative views in normative ethics put forward by Parfit (2003). (Thanks to Elliott Thornley for pointing out this parallel.) The non-aggregative views in question claim that for a sufficiently large benefit B and a sufficiently small benefit b, it is better to provide B to one person than to provide b to any number of people. But suppose that n instances of b add up to a benefit greater than B. (For instance, in Parfit's example, B and b are an additional year and minute of life respectively. In this case n instances of b add up to a greater benefit than B as long as $n > 525{,}600$.) Then imagine a population of n individuals and a sequence of choices between providing B to one individual or b to every individual. Non-aggregationism seems to require making the former choice in each instance, even though making the latter choice in each instance would leave everyone better off.

An important difference between this argument and the cyclicity argument against anti-fanaticism is that the non-aggregationist has the option of denying that many instances of b can add up to a benefit greater than B. (This is undeniable in Parfit's example, but the non-aggregationist gets to choose the B and b to which their thesis applies. They might

This is, it seems to me, a *very* strong objection to General Anti-Fanaticism. In particular, it relies on much weaker premises than extant arguments for fanaticism (like those discussed in Wilkinson (2022), Russell (2023), and Beckstead and Thomas (2024)), which rely on axiological separability assumptions or substantive constraints on an agent's risk attitudes. No Best/Worst Outcome, Minimal Dominance, and Acyclicity are all intuitively plausible and supported by compelling arguments. To my mind, the easiest principle for the anti-fanatic to give up is Acyclicity. The cost of this move is somewhat mitigated by the fact that the cycles into which General Anti-Fanaticism leads us may be *extremely* long. For instance, an agent who ignores probability differences smaller than 10^{-15} need not have any preference cycles shorter than 10^{15} steps. (This will, among other things, make her fairly difficult to money pump—it will require a setup in which she faces, or believes herself to face, a potential sequence of choices at least 10^{15} nodes long.) Still, we should not let the juxtaposition with other, even-less-deniable principles fool us: Acyclicity is an extremely plausible principle, and denying it is a very serious cost. I conclude, therefore, that we should reject General Anti-Fanaticism.

6 Bounded expected utility

Let's now consider what the preceding argument tells us about two views in normative decision theory that have been treated as paradigmatic forms of anti-fanaticism: bounded expected utility maximization (Bounded EU) and small-probability discounting.

Expected utility theory claims that you should evaluate prospects by first assigning a numerical *utility* to each possible outcome, and then ranking each prospect according to the expectation (probability-weighted sum) of the utilities of its possible outcomes. (Or more properly, expected utility theory claims that your preferences should be such that we can *represent* them as maximizing the expectation of some utility function. This utility function need not line up with any antecedently specified moral or prudential value function—this is what distinguishes expected *utility* maximization from expected *value* maximization.) Bounded EU adds the claim that your utilities should be bounded: There should be some real numbers $\overline{u}$ and $\underline{u}$ such that no outcome receives a utility greater than $\overline{u}$ or less than $\underline{u}$. This need

deny, for instance, that any number of lollipop licks can add up to a greater benefit than one close friendship.) The parallel move for the anti-fanatic would be to claim that, for any $p' \leq p$, no number of p' probability shifts to an astronomically good outcome o^{*+} can add up to certainty of o^{*+}. But this is obviously false, given that $p > 0$.

not imply that there is a highest-utility or a lowest-utility outcome, since the utilities of outcomes may approach the bounds without ever reaching them. But it does satisfy Narrow Anti-Fanaticism: For instance, given a choice between certainty of an outcome with near-maximal utility and a risky prospect that is almost certain to yield a significantly worse outcome, any long-shot reward will be at most slightly better than the sure thing, and so insufficient to justify the risk.

There are various motivations for the requirement of boundedness in expected utility theory. Unbounded utilities allow for prospects with *infinite* expected utility (generalizations of the St. Petersburg game), which have various paradoxical properties and are in tension with aspects of expected utility theory.[13] But in addition to these more technical arguments, some expected utility theorists have seen the avoidance of counterintuitive fanaticism as a sufficient justification for boundedness.[14]

The upshot of the preceding discussion, however, is that despite appearances (and despite satisfying Narrow Anti-Fanaticism), Bounded EU in its standard form is not generally anti-fanatical: Since Bounded EU is acyclic, Theorem 1 implies that, given No Best Outcome, No Worst Outcome, and Minimal Dominance, it will not satisfy General Anti-Fanaticism.

The intuitive explanation for this fact (focusing, as usual, on the positive case) is that as we approach the upper bound of the utility function, the marginal utility of any given improvement must decrease very rapidly. Consider, for instance, a choice between shifting a small amount of probability from a mediocre outcome (near the middle of the utility function) to an

[13] In particular, prospects with infinite expected utility violate both Continuity and an infinitary generalization of Independence, the essential characteristic principle of expected utility theory. Menger (1934) was the first to observe that any expected utility maximizer with an unbounded utility function would be susceptible to such prospects (though he did not see boundedness as the solution, preferring instead a form of small-probability discounting). Russell and Isaacs (2021) give a cutting-edge presentation of the case for boundedness based on the paradoxical features of St. Petersburg-like lotteries. (As they point out, however, without the questionably-motivated requirement of Continuity, what these arguments support is not strictly boundedness but a weaker property they call "limitedness", which permits lexicographic utilities, but requires utilities in any lexicographic equivalence class to be bounded.) On the other hand, Goodsell (2024) shows that it is possible to construct a consistent decision theory that allows unbounded utility and satisfies many of the core principles of expected utility theory, apart from the infinitary form of Independence.

[14] For instance, Aumann writes: "Unbounded utility would lead to counterintuitive conclusions even without the St. Petersburg paradox. If Paul's utility were unbounded, then for any fixed prospect x (e.g., a long, happy, and useful life), there would be a prospect y with the property that Paul would prefer a lottery yielding y with probability $\frac{1}{10^{100}}$ and death with the complementary probability to the prospect x. This, I think, is about as hard to swallow as the idea of infinite utility" (Aumann, 1977, p. 444). For similar sentiments, see Machina (1982, pp. 283-4).

astronomically good outcome (near the upper bound), or else applying some fixed improvement to every outcome. If the baseline probability of the astronomically good outcome is already high, then the increment to *expected* utility from the latter option will be very small, since the improvement is most likely to be applied to an outcome that already has nearly maximal utility.

As a concrete illustration, consider a choice between slightly reducing the risk of premature human extinction and creating some fixed benefit, e.g. saving a life. Suppose you think that the baseline risk of premature extinction is fairly low, and that if we avoid it, we are very likely to achieve a very good future. Then, as a bounded expected utility maximizer, you will be likely to prefer even a tiny reduction in the already-small risk of human extinction to a sure improvement that is very likely to improve an already-very-good world, and so counts for very little.[15]

It is *possible* for a bounded expected utility maximizer to satisfy General Anti-Fanaticism, by assigning some outcome(s) maximal utility and some outcome(s) minimal utility. (We might call this *compact* expected utility, since it requires the range of the utility function to be not just bounded but compact.) General Anti-Fanaticism is then satisfied because (focusing on the positive case) we can find an improvement that maps every outcome to an outcome with maximal utility. Applying this improvement to every outcome in a baseline prospect yields a prospect with maximal expected utility, which is weakly preferred to any fanatical alternative, and strictly preferred unless the alternative prospect consists entirely of maximal outcomes.

Unlike the standard form of Bounded EU, however, this view must give up either No Best/Worst Outcome or Minimal Dominance. While one might deny No Best/Worst Outcome, it seems backward to give up these principles for the sake of avoiding fanaticism. (The value of outcomes is, it seems to me, a matter prior to and independent of the ranking of risky prospects; it would be strange to deny that, for instance, saving or improving a life always makes the world a better place merely because this denial makes things easier for decision theorists.) So someone who wants to satisfy General Anti-Fanaticism within the confines of expected utility theory should, I think, deny Minimal Dominance. In particular, they should hold that one's utility function ought to be strictly increasing in the value of outcomes

[15] Beckstead and Thomas (2024, §3.4) make a related point, that for a concrete improvement like benefiting some fixed set of people, Bounded EU can prefer a prospect that offers an arbitrarily small probability of that improvement to one that offers a much greater probability, if the latter yields the improvement in states where the world is already very close to the upper bound of utility, while the former yields the improvement in states where it is far from the bound. This is another illustration of Bounded EU's potential fanaticism.

(assigning greater utility to better outcomes) within some middle range of outcomes, but constant above and below that range. Given No Best/Worst Outcome, this would mean that infinitely many outcomes have maximal utility, and infinitely many outcomes have minimal utility. For instance, perhaps any outcome at least as good as one trillion people living long, happy lives has maximal utility, and any outcome at least as bad as one trillion people living long, miserable lives has minimal utility. Differences above and below these thresholds would then simply count for nothing, from the point of view of rational decision-making. This amounts to a strong sort of rational satisficing for very good outcomes, and a symmetrical property ("desensitizing"?) for very bad outcomes. I don't find this view at all plausible.[16] But someone who was initially attracted to Bounded EU and so willing to accept that apparently-significant improvements to very good outcomes (and worsenings of very bad outcomes) have *vanishingly little* marginal utility might be prepared to go a step further and assert that they have *zero* marginal utility.

7 Small-probability discounting

The other commonly proposed strategy for avoiding both fanaticism and the paradoxical implications of St. Petersburg-like prospects is to simply *ignore* small probabilities. This idea has a long history, going back at least to a 1714 letter from Nicolaus Bernoulli to Pierre Rémond de Montmort. And it has recently experienced something of a revival, being advocated by Smith (2014, 2016) and Monton (2019), and seriously entertained by Buchak (2013, pp. 73–4) and Hong (2024).[17]

Small-probability discounting can take many forms. The simplest forms are *state discounting*, which ignores very improbable states, and *outcome discounting*, which ignores very improbable outcomes. The *very* simplest versions of these approaches, which tell us to ignore all states or outcomes

[16] Among other things, while I think Compact EU should clearly count as anti-fanatical, it is in tension with some of the intuitions that might motivate anti-fanaticism, like the intuition that certain concrete improvements to the world (e.g., saving a life) carry substantial unconditional normative weight. The claim that saving a life counts for literally nothing if the world is already sufficiently good will, I suspect, hold no more intuitive appeal for anti-fanatics than for anyone else.

[17] For a detailed history of the idea of small-probability discounting, see Monton (2019). There is also a parallel literature, of more recent origin, that discusses the idea of small-probability discounting under the name of "*de minimis* risk", though in this literature it tends to be understood more as a heuristic for policymakers and regulators than as a basic principle of rationality. For a succinct and representative statement of the small-probability discounting view in this literature, see Comar (1979).

below some fixed threshold t, seem unworkable, since there will be situations where *all* states and outcomes have probabilities below that threshold (Arrow, 1951, pp. 414–5). But more sophisticated versions of state/outcome discounting avoid this problem, e.g. by ignoring all states/outcomes that are at least n times less probable than the *most* probable state/outcome, or ignoring the *least* probable states/outcomes up to a total probability of t. Even with these amendments, though, state and outcome discounting are subject to powerful objections. They are implausibly sensitive to small differences that can turn a single high-probability state/outcome into many low-probability states/outcomes (Beckstead and Thomas, 2024, §2.3). And they can easily run afoul of dominance principles (Isaacs, 2016; Kosonen, 2022, pp., 151–64; Beckstead and Thomas, 2024, §2.3).

A more promising way of ignoring small probabilities is *tail discounting*.[18] Roughly, tail discounting tells us to ignore the very-worst-case and very-best-case outcomes of every prospect, up to a certain probability (say, 0.01%)—either simply removing those worst-case and best-case outcomes from each prospect altogether (and renormalizing the remaining probabilities), or rounding those outcomes up/down (so that, for instance, any possible outcomes better than the 99.99th percentile of possible outcomes are "rounded down" to the 99.99th percentile outcome, and any possible outcomes worse than the 0.01st percentile are "rounded up" to the 0.01st percentile outcome). We can then apply expected value maximization, or another decision rule, to these "truncated" prospects.

Tail discounting has significant advantages over state and outcome discounting. In particular, its verdicts do not depend on how we individuate states or outcomes, and it therefore avoids extreme sensitivity to small differences between outcomes. Moreover, it can be straightforwardly reconciled with statewise and stochastic dominance principles, by stipulating that the truncated portions of a prospect are used as tiebreakers (Beckstead and Thomas, 2024, §2.3).

But as with Bounded EU, the arguments of §§3–5 imply that whatever its other merits and despite initial appearances, tail discounting does not fully satisfy our anti-fanatical intuitions. This is easier to see for tail discounting than for Bounded EU. If, for instance, we are considering how much we should be willing to sacrifice to reduce the probability of near-term human extinction from 0.5 to $0.5-\varepsilon$, tail-discounted expected value maximization

[18]Versions of this idea are discussed, though not endorsed, by Buchak (2013, pp. 73–4), Kosonen (2022, pp. 164–78), and Beckstead and Thomas (2024, §2.3). For an argument against tail discounting (along with other forms of small-probability discounting), see Kosonen (2024).

will be just as fanatical as ordinary expected value maximization.[19]

As with Bounded EU, it is possible to devise a version of small-probability discounting that satisfies General Anti-Fanaticism. I will describe one such version, which is based on representing prospects by their *quantile functions*. For a prospect P_i on a set of totally ordered outcomes, the quantile function of Q_i of P_i is a function from probabilities to outcomes, mapping any probability $p \in (0,1)$ to the worst outcome o such that the probability of an outcome no better than o is greater than or equal to p.[20] Thus, for instance, for p sufficiently close to zero, $Q_i(p)$ gives the worst possible outcome of P_i; for p sufficiently close to 1, $Q_i(p)$ gives the best possible outcome; $Q_i(0.5)$ gives the median outcome, $Q_i(0.75)$ the 75th percentile outcome, and so on.

Figure 1 gives an example, comparing the quantile functions of (i) a baseline prospect, (ii) a non-fanatical prospect that applies an improvement to every outcome in the baseline, and (iii) a fanatical prospect that instead shifts some probability from a much worse to a much better outcome. The characteristic feature of the "Pascalian" choice situations in which the question of fanaticism arises is that the quantile function of the non-fanatical option is *slightly* greater (that is, yields a slightly better outcome) almost everywhere, while the quantile function of the fanatical option is greater for only a small range of p, but *much* greater (that is, yields a much better outcome) at least somewhere in that range.

This concept in hand, we can now describe a version of small-probability discounting that satisfies General Anti-Fanaticism:

Quantile Discounting For any prospects P_i and P_j with quantile functions Q_i and Q_j, if the range of quantiles for which Q_j exceeds Q_i (formally, the Lebesgue measure of the set $\{p \in (0,1) : Q_j(p) > Q_i(p)\}$) is less than or equal to some $\underline{t}$ and the range of quantiles where Q_i exceeds Q_j is greater than or equal to a further threshold $\overline{t} \leq 1 - \underline{t}$, then $P_i \succ P_j$.[21]

[19] For further illustrations of this point, see Kosonen (2022, Ch. 6) and Cibinel (2023). Both Kosonen and Cibinel make many of the same points about small-probability discounting in particular that §§3–5 have made about anti-fanaticism in general.

[20] The definition is a bit more complicated for continuous prospects, but we are restricting our attention here to discrete prospects with only finitely many possible outcomes.

[21] This isn't a fully spelled-out decision rule. We could turn it into one by integrating it with expected value maximization, perhaps along the following lines: Let the values of the quantile function be real numbers representing the values of outcomes. For two prospects P_i and P_j, the difference in expected value between them is then given by the integral of the difference of their quantile functions: $\int_0^1 Q_i(x) - Q_j(x)\,dx$. Quantile-discounted MEV might then compare P_i and P_j by taking this integral, while excluding one or more intervals of total length $\underline{t}$ chosen to maximize $Q_i - Q_j$, and likewise intervals of total length $\underline{t}$ chosen

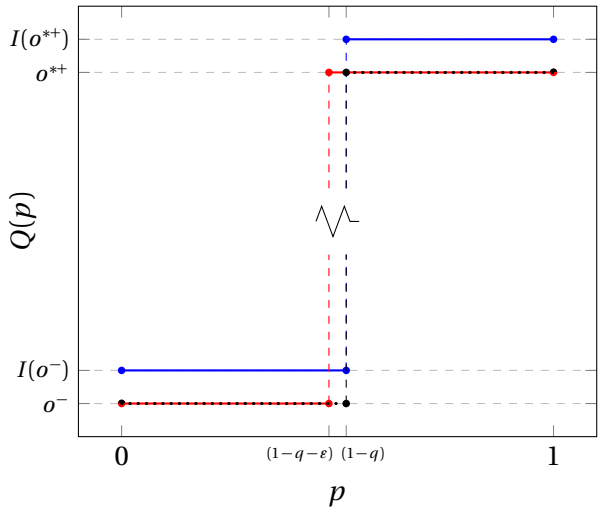

FIGURE 1: Quantile functions of a baseline prospect (dotted black), a non-fanatical prospect (blue) that slightly improves both possible outcomes of the baseline prospect, and a fanatical prospect (red) that increases the probability of the more desirable outcome by ε.

This view is, I think, the most natural and plausible form of General Anti-Fanaticism. It straightforwardly reflects and vindicates intuitions such as, for instance, that we should always prefer a sure improvement to the world over a sufficiently small reduction in the probability of existential catastrophe, no matter the baseline probability of catastrophe.[22] But as we have seen, given No Best/Worst Outcome and Minimal Dominance, satisfying General Anti-Fanaticism must come at the cost of preference cycles. In the case illustrated by Table 2, for instance, quantile discounting will strictly prefer P_{i+1} to P_i for every $i < n$. Again, I don't want to totally rule out the possibility of embracing cyclicity—as described in §5, there is the mitigating circumstance that quantile discounting with a small enough probability threshold only generates very *long* cycles. But it is a steep cost

to minimize $Q_i - Q_j$. P_i is weakly preferred to P_j iff this restricted integral is weakly positive (≥ 0). In the example depicted in Figure 1, for instance, this rule is guaranteed to strictly prefer the non-fanatical prospect (blue) as long is $\varepsilon < \underline{t}$.

[22] A small caveat: The concept of quantiles, and therefore both tail discounting and quantile discounting, are tricky to generalize to cases where outcomes are not totally ordered. In this context, quantile discounting may not satisfy General Positive Anti-Fanaticism if, for every improvement I, it is possible to find a pair of outcomes o^{*+} and o^- such that o^{*+} is strictly better than o^- but incomparable with $I(o^-)$. We could fix this issue by restricting General Positive Anti-Fanaticism to pairs of outcomes o^{*+} and o^- such that $o^{*+} \succ_{\mathbb{O}} I(o^-)$. (As usual, analogous remarks apply to the negative case.)

to pay in order to satisfy our anti-fanatical intuitions.[23]

8 An argument for fanaticism?

We have seen that there is a strong argument against General Anti-Fanaticism. Is this also an argument for fanaticism?

On the one hand, it is possible to extend the argument from §5 into an argument for the following weak version of fanaticism:

Weakly General Positive Fanaticism It is rationally required that, for every improvement I and probability $p > 0$, there are outcomes $o^{*+} \succ_{\mathbb{O}} o^{-}$ and probabilities $q < 1$, $p' \leq p$ such that the prospect $\langle o^{*+}, q + p', o^{-} \rangle$ is strictly preferred to $\langle I(o^{*+}), q, I(o^{-}) \rangle$.

Weakly General Negative Fanaticism It is rationally required that, for every worsening W and probability $p > 0$, there are outcomes $o^{+} \succ_{\mathbb{O}} o^{*-}$ and probabilities $q < 1$, $p' \leq p$ such that the prospect $\langle W(o^{*-}), q, W(o^{+}) \rangle$ is strictly preferred to $\langle o^{*-}, q + p', o^{+} \rangle$.

Weakly General Fanaticism (the conjunction of these theses) weakens Narrow Fanaticism in several ways, by existentially quantifying over the

[23] An alternative strategy for capturing more of our anti-fanatical intuitions within the general framework of small-probability discounting is *difference-making* tail discounting. Roughly speaking, where tail discounting ignores the most extreme outcomes of a prospect, difference-making tail discounting ignores the outcomes that *differ* most from the outcome of the baseline prospect in the same state of nature. This form of tail discounting is discussed in Kosonen (2022, Ch. 4) and in unpublished work by Jacob Barrett (Barrett, ms), who is sympathetic to the approach. On the other hand, Greaves et al. (2024) make a general case against any view that, like difference-making tail discounting, combines a primary concern with the difference one makes to the world with a non-neutral attitude toward risk.

I won't attempt a full discussion of the difference-making approach here, but will just make one point: Although it comes closer than simple tail discounting to capturing our anti-fanatical intuitions, I don't think it gets all the way there. Consider the case of existential risk mitigation. As I suggested earlier (fn. 3), it might well be that the course of history is extremely sensitive to small changes, such that any action you take carries substantial probabilities both of causing and of preventing some future existential catastrophe. An action meant to mitigate existential risk, then, does not carry a tiny probability of preventing an existential catastrophe (and thereby making a very large difference to the baseline outcome)—rather, what is tiny is the *difference* between the probability of that action preventing existential catastrophe and the probability of any given alternative action preventing existential catastrophe. If the world in fact works this way, then difference-making tail discounting will do little if anything to blunt expected value arguments for existential risk mitigation, since the probability of making an astronomically large difference is not small enough to discount. But I don't think that those expectational arguments are any less counterintuitive in these circumstances.

baseline probability q, probabilities $p' \leq p$, and the "or else" outcome (o^- in the positive case, o^+ in the negative case). (Narrow Fanaticism concerns $q = 0$ and all $p' \geq p$, and universally quantifies over "or else" outcomes.)

To get an argument for Weakly General Fanaticism, we have to strengthen Acyclicity to Transitivity (the requirement that, if $P_i \succsim P_j$ and $P_j \succsim P_k$, then $P_i \succsim P_k$), and add one new premise:

Completeness It is rationally required that, for any prospects P_i and P_j, either $P_i \succsim P_j$ or $P_j \succsim P_i$ (or both).

We can then prove the following result:

Theorem 2. *Completeness, Transitivity, Minimal Dominance, and No Best Outcome imply Weakly General Positive Fanaticism. Completeness, Transitivity, Minimal Dominance, and No Worst Outcome imply Weakly General Negative Fanaticism.*[24]

Here is the proof (again focusing on the positive case): Let I be any improvement and p any positive probability. Choose an n such that $\frac{1}{n} \leq p$. Then consider the set of prospects described in Table 3. This is a slight modification of the case in Table 2, with the astronomical improvement now consisting of $n+1$ iterations of I, so that after n instances of choosing certainty of a small improvement over a $\frac{1}{n}$ chance of an astronomical improvement, we end up not back where we started but with a strictly worse prospect—certainty of $I^{n-1}(o)$ rather than $I^n(o)$. No Best Outcome implies that improvements are always strict, so $I^n(o) \succ_{\mathbb{O}} I^{n-1}(o)$. By Completeness, for every $i < n$, either $P_{i+1} \succsim P_i$ or $P_i \succ P_{i+1}$. Suppose first that $P_{i+1} \succsim P_i$ for every $i < n$. Then, by Transitivity, $P_n \succsim P_0$. But this contradicts Minimal Dominance, which requires that $P_0 = \langle I^n(o) \rangle \succ P_n = \langle I^{n-1}(o) \rangle$. So there must be some i for which $P_i \succ P_{i+1}$. And this vindicates Weakly General Positive Fanaticism: We have found outcomes o^- (namely, $I^{i-1}(o)$) and o^{*+} (namely, $I^{n+i}(o)$), a baseline probability q (namely, $\frac{n-i-1}{n}$), and a $p' \leq p$ (namely, $\frac{1}{n}$) such that increasing the baseline probability of o^{*+} by p' (yielding P_i)is preferred to applying improvement I to both outcomes (yielding P_{i+1}).

But this argument is substantially less compelling than the argument against General Anti-Fanaticism. While I have no objection to Transitivity, it is significantly stronger and more controversial than Acyclicity. Completeness is more controversial still—and for my part, I am inclined to reject it.

[24] Replacing Minimal Dominance with Binary Monotonicity (fn. 11) would let us derive a version of Weakly General Fanaticism that universally quantifies over $p' \geq p$.

	$s_1\,(\frac{1}{n})$	$s_2\,(\frac{1}{n})$	$s_3\,(\frac{1}{n})$	$\cdots$	$s_n\,(\frac{1}{n})$
P_0	$I^n(o)$	$I^n(o)$	$I^n(o)$	$\cdots$	$I^n(o)$
P_1	o	$I^{n+1}(o)$	$I^{n+1}(o)$	$\cdots$	$I^{n+1}(o)$
P_2	$I(o)$	$I(o)$	$I^{n+2}(o)$	$\cdots$	$I^{n+2}(o)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
P_n	$I^{n-1}(o)$	$I^{n-1}(o)$	$I^{n-1}(o)$	$\cdots$	$I^{n-1}(o)$

TABLE 3: An illustration of the argument for Weakly General Fanaticism

Moreover, even if the argument succeeds, Weakly General Fanaticism is a weak enough thesis that it hardly seems like a vindication of fanaticism. For instance, unlike Narrow Fanaticism, Weakly General Fanaticism is compatible with Bounded EU—which may not be comprehensively anti-fanatical, but is not intuitively fanatical either. Moreover, no similar argument for Narrow Fanaticism will be forthcoming: Since Bounded EU (for example) satisfies Completeness and Transitivity and is compatible with No Best/Worst Outcome and Minimal Dominance, but does not satisfy Narrow Fanaticism, the premises that imply Weakly General Fanaticism cannot imply this stronger thesis.

Of course, the arguments in §3 suggest that our real focus should not be on Narrow Fanaticism, but on a version of fanaticism that applies more generally to the context of uncertain baseline prospects. If we generalize Narrow Fanaticism to this context in the same way we did for Narrow Anti-Fanaticism, by universally quantifying over the baseline probability q, we get the following theses:

General Positive Fanaticism It is rationally required that, for any outcome o^-, improvement I, and probability $p > 0$, there is some outcome o^{*+} such that the prospect $\langle o^{*+}, q + p', o^- \rangle$ is strictly preferred to $\langle I(o^{*+}), q, I(o^-) \rangle$ for all probabilities $p' \geq p$ and $q \leq 1 - p'$.

General Negative Fanaticism It is rationally required that, for any outcome o^+, worsening W, and probability $p > 0$, there is an outcome o^{*-} such that the prospect $\langle W(o^{*-}), q, W(o^+) \rangle$ is strictly preferred to $\langle o^{*-}, q + p', o^+ \rangle$ for all probabilities $p' \geq p$ and $q \leq 1 - p'$.

Just as General Anti-Fanaticism is stronger than Narrow Anti-Fanaticism, so General Fanaticism is stronger than Narrow Fanaticism.[25] Since the

[25]The difference in strength is somewhat less interesting, though: I'm not aware of any plausible decision theory that would satisfy Narrow but not General Fanaticism.

preceding arguments do not support Narrow Fanaticism, therefore, they will not support General Fanaticism either.

9 Conclusion

I have argued that the debate between fanaticism and anti-fanaticism should focus not just on simple choices between risk and certainty, but on a more general setting of choices between a sure improvement or a small probability shift in a risky baseline prospect. This setting more fully captures the question of how much weight we should give to "small probabilities" (or more appropriately, small probability differences), and the important real-world cases where this question arises. And only a version of anti-fanaticism formulated in this more general setting can fully capture our intuitive resistance to fanaticism.

We have then seen a strong argument against General Anti-Fanaticism, which is not matched by an equally strong argument for General (or Narrow) Fanaticism. What should we make of this gap in argumentative strength? The takeaway, I think, is that those who find fanaticism counterintuitive should favor not anti-fanaticism but permissivism. More specifically, they should favor a version of permissivism that permits *incomplete* preferences that are neither fanatical nor anti-fanatical.[26]

An agent with incomplete preferences *could* behave, in practice, very much like one who satisfies General Anti-Fanaticism, while satisfying both Minimal Dominance and Acyclicity, and avoiding sure losses. For instance, consider an agent facing a sequence of choices among the prospects from Table 3. She might start off with no preferences except for the dominance-based preference for P_0 over P_n, but treat each choice as an update to her preferences. Specifically, if she chooses P_i when P_j is available, she adopts the preference $P_i \succsim P_j$ and all that it entails by transitive closure. (So, for instance, if she already strictly preferred P_j to P_k, she will now also strictly prefer P_i to P_k.) And when she has preexisting preferences among the available prospects, she chooses according to those preferences. This policy prevents her from making the full dominated sequence of trades from P_0 to P_1, P_1 to P_2, and so on until she is left with P_n: By the time she has traded P_{n-2} for P_{n-1}, she has adopted the preferences $P_{n-1} \succsim P_{n-2} \succsim \ldots \succsim P_1 \succsim$

[26] I defend one such version of permissivism in Tarsney (2020). Note that, while the permissivist might also allow an agent to form complete preferences of a more or less fanatical character, the cyclicity argument in §5 tells against any view that even permits preferences to satisfy General Anti-Fanaticism, since any such view must deny at least one of Acyclicity, Minimal Dominance, or No Best/Worst Outcome.

$P_0 \succ P_n$, and so will prefer and choose P_{n-1} over P_n. But, as the example illustrates, this policy *does* allow her to act like a General Anti-Fanatic for quite a long time, if she likes. That is, she can prefer sure improvements to small probability shifts until the preferences implied by her past choices force her to do otherwise. And that might not happen until she has made a truly enormous number of choices—perhaps more choices than an ordinary human being will face in a lifetime.

Of course, I have given no reason to prefer the permissivist view over simply embracing fanaticism, apart from emphasizing the counterintuitiveness of fanaticism. And I have not tried to answer the various arguments for fanaticism in the recent literature. These are topics for another occasion. I have only pointed out that we should not equate *the rejection of fanaticism* with *anti-fanaticism*, and that there is both logical room and argumentative motivation for a middle ground.[27]

References

Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. *Econometrica: Journal of the Econometric Society 19*, 404–437. doi: 10.2307/1907465.

Aumann, R. J. (1977). The St. Petersburg paradox: A discussion of some recent comments. *Journal of Economic Theory 14*, 443–445. doi: 10.1016/0022-0531(77)90143-0.

Balfour, D. (2021). Pascal's mugger strikes again. *Utilitas 33*, 118–124. doi: 10.1017/S0953820820000357.

Barrett, J. In defense of moderation. Unpublished manuscript.

Beckstead, N. (2019). A brief argument for the overwhelming importance of shaping the far future. In H. Greaves and T. Pummer (Eds.), *Effective Altruism: Philosophical Issues*, pp. 80–98. Oxford: Oxford University Press. doi: 10.1093/oso/9780198841364.003.0006.

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

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy 4*, 15–31. doi: 10.1111/1758-5899.12002.

[27] For extremely helpful comments on earlier versions of this paper, I am grateful to Jacob Barrett, Jeffrey Russell, Teruji Thomas, and participants in work-in-progress seminars at the Global Priorities Institute and UT Austin.

Buchak, L. (2013). *Risk and Rationality*. Oxford: Oxford University Press.

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

Comar, C. (1979). Risk: A pragmatic de minimis approach. *Science 203*, 319–319. doi: 10.1126/science.203.4378.319.

Cowen, T. (2007). Caring about the distant future: Why it matters and what it means. *University of Chicago Law Review 74*, 5–40.

Goodsell, Z. (2024). Decision theory unbound. *Noûs 58*, 669–695. doi: 10.1111/nous.12473.

Greaves, H. and W. MacAskill (2021). The case for strong longtermism. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 5-2021.

Greaves, H., T. Thomas, A. Mogensen, and W. MacAskill (2024). On the desire to make a difference. *Philosophical Studies 181*, 1599–1626. doi: 10.1007/s11098-024-02102-0.

Gustafsson, J. E. (2022). *Money-Pump Arguments*. Cambridge: Cambridge University Press. doi: 10.1017/9781108754750.

Hong, F. (2024). Know your way out of St. Petersburg: An exploration of "knowledge-first" decision theory. *Erkenntnis 89*, 2473–2492. doi: 10.1007/s10670-022-00639-2.

Isaacs, Y. (2016). Probabilities cannot be rationally neglected. *Mind 125*, 759–762. doi: 10.1093/mind/fzv151.

Kosonen, P. (2022). *Tiny Probabilities of Vast Value*. Ph. D. thesis, Worcester College, University of Oxford.

Kosonen, P. (2024). Probability discounting and money pumps. *Philosophy and Phenomenological Research*. Advanced online publication. doi: 10.1111/phpr.13053.

Machina, M. J. (1982). "Expected utility" analysis without the independence axiom. *Econometrica: Journal of the Econometric Society 50*, 277–323. doi: 10.2307/1912631.

Menger, K. (1979 (1934)). The role of uncertainty in economics. In H. L. Mulder (Ed.), *Selected Papers in Logic and Foundations, Didactics, Economics*, pp. 259–278. Dordrecht: D. Reidel. doi: 10.1007/978-94-009-9347-1_25.

Monton, B. (2019). How to avoid maximizing expected utility. *Philosophers' Imprint 19*, 1–25.

Parfit, D. (2003). Justifiability to each person. *Ratio 16*, 368–390. doi: 10.1046/j.1467-9329.2003.00229.x.

Russell, J. S. (2023). On two arguments for fanaticism. *Noûs 58*, 565–595. doi: 10.1111/nous.12461.

Russell, J. S. and Y. Isaacs (2021). Infinite prospects. *Philosophy and Phenomenological Research 103*, 178–198. doi: 10.1111/phpr.12704.

Smith, N. J. J. (2014). Is evaluative compositionality a requirement of rationality? *Mind 123*, 457–502. doi: 10.1093/mind/fzu072.

Smith, N. J. J. (2016). Infinite decisions and rationally negligible probabilities. *Mind 125*, 1199–1212. doi: 10.1093/mind/fzv209.

Tarsney, C. (2020). Exceeding expectations: Stochastic dominance as a general decision theory. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 3-2020.

Wilkinson, H. (2022). In defense of fanaticism. *Ethics 132*, 445–477. doi: 10.1086/716869.