On two arguments for Fanaticism

Jeffrey Sanford Russell (University of Southern California)

Global Priorities Institute | October 2021

GPI Working Paper No . 17-2021

On Two Arguments for Fanaticism*

Jeffrey Sanford Russell

October 2021 draft

Abstract

Should we make significant sacrifices to ever-so-slightly lower the chance of extremely bad outcomes, or to ever-so-slightly raise the chance of extremely good outcomes? *Fanaticism* says yes: for every bad outcome, there is a tiny chance of extreme disaster that is even worse, and for every good outcome, there is a tiny chance of an enormous good that is even better. I consider two related recent arguments for Fanaticism: Beckstead and Thomas's argument from *strange dependence on space and time*, and Wilkinson's *Indology* argument. While both arguments are instructive, neither is persuasive. In fact, the general principles that underwrite the arguments (a *separability* principle in the first case, and a *reflection* principle in the second) are *inconsistent* with Fanaticism. In both cases, though, it is possible to rehabilitate arguments for Fanaticism based on restricted versions of those principles. The situation is unstable: plausible general principles tell *against* Fanaticism, but restrictions of those same principles (with strengthened auxiliary assumptions) *support* Fanaticism. All of the consistent views that emerge are very strange.

> Not madness but the mathematics of eternity drove them.
>
> Mary Doria Russell, *The Sparrow*

*Thanks to Zach Goodsell, John Hawthorne, Frank Hong, Harvey Lederman, Christian Tarsney, and other participants in the 6th Oxford Workshop on Global Priorities Research. Special thanks to the Global Priorities Institute's Philosophy Work In Progress group for their extensive feedback and discussion.

1 Almost certainly pointless, but good?

> Kayla has a minor cough. She knows it is probably nothing serious—very likely it's just her seasonal allergies—but there is a small chance that it is COVID-19. She decides that she had better skip her mother's birthday celebration this year. It will already be a much smaller gathering this year than usual, but her grandfather will be there, and she doesn't want to risk giving him COVID. Still, the choice breaks Kayla's heart; she knows her mother will be deeply disappointed. Part of what makes it so frustrating is that she knows that her cough is probably nothing—so she feels like she is giving up something important for nothing.

Choices like Kayla's are frustrating; even so, sometimes it is best to give up something morally important, even though *very probably* no good will come of the sacrifice. This can happen when the stakes are high enough. Disappointing her mother is a bad thing, but not nearly as bad as giving her grandfather a deadly disease.

How far can such trade-offs take us? What if the probability of losing something important seems negligibly tiny, but what would be lost is unbearably immense—trillions of lives, whole worlds of good? Can it be worthwhile to make weighty sacrifices to avoid such risks, even though the sacrifices are almost certainly pointless?

Bostrom (2003) considers space settlement. He argues:

> [T]he Virgo Supercluster could contain 10^{23} biological humans … What matters for present purposes is not the exact numbers but the fact that they are huge. (p. 309)

Thus, according to Bostrom, it is of paramount importance that we ensure that humanity reaches as much of the potentially habitable universe as possible. One way this could be thwarted is if an existential catastrophe were to "either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." (p. 310) Bostrom's "most conservative" estimate suggests that reducing existential risk by just one in a million million million millon would be better than saving a hundred thousand human lives directly. He concludes:

> For standard utilitarians, priority number one, two, three and four should consequently be to reduce existential risk.

Many others have been taken with this line of thought; they have also worked to generalize similar lessons beyond the concerns of just "standard utilitarians" to a broader range of ethical starting points (Ord 2020; Greaves and MacAskill 2021). But this conclusion is troubling. Making such a sacrifice would have a 99.9999999999999999999999% chance of doing no good at all—in which case it amounts to sacrificing a hundred thousand lives for naught. This seems *fanatical*.[1]

Are such sacrifices good? Or is there some limit to how tiny the probability of doing any good can be for a given sacrifice to be warranted? The general question is whether the following principle is true.[2]

(Positive) Fanaticism. For any finite good x and for any non-zero probability p, there is some finite good y such that it is better to have y with probability p than to have x for sure.

Bostrom's case involves sacrificing something good for a small chance of great gain. In contrast, Kayla's case involves sacrificing something good to *avoid* a small chance of great loss. This corresponds to a dual form of Fanaticism:

Negative Fanaticism. For any finite loss x and any non-zero probability p, there is some finite loss y such that it is better to have x for sure than to have y with probability p.

It is natural, though not logically inevitable, that the two theses should go together. If the goods that warrant positive fanaticism involve large numbers of happy people, or long ages of flourishing, then there are corresponding evils involving large numbers of suffering people, or eons of despair. In what follows we will primarily focus on the positive thesis, for economy of presentation.

There are two ideas built into Fanaticism: an *axiological* idea about how good things are, and a *decision-theoretic* idea about how to weigh risks. The axiological idea is that some things are extremely valuable—for Bostrom's argument, these are extremely large happy populations. The decision-theoretic idea is that an absurdly small chance of something extremely valuable can be worth a large cost. Accordingly, intuitively there are two different ways in which one might *reject* Fanaticism. One might reject the axiological thought, and hold that value is *bounded*. Nothing is

[1] This is Bostrom's term from another context (2011), which seems to have caught on.
[2] See appendix A for technical background assumptions.

good enough to warrant some sacrifices at small odds. Or one might reject the decision-theoretic thought, and hold that when it comes to "fanatical" trade-offs, the option that maximizes expected value is not best. One might say that cardinal *value* comes apart from the cardinal *utility* function whose expectation ought to be maximized—and furthermore, this utility function is bounded.[3] Or one might hold that one should *discount* sufficiently small probabilities, rounding them down to zero (see Smith 2014; Monton 2019; Beckstead and Thomas 2020, sec. 2.3). Even if some things are extremely good, perhaps a very small chance of such a thing is not good in proportion to its chance.

It is not always easy to separate the axiological idea from the decision-theoretic idea: that requires us to make sense of a cardinal scale of value *apart* from how values are weighed in trading off risks. Some theories allow us to make sense of such a scale, but not all do. But we do not have to disentangle the two ideas in order to evaluate Fanaticism.

Why would anyone be tempted to Fanaticism? One way in is from what Bostrom called a "standard utilitarian" starting point. This package includes a *totalist* axiology, according to which very large happy populations are extremely good. It also includes an *expectational* decision theory, according to which the value of a chance p at getting an outcome x is given by multiplying the value of x by p. A huge number multiplied by a tiny number can still be very big.

But Fanaticism itself is not tied to either part of this specific picture. Not all ways of rejecting totalism or expectationalism are ways of escaping Fanaticism. And there are much more general arguments for Fanaticism that do not rely on Bostrom's starting point. In this essay I will be focus on two closely related, interesting, and powerful arguments for Fanaticism: the argument from *strange dependence on distant space and time* from Beckstead and Thomas (2020), and the *Indology* argument from Wilkinson (forthcoming).

The Fanaticism thesis concerns very large finite values. We can contrast fanatical wagers with properly *Pascalian* wagers, which intuitively have *infinite* values at stake. But it turns out that there are many tight connections between large finite values and infinite values. This essay is about those connections—and specifically how infinite lotteries make trouble for certain

[3]Or one might take on a permissive theory of risk of the kind advocated by Tarsney (2020), which says that no particular cardinal utility function is mandatory. More on this in section 3.

arguments for Fanaticism. Cases involving infinitely many possibilities or infinite values raise many paradoxes, but I am convinced that this is not a reason to ignore such cases, but rather to take them seriously, and pay attention to what they can teach us. What can the basic principles of value be like, if they are not to fall into contradiction? (Furthermore, one of the two arguments I am addressing—Wilkinson's Indology argument—already relies on infinite cases.)

I do not ultimately find either of the two arguments for Fanaticism persuasive. But let's be clear: my aim here is not to settle the question of whether Fanaticism is true. I don't know. Whatever the truth of the matter, the ethics of huge numbers is deeply weird and full of surprises. This is something we must face up to. Some paradoxes of infinity are mere intellectual curiosities, but these puzzles are of real practical importance. As Hutchinson (2021) puts it :

> The future of sentient beings is potentially unimaginably large. That means if we have only a very small chance of affecting it in a lasting and positive way, taking that chance is worth it.

We have actions available that amount to taking such chances, but which involve substantial sacrifices of other important goods. For example, we might choose to divert resources that could prevent thousands of cases of malaria to instead *very slightly* reduce the risk of extreme catastrophes from climate change, or artificial intelligence, or pandemics. Since we face genuine options like this, our actual moral predicament is puzzling and troubling.

2 The argument from strange dependence on distant space and time

Beckstead and Thomas (2020, sec. 3.2) argue that if Fanaticism is false, then it turns out that which prospects are best depends on what is going on in far away places and times in weird ways.

Their argument is framed in terms of bringing into existence large numbers of happy people—though as we will see below, it generalizes considerably. For now, let us make the simplifying assumption that all that matters in each outcome is the total number of happy lives: any two outcomes that agree on this number are equally good. (We suppose there is no inequality: all of the different lives in question involve the same amount of happiness.)

It will be helpful to introduce some notation. If p is a probability and n is a number, write $p * n$ for a prospect that results in n happy lives with probability p, and otherwise zero lives with probability $1-p$. For prospects X and Y, we'll use the notation $X \succ Y$ to mean that X is strictly better than Y. In this setting, we can rewrite Fanaticism like this:

> For any number of happy lives n, and for any probability $p > 0$, there is some number of happy lives N such that $p * N \succ 1 * n$.

(Beckstead and Thomas call this conclusion "Recklessness.")

The argument is based on three ideas. The first idea is simple: it is better to have a *much higher* chance of *many more* happy lives, than a *smaller* chance of *fewer.* In symbols,[4]

More is Better. For probabilities $p \gg q$ and numbers $N \gg n$,

$$p * N \succ q * n$$

This seems hard to argue with. It follows from the idea that a much larger happy populations is at least as good as a smaller happy population, together with very modest principles about risk.

The second idea is that the first idea is still true even if you don't know how many happy people there are in distant galaxies. For prospects X and Y, let $X + Y$ be a prospect that, in any state of nature s, results in the happy people that result from X in s *as well* as the happy people that result from Y in s—intuitively, with the Y people all living in some distant galaxy that we have no way of affecting.

The trick is then to consider "nearby" prospects of the form $p * N$ and $q * n$, together with a "distant galaxy" prospect whose uncertainty lines up with the local uncertainty in the right way: see table 1. In the first row we have a prospect of the form $p * N$, and in the second row we have a prospect of the form $q * n$. In the third row we have the distant galaxy prospect, which results in the same smaller number of happy lives n only in the case where

[4] The "much greater than" notation $\gg$ implicitly builds in some non-obvious quantificational structure, which is a bit complicated to spell out. But let's not quibble: I am happy to grant a simpler, stronger premise:

> For any probabilities $p > q$ and any number n, there is some number N such that $p * N \succ q * n$.

This is a consequence of Stochastic Dominance, discussed below.

the $p * N$ prospect succeeds. The final two rows show the result of adding together each local prospect with the fixed distant galaxy prospect.

Table 1: The "local" prospects X and Y are combined with a "distant galaxy" prospect A.

Prospect	p	q	$1-p-q$
X	N	0	0
Y	0	n	0
A	n	0	0
$X+A$	$n+N$	0	0
$Y+A$	n	n	0

Comparing the two combined prospects $X+A$ and $Y+A$, we observe that $X+A$ has a *slightly lower* probability p of a *much larger* number of people $n+N$. So the two ideas so far imply:

Anti-Timidity. For any probabilities $p \gg q$ and numbers $N \gg n$,

$$p * (n+N) \succ (p+q) * n$$

In other words, a chance at a sufficiently large number of happy lives is better than a *slightly higher* chance of a *much smaller* number of happy lives.

The *third* idea of the argument is that Anti-Timidity implies Fanaticism. This is the central observation of Beckstead and Thomas's rich paper, which is based on a "continuum" argument (see also Wilkinson, forthcoming, sec. 4). Here is the basic idea. Suppose you are about to make a world with n happy lives, for sure. Then you are offered a trade: instead of just n lives for sure, you can create a *much larger* number of lives *almost* for sure. Anti-Timidity says that this is good trade: better to have chance $p = 1 - \varepsilon$ of many more happy lives than the slightly higher chance 1 of just n lives. So you take it. Then you are offered another trade: what about the slightly smaller chance $1 - 2\varepsilon$ of even *more* lives? Anti-Timidity recommends this trade, too. And so on, until you are left with a ridiculously *tiny* probability of a truly *enormous* number of lives. By transitivity, this absurdly long odds gamble is better than the sure thing you began with.

(Some deny that betterness is transitive (for example, Temkin 2012), which would block this argument. In order to keep things under control, I will not take up this idea here. Throughout this essay I will assume without further comment that goodness is *ordered*: in particular, *at least as good* is transitive and reflexive, and *better*, *worse*, and *equally good* are related to *at least as good* in the usual ways. I do not generally assume betterness is a *complete* order—different goods may be incomparable, as we will discuss later.)

You might wonder: why so much bother in order to argue for Anti-Timidity, by way of considerations about distant galaxies? The principle already sounds very plausible without all that. The trouble is that (if betterness is an order) consistency demands either Timidity or Fanaticism—and *either one* of these is quite implausible. We face counterintuitive consequences no matter what we say. To make progress, Beckstead and Thomas explore various costs on each side. The argument we are here considering shows that strange dependence on distant space and time is one of the costs of Timidity.

The step from Anti-Timidity to Fanaticism is just math. More is Better—the idea that a much larger probability of many more happy lives is better than a much smaller probability of fewer happy lives—seems pretty unimpeachable. So the key step to examine is the move from More is Better to Anti-Timidity. What motivated this was the idea that if More is Better, then more is *still* better given arbitrary uncertainty about what is going on far away. Here is a general principle that would underwrite such reasoning:

Separability. For any prospects X, Y, and A,

$$X \succ Y \quad \text{iff} \quad X + A \succ Y + A$$

More is Better told us that $p * N \succ q * n$; then Separability tells us that $p * N$ added together with the additional gamble $p * n$ far away is likewise better than $q * n$ added together with same additional gamble—which amounts to Anti-Timidity. We can sum up Beckstead and Thomas's core idea as follows.

Theorem 1 (Beckstead and Thomas)**.** *If all that matters is the number of happy lives, More is Better and Separability together imply Fanaticism.*

Separability is a highly plausible principle. How could distant lives completely unconnected to our actions make any difference to what it is best to do about the here and now? More is Better is also hard to argue with, and the assumption that the numbers are all that matter looks like a harmless idealization. So this seems like a very strong argument for Fanaticism.

Nonetheless, I am convinced that it is unsound: either Separability is false, or else (of necessity) the numbers are not all that matters. This is due to a striking result originally proved by Seidenfeld, Schervish, and Kadane (2009), and applied to population ethics by Goodsell (forthcoming). The argument I'll give below is an application of their proofs with minor modifications. It is closely related to the St. Petersburg paradox. As with Beckstead and Thomas's argument, for now we'll hold onto the simplifying assumption that all that matters is the number of happy lives, and so good outcomes can just be thought of as natural numbers. Later we will generalize.

Consider a lottery for happy lives W (table 2). A fair coin is flipped until it comes up heads. If it's heads on the first flip, there are two happy lives. If it's heads on the second flip, there are four; if the third, eight; and so on.

Now consider another lottery X which has the same probabilities as W, but slightly worse outcomes: where W has probability 2^{-n} of 2^n happy lives, X has probability 2^{-n} of $2^n - 1$ happy lives. It seems clear that W is better than X.

Consider also a third lottery Y which is isomorphic to X. Then W is better than Y as well.

Table 2: Three population lotteries. The lottery W is better than either X or Y.

	$1/2$	$1/4$	$1/8$	$\cdots$
W	2	4	8	$\cdots$
X	1	3	7	$\cdots$
Y	1	3	7	$\cdots$

Now Separability tells us that two *copies* of W—one around here and the other in a distant galaxy—is better than X around here and Y in a distant galaxy. Since $W \succ X$, it is better nearby, and since $W \succ Y$, it is better far off; so $W + W$ is better than $X + Y$ all around.[5]

Now here's the trick. I told you what probabilities these lotteries assigned to their various outcomes, but I *didn't* tell you how the outcomes were arranged across states of nature. We do it in a tricky way (see table 3). As we said, the outcome of lottery W is based on flipping a fair coin until it first lands heads. In addition to those coin flips, we also flip one extra coin—the

[5]To be explicit, we can do this in two steps. First, since $W \succ X$, by Separability $W + W \succ X + W$. Second, since $W \succ Y$, by a second application of Separability $X + W \succ X + Y$.

"bonus coin." For W, we simply ignore the bonus coin and get the same result however it comes up.

For X, if the bonus coin comes up heads you just get one happy life for sure, ignoring all the other coin flips. If the bonus coin comes up tails you get *twice* the outcome of W, minus one happy life. Note that this agrees with the probabilities in table 2: probability $1/2$ of 1 life, $1/4$ of 3, $1/8$ of 7, and so on.

For Y, if the bonus coin comes up *tails* you just get one happy life, and if it comes up *heads*, then you get twice the outcome of W minus one happy life. Again, this agrees with the probabilities in table 2.

Table 3: How the outcomes are arranged across different states. H, n is the event where the bonus coin comes up heads, and the first heads in the St. Petersburg sequence is on the nth flip. Likewise T, n means the bonus coin comes up tails.

	$H,1$	$H,2$	$H,3$	$\cdots$	$T,1$	$T,2$	$T,3$	$\cdots$
W	2	4	8	$\cdots$	2	4	8	$\cdots$
X	1	1	1	$\cdots$	3	7	15	$\cdots$
Y	3	7	15	$\cdots$	1	1	1	$\cdots$
$X+Y$	4	8	16	$\cdots$	4	8	16	$\cdots$

Now, when these lotteries are lined up this way, what happens if you get *both* X and Y? It's the same as $W + W$! In every state, $X + Y$ and $W + W$ result in precisely the same number of happy lives. In short, since $W \succ X$ and $W \succ Y$, Separability tells us:

$$W + W \succ X + W \succ X + Y \sim W + W$$

But this cannot be.

This argument relied on some reasoning that I did not make explicit: it initially seemed clear that $W \succ X$ and $W \succ Y$—but why? The key thing is that, for any outcome you might get from X, the lottery W is at *least as likely* to give you an outcome that good or better. W is as likely as X to give you at least one happy life, more likely than X to give you at least two, as likely to give you at least three, and so on. However well X might go, W

is at least as likely to turn out so well, and in some cases even more likely. So W seems clearly better than X. The same goes for Y.

Here is the general principle that underlies this reasoning. In general, for a prospect X and an outcome x, let $P[X \succsim x]$ be shorthand for the probability that X turns out at least as well as x.

Stochastic Dominance. Let X and Y be prospects. If for every outcome x,

$$P[X \succsim x] \leq P[Y \succsim x]$$

then $X \precsim Y$. If, furthermore, for some outcome x,

$$P[X \succsim x] < P[Y \succsim x]$$

then $X \prec Y$.

Stochastic Dominance is a fairly uncontroversial principle of decision theory—even among those who reject other parts of standard expected decision theory (such as Quiggin 1993; Buchak 2013), and even in settings where other parts of standard expectational decision theory give out (see for example Easwaran 2014).[6]

Here is another way of putting the point. Consider another lottery X', which results in $2^n - 1$ happy lives if the first heads in the St. Petersburg sequence of coin flips comes on the nth flip (see table 4). Then X and X' are just rearrangements: they each have exactly the same probability of resulting in any particular outcome. That is to say, X and X' are *stochastically equivalent* prospects. So it seems clear that X and X' are equally good.

But also, W is clearly better than X': for in fact it is *sure* to turn out better than X' no matter what happens. That is, W (strictly) *statewise dominates* X. However the coin flips turn out, X' gives you $2^n - 1$ happy lives, while W gives you one more than that. So W seems clearly better than X'.

[6] For other defenses of Stochastic Dominance, on which I here draw, see Tarsney (2020, 8); Wilkinson (forthcoming, 10); Bader (2018).

Table 4: The lottery W strictly dominates a lottery X' which is stochastically equivalent to X.

	$H,1$	$H,2$	$H,3$	$\cdots$	$T,1$	$T,2$	$T,3$	$\cdots$
W	2	4	8	$\cdots$	2	4	8	$\cdots$
X'	1	3	7	$\cdots$	1	3	7	$\cdots$

More generally, if a prospect Y stochastically dominates a prospect X, then there is some prospect X' which is *stochastically equivalent* to X, while Y *statewise dominates* X'.[7] So Stochastic Dominance follows from two ideas:

Stochastic Equivalence. Stochastically equivalent prospects are equally good.

Statewise Dominance. If Y statewise dominates X, then Y is better than X.

Could either of these ideas go wrong?

What if stochastically equivalent prospects are not equally good? This is the lesson Seidenfeld, Schervish, and Kadane (2009) drew.[8] This would mean that something else must matter for how good prospects are besides the probabilities they assign to each outcome. In fact, I do think we can imagine cases where this is plausible.[9] But in the case at hand, it is hard to think of

[7]This relies on outcomes having certain structure. The structure of the real numbers suffices, and the principle is usually stated in a context where this is assumed. But we will not generally assume this here. In the general case, the principle requires reformulation. Even so, all of the appeals to Stochastic Dominance in this essay *can* be replaced by appeals to Stochastic Equivalence and Statewise Dominance. So we can ignore this detail.

I also note that the standard version of Stochastic Dominance stated here corresponds to *strong* statewise dominance. (Y strongly dominates X iff Y is sure to be at least as good as X, and might be strictly better.) But for the arguments that follow, *strict* statewise dominance suffices—and this distinction may be philosophically important. (Y strictly dominates X iff Y is sure to be strictly better than X.) We can give an alternative formulation of Stochastic Dominance that corresponds to this weaker condition, and which would also suffice for the arguments in this essay.

> If for every outcome x there is some outcome $y \succ x$ such that $P[X \succsim x] \leq P[Y \succsim y]$, then $X \prec Y$.

[8]It is further developed in Lauwers and Vallentyne (2016); Lauwers and Vallentyne (2017). Bales, Cohen, and Handfield (2014) gives a different argument against Stochastic Equivalence, which I discuss in section 3 below.

[9]For example, the probability of an ideally sharp dart hitting a particular point may be

what that *something else* could be. We would need to say that when it comes to gambles that only depend on the outcome of coin flips, it can make a moral difference *which* coin flips they are. The prospect X' results in $2^n - 1$ happy lives if the first heads in the sequence of St. Petersburg coin flips is on the nth flip. But we can also consider the *extended sequence*, which starts with the bonus coin flip, and is followed by the usual St. Petersburg sequence. The prospect X results in $2^n - 1$ happy lives if the first heads in the *extended sequence* is on the nth flip. So if X and X' are not equally good, then it matters morally which of these two sequences of coin flips we use. This seems untenable. I am open to the idea that some kind of *isomorphism* between prospects which is stronger than stochastic equivalence is required to ensure that prospects are equally good. But my guess is that when we spell this notion out in any plausible way, the prospects X and X' will still count as isomorphic even in the stronger sense.

What if Statewise Dominance fails? In that case, I'm not sure what we're doing when we compare how good prospects are. As many others have emphasized (for example, Schoenfield 2014, 268), what we ultimately care about is how well things turn out; choosing better prospects is supposed to guide us toward achieving better outcomes. In light of this, if dominance reasoning is wrong, then I don't want to be right. If A is sure to turn out better than B, then this tells us precisely the thing that betterness-of-prospects is supposed to be a guide to. A guide that does not lead us to our destination, when we already know exactly how to get there, is not worth following.[10]

We should not utterly foreclose giving up Stochastic Dominance—we are facing paradoxes, so some plausible principles will have to go—but I do not think this is a very promising direction. In what follows, I will take Stochastic Dominance for granted.

Now we can sum up the argument I just outlined.

Theorem 2. *If all that matters is the number of happy lives, Stochastic Dominance and Separability are jointly inconsistent.*

zero—but the prospect of sparing a child from malaria if the dart hits that point (and otherwise nothing) may still be better than the prospect of getting nothing no matter what. But these two prospects are stochastically equivalent. Perhaps what is best depends on what features of its outcomes are *sure*—where in general this can come apart from what is *almost sure*—that is, has probability one.

[10]Note, though, that this is a defense of *strict* dominance, while Stochastic Dominance as stated is tantamount to *strong* dominance. This is one reason why the alternative formulation of stochastic dominance in footnote 7 might be preferred.

This looks like very bad news for Separability.

Where does this leave Beckstead and Thomas's argument for Fanaticism? In theorem 1, I stated a version of their argument which has Separability as a premise. If Separability is false, that version is unsound. But Beckstead and Thomas are more cautious. They write:

> The argument … is closely related to well-known arguments from 'separability' for totalist views in population ethics (see Broome 2004). However, the issue for us is not separability in general—perhaps modest violations of separability would be acceptable—but the particular dramatic violations to which timidity leads. (2020, footnote 15 on p. 17)

I'm not sure what they have in mind when they speak of "modest" versus "dramatic" violations of separability. But it is entirely true that it might turn out that, while Separability has counterexamples, the cases that Beckstead and Thomas's argument relies on are not among them. Still, I think if we conclude that Separability simply *can't* be true in general, we should lose much of our confidence in the particular judgments as well. That would tell us that what is better than what really *does* depend in strange ways on what is going on in distant space and time. Given a choice between the lotteries W and X, it *matters* whether you think there is another St. Petersburg population lottery going on in a distant galaxy. This is bizarre—but Stochastic Dominance tells us that it is true. So our intuitions about separability, while admittedly strong, are not to be trusted.

Is there a more restricted principle than Separability which has better hope of being true, and which still can underwrite Beckstead and Thomas's argument? Here is something to try. A *simple* prospect is one that has only finitely many possible outcomes.

Simple Separability. For any simple prospects X, Y, and A, $X \succ Y$ iff $X + A \succ Y + A$.

This very restricted principle is consistent with Stochastic Dominance. And Simple Separability can do the same work as Separability in Beckstead and Thomas's argument.

But is it true? It might be, but it is hard to be confident of this. What would the motivation be for it that is not also motivation for the unrestricted principle? It can't be simply the idea that if what is going on in distant space and time is the same for both of two options, then it is irrelevant to which is better. That idea supports full-fledged Separability. So is there something special about *simple* prospects that makes their value insensitive to what is going on in distant space and time? I leave this question open.

There is one other possible response: perhaps the idealizing assumption that the number of happy lives is all that matters is not an innocent simplification. Both theorem 1 and theorem 2 rely on the drastic idealization that lives can be freely rearranged between near and far galaxies without affecting anything of value—for instance, there are no morally important relationships between people in the same galaxy. Could this be where we went wrong? This would not save Beckstead and Thomas's argument, but it might allow us to salvage Separability.

In fact, both arguments can be generalized to avoid relying on this drastic idealization. I have talked about adding up *numbers* of happy lives—but all the arguments really require is a much more general sense in which outcomes can be "added up." The key idea is that each finite outcome can be split up into two parts: a *near* part, concerning what is going on around here in the part of the world we might make any difference to, and a *far* part, concerning what is going on in Beckstead and Thomas's "distant galaxy." We can "add up" a *near outcome* x and a *far outcome* y to get a combined outcome which we'll call $x \oplus y$. We can similarly talk about *near prospects* and *far prospects*, which can be "added up" outcome by outcome.

Separability can be restated in these more general terms.[11]

Separability. For any near prospects X and Y, and any far prospect A,

$$X \succ Y \quad \text{iff} \quad X \oplus A \succ Y \oplus A$$

For any far prospects X and Y, and any near prospect A,

$$X \succ Y \quad \text{iff} \quad A \oplus X \succ A \oplus Y$$

The strong idealizing assumption that all that matters is the number of happy lives can then be replaced with much weaker structural assumptions about adding up outcomes. If all that matters is the number of happy lives, then we can freely rearrange happy lives between the near and far parts of the world without losing any value: an outcome with m happy lives nearby and n happy lives far away is just as good as an outcome with *zero*

[11]We have added a second clause because our new notion of "adding up" outcomes or prospects need not be commutative.

happy lives nearby and $m + n$ happy lives far away. In our more general setting, the key principle is that we can always *compensate* somehow for making things worse nearby, by making things sufficiently better far away (and vice versa). We will call this assumption *(Positive) Compensation*; it is stated precisely in appendix A.

We could also restate the other premise of Beckstead and Thomas's argument, More is Better, in these more general terms. But if we help ourselves to the stronger assumption of Stochastic Dominance, we can avoid some complications in the general statement of the theorem and also give a substantially simpler proof.

Theorem 3. *Stochastic Dominance, Simple Separability, and Positive Compensation together imply Fanaticism.*

This generalization of theorem 1 does not provide a new way of defending Beckstead and Thomas's argument for Fanaticism: for we can also generalize theorem 2.

Theorem 4. *Stochastic Dominance, Separability, and Positive Compensation are jointly inconsistent.*

(Proofs are given in appendix A.)

So to argue for Fanaticism on the basis of theorem 3, we would again have to find a way to motivate *Simple* Separability without going all the way to full-fledged Separability.

But the generalization also clarifies how full-fledged Separability might still be true after all: it might be *Compensation* that fails, instead. There is—of necessity—more that matters morally than just the total number of happy lives; lives can not be freely rearranged without any effect on value.

Note that, unlike Stochastic Dominance or Separability, Compensation is not a principle about risk, but purely about what ways for the world to be are best. Giving up Compensation imposes constraints on axiology. Rejecting Compensation will be strange—but theorem 4 ensures that *every* consistent view is strange. One way of developing this idea is to say that eventually the value of a galaxy is "saturated," so no further vast improvements are possible—and in particular, no further improvements would suffice to make up for a large loss of value in another galaxy. Adding more happy lives to a galaxy far away that already contains some huge number of happy lives simply cannot make up for eliminating many happy lives nearby.

Here is a simple model. "Near" value and "far" value are represented by two bounded utility functions. The total utility of an outcome is given by the sum of its near utility and its far utility, and the best prospect is that which maximizes expected utility. In this model, if "far utility" is already close to its upper bound, there will be no way of improving it enough to compensate for a large loss in "near utility." This model satisfies Stochastic Dominance and Separability, but not Compensation. Fanaticism also fails in this model: since total utility is bounded, there are no goods immense enough to warrant extremely long-odds gambles.

It turns out that *every* natural model is going to work out in basically the same way. It is possible to show that Stochastic Dominance and Separability, together with a few other auxiliary assumptions, imply that Fanaticism is *false.* But the precise statement and proof of this result are a bit more complicated than the others in this essay, so I will omit it.

Where does this leave us? (I'll hold Stochastic Dominance and the order axioms fixed.) We find ourselves in an unstable dialectical situation. The basic idea, remember, was that the value of a prospect shouldn't depend in strange ways on distant space and time: if two prospects are exactly alike in terms of the probabilities they assign to distant goings-on, then it seems we should be able to "subtract" the distant part and compare the prospects just based on what they say about the part of the world where the choice between them might conceivably make a difference. What we have found is that a *restricted* version of this idea—*Simple* Separability—together with the "value rearrangement" Compensation principle, implies that Fanaticism is true. But Separability *in general* is *inconsistent* with Compensation, and in fact (with other auxiliary assumptions, I have claimed) it implies that Fanaticism is *false*.

3 The Indology argument

Wilkinson (forthcoming) gives another argument for Fanaticism which is closely related to Beckstead and Thomas's. This argument is the last of three in Wilkinson's rich paper, and the one that Wilkinson reckons the most compelling (p. 6). (My presentation generalize Wilkinson's.)[12]

[12] Wilkinson's presentation presupposes totalism, which allows him to state Background Independence in terms adding the *(cardinal) values of outcomes* together—understood as real numbers. I am generalizing his argument to a more axiologically neutral setting.

This argument also has three steps. Once again, we will consider outcomes that can be split into two parts, which we'll call "near" and "far."

Step 1. Consider a different way or restricting Separability, where we require that the "background" prospect involves no uncertainty.

Background Independence. For any near prospects X and Y and any far outcome a,

$$X \succ Y \text{ iff } X \oplus a \succ Y \oplus a$$

Say a is an outcome where there are a million happy lives, far away and long ago. Then $X \oplus a$ is a prospect that is sure to result in a million *additional* happy lives (far away) besides what results from X (around here). Background Independence is the basis for the classic "Egyptology" objection to the average view in population ethics (McMahan 1981, 115; Parfit 1984, 420). It would be very strange if, when choosing between two policies, our decision might turn on whether an additional million people thrived in ancient Egypt utterly unaffected by our choice.

We might worry about whether this principle is true, for the kind of reasons discussed in the previous section. But at least it does not give rise to the exact same kind of problems as the general Separability principle. Since the background outcome a does not build in any uncertainty of its own, there is no worry about this interacting with X and Y in different ways in different states, which might compensate for Y's bad cases without compensating for X's bad cases.[13]

Step 2. Wilkinson gives a different argument for a kind of separability failure, based on Tarsney (2020). Suppose Fanaticism is false. Then it can be shown that there is a *risky lottery* X and a *safe lottery* Y such that $X \not\succ Y$, but for some *background prospect* B, which is independent of both X and Y, $X \oplus B$ stochastically dominates $Y \oplus B$. Then Stochastic Dominance tells us:

$$X \oplus B \succ Y \oplus B$$

Instead of imagining extra lives in ancient Egypt, Wilkinson imagines the background prospect concerning what went on in the ancient Indus valley, about which we have a great deal of uncertainty.

[13]Furthermore, unlike Separability, Background Independence is consistent with Stochastic Dominance, even if only the numbers matter. Thanks to Zach Goodsell for discussion.

> We happen to know even less about what happened in the ancient Indus Valley than in ancient Egypt—archaeological research and excavations of key sites in India began centuries later than similar work in Britain, Italy, and Egypt. So there is likely plenty left to learn in Indology. (Wilkinson, forthcoming, 26)

Step 3. Now consider how things might go once we *remove* this uncertainty. The effect would be to replace the uncertain prospect B with some particular outcome b. But since $X \not\succ Y$, Background Independence tells us

$$X \oplus b \not\succ Y \oplus b$$

This is strange!

> From Background Independence we know that, whatever you might uncover in your research, you would conclude that the risky lottery is no better than the safe lottery. … So you know what judgement you would make if you simply learned more, no matter what it is you would actually learn. So why do the many years of research?
>
> … Surely we can sidestep those years of research into how B turns out, and make the judgement required by every possible value of b. Surely rationality requires that we do so, rather than require that we do not. But, if we deny Fanaticism, we must accept this inconsistency — an inconsistency which, to me, seems far more absurd than simply accepting Fanaticism and even more absurd than the Egyptology Objection (Wilkinson, forthcoming, 27–28)

I take it that Wilkinson's judgment here is based on the following principle.

Negative Reflection. For prospects X and Y and a question Q, if X is not better than Y conditional on any possible answer to Q, then X is not better than Y unconditionally.[14]

This is a kind of "no regret" principle (compare Arntzenius 2008; see Russell and Isaacs 2020, sec. 2, and references therein). If Negative Reflection is

[14] We model a *question* as a *regular partition*, a set of mutually exclusive and jointly exhaustive events each of which has positive probability. (Following Tarsney, Wilkinson's original way of running the argument used a continuous distribution B, but we can just as well use a discrete distribution for simplicity.)

violated, then you may choose the better prospect, go on to do your research, and then find that your chosen option isn't better after all—no matter what your research turns up. That does seem weird.[15]

As with Beckstead and Thomas's argument, Wilkinson's argument can be summarized by a theorem. Before I state it in general terms, there is one structural point to explain. An important feature of the background prospect B that figures in the Indology argument is that it has "heavy tails," allocating substantial probability to very good outcomes, and also to very *bad* possible outcomes. The results in section 2 only relied on very good outcomes; Positive Compensation ensured that there were sufficiently good outcomes around. For the Indology argument we additionally need a *Negative* Compensation principle, which ensures that are sufficiently *bad* outcomes. It is spelled out in appendix A.[16]

Here is the generalization of Wilkinson's main result.

Theorem 5. *Stochastic Dominance, Negative Reflection, Background Independence, and Positive and Negative Compensation together imply Fanaticism.*

These four principles seem very plausible, and the Indology argument seems very strong. But there are two problems with it—one big, and one smaller.

Here is the big problem:

Theorem 6. *Stochastic Dominance and Negative Reflection together imply that Fanaticism is false.*

This tells us that the premises of the Indology argument—which include Stochastic Dominance and Negative Reflection—are in fact jointly inconsistent, and so the argument cannot be sound.

Here is the basic idea behind theorem 6. (Details are again in appendix A.) First, Fanaticism implies the existence of *generalized St. Petersburg prospects*

[15]Negative Reflection is a variant for "not better" of what Russell and Isaacs (2020) call Countable Independence, and which is also called Dominance* in other places (for example, Lee 2013).

[16]Wilkinson appealed to the stronger principle that composing near and far outcomes just amounts to adding up real numbers—which follows from his totalist assumptions. Tarsney's theorem, which Wilkinson's argument relies on, uses the technical assumption that outcomes have *additive conjoint structure* (see Tarsney 2020, 11; Krantz, Suppes, and Luce 1971, 245–66). Positive and Negative Compensation both follow from one of the four axioms that characterize such structures.

(see Beckstead and Thomas 2020, sec. 4). We can find a sequence of outcomes that get better very fast, and use these to construct a prospect which is *strictly better than any of its outcomes*; Russell and Isaacs (2020) call such prospects *improper*. The existence of improper prospects is a particularly unsettling consequence of Fanaticism.

Second, improper prospects violate Negative Reflection.[17] We can play *two* generalized St. Petersburg games X and Y, independently. We can choose Y so its outcomes are a little better than X's—and so Y stochastically dominates X—but still, none of Y's outcomes are as good as the *prospect* X. Conditional on any way Y could turn out, Y is only as good as one of its mundane outcomes, and so no better than X. But Y is unconditionally better than X, contradicting Negative Reflection.

There is also a second, less serious problem for the Indology argument. Many people hold that some outcomes are *incomparable* to one another: neither is strictly better than the other, but they are still not equally good (for instance, Chang 2002). Is it better to experience a profoundly moving improvised one-person show, or to take a meditative three-day wilderness hike in the Sierra Nevada? Neither seems clearly better than the other. One way of arguing that they are also not equally good is that *sweetening* either option, say by throwing in a free sticker, still does not make one option seem better than the other.

But the combination of Stochastic Dominance and Negative Reflection rules out such cases (see Hare 2010; Schoenfield 2014; Bales, Cohen, and Handfield 2014; Bader 2018). The argument is based on so-called "opaque sweetening" (see table 5). You flip a coin: Heads, you take the hike; Tails, the show. Call this prospect X. Then consider a "sweetened" prospect Y: Heads, you take the hike *and* get a sticker; Tails you do the show. The sweetened option Y dominates X, so Y is better than X. But then you think, what's so special about Heads? Consider the prospect X': Heads, you go to the show; Tails, you take a hike. Then X and X' are stochastically equivalent—so equally good. In short, Y *stochastically dominates* X'—so Y is better than X'. Finally, though, by assumption Y is *not better* than X' given Heads, nor is it better given Tails. So Negative Reflection implies that Y is not better than X'. We have a contradiction.[18]

[17] The analogous point for *Positive* Reflection (below) was one of the lessons of the two-envelope paradox (see Broome 1995; Chalmers 2002).

[18] Note also that this argument does not turn on the infinite partitions that figure in both theorem 5 and theorem 6.

Table 5: Y dominates X, X is stochastically equivalent to X', but by Negative Reflection, Y is not better than X'.

	Heads	Tails
X	Hike	Show
Y	Hike + Sticker	Show
X'	Show	Hike

So the second, smaller problem for the Indology argument is that two of its premises—Stochastic Dominance and Negative Reflection—are inconsistent with incomparability between outcomes.

This consequence is especially pressing in the context of an argument for Fanaticism: for Fanaticism provides special reasons to suspect that some *prospects* are incomparable to one another—even in a context where all possible *outcomes* are totally ordered; but a very similar argument from Negative Reflection rules this out as well.[19] Lauwers (2016) shows that if outcomes are represented by unbounded real-valued utilities, then there is no *constructible* total ordering of prospects that obeys an independence axiom (similar to the Sure Thing Principle, discussed below). We could never hope to write down a decision theory, extending the standard theory of expected value, that told us how to rank every pair of prospects—the existence of such orders in platonic heaven depends on the Axiom of Choice.[20]

Despite this, my preferred response to the "opaque sweetening" argument is to reject incomparability (for defense see Dorr, Nebel, and Zuehl, Manuscript)—but I do not wish to take this controversial view for granted in this essay. Others have wielded this as an argument against Stochastic Equivalence (Schoenfield 2014; Bales, Cohen, and Handfield 2014). A more standard view gives up Negative Reflection (for example, Aumann 1962; see also Bader 2018).

In contrast, consider the analogous principle for *at least as good* rather than *not better*:

[19] The generalized argument requires an additional premise: the Sure Thing Principle (below) suffices.

[20] Another consideration is that Askell (2018) shows that for *infinite* populations, no total ordering of outcomes is compatible with both the principle that what is better for everyone is better overall, and the principle that betterness is invariant under arbitrary permutations of welfare.

Positive Reflection. For prospects X and Y and a question Q, if X is at least as good as Y conditional on any possible answer to Q, then X is at least as good as Y unconditionally.

Unlike Negative Reflection, Positive Reflection is perfectly compatible with incomparability (whether between outcomes or prospects). But Positive Reflection still *does* rule out Fanaticism, given Stochastic Dominance. The argument is essentially the same as for theorem 6.

To sum up, not only do Reflection principles fail to support Fanaticism, but in fact they undermine it.

Still, I agree with Wilkinson that Reflection principles have much to be said in their favor. One argument is based on the *value of information*.[21] Suppose you have two basic options—*Take It* or *Leave It*. Then consider a third option: *Take It If the Taking Is Good.* That is, you can commit to Taking It *only* if, given the additional information of how many people there are in the universe, Taking It turns out to be as good as Leaving It. This third option seems like it should be at least as good as either of the other two. It amounts to doing what is best given *more* information, rather than less—and how could this be bad? Betterness-for-prospects is supposed to be a guide to outcomes that really *are* better, given *all* the information about how things turn out. Betterness-given-more-information should be as good or better a guide than betterness-given-less-information. As Broome (1991, 129) puts it:

> [P]robabilities derived from more information have a higher status than those derived from less. … At the extreme, what would actually happen has the highest status of all. … This at least is true: you ought not to found your judgements on lower-status probabilities when higher-status probabilities are available.

But if Positive Reflection fails, then it can turn out that Taking It If the Taking Is Good is *not as good* as just Leaving It. For if Positive Reflection fails, then it could be that *Take It* is as good as *Leave It* given *any* number of people—and so *Take It If the Taking Is Good* simply amounts to the same thing as *Take It*—and yet *Take It* is *not* as good as *Leave It*. This seems absurd.

[21]This is based on one of several arguments Russell and Isaacs (2020) give in defense of Positive Reflection, generalizing standard arguments for orthodox expectational decision theory in finite cases. Their arguments are about rational preference, but they can be easily adapted to arguments about which prospects are morally best. Note that while this argument is structurally related to standard *dynamic consistency* arguments, it is not an argument about sequential decision-making through time.

Such arguments for Positive Reflection amount to arguments *against* Fanaticism. Indeed, Positive Reflection, together with other standard axioms of decision theory, presses us to a specific kind of anti-fanatical view: expected utility maximization with a bounded utility function (see Hammond 1998; for a generalization see Russell 2020). This kind of theory keeps Stochastic Dominance and both Positive and Negative Reflection. Which of the other two propositions in the inconsistent tetrad we keep—Background Independence or Compensation—depends on how we fill in details. As in section 2, we could let the utility of an outcome $x \oplus y$ be given as the sum of two bounded utility functions $u_1(x) + u_2(y)$—one "near" and one "far." This version satisfies Background Independence (and indeed Separability), but not Compensation. The alternative is to use a utility function that is not additively separable in this way: for example, instead of taking a sum of two bounded functions, we could use a bounded function of the *sum* of the two parts—if we have functions f_1 and f_2 that represent near and far outcomes with numbers, respectively, then the utility would be $u(f_1(x) + f_2(y))$ for some bounded function u. This kind of model keeps Compensation, while giving up Background Independence.

But even though Reflection principles have much in their favor, we should not be too hasty to accept them, along with their anti-Fanatical consequences. For these consequences are also very strange.

As in section 2, we should ask whether we can repair Wilkinson's Indology argument by substituting some more restricted reflection principle that can do the same work. Here is something to try.[22]

The Sure Thing Principle. If E has positive probability and prospects X and Y are equally good conditional on not-E, then X is at least as good as Y conditional on E iff X is at least as good as Y unconditionally.

This principle can be motivated by very similar reflection considerations.[23] Say you are again choosing between *Take It* or *Leave It*. And you are about to learn whether it's *Hot* or *Cold.* Suppose that, given Hot, Take It is at least as good as Leave It; and given Cold, Take It is just as good as Leave It. Then

[22] There are several non-equivalent principles that are called "The Sure Thing Principle"—see Schlee (1997). But I think this is one reasonable candidate for that label. In Savage's framework probabilities are not given; we can replace the condition that E has positive probability with the condition that E is "non-null" in the sense that some prospects that agree on $\neg E$ are not equally good.

[23] Indeed, this is how Savage (1954, sec. 2.7) motivated it.

whatever you learn, you will conclude that Take It is at least as good as Leave It. So it seems you ought to be able to conclude *in advance* that Take It is at least as good as Leave It—this is the same kind of reasoning as in Wilkinson's Indology story. Similar reasoning supports the converse direction of the Sure Thing Principle as well.

The Sure Thing Principle is somewhat controversial (more so than Stochastic Dominance). Theories that allow risk aversion violate it (Quiggin 1993; Buchak 2013); theories that tell us to ignore small-probability outcomes do as well.[24] But, while it is in the same spirit as Positive and Negative Reflection, it is much less demanding. Repeated applications of the Sure Thing Principle can tell us that if one option is better than another conditional on each event in a *finite* partition, then it is better unconditionally. But Positive and Negative Reflection apply even to *infinite* partitions—and on these the Sure Thing Principle is silent. Unlike the infinitary principles, the Sure Thing Principle is consistent with Fanaticism. It is also consistent with incomparable prospects.

Can the Sure Thing Principle, together with Background Independence, serve as the basis for a new "Egyptology" style argument for Fanaticism? Not in the same way as in Wilkinson's Indology argument, no. The background prospect B that figures in Tarsney's construction, which Wilkinson's argument is based on, has infinitely many possible outcomes—recall its "heavy tails." (See appendix A for more explanation.) But we can give a different argument. The Sure Thing Principle restricts the space of possible decision theories quite a bit. Meanwhile, Background Independence restricts the space of possible axiologies quite a bit. Together, these squeeze us toward Fanaticism.

I noted earlier that the Indology argument relies on *bad* outcomes, in a way that the Separability arguments in section 2 did not. We can show that the argument would not work without them, with a model. Suppose that all that matters is the number of happy lives. Then we can construct an expected utility model with a bounded utility function (which takes values between -1 and 0):

$$U(n) = -2^{-n}$$

[24]Suppose E is an event with positive but negligibly small probability. Let X and Y be prospects that always have the same outcome, except in E, in which case Y is sure to turn out strictly better than X. Then theories that neglect negligibly small probabilities will say that X and Y are equally good unconditionally—so $X \succsim Y$. But while X and Y are equally good conditional on not-E, X is not at least as good as Y conditional on E.

This is a theory according to which additional happy people have diminishing marginal utility. Any bounded expected utility model satisfies Stochastic Dominance and both Positive and Negative Reflection (see Hammond 1998). Moreover, *this* utility function has the nice feature that, for a prospect X,

$$EU(X+n) = 2^{-n} \cdot EU(X)$$

This is an order-preserving transformation of $EU(X)$, which guarantees Background Independence. Since the utility function is bounded, this model does not satisfy Fanaticism.

But we cannot simply extend this utility function down to negative numbers representing unhappy lives. Note that $U(n)$ bends downward: for instance,

$$U(1) > \frac{1}{2}U(0) + \frac{1}{2}U(2)$$

This means that in this model, getting the one happy life for sure is better than flipping a coin to decide whether you get zero or two. In other words, the model endorses *risk aversion* with respect to happy lives. But Background Independence tells us that we can "shift" this risk-averse preference down, by adding any negative value to each of the outcomes. So we must also have, for example,

$$U(0) > \frac{1}{2}U(-1) + \frac{1}{2}U(1)$$

(See figure 1.) In fact, the utility function must continue to bend downward no matter how far you go to the left. This means that its extension to negative values must be unbounded below—which results in *Negative* Fanaticism, where an arbitrarily small risk of a *very bad* outcome is worse than a certain outcome which is still quite bad.

This idea generalizes. As it turns out, the general argument needs one more premise. As I noted earlier, we have not generally assumed that prospects are comparable: there may be prospects such that neither is better than the other, and yet they are not equally good. For the present argument we still do not need to assume comparability in general—but we do require some especially simple value comparisons.

Simple Comparability. If x is a good outcome and y is a bad outcome, then a fair lottery with outcomes x or y is good, bad, or neutral.

Now we can state the new "Egyptology" argument for Fanaticism.

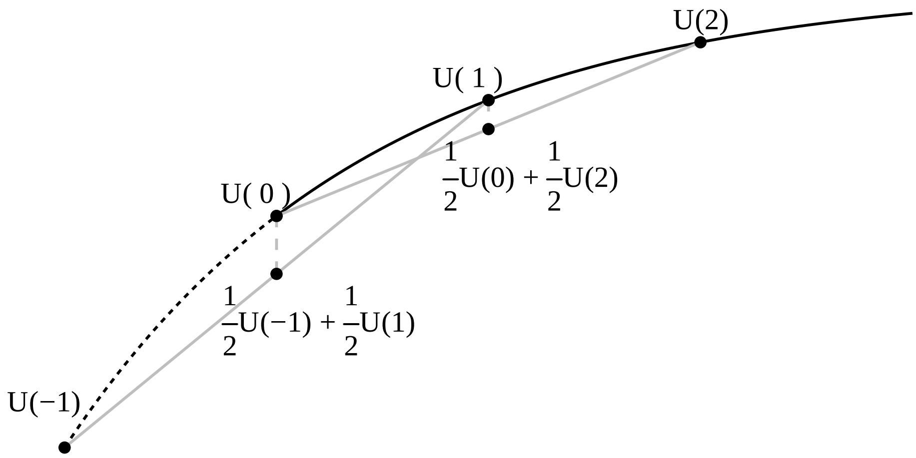

Figure 1: If U is concave for positive values, Background Independence requires that U is also concave for negative values, which leads to Negative Fanaticism.

Theorem 7. *The Sure Thing Principle, Stochastic Dominance, Background Independence, Positive and Negative Compensation, and Simple Comparability together imply that at least one of Positive Fanaticism or Negative Fanaticism is true.*

A proof is given in appendix A.

What are we to make of this? Once again, the dialectical situation is unstable. A strong reflection principle implies that Fanaticism is false. But a weak reflection principle—the Sure Thing Principle—together with Background Independence (and some auxiliary principles) implies that Fanaticism is true. Facing both opposing arguments, I do not think we should place our trust in either yet. Whatever the truth, it is very strange, and we are still far from understanding it.

What are our options? (In order to keep this list under control, I will hold Stochastic Dominance and the order axioms fixed.)

1. Accept at least one of Positive or Negative Reflection, and thereby reject Fanaticism. Then we must also give up one of Background Independence or Compensation; things are strange either way. If Background Independence fails, then we have even *stranger* dependence on distant space and time than before—since now even if we *know*

precisely what is going on out there, it *still* can make a difference to what prospects for nearby matters are best. If Compensation fails, then the universe can mysteriously run out of room for more value (or disvalue), and rearranging populations between near and far parts of the world can make a moral difference.

2. Reject Positive and Negative Reflection, but keep the Sure Thing Principle. This is strange, first because it seems unprincipled, when the arguments in support of each are so similar—including arguments based on regret or the value of information. Second, because again we must either reject Background Independence or Compensation, or else accept Fanaticism—each of which is strange.

3. Reject Reflection *and* the Sure Thing Principle, perhaps going for some less standard decision theory (such as risk-weighting or discounting small probabilities). This allows for regret and negative value of information even in simple cases.[25] Fanaticism remains an open question.

4. Give up Simple Comparability. This might sound desperate—but Tarsney (2020) defends the austere view that Stochastic Dominance is the *only* normative principle of decision theory. This theory allows rampant incomparability between prospects. Combined with a suitable axiology, it satisfies all of the premises of theorem 7 *except* Simple Comparability, while upholding neither Positive nor Negative Fanaticism. (Note, incidentally, that it also does not satisfy Separability.) However, it must be noted that Tarsney's rejection of Fanaticism is half-hearted. While extremely risky gambles are not deemed *better* than safe options, neither are they deemed *worse*: accepting a fanatical wager and rejecting it are also incomparable prospects, in this theory.

4 Taking stock

One of the main routes to Fanaticism is via expectational total utilitarianism (what Bostrom called "standard utilitarianism"). One of the attractions of that package is that it seems to make for a very elegant moral universe. You can have clean principles like Separability and the Sure Thing Principle, and there are powerful theorems about how such principles constrain mathematical representations of betterness (see Broome 2004). The theory does make some counterintuitive predictions, but these may be worth tak-

[25] See also Briggs (2015)

ing in stride, since the theory is so tidy and principled … in simple cases where there are only finitely many different states of nature, and there are at most finitely many people.

But as soon as you consider prospects with infinitely many states and unbounded populations, everything comes apart. In this general setting we find that expectational totalism is a radically incomplete theory (see Hájek and Nover 2008; Lauwers 2017). There is no way of extending expectational totalism to infinite cases with the same elegance that we are used to from finite cases. Elegant principles must be given up, including Separability and Reflection.

Expectational totalism does still seem to be a consistent live option, and with it Fanaticism. It satisfies modified and truncated versions of the original elegant principles. But the modifications and truncations seem like they might be telling us that there is something wrong about the underlying ideas. Arguments for Fanaticism based on these modified and truncated principles don't have nearly the same immediate grip as the ones involving the sweeping clean versions. So even though the Fanatical position is still *there*, it no longer stands out as the bold, austere, and systematic ethical framework that it once seemed.

Furthermore, what Fanaticism has lost in elegance, anti-Fanatical theories have gained. For in fact, we *can* have sweeping, clean principles like Separability or Reflection, as long as we *give up* Fanaticism, along with some of the premises that got us there.

Here are three broad versions of this. One approach keeps both Reflection and Separability, while cutting our axiology down to size by giving up Compensation. A second keeps Reflection and Compensation, and gives up both Separability and Background Independence. Both of these two approaches can be represented using standard expected utility theory with bounded utilities; they only differ in the structure of their utility functions. (In the first case only, the utility function is additively separable.) A third elegant vision is Tarsney's theory, which keeps most of these principles (though not Separability) by giving up many comparisons of value.

All of these options are strange. I don't know which is true, and I think it is premature to be confident in any of them.

The paradoxes of large values and small probabilities are deeply weird. But they aren't outlandish. In our *actual situation*, I take it that there are infinitely many live possibilities for what our universe is like, and we should assign significant probability to very good and very bad outcomes. Our own species might thrive for a very long time, and for all we know there may be many others in the universe. These paradoxes are not just brain-teasers that can be ignored when we are doing serious practical ethics—they raise difficult questions that we must answer, if we are to do as much good as we can.

A Theorems and Proofs

First, the background technical framework. There is a set of *states* equipped with a σ-algebra of *events*, and and a set of *(finite) outcomes*. A *prospect* is a measurable function from states to outcomes. (We will focus on discrete prospects.) We fix in the background some probability measure P on states; we assume this is suitably rich (e.g., non-atomic). There is a relation $\succsim$ (*at least as good*) that holds between prospects; this is assumed to be transitive and reflexive. Strict betterness $\succ$ and indifference $\sim$ are defined in terms of $\succsim$ in the usual way. We will generally treat outcomes interchangeably with their corresponding constant prospects.

For convenience of exposition, we take as fixed some baseline "zero" outcome 0. Outcomes better than this are *good* or *gains*, and outcomes worse than this are *bad* or *losses*; outcomes exactly as good as the baseline are *neutral*. We take for granted that there is at least one good outcome. For an outcome x, we let $p * x$ stand for an arbitrary prospect with probability p of outcome x and probability $1 - p$ of the outcome 0.

(Positive) Fanaticism. For any probability $p > 0$ and any finite outcome $x \succ 0$, there is a finite outcome y such that $p * y \succ x$.

That is to say, any prospect which has probability p of outcome y and probability $1 - p$ of 0 is strictly better than the constant prospect which has outcome x in every state.

Negative Fanaticism. For any probability $p > 0$ and any finite outcome $x \prec 0$, there is a finite outcome y such that $p * y \prec x$.

There is also a set of *near outcomes* and a set of *far outcomes*. Each finite outcome has a *near* component and a *far* component. It will simplify our reasoning if we suppose that near and far outcomes are freely recombinable: for each near outcome x and far outcome y, there is a combined outcome $x \oplus y$ with those components. (This assumption could be weakened.) Without too much risk of confusion, we also use the notation 0 for the near and far components of the baseline outcome, so $0 = 0 \oplus 0$.

Near/far prospects are measurable functions from states to near/far outcomes. For a near prospect X and a far prospect A, we define the combined prospect statewise:

$$(X \oplus A)(s) = X(s) \oplus A(s)$$

If X and Y are near prospects, let $X \succ Y$ abbreviate $X \oplus 0 \succ Y \oplus 0$, and analogously for far prospects.

We will restate the principles that figure in the following theorems for easy reference.

Stochastic Dominance. Let X and Y be prospects. If for every outcome x,

$$P[X \succsim x] \leq P[Y \succsim x]$$

then $X \precsim Y$. If, furthermore, for some outcome x,

$$P[X \succsim x] < P[Y \succsim x]$$

then $X \prec Y$.

Separability. For any near prospects X and Y, and any far prospect A,

$$X \succ Y \quad \text{iff} \quad X \oplus A \succ Y \oplus A$$

For any far prospects X and Y, and any near prospect A,

$$X \succ Y \quad \text{iff} \quad A \oplus X \succ A \oplus Y$$

Simple Separability. For any simple prospects X, Y, and A, $X \succ Y$ iff $X + A \succ Y + A$.

It is also useful to distinguish this special case:

Outcome Separability. For any near outcomes x and y and any far outcome a,

$$x \prec y \quad \text{iff} \quad x \oplus a \prec y \oplus a$$

For any far outcomes x and y and any near outcome a,

$$x \prec y \quad \text{iff} \quad a \oplus x \prec a \oplus y$$

(Positive) Compensation. For any near good x and far good y, there is a far good z such that $x \oplus y \sim 0 \oplus z$, and there is a near good w such that $x \oplus y \sim w \oplus 0$.

Even if we zero out all of the near value, we can offset this by improving things enough in a distant galaxy.

Say that a sequence of near outcomes $x_1, x_2, \ldots$ form an *arithmetic progression*, with difference z (which is a far good), iff

$$x_n \oplus z \sim x_{n+1} \oplus 0 \quad \text{for each } n$$

Positive Compensation implies that for any near good x_1 and any far good z, x_1 is the start of an infinite arithmetic progression with difference z. Furthermore, Outcome Separability implies that $x_1 \prec x_2 \prec \cdots$. We can say the same things about arithmetic progressions of far goods, where the difference is a near good.

Theorem 3. *Stochastic Dominance, Simple Separability, and Positive Compensation together imply Fanaticism.*

We first prove a lemma.

Lemma 1. *Positive Compensation and Outcome Separability imply that, for any near good x and any probability $p > 0$, there is a near good y, a prospect $Y = p * y$, and a simple prospect A such that $x \oplus A$ is stochastically dominated by $Y \oplus A$.*

This lemma suffices for theorem 3: Stochastic Dominance ensures that $x \oplus A \prec Y \oplus A$, and then by Simple Separability $x \prec Y$, QED.

Proof of lemma 1. Assume $p < 1$. Consider a partition of $n + 1$ events: the first has probability p, and the rest each have probability $q = (1 - p)/n$, with n chosen to be large enough so that $q < p$.

By Positive Compensation, we can let $0, z_1, z_2, \ldots, z_n$ be an arithmetic progression of far goods with difference x: that is, for each k,

$$x \oplus z_k \sim 0 \oplus z_{k+1}$$

We can also choose a near good y such that

$$x \oplus z_n \sim y \oplus 0$$

By Outcome Separability,

$$x \oplus 0 \prec x \oplus z_1 \prec x \oplus z_2 \prec \cdots \prec x \oplus z_n \sim y \oplus 0$$

Then we construct prospects X, Y, and A as in table 6: X is sure to have outcome x, Y is a gamble with chance p of y, and A yields nothing if y pays off, and otherwise yields the result of a fair lottery between each of the outcomes $z_1, z_2, \ldots, z_n$.

Table 6: A simpler Separability argument, using Stochastic Dominance.

	p	q	q	$\cdots$	q
x	x	x	x	$\cdots$	x
Y	y	0	0	$\cdots$	0
A	0	z_1	z_2	$\cdots$	z_n

It can be checked that $Y \oplus A$ stochastically dominates $x \oplus A$. □

Theorem 4. *Stochastic Dominance, Separability, and Positive Compensation are jointly inconsistent.*

Proof. First a basic fact: for each outcome $x \oplus 0$, there is some strictly better outcome $y \oplus 0$ (and similarly for far outcomes). We have assumed that there is some good outcome; by Positive Compensation some outcome $0 \oplus z$ is good. Then Positive Compensation and Outcome Separability tell us that there is some $y \oplus 0 \sim x \oplus z \succ x \oplus 0$.

Now we recursively construct two sequences of finite goods as follows. For the base case, let $x_0 \oplus y_0 = 0 \oplus 0$. For the recursive step, for each n, we can find w_n and z_n such that

$$x_{n-1} \oplus 0 \prec z_n \oplus 0$$
$$0 \oplus y_{n-1} \prec 0 \oplus w_n$$

By Compensation we can then find x_n and y_n such that

$$w_n \oplus z_n \sim 0 \oplus y_n$$
$$w_n \oplus z_n \sim x_n \oplus 0$$

Next, we will use these outcomes to construct four prospects X, Y, W, Z as in table 7. Choose events $E_1, E_2, \ldots$ and $F_1, F_2, \ldots$, where $P(E_n) = P(F_n) = 2^{-(n+1)}$ (just as in section 2).

Table 7: The generalized counterexample to Separability.

	E_1	E_2	E_3	$\cdots$	F_1	F_2	F_3	$\cdots$
X	0	0	0	$\cdots$	x_1	x_2	x_3	$\cdots$
Y	y_1	y_2	y_3	$\cdots$	0	0	0	$\cdots$
W	w_1	w_2	w_3	$\cdots$	w_1	w_2	w_3	$\cdots$
Z	z_1	z_2	z_3	$\cdots$	z_1	z_2	z_3	$\cdots$

By construction, $X \oplus Y$ is equally good as $W \oplus Z$ in every state. But also, $W \oplus 0$ stochastically dominates $X \oplus 0$, and $0 \oplus Z$ stochastically dominates $0 \oplus Y$. So $X \oplus 0 \prec W \oplus 0$ and $0 \oplus Y \prec 0 \oplus Z$, and thus Separability tells us:

$$X \oplus Y \prec W \oplus Y \prec W \oplus Z \sim X \oplus Y$$

This cannot be. □

Background Independence. For any near prospects X and Y and any far outcome a,

$$X \succ Y \text{ iff } X \oplus a \succ Y \oplus a$$

We will also consider *conditional* prospects: if X is a prospect and E is an event with positive probability, then $X|E$ is the restriction of X to E. We take the betterness relation to apply to conditional prospects as well. Stochastic Dominance (and Stochastic Equivalence) will be understood to also apply to conditional prospects, where $P[(X|E) \succsim x]$ is understood as the conditional probability $P[X \succsim x \mid E]$. Let $X \succsim_E Y$ mean that $X|E \succsim Y|E$.

Negative Reflection. For any prospects X and Y and any regular partition $\mathcal{E}$, if $X \not\succ_E Y$ for each $E \in \mathcal{E}$, then $X \not\succ Y$.

Negative Compensation. For any near good x and any far outcome y, there is a far outcome z such that $x \oplus z \sim 0 \oplus y$. For any near outcome x and any far good y, there is a near outcome w such that $w \oplus y \sim x \oplus 0$.

This lets us run compensation "downward": not only can we offset making nearby things worse by making far away things sufficiently better, but also we can offset making nearby things *better* by making far away things sufficiently *worse*.

Negative Compensation implies that we can also extend arithmetic progressions of outcomes *downward*: so Positive and Negative Compensation together tell us that for any pair of near (far) outcomes x_0, x_1 can be extended to an arithmetic progression ..., $x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots$. Outcome Separability implies that if $x_0 \prec x_1$ then each outcome in the sequence is better than those before it.

Theorem 5. *Stochastic Dominance, Negative Reflection, Background Independence, and Positive and Negative Compensation together imply Fanaticism.*

The theorem follows from the following lemma.

Lemma 2. *Positive and Negative Compensation and Outcome Separability together imply that for any good x and any probability $p > 0$, there is a prospect $Y = p * y$ and an independent prospect B such that $x \oplus B$ is stochastically dominated by $Y \oplus B$.*

(The key difference from lemma 1 is that the background prospect B is now required to be *independent* of Y; but it can no longer be guaranteed to be a *simple* prospect.)

From lemma 2, we can apply the same reasoning as in section 3: by Stochastic Dominance, $x \oplus B \prec Y \oplus B$; by Negative Reflection, B must have some possible outcome b such that $x \oplus b \prec Y \oplus b$; so by Background Independence $x \prec Y$.

Proof sketch for lemma 2. This is essentially the same as Tarsney's "Sufficiency Theorem" (2020), adapted to our more general setting, so we will not go into details. Here is the main idea.

Let x be a near good. Our Compensation principles tell us that there is an infinite arithmetic progression of far goods ... , $z_{-2}, z_{-1}, z_0, z_1, z_2 \ldots$, with difference x, where $z_0 = 0$. We let these be the possible outcomes of the background prospect B. We assign these outcomes probabilities that are sufficiently spread out. A *discrete Laplace distribution* will do:

$$\beta(n) = \alpha\, 2^{-|qn|}$$

where α is a normalization constant and $q < p/2$. The smaller the probability p is, the wider this distribution will be. Then we let y be a near good which is as good as some good enough far outcome z_N ($N > 1/q$). It can be shown that if $Y = p * y$, and Y and B are independent, then $Y \oplus B$ stochastically dominates $x \oplus B$. □

Theorem 6. *Stochastic Dominance and Negative Reflection together imply that Fanaticism is false.*

Proof. First, by Fanaticism we can construct a fast-growing sequence of outcomes: $x_1 = 0$, and for each $n > 1$, x_n is an outcome such that $2^{-n} * x_n \succ x_{n-1}$.

We flip two coins—Coin A and Coin B—until each of them has landed heads. Let A_n be the event where Coin A first comes up heads on the nth flip, and similarly for B_n. So

$$P(A_m B_n) = 2^{-(m+n)} \quad \text{for each } m, n \geq 1$$

Let X be a prospect that has outcome x_m in each event A_m. Then (by Stochastic Dominance) for each $n > 1$,

$$X \succ 2^{-n} * x_n \succ x_{n-1}$$

This also holds conditional on each event B_n (since X is independent of these events).

Let Y be a prospect that has outcome x_{n+1} in each event B_n. Then Y stochastically dominates X; so $Y \succ X$. But the events B_n form a regular partition, and conditional on each event B_n, Y is just as good as x_{n+1}, while X is strictly better than x_{n+1}. So Y is not better than X conditional on any B_n, which contradicts Negative Reflection. □

The Sure Thing Principle. For any event E such that $P(E) > 0$, for any prospects X and Y such that $X \sim_{\neg E} Y$,

$$X \precsim Y \quad \text{iff} \quad X \precsim_E Y$$

We'll use the notation $(p * X,\ (1-p) * Y)$ for an arbitrary prospect such that the chance of any outcome x is

$$p \cdot P[X = x] + q \cdot P[Y = x]$$

We simplify $(p * X,\ (1-p) * 0)$ to $p * X$.

The Sure Thing Principle and Stochastic Equivalence together imply:

Independence. For any prospects X, Y, Z, and any probability $p > 0$,

$$X \precsim Y \quad \text{iff} \quad (p * X,\ (1-p) * Z) \precsim (p * Y,\ (1-p) * Z)$$

We also suppose:

Simple Comparability. For any good outcome x and bad outcome y,

$$\left(\frac{1}{2} * x,\ \frac{1}{2} * y\right) \succsim 0 \quad \text{or} \quad \left(\frac{1}{2} * x,\ \frac{1}{2} * y\right) \precsim 0$$

Theorem 7. *The Sure Thing Principle, Stochastic Dominance, Background Independence, Positive and Negative Compensation, and Simple Comparability together imply that at least one of Positive Fanaticism or Negative Fanaticism is true.*

Proof. The proof has three steps.

Step 1. If Negative Fanaticism is false, there exists a bad outcome x such that for all y, $1/2 * y \not\prec x$.

If Negative Fanaticism is false, this means that there is a bad outcome x such that for some probability $p > 0$, for all bad outcomes y, $p * y \not\prec x$. Call a probability p *nice* iff it has the property that $p * y \not\prec x$ for all bad outcomes y. There are two cases to consider.

1. Some $p \geq 1/2$ is nice. Suppose y is a bad outcome such that $\frac{1}{2} * y \prec x$. Then by Stochastic Dominance (or Independence),

$$p * y \precsim \frac{1}{2} * y \prec x$$

   This contradicts the assumption that p is nice. So there is no such y.

2. No $p \geq 1/2$ is nice. In that case there is some nice $p < 1/2$ such that $2p$ is not nice: that is, there is some bad outcome x' such that $2p * x' \prec x$. Now suppose that y is a bad outcome such that $\frac{1}{2} * y \prec x'$. By Independence,

$$p * y \precsim 2p * x' \prec x$$

   This contradicts the assumption that p is nice.

Step 2. For any good outcome x, there is an arithmetic progression $x^-, 0, x^+$ such that $x \precsim x^+$, and

$$\left(\frac{1}{2} * x^-,\ \frac{1}{2} * x^+\right) \succsim 0$$

Let y^- be a bad outcome as given by Step 1; without loss of generality we can suppose y^- is a *far* outcome (by Exact Compensation). Exact Compensation ensures that there is an arithmetic progression $y^-, 0, y^+$ with difference d. For any outcome z, we can choose a far outcome z' such that $z \sim d \oplus z'$. The property from Step 1 tells us:

$$\left(\frac{1}{2} * z',\ \frac{1}{2} * 0\right) \not\precsim y^-$$

By Background Independence, we can add d to each outcome, which yields:

$$\left(\frac{1}{2} * z,\ \frac{1}{2} * y^+\right) \not\precsim 0$$

By Simple Comparability, then, for any bad outcome z,

$$\left(\frac{1}{2} * z,\ \frac{1}{2} * y^+\right) \succsim 0$$

Now let $x^+ = x \oplus y^-$. This is better than either x or y^+, since they are each good (using Outcome Separability). And there is an arithmetic progression $x^-, 0, x^+$, such that (by Independence),

$$\left(\frac{1}{2} * x^-,\ \frac{1}{2} * x^+\right) \succsim \left(\frac{1}{2} * x^-,\ \frac{1}{2} * y^+\right) \succsim 0$$

Step 3. Deduce Positive Fanaticism.

We can show by induction that for each $n > 0$, there is some (near) good outcome y such that

$$\frac{1}{n} * y \succsim x^+$$

The base case is clear. For the inductive step, suppose this holds for n. By Independence,

$$\left(\frac{1}{2} * \frac{1}{n} * y,\ \frac{1}{2} * x^-\right) \succsim \left(\frac{1}{2} * x^+,\ \frac{1}{2} * x^-\right) \succsim 0$$

The left hand side can be rewritten:

$$\frac{n+1}{2n} * \left(\frac{1}{n+1} * y,\ \frac{n}{n+1} * x^-\right)$$

(and 0 can be rewritten as $\frac{n+1}{2n} * 0$), so by Independence,

$$\left(\frac{1}{n+1} * y,\ \frac{n}{n+1} * x^-\right) \succsim 0$$

Background Independence lets us add z to each outcome.

$$\left(\frac{1}{n+1} * (y \oplus z),\ \frac{n}{n+1} * 0\right) \succsim x^+$$

Finally, by Exact Compensation, there is a near outcome y' such that $y' \oplus 0 \sim y \oplus z$, and we have

$$\frac{1}{n+1} * y' \succsim x^+$$

completing the induction.

Finally, for any probability $p > 0$, we can choose n such that $1/n < p$. For some good y,

$$p * y \succ \frac{1}{n} * y \succsim x^+ \succsim x \qquad \square$$

References

Arntzenius, Frank. 2008. "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68 (2): 277–97. https://doi.org/10.1007/s10670-007-9084-8.

Askell, Amanda. 2018. "Pareto Principles in Infinite Ethics." New York University. https://askell.io/publication/pareto-principles-in-infinite-ethics.

Aumann, Robert J. 1962. "Utility Theory Without the Completeness Axiom." *Econometrica: Journal of the Econometric Society*, 445–62.

Bader, Ralf M. 2018. "Stochastic Dominance and Opaque Sweetening." *Australasian Journal of Philosophy* 96 (3): 498–507. https://doi.org/10.1080/00048402.2017.1362566.

Bales, Adam, Daniel Cohen, and Toby Handfield. 2014. "Decision Theory for Agents with Incomplete Preferences." *Australasian Journal of Philosophy* 92 (3): 453–70. https://doi.org/10.1080/00048402.2013.843576.

Beckstead, Nicholas, and Teruji Thomas. 2020. "A Paradox for Tiny Probabilities and Enormous Values." *Global Priorities Institute Working Papers Series*, no. 4.

Bostrom, Nick. 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308–14.

———. 2011. "Infinite Ethics." *Analysis and Metaphysics* 10: 9–59.

Briggs, R. A. 2015. "Costs of Abandoning the Sure-Thing Principle." *Canadian Journal of Philosophy* 45 (5): 827–40.

Broome, John. 1991. *Weighing Goods: Equality, Uncertainty and Time*. Wiley-Blackwell.

———. 1995. "The Two-Envelope Paradox." *Analysis* 55 (1): 6–11.

———. 2004. *Weighing Lives*. Oxford University Press.

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

Chalmers, David J. 2002. "The St. Petersburg Two-Envelope Paradox." *Analysis* 62 (2): 155–57.

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112 (4): 659–88. https://doi.org/10.1086/339673.

Dorr, Cian, Jacob M Nebel, and Jake Zuehl. Manuscript. "The Case for Comparability."

Easwaran, Kenny. 2014. "Decision Theory Without Representation Theorems." *Philosophers' Imprint* 14.

Goodsell, Zach. Forthcoming. "A St. Petersburg Paradox for Welfare Aggregation." *Analysis*.

Greaves, Hilary, and William MacAskill. 2021. "The Case for Strong Longtermism." *Global Priorities Institute Working Papers Series*, no. 5.

Hájek, Alan, and Harris Nover. 2008. "Complex Expectations." *Mind* 117 (467): 643–64. https://doi.org/10.1093/mind/fzn086.

Hammond, Peter J. 1998. "Objective Expected Utility: A Consequentialist Perspective." In *Handbook of Utility Theory*, edited by Peter J. Hammond, Salvador Barberà, and Christian Seidl. Vol. I. Kluwer Academic.

Hare, Caspar. 2010. "Take the Sugar." *Analysis* 70 (2): 237–47. https://doi.org/10.1093/analys/anp174.

Hutchinson, Michelle. 2021. "Why I Find Longtermism Hard, and What Keeps Me Motivated." 80,000 Hours. February 21, 2021. https://80000hours.org/2021/02/why-i-find-longtermism-hard/.

Krantz, David H, Patrick Suppes, and R Duncan Luce. 1971. *Foundations of Measurement, Volume 1: Additive and Polynomial Representations*. Academic Press.

Lauwers, Luc. 2016. "Why Decision Theory Remains Constructively Incomplete." *Mind* 125 (500): 1033–43. https://doi.org/10.1093/mind/fzv174.

———. 2017. "Infinite Lotteries, Large and Small Sets." *Synthese* 194 (6): 2203–9. https://doi.org/10.1007/s11229-016-1052-x.

Lauwers, Luc, and Peter Vallentyne. 2016. "Decision Theory Without Finite Standard Expected Value." *Economics and Philosophy* 32 (3): 383–407.

https://doi.org/10.1017/S0266267115000334.

———. 2017. "A Tree Can Make a Difference." *Journal of Philosophy* 114 (1): 33–42. https://doi.org/10.5840/jphil201711412.

Lee, Chunghyoung. 2013. "The Two-Envelope Paradox: Asymmetrical Cases." *Mind* 122 (485): 1–26. https://doi.org/10.1093/mind/fzt023.

McMahan, Jefferson. 1981. "Problems of Population Theory." *Ethics* 92 (1): 96–96. https://doi.org/10.1086/292301.

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

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. First edition. New York: Hachette Books.

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

Quiggin, John. 1993. *Generalized Expected Utility Theory: The Rank-Dependent Model*. Springer Netherlands. https://doi.org/10.1007/978-94-011-2182-8.

Russell, Jeffrey Sanford. 2020. "Non-Archimedean Preferences over Countable Lotteries." *Journal of Mathematical Economics* 88 (May 2020): 180–86.

Russell, Jeffrey Sanford, and Yoaav Isaacs. 2020. "Infinite Prospects." *Philosophy and Phenomenological Research*.

Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.

Schlee, Edward E. 1997. "The Sure Thing Principle and the Value of Information." *Theory and Decision* 42 (1): 21–36.

Schoenfield, Miriam. 2014. "Decision Making in the Face of Parity." *Philosophical Perspectives* 28 (1): 263–77. https://doi.org/10.1111/phpe.12044.

Seidenfeld, Teddy, Mark J Schervish, and Joseph B Kadane. 2009. "Preference for Equivalent Random Variables: A Price for Unbounded Utilities." *Journal of Mathematical Economics* 45 (5-6): 329–40.

Smith, Nicholas J. J. 2014. "Is Evaluative Compositionality a Requirement of Rationality?" *Mind* 123 (490): 457–502. https://doi.org/10.1093/mind/fzu072.

Tarsney, Christian. 2020. "Exceeding Expectations: Stochastic Dominance as a General Decision Theory." *Global Priorities Institute Working Papers Series*, no. 3.

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

Wilkinson, Hayden. Forthcoming. "In Defense of Fanaticism." *Ethics*.