In defence of fanaticism

Hayden Wilkinson

Global Priorities Institute | September 2020
GPI Working Paper No. 4-2020
In defence of fanaticism

Hayden Wilkinson*

August, 2020
Comments welcome: hayden.wilkinson@anu.edu.au

Abstract: Consider a decision between: 1) a certainty of a moderately good outcome, such as one additional life saved; 2) a lottery which probably gives a worse outcome, but has a tiny probability of a far better outcome (perhaps trillions of blissful lives created). Which is morally better? Expected value theory (with a plausible axiology) judges (2) as better, no matter how tiny its probability of success. But this seems fanatic. So we may be tempted to abandon expected value theory.

But not so fast - denying all such verdicts brings serious problems. For one, we must reject either: that moral betterness is transitive; or even a weak tradeoffs principle. For two, we must accept that judgements are either: ultra-sensitive to small probability differences; or inconsistent over structurally-identical pairs of lotteries. And, for three, you must sometimes accept judgements which you know you would reject if you learned more. Better to accept fanaticism than these implications.

*School of Philosophy, Australian National University
1 Introduction

Suppose you face the following moral decision.

Dyson’s Wager

You have $2,000 to use for charitable purposes. You can donate it to either of two charities.

The first charity distributes bednets in low-income countries in which malaria is endemic.\(^1\) With an additional $2,000 in their budget, they would prevent one additional death from malaria in the coming year. You are certain of this.

The second organisation does speculative research into how to do computations using ‘positronium’ - a form of matter which will be ubiquitous in the far future of our universe. If our universe has the right structure (which it probably does not), then in the distant future we may be able to use positronium to instantiate all of the operations of human minds living blissful lives, and thereby allow morally valuable life to survive indefinitely.\(^2\) From your perspective as a good epistemic agent, there is some tiny, non-zero probability that, with (and only with) your donation, this research would discover a method for stable positronium computation and would be

---

\(^1\)I have in mind the Against Malaria Foundation. As of 2019, the charity evaluator GiveWell estimated that the Against Malaria Foundation prevents the death of an additional child under the age of 5 for, on average, every US$3,710 donated (GiveWell 2020). Including other health benefits, a total benefit roughly equivalent to that is produced for, on average, every US$1,690 donated. Of course, in reality, a donor can never be certain that their donation will result in an additional life saved. This assumption of certainty is for the sake of simplicity.

\(^2\)Dyson (1981) was the first to suggest positronium as a medium for computation and information storage. This follows Dyson (1979), wherein it is argued that an infinite duration of computation could be performed with finite energy if the computation hibernates intermittently, and if the universe has a particular structure. Tipler (1986) suggests an alternative method which may work if the universe has a different structure. Sandberg (n.d.) argues that both Dyson and Tipler’s proposals are unlikely to work, as our universe appears to match neither structure. Nonetheless, it is still epistemically possible that the universe has the right structure for Dyson’s proposal. And possibility is sufficient for my purposes.

\(^3\)Would such artificially-instantiated lives hold the same moral value as lives led by flesh-and-blood humans? I assume that they would, if properly implemented. See Chalmers (2010) for arguments supporting this view. And note that, for the purposes of the example, all that’s really needed is that it is epistemically possible that the lives of such simulations hold similar moral value.
used to bring infinitely many (or just arbitrarily many) blissful lives into existence.  

What ought you do, morally speaking? Which is the better option: saving a life with certainty, or pursuing a tiny probability of bringing about arbitrarily many future lives?

A common view in normative decision theory and the ethics of risk - *expected value theory* - says that it’s better to donate to the speculative research. Why? Each option has some probability of bringing about each of several outcomes, and each of those outcomes has some value, specified by our moral theory. Expected value theory says that the best option is whichever one has the greatest probability-weighted sum of value - the greatest *expected value*. Here, the option with the greatest expected value is donating to the speculative research (at least on certain theories of value - more on those in a moment). So, plausibly, that’s what you should do.

This verdict is counterintuitive to many. All the more counterintuitive is that it’s still better to donate to speculative research *no matter how low* the probability is (short of being 0). For instance, the odds of your donation actually making the research succeed could be 1 in $10^{100}$. ($10^{100}$ is greater than the number of atoms in the observable universe). The chance that the research yields nothing at all would be 99.99... percent, with another 96 nines after that. And yet you ought to take the bet, despite it being *almost guaranteed* that it will actually turn out worse than the alternative; despite the fact that you will almost certainly have let a person die for no actual benefit? Surely not, says my own intuition. On top of that, suppose that $2,000 spent on preventing malaria would save more than one life. Suppose it would save a billion lives.

---

4I have deliberately chosen a case involving many separate lives rather than a single person’s life containing enormous value. Why? You might think that one individual’s life can contribute only some bounded amount of value to the value of the world as a whole - you might prefer for 100 people to each obtain some finite value than for one person to obtain infinite value. But whether this verdict is correct is orthogonal to the issue at hand, so I’ll focus on large amounts of value spread over many people.

5Note that expected *value* is distinct from the frequently-used notion of expected *utility*, and expected value theory distinct from expected *utility* theory (which says much the same thing, but with expected utility replacing expected value). Expected value is the expectation of value itself, as given by our theory of value (see Section 3). Meanwhile, expected *utility* can be given by *any* increasing function of value - perhaps a concave function, such that additional value contributes less and less additional utility. The utility of an outcome may even be bounded, such that arbitrarily large amounts of additional value contribute arbitrarily little additional utility. As a result, expected utility theory can avoid the fanatical verdict described here. But, if it does, it faces the objections raised in Sections 4, 5, and 6. Where relevant, I will indicate in notes how the argument applies to expected utility theory.
or any enormous finite value. Expected value theory would say that it’s still better to take the risky bet - that it would be better to risk those billion or more lives for a miniscule chance at much greater value. But endorsing that verdict, regardless of how low the probability of success and how high the cost, seems *fanatical*.

That verdict does depend on more than just our theory of instrumental rationality, expected value theory. It also requires that our moral theory endorses *totalism*: that the ranking of outcomes can be represented by a total (cardinal) value of each outcome; and that this total value increases linearly, without bound, with the sum of value in all lives that ever exist. Then the outcome containing vastly more blissful lives is indeed a *much* better one than that in which one life is saved. And, as we increase the number of blissful lives, we can increase *how much better* it is without bound. No matter how low the probability of those many blissful lives, there can be enough such lives that the expected total value of the speculative research is greater than that of malaria prevention. But this isn’t a problem unique to totalism. When combined with expected value theory, analogous problems face most competing views of value (*axiologies*), including: averageism, pure egalitarianism, maximin, maximax, and narrow person-affecting views. Those axiologies all allow possible outcomes to be unboundedly good, so it’s easy enough to construct cases like Dyson’s Wager for each. I’ll focus on totalism here for simplicity, and also because it seems to me far more plausible than the others. But suffice it to say that just about any plausible axiology can deliver fanatical verdicts when combined with expected value theory.

A little more generally, we face fanatical verdicts if our theory of instrumental rationality (in conjunction with our theory of value) endorses *Fanaticism*. And to avoid fanatical verdicts it must, at minimum, avoid Fanaticism.

**Fanaticism:** For any (finite) probability $\epsilon > 0$ (no matter how low), and for any finite value $v$ on a cardinal scale, there is some value $V$ which is large enough that: we are

---

6Take (standard, welfarist) averageism, for instance. A population containing at least one blissful life of infinite (or arbitrarily long) duration will have average value greater than any finite value we choose. And so, to generate an averageist analogue of Dyson’s wager, we can substitute an outcome containing this population for the outcome of infinitely many lives in the original wager.

7Each of the other axiologies listed falls prey to devastating objections. See Arrhenius (2000), Huemer (2008), Greaves (2017), and chapters 17-19 of Parfit (1984).

8This use of the term ‘fanaticism’ seems to originate with Bostrom (2011) and Beckstead (2013: chapter 6). My formulation is slightly stronger than each of theirs but also, unlike theirs, applicable even if infinite total value cannot exist.
rationally required to choose the lottery $L_{\text{risky}}$ over $L_{\text{safe}}$.

$L_{\text{risky}}$: (an outcome with) value $V$ with probability $\epsilon$; value 0 otherwise\(^9\)

$L_{\text{safe}}$: value $v$ with probability 1

The comparison of lotteries $L_{\text{risky}}$ and $L_{\text{safe}}$ resembles Dyson’s Wager: one option gives a slim chance of an astronomical value $V$; the other a certainty of some modest value $v$. $V$ need not be infinite, in case you think infinite value impossible. But $v$ must be finite. And Fanaticism in this form implies the fanatical verdict in Dyson’s Wager, if we choose sufficiently many (perhaps infinitely many) blissful lives. Likewise, to reject the fanatical verdict in Dyson’s Wager, we must reject Fanaticism.

You might think it easy enough to reject Fanaticism. Many philosophers have done so, in the domains of practical reason and also moral decision-making. For instance, Bostrom (2009) presents a compelling *reductio ad absurdum* for all fanatical views in the prudential context. Bostrom (2011), Beckstead (2013), and Askell (2019) treat (a weak form of) Fanaticism as itself a *reductio* for theories in the moral context. Others propose theories of rationality by which we simply ignore small enough probabilities (e.g., D’Alembert 1761; Buffon 1777; Smith 2014\(^10\); Monton 2019). And Tarsney (n.d.) goes to great lengths to develop a theory resembling expected value theory which specifically avoids Fanaticism.

Meanwhile, there are few defenders of Fanaticism. No philosopher I know of has *explicitly* defended it in print. And few philosophers have defended fanatical verdicts in cases like Dyson’s Wager, with the exception of Pascal (1669) himself and those who, I suspect reluctantly, endorse his conclusion.\(^{11}\) And even they accept it only as a consequence of expected value theory, not

\(^9\)Note that this outcome with value 0 need not be what we typically think of as situations with no value: a world devoid of all value and disvalue; or one in which value and disvalue perfectly cancel one another out. These values are on merely a *cardinal* scale. They each might contain enormous amounts of value and/or disvalue, but the differences between them are *proportional* to $V$, $v$, and $V - v$. This will be important later.

\(^{10}\)Smith also suggests a version of his proposal which does not rule out Fanaticism. By that version, we still ignore events with probability below some threshold but, in any lotteries over finitely many different outcomes, that threshold is set lower than that of the least probable outcome. This still allows for Fanaticism as defined above, and also allows Smith to avoid the counterintuitive judgements in the St Petersburg and Pasadena games in which he is more interested. But a more natural version of Smith’s proposal which matches that of, e.g., Buffon (1777), states that there is some fixed probability threshold, and so rules out Fanaticism.

\(^{11}\)Parfit (1984: §3.27) can be interpreted as endorsing Fanaticism too.
because it has good independent justification. Even to most who endorse them, I suspect that
fanatical verdicts are seen as unfortunate skeletons in the closet of expected value theory.

I think that this situation is unfortunate. We have good reason to accept Fanaticism beyond
just expected value theory. As I hope to show, there are compelling arguments in favour of
Fanaticism in the moral context. As those arguments show, if we reject Fanaticism then we face
disturbing implications.

The paper proceeds as follows. Section 2 addresses common motivations for rejecting Fanaticism. Section 3 introduces the necessary formal framework for what follows. Sections 4 through 7 present arguments in favour of Fanaticism, each premised on weaker claims than expected value theory, and each (by my reckoning) more compelling than the last. The first is a basic continuum argument. The second is driven by a basic assumption that we can put at least some value on each lottery we face. The third is that: to deny Fanaticism we must accept either ‘scale-inconsistency’ or an absurd sensitivity to small probability differences, both of which are implausible. And the final nail in the coffin is what I will call the Indology Objection (a cousin of Parfit’s classic Egyptology Objection) by which those who deny Fanaticism must make judgements which appear deeply irrational. Section 8 is the conclusion.

2 The case against fanaticism

You might think that Fanaticism is simply absurd, and nothing more needs saying about it. I have
often heard this response. So, before presenting arguments in favour of Fanaticism, let us address
the most common objections: the objection from intuition; the objection from impossibility; and
the objection from tolerance. Although each objection seems initially plausible, I don’t think
any of them turn out to be compelling. And even if you find them more compelling than I
do, they certainly do not rule decisively against Fanaticism; the arguments for it are still worth
hearing.

\footnote{There are others which I won’t address here, both for brevity and because I find them even less compelling. See Monton (2019) and Smith (2014) for other objections which, with some modification, can be applied to Fanaticism. See also Hájek (2014; n.d.) for compelling responses.}
2.1 The objection from intuition

The first, and most common, objection is this: Fanaticism is frightfully counterintuitive.\footnote{In effect, this is a core point of Bostrom’s (2009) objection. Meanwhile, Bostrom (2011), Beckstead (2014), and Tarsney (n.d.) all treat Fanaticism as implausible for, it seems, the same reason.} It seems so very implausible that I should give up a certainty of a good payoff for a tiny chance of something better, no matter how tiny the chance. Rather, intuition says that an event with such a tiny chance can never tip the scales against the sure payoff.

But the psychology literature tells us that our initial intuitions about matters of probability are often misguided. We frequently commit the Conjunction Fallacy (Tversky & Kahneman 1983), the Gambler’s Fallacy (Chen et al. 2016), the Hot Hand Fallacy (Gilovich et al. 1985), and the Base Rate Fallacy (Kahneman & Tversky 1982). Perhaps worst of all, we are prone to simply ignoring facts of probability - many of us choose based entirely on how we anticipate things will actually turn out, with no regard for risks of loss or chances of gain (Baron et al. 1993; Gurmankin & Baron 2005). Even when we do account for probability, we often radically overestimate some probabilities due to availability bias (Tversky & Kahneman 1974), and underestimate others out of indefensible optimism (Hanoch et al. 2019).

And our intuitions are no better with decision-making in the face of low-probability events. For instance, jurors are just as likely to convict a defendant based primarily on fingerprint evidence if that evidence has probability 1 in 100 of being a false positive as if it were 1 in 1,000, or even 1 in 1 million (Garrett et al. 2018). In another context, when presented with a medical operation which posed a 1% chance of permanent harm, many respondents considered it no worse than an operation with no risk at all (Gurmankin & Baron 2005). And in yet another context, subjects were unwilling to pay any money at all to insure against a 1% chance of catastrophic loss (McClelland et al. 1993). So it seems that many of us, in many contexts, treat events with low probabilities as having probability 0, even when their true probability is as high as 1% (which is not so low). Upon reflection, this is clearly foolish. We may have the raw intuition that we may (or should) ignore all outcomes with probability 1% but it does not hold up to scrutiny, and so is likely mistaken. And given how widespread these mistakes are, we should give little weight to our intuitions about what we should do in cases of low probability, including those which lead us to recoil from fanatical judgements - with a little more scrutiny, those intuitions may appear foolish too.
2.2 The objection from impossibility

Here is another objection to Fanaticism. We might confidently claim that: ‘any event with a sufficiently low probability will not happen’. This claim is known as Cournot’s Principle, and was endorsed by many leading figures of probability theory, including not just Cournot (1843) but also Markov (1900), Levy (1925), Kolmogorov (1950), and Borel (1962). If they’re right that such events simply won’t happen, then it may seem fine to ignore them in our decision-making.

Why might we believe Cournot’s Principle? One reason might be that we should believe to be false any proposition which has probability sufficiently close to 0 (but not necessarily equal to 0) of being true. So, for any sufficiently unlikely event, we should believe that it won’t happen. But this isn’t reason enough. A low probability of an event may be enough to believe that it will not happen, but belief - at least in a binary form like this - is not enough for decision-making. We can believe a proposition without being entirely confident in it. And it would be foolish to ignore all possibilities which we believe won’t occur, but nonetheless have reasonably high probability of occurring.

A second reason to accept it is this. Suppose there are only finitely many distinct ways that the full history of the universe could turn out. You might think that this is the case if all events can be fully specified by the properties of finitely many discrete particles at finitely many discrete times, and if all such properties admit discrete values (as do length and duration, in multiples of the Planck length and Planck time). Then, since there are only finitely many possible histories of the universe, in any given lottery there can be only finitely many outcomes with chance greater than zero. So there must be some minimum probability $\epsilon$. And indeed no event with probability less than this will ever happen. There are no such events!

But this justification is unpersuasive. For one, for it to hold, every single physical property must admit discrete values, rather than continuous values. Perhaps that is true, but it quite plausibly isn’t. For two, the universe would need to contain only finitely many particles, finitely many distinct times, and finitely many spatial positions. For three, not only must the universe be finite in all of these respects, but it must also be guaranteed to not to exceed some specific finite bound in each respect. It might not only be certain that the universe has, say, some finite number of particles, de dicto, but it may not be certain that it has some specific finite number of particles, de re. There could still be infinitely many possible numbers of particles just as there are infinitely many integers, even though each of them is finite. And that would give us
infinitely many possible outcomes, thereby quashing the justification above. And finally, for three, Fanaticism wouldn’t be ruled out even if we granted all of these required assumptions. After all, it’s a claim about what we should do if we faced such a decision. The mere fact that we do or will not does not imply anything about what we should do if we did. A fully general normative theory - one that applies to a broader range of possible outcomes and probabilities - could well still be fanatical.

But no matter its justification, we can reject Cournot’s Principle and the objection from impossibility. The normatively relevant notion of probability, I should think, is the agent’s subjective degree of belief (their credence) in the outcome.\textsuperscript{14} And there are events that do occur and to which we ought to have assigned arbitrarily low credence. For instance, consider the outcome that the Sydney Opera House is presently the exact distance that it actually is \textit{(de re)} from the Golden Gate Bridge.\textsuperscript{15} As an agent myself, I am wildly uncertain of what that distance is. Perhaps I know that it must be an integer multiple of the Planck length, based on my understanding of physics. But I am uncertain even of exactly how long a Planck length is. (I admit, I had to look it up.) You should be too, since there is no way we can measure the Planck length down to an arbitrarily precise value. An agent like you or I, even if we are epistemically rational, should place arbitrarily low credence on the exact distance between the two monuments being within an arbitrarily small interval of length. But there is some such precise length between the two sites, which lies within some arbitrarily small interval to which we’ve given credence arbitrarily close to zero. And yet that outcome has happened! So Cournot’s Principle is demonstrably false.

2.3 The objection from tolerance

Here is one last objection to Fanaticism, from Smith (2014).

Any plausible decision theory says that we can ignore at least some outcomes - that improving those outcomes makes no change to the verdict. Those are the outcomes with probability precisely 0. But decision-making is a practical activity, Smith points out; we humans can only do it within some (perhaps small) margin of tolerance. We often cannot pin down precise probabilities for

\textsuperscript{14}Alternatively, it might be the evidential probability of the outcome, which carries the same implications.

\textsuperscript{15}To avoid some of the possible vagueness here, we might replace ‘the Sydney Opera House’ with ‘the centre of gravity of the Sydney Opera House’, and similarly for the Golden Gate Bridge.
outcomes to within a margin of less than some $\epsilon$. So what should we do when an outcome has probability near 0, within $\epsilon$? It seems we are permitted (or perhaps required) to ignore it, just as we would if the probability were precisely 0. And so Fanaticism won’t hold - no matter how good that outcome, there is some probability $\epsilon$ below which it need not count at all.

Smith’s argument has faced various counter-objections (see Hájek 2014; Isaacs 2016). Here is my preferred one. The requirement of tolerance suggests that we may (or should) treat a probability of $\frac{\epsilon}{2}$ as 0, but no more than it suggests that we may (or should) treat it as $\epsilon$. Whatever that minimum margin of accuracy/inaccuracy, tolerance by itself doesn’t imply anything about whether we should adjust those probabilities down or up. If we interpret Smith as concluding that we are required to approximate the probability to 0, then that conclusion is unjustified. If we instead interpret him as concluding that we are permitted to approximate to 0, then perhaps that is plausible, but we are also permitted to approximate up to $\epsilon$. And this permissibility yields strange verdicts, more fanatical even than those that I had in mind - we are permitted to overvalue some low-probability outcomes! Even if we only allow ourselves to round to the nearest multiple of $\epsilon$, then outcomes with probability as low as $\frac{\epsilon}{2}$ may be overvalued, by up to a factor of 2.\(^{16}\) So our judgements can be swayed by low-probability events even more often. If you recoil from Fanaticism, I expect you will recoil from this implication too.\(^{17}\)

In short, none of those three objections give us a rock-solid case against Fanaticism. At the very least, that case is wobbly enough to seriously consider whether Fanaticism might be true.

### 3 Background assumptions

Before I make a case for Fanaticism, here are my basic assumptions.

First, for any decision problem, there is some set of epistemically possible outcomes $O$. Some outcomes are better or worse than others. So assume that there exists some binary relation on $O$, denoted by $\succeq_O$, which represents an ‘at least as good as’ relation between outcomes.

\(^{16}\)Alternatively, if we can just approximate probabilities anywhere in $[0, \epsilon]$ down to 0 or up to $\epsilon$, then our verdicts are even stranger - an arbitrarily low probability could be treated as high as $\epsilon$.

\(^{17}\)As I’ll show in Section 5, if we reject Fanaticism then we may be required to be even more ‘intolerant’ of arbitrarily small changes in probability. I would claim that the intolerance I describe there is worse than simply treating probability $\epsilon$ as $\epsilon$.  

10
As mentioned earlier, I’m assuming a totalist view of value. So $\succeq_O$ is already given: for any possible outcomes $O_1$ and $O_2$, we have $O_1 \succeq_O O_2$ if and only if $O_1$ contains at least as much total value as $O_2$. Total values are totally ordered, so we know that $\succeq_O$ will be reflexive, transitive, and complete on $\mathcal{O}$. And the total value of outcomes can be represented with a cardinal value function $V : \mathcal{O} \to \mathbb{R}$, at least when those outcomes have finite differences in value.\footnote{The differences in value for two pairs of outcomes are treated as finite precisely when one can be represented as a real multiple of the other.} And that function $V$ is unique, at least up to affine transformations - for any $V_1$ or $V_2$ we might use here, $V_2(O) = a \times V_1(O) + b$ for some positive $a$ and real $b$.

We may not always be able to give a real representation of the total value for every outcome. If an outcome $O_\infty$ contains infinitely more value than others - as when we create infinitely many blissful lives in Dyson’s Wager - then it may fall beyond the scope of $V$. Then $V(O_\infty)$ won’t be defined, but still $O_\infty \succeq_O O$ for each finitely-valued outcome $O$. And that’s fine - a real-valued total value function $V$ won’t represent $\succeq_O$ in all cases, but it will do fine in purely finite cases. I’ll use $\mathcal{O}_\mathbb{R} \subset \mathcal{O}$ to denote any given set of outcomes over which $\succeq_O$ admits a real-valued representation $V$. The arguments that follow can be run entirely in terms of $\mathcal{O}_\mathbb{R}$.

What we are really interested in is lotteries. I’ll assume that the relevant aspects of a lottery can be fully described by a probability measure $L$ on a set of outcomes. I’ll also assume that this probability represents either the agent’s subjective degree of confidence or the evidential probability of each outcome arising. Almost all of what follows can be read in terms of either. With minor changes, what follows could also be read in terms of objective chance or whatever notion of probability you consider relevant for decision making. Whichever notion it is, let $\mathcal{L}$ be the set of all lotteries on $\mathcal{O}$, and $\mathcal{L}_\mathbb{R}$ the set of all lotteries on a given subset $\mathcal{O}_\mathbb{R}$. So each lottery $L$ is a function which maps sets of possible outcomes to some value in the interval $[0,1]$, and this function must obey the standard probability axioms.

To keep things brief, I’ll use the following each as a shorthand. When we are interested in the probability of a single outcome, I’ll shorten $L(\{O\})$ to $L(O)$. When a lottery $L$ gives a certainty of an outcome $O$ - when $L(O) = 1$ - I’ll write $L$ as $O$. To represent a lottery with the same probabilities as $L$, but outcomes which have $k$ times the value (on that same cardinal scale), I’ll write $k \cdot L$.\footnote{Formally, for any real $k$ and $L \in \mathcal{L}$, define $k \cdot L$ as a probability measure on $\mathcal{O}$ such that, for all $O$ in $\mathcal{O}$, $k \cdot L(O_k) = L(O)$, where $O_k$ is an outcome in $\mathcal{O}$ such that $V(O_k) = k \times V(O)$.} And I’ll write $L_a + L_b$ to represent the lottery obtained from adding together the
values of the outcomes of lotteries $L_a$ and $L_b$, run independently.\footnote{This can be made more precise. Let $V_i$ denote the random variable corresponding to a given lottery $L_i$, which outputs the total value of the outcome - equivalently, the probability that $V_i$ takes on a value in the interval $[a, b]$ is given by $L_i$\{ $O \mid V(O) \in [a, b]$\}). Then, for any $L_a, L_b \in \mathcal{L}$, define $L_a + L_b$ as some lottery on $\mathcal{O}$ - there may be several - which corresponds to the random variable $(V_a + V_b)$. Equivalently, $L_a + L_b$ is some lottery such that $L_a + L_b$\{ $O \mid V(O) \in [c, d]$\}) is equal to the probability that $(V_a + V_b)$ takes on a value in the interval $[c, d]$.}

Fanatical verdicts, and their rejection, consist in comparisons of lotteries. So we need another ‘at least as good relation’. Let $\succsim$ be a binary relation on $\mathcal{L}$ (and hence also on the subset $\mathcal{L}_\mathbb{R}$). Strict betterness ($\succ$) and equality ($\sim$) are defined as the asymmetric and symmetric components, respectively.

I won’t assume that $\succsim$ is complete on $\mathcal{L}$. But I will assume that it is reflexive: that $L \succsim L$ for all $L \in \mathcal{L}$. I’ll also assume that it is transitive: that, for all $L_a, L_b, L_c \in \mathcal{L}$, if $L_a \succsim L_b$ and $L_b \succsim L_c$, then $L_a \succsim L_c$. Both of these properties are highly plausible. If either of them do not hold, then betterness is a peculiar thing.\footnote{Some argue that moral betterness (and presumably also instrumental moral betterness) is not transitive, e.g. Temkin (2012). But the most compelling of these arguments assume pluralism with respect to moral value. But I’m focusing on monistic theories of value here, so transitivity remains a compelling principle.}

I also want to assume another highly plausible principle of instrumental rationality, which will be useful in what follows: Stochastic Dominance. This principle says that if two lotteries have exactly the same probabilities of exactly the same (or equally good) outcomes as each other, then they are equally good. And if you improve an outcome in one of those two lotteries, keeping the probabilities the same, then you improve that lottery. And that’s hard to deny!

We can express Stochastic Dominance as follows. Here, $\mathcal{O}_{\succ O'}$ denotes the set of outcomes in $\mathcal{O}$ that are at least as good as $O'$.

**Stochastic Dominance:** For any $L_a, L_b \in \mathcal{L}$, if $L_a(\mathcal{O}_{\succ O'}) \geq L_b(\mathcal{O}_{\succ O'})$ for all $O' \in \mathcal{O}$, then $L_a \succ L_b$.

If, as well, $L_a(\mathcal{O}_{\succ O'}) > L_b(\mathcal{O}_{\succ O'})$ for some $O' \in \mathcal{O}$, then $L_a \succ L_b$. 

I find Stochastic Dominance overwhelmingly plausible. In this setting, it is far weaker than (but necessary for) expected value theory - unlike the stronger theory, it does not rule out risk aversion, nor Allais preferences. To deny it is to accept either: that you can swap an outcome
in a lottery for a better one without making the lottery better; or that evaluations of lotteries are dependent on something other than the values of the outcomes and their probabilities. Both are implausible.\textsuperscript{22}

And Stochastic Dominance in a setting like this is fairly uncontroversial among decision theorists, including some who reject expected value theory. As far as I know, no serious proposal has been made in normative decision theory which violates Stochastic Dominance (at least in this setting).\textsuperscript{23}

\section{A continuum argument}

You might argue for Fanaticism in the following way: expected value theory is true; expected value theory implies Fanaticism; therefore, Fanaticism is true. And likewise for verdicts which seem fanatical, like that which the expected value theorist must accept in Dyson’s Wager. Such verdicts are rarely defended on any grounds other than ‘That’s what expected value theory says’.\textsuperscript{24} But we can do better than this. Fanaticism is far weaker than expected value theory in all its strength (and weaker even than expected utility theory), and so it should be easier to justify.

In this and the next two sections, I’ll give four arguments for Fanaticism but not for expected value theory. Here is the first.

Take the lottery $L_0$, by which one stranger’s life is saved with near certainty.

$L_0$: value 1 with probability 0.99; value 0 otherwise

And take another lottery $L_1$ by which vastly more strangers are saved, with only slightly lower probability of success - perhaps $10^{10}$ lives saved with probability $0.99 \times 0.999999$.

$L_1$: value $10^{10}$ with probability $0.99 \times 0.999999$; value 0 otherwise

\textsuperscript{22}For compelling arguments in favour of Stochastic Dominance, see Easwaran (2014) and Bader (2018).
\textsuperscript{23}Schoenfield (2014) rejects a similar principle, but only for lotteries over outcomes for which we don’t have a complete ordering. With our complete $\succeq_O$ relation on $O$, even Schoenfield would endorse Stochastic Dominance here.
\textsuperscript{24}One refreshing exception is Beckstead (2013: 142-3). See also Isaacs (2016).
Intuitively, $L_1$ seems better. Accepting a *slightly* lower probability of success for a *vastly* greater payoff seems a good trade. But then consider $L_2$, which has a slightly lower probability of success but, if successful, even more lives would be saved.

$L_2$: value $10^{10^{10}}$ with probability $0.99 \times 0.9999999^{2}$; value 0 otherwise

Again, this seems better than $L_1$, or at least it seems that there is some number of lives high enough that it would be better. And so we could continue, with $L_3$, $L_4$, and so on until some $L_n$, such that $0.99 \times 0.9999999^n$ is less than $\epsilon$, for any arbitrarily low $\epsilon$ you want. Intuition suggests that a *vastly* greater payoff trumps a *slightly* lower probability; that each lottery in the sequence is better than the one that precedes it. So the final lottery in the sequence must be better than the first. But the final lottery has a probability less than $\epsilon$ of any positive payoff at all. So we have Fanaticism.\(^{25}\)

This argument rests on two intuitively plausible principles: the first, transitivity of $\succeq$; the second, *Minimal Tradeoffs*.

*Minimal Tradeoffs*: There is some ratio $r < 1$ such that, for any real value $v$ and any probability $p$, there is some real $r^*$ such that $L_a$ is better than $L_b$ (as defined below).

$L_a$: value $v$ with probability $p$; 0 otherwise

$L_b$: value $r^* \times v$ with probability $r \times p$; 0 otherwise

This principle has intuitive force. For any given lottery with the form of $L_a$, *surely* there is some lottery with ever-so-slightly lower probability of success that is better. Intuition suggests that we can always make at least some tradeoff between probability and value: we can compensate for a $0.000001\%$ lower probability of success with a vastly greater payoff.

But the strength of your intuition in favour of Minimal Tradeoffs may not be as strong as your intuition that Fanaticism is false. And other than clashing with intuition, there is nothing immediately problematic about denying Minimal Tradeoffs. Perhaps we should just reject it; perhaps there is at least one threshold of probability $p'$ between $p$ and 0 (which might be unknown or vague) at which we have a discontinuity. Some value with probability above $p'$ is no worse

\(^{25}\)This argument originates in Beckstead (2014: 139-47).
than any value with probability below $p'$. One way to do this is by adopting expected utility theory with a bounded utility function, which would imply that the former lottery is better than the latter. But, however we reach that conclusion, there is some point at which effectively we can no longer trade off probability against value, no matter how great the value. And what of it? It seems a little counterintuitive but, beyond that, it is not obviously problematic.

So you might reject Minimal Tradeoffs, along with the continuum argument it generates. To deny Fanaticism, you must do so.

5 A dilemma for the unfanatical

The next argument for Fanaticism comes in the form of a nasty dilemma which you must face if you deny it. In brief, they who deny Fanaticism must accept at least one of: abandoning Scale Consistency, as defined below; and allowing comparisons of lotteries to be absurdly sensitive to tiny changes.

*Scale Consistency:* For any lotteries $L_a$ and $L_b$, if $L_a \succeq L_b$, then $k \cdot L_a \succeq k \cdot L_b$ for any positive, real $k$.

I find Scale Consistency hard to deny. Take any two lotteries, perhaps the $L_1$ and $L_2$ given below.

$L_1$: value 1 with probability 1

$L_2$: value 2 with probability $\frac{1}{2}$; value 0 otherwise

I’ll remain agnostic on which is better; $\succeq$ might recommend either. But suppose we scale up each lottery by 2, to obtain $2 \cdot L_1$ and $2 \cdot L_2$ - lotteries with the same probabilities as above but payoffs doubled (at least on that same cardinal scale). However we might rank the original lotteries, intuitively, that ranking shouldn’t change as we double the payoffs. After all, the scaled-up lotteries have exactly the same structure as the originals. Whatever general principles lead us to say whether $L_1$ is better than $L_2$ should also lead us to say the same about $k \cdot L_1$ and $k \cdot L_2$, for any positive $k$ you like.
On top of this, recall that value may be represented only cardinally (rather than perhaps on a ratio or unit scale). In value theory, it is often assumed that this is all the structure there is, since measures of value are often just representations of the (more fundamental) betterness relation \( \succeq_O \). This is much like cardinal measures of temperature, by which it doesn’t matter whether we measure with degrees Celsius, Farenheit, or Kelvin. Likewise for financial value (which admits a ratio-scale representation): the same underlying value can be represented in units of Australian dollars, British pounds, or Chinese RMB. As with these other phenomena with merely cardinal- or ratio-scale representations, the value of worlds might be represented with units of one size, or units that are \( \frac{1}{k} \) of the size. So \( L_1 \) and \( k \cdot L_1 \) can be representations of exactly the same underlying lottery, and \( L_2 \) and \( k \cdot L_2 \) the corresponding representations of another lottery. So to reject Scale Consistency is to compare lotteries on the basis of some factor beyond just their probabilities and (cardinal) representations of value. But that should be all we need to compare lotteries, and so Scale Consistency should hold.

I’ll illustrate below what I mean by ‘sensitivity to tiny changes’ shortly but, for now, be assured that it’s just as implausible as Scale Inconsistency.

The argument goes like this. First, for Fanaticism to be false, there must be some value \( v \) and some probability \( \epsilon > 0 \) such that \( L_{\text{risky}} \) is no better than \( L_{\text{safe}} \), no matter how big \( V \) is.

\[
L_{\text{risky}}: \text{value } V \text{ with probability } \epsilon; \text{ 0 otherwise}
\]

\[
L_{\text{safe}}: \text{value } v \text{ with probability 1}
\]

But what about values below \( v \)? How do outcomes with lower value compare to an \( \epsilon \)-probability of arbitrarily enormous \( V \)? There are only two distinct possibilities (at least if we maintain Stochastic Dominance and transitivity). The first is: there is some smaller \( v^- > 0 \) for which \( L_{\text{risky}} \) is better than a sure outcome with value of \( v^- \), for sufficiently large \( V \). The second is: for any such \( v^- \), \( L_{\text{risky}} \) isn’t better than any sure outcome with value \( v^- \), no matter how large \( V \) is.\(^{26}\) As we’ll see, each possibility impales us on a respective horn of the dilemma.

Assume that the first possibility holds. \( L_{\text{risky}} \) is better than a sure outcome with value \( v^- \), for some \( 0 < v^- < v \) and some large \( V \). Still, \( L_{\text{risky}} \) won’t be better than a sure outcome with

\(^{26}\)This difference is analogous to the difference between weak and strong superiority in Arrhenius & Rabinwocz (2015).
value \( v \) (i.e., \( L_{\text{safe}} \)), or any sure outcome with even greater value (by Stochastic Dominance). But somewhere below \( v \), this changes; for some positive \( v^- \), the same doesn’t hold.

Since that \( v^- \) is a positive real number, we know that there will be some real \( k \) such that \( k \times v^- \geq v \). And that means that a sure outcome with value \( k \times v^- \) is no worse than \( L_1 \), no matter how great \( V \) is. But here is the problem: a certainty of \( k \times v^- \) is the scaled-up counterpart of \( v^- \); and the counterpart of \( L_{\text{risky}} \), which is \( k \cdot L_{\text{risky}} \), is just \( L_{\text{risky}} \) with \( V \) replaced by another value which is \( k \) times higher. That still has the same form as \( L_{\text{risky}} \). So it cannot be better than the sure outcome, not if we reject Fanaticism. But then our verdict for \( v^- \) versus \( L_{\text{risky}} \) doesn’t match our verdict for \( k \times v^- \) versus \( k \cdot L_{\text{risky}} \). So we violate Scale Consistency.\(^{27}\)

But there was a second possibility: that there is no such \( v^- > 0 \); that \( L_{\text{risky}} \) is no better than a certainty of any \( v^- > 0 \). If so, we can avoid the problem of scale inconsistency: there’s no inconsistency between the judgements for \( v \) and judgements for any ‘scaled-down’ counterpart \( v^- \), since all such \( v^- \) compare the same way to lotteries like \( L_{\text{risky}} \).\(^{28}\) But we then face another serious problem.

Take the probabilities \( \epsilon \) and 1. We can give a sequence of increasing probabilities in between \( \epsilon \) and 1, spaced evenly apart and as finely as we want: \( \epsilon < p_1 < p_2 < \ldots < p_n < 1 \). By assumption, no amount of value with probability \( \epsilon \) (and value 0 otherwise) is better than any (positive) amount of value with probability 1. But then there must be some pair of successive \( p_i, p_{i+1} \) such that no amount of value at \( p_i \) is better than any amount of value with probability \( p_{i+1} \). If there were no such pair, we could give a sequence of better and better, and less and less likely, lotteries much like that in the previous section.

Crucially, that \( p_i \) and \( p_{i+1} \) can be arbitrarily close together, since the same result holds no matter how finely spaced the sequence was. No amount of value at \( p_i \) is better than any amount of value at \( p_{i+1} \). We could have some astronomical value at \( p_i \), and an imperceptibly small amount of value at \( p_{i+1} \), and the latter would still be better. So we have a radical discontinuity in how we value lotteries.

\(^{27}\)If we adopt expected utility theory with a bounded utility function, we are impaled on this horn of the dilemma - \( L_{\text{risky}} \) will still be better than a certainty of some lower value \( v^- \), and so we will violate Scale Consistency (see Rabin 2000).

\(^{28}\)The comparison won’t be the same for values 0 and below, but that’s okay. Such values cannot be scaled up to \( v \); multiply 0 or any negative number by any real positive value you want and you still won’t reach \( v \).
This will make our judgements of betterness *absurdly sensitive* to tiny changes in probability. For instance, consider the following four lotteries. Here, $\epsilon' > 0$ is some arbitrarily small number.

$L_0$: value 0 with probability 1  
$L_1$: value $10^{10^{10}}$ with probability $p_i$; 0 otherwise  
$L_2$: value $\epsilon'$ with probability $p_{i+1}$; 0 otherwise  
$L_3$: value $10^{10^{10}}$ with probability $p_{i+1}$; 0 otherwise

By the above, we are forced to rank these lotteries as $L_0 \prec L_1 \prec L_3$ and $L_0 \prec L_2 \prec L_3$, but $L_1 \not\prec L_2$. And that’s a peculiar ranking. $L_1$ and $L_3$ are almost indistinguishable; their probabilities can differ by an arbitrarily tiny amount. So are $L_0$ and $L_1$; their payoffs differ similarly. For each pair, it makes sense that one is better than the other, but not that it is *much* better. But here, in a sense, $L_3$ is *much* better than $L_1$, as is $L_2$ much better than $L_0$. How so? Between $L_1$ and $L_3$ comes $L_2$. It has an *astronomically* smaller payoff than $L_3$ and so is far worse than $L_3$, and yet $L_1$ still is not better than it. Effectively, it is no worse to lose almost all of the potential value via $L_2$ than it is to lose that sliver of probability via $L_1$. And similarly for $L_0$ and $L_2$: between them comes $L_1$. That lottery $L_1$ has astronomically greater potential payoff than $L_0$ (and much higher probability of positive payoff too), yet it’s still not better than $L_2$. Effectively, it is no better to gain a $p_i$-probability of enormous value than it is to gain that tiny sliver of slightly-more-probable value in $L_2$. So those tiny changes in probability, and payoffs, make those lotteries in an intuitive sense *much* better or worse. And this seems absurd. Evaluations of lotteries should not change quite so wildly - so discontinuously - with arbitrarily tiny changes in probability or payoff.

Beyond its intuitive implausibility, this sensitivity leads to practical difficulties. We human agents have access to subjective degrees of belief and to evidential probabilities. It’s awfully difficult to pin either of these down with arbitrary precision, at least not with our merely finite capacity for calculation. But we’ll often need arbitrary precision, at least by this horn of the dilemma. If not, in cases like that immediately above, we will have no idea whether a lottery is as good as $L_3$ or as bad as $L_1$ with its imperceptibly less probably payoff. We must either require (unrealistically) that agents be arbitrarily precise in their probabilities, or else accept that in practice our decision theory will be unhelpful for agents choosing between lotteries like...
This is the dilemma we face if we reject Fanaticism. We must accept either scale inconsistency, by which our judgements of lotteries can vary without any structural reasons for doing so, or that those judgements are sensitive to imperceptibly small differences in either probability and value. Both options seem absurd to me, indeed more absurd than Fanaticism itself.

6 The Indology Objection

But it gets worse. If you reject Fanaticism, you also face what I’ll call the Indology Objection. For context, Indology is the study of the history and culture of India. In particular, classical Indology includes the study of the Indus Valley Civilisation, a Bronze Age civilisation which lasted from 3300 BCE to 1300 BCE. Although less well-known and understood than its contemporaries - ancient Egypt and Mesopotamia - it spanned a greater area than either and rivalled them in population, technology, and urban infrastructure (see Wright 2009).

What does this have to do with normative theory? The objection I have in mind bears a close resemblance to the classic Egyptology Objection from population ethics, which is sometimes made against views like averageism and egalitarianism.

As a brief refresher, the Egyptology objection goes like this. Suppose you are making a moral decision here and now in the 21st Century, and the available outcomes differ only by some small-scale changes in the very near future. Your actions will not change what happens/happened in distant galaxies or in, say, ancient Egypt. Then, intuitively, what you ought to do should depend only on the changes your choices would make - those small-scale, short-term effects. It should not depend on events in distant galaxies or in ancient Egypt, at least not if your actions do not change them. Intuitively, it seems implausible that those unaltered events make any difference to your evaluation. And such dependence brings practical problems too. If what you ought to do is not dependent on such events - not just in the remote future but also the remote past, of which we are often clueless - then we have little guidance for what we ought to do. This seems a major disadvantage for a moral view.

Recall Smith’s (2014) argument against Fanaticism - that, if we accept it, our decision theory is intolerant of human imprecision. But rejecting Fanaticism, by this horn of the dilemma, brings similar intolerance.

This argument appears earliest in MacMahan (1981: 115) but is often attributed to Parfit (1984: 420).
But, under some views of value, what you ought to do in such cases is dependent on unaltered, remote events. Take a (standard, welfarist) averageist view. Here and now, should you bring an additional person into existence? It depends - will that additional life raise or lower the average value of lives that ever exist? Sometimes yes, sometimes no, depending on what actually happened in ancient Egypt, in distant galaxies, and everywhere else. A standard averageist view will thus imply that what happened in ancient Egypt can affect whether it is better to bring an additional person into existence now than to not do so. And, further, to be confident of which is the better outcome, you may need to do some serious research into Egyptology (along with plenty of other historical studies). But this is implausible. By intuition, we can ignore events that are unaffected by our actions and took place millennia ago, as we do under axiologies like totalism, aggregative person-affecting views, and any other axiology which admits an additively separable representation. To me, this is one of the main appeals of those views.

That’s the traditional Egyptology objection. What does it have to do with Fanaticism? It turns out that, if you reject Fanaticism, you fall prey to a similar objection. For one, your comparisons of lotteries will sometimes depend on what you know about events in ancient civilisations - such as ancient Egypt, Mesopotamia, and the Indus Valley. But, even worse than this dependence, you must sometimes make judgements which you know to be inconsistent with what you’d judge if you learned more. It may be that, no matter what you learn, you know you would settle on a particular judgement (de re). Yet, you still adopt a different judgement, which seems a deeply irrational thing to do.

The Indology Objection requires some setup. First, recall the lotteries from the definition of Fanaticism: \( L_{\text{risky}} \) and \( L_{\text{safe}} \). If we reject Fanaticism then, for some real \( v \) and for tiny enough \( \epsilon > 0 \), \( L_{\text{safe}} \) is no better than \( L_{\text{risky}} \), no matter how great \( V \) is.

\[
L_{\text{risky}}: \text{value } V \text{ with probability } \epsilon; \text{ 0 otherwise} \\
L_{\text{safe}}: \text{value } v \text{ with probability 1}
\]

We can imagine these lotteries representing the options in Dyson’s Wager: donate to combating malaria and you’ll produce \( L_{\text{safe}} \), saving one life with some modest \( v \); or donate to positronium research and you’ll produce \( L_{\text{risky}} \), producing some very large additional value \( V \). Those are the only differences among the possible outcomes of your choice, at least if we hold fixed all of the events which your choice won’t affect. When we hold fixed those unaffected events and value
outcomes only by their differences in value that the agent causes (e.g., 0, v, and V), call those values the simple payoffs.

But there is a great deal of value (and disvalue) in the other events in each outcome - the events which you're actions don’t alter, such as those in distant galaxies or in the ancient Indus Valley. I’ll call the value across those unaffected events - the value in the rest of the world - the background payoff. But, if you’re neither an Indologist nor omniscient, you’ll be uncertain as to what that background payoff is (on any given cardinal scale). Call this your background uncertainty, which forms a lottery $B$ over the possible outcomes of the rest of the universe. This lottery $B$ is independent of your choice, and of the lottery over simple payoffs which follows your choice. You cannot affect it - it will run no matter what you choose, and the actual background payoff will be the same either way.

With the distinction of simple and background payoffs in place, recall Stochastic Dominance. This overwhelmingly plausible principle states that: if a lottery $L_a$ gives at least as high a probability as $L_b$ of resulting in an outcome which is at least as good as $O$, for every possible outcome $O$, then $L_a$ is at least as good a lottery. And if $L_a$ also gives a strictly greater probability of turning out at least as good as some $O$, then it is strictly better.

Stochastic dominance is easy to spot graphically. To illustrate, consider the cumulative probability graphs of the following two lotteries, $L_1$ and $L_2$.

$L_1$: value 1 with probability $\frac{1}{2}$; value 0 otherwise
$L_2$: value 2 with probability $\frac{1}{3}$; value 1 with probability $\frac{1}{3}$; value 0 otherwise
On a graph like this, one can easily see when Stochastic Dominance says that $L_2 \succeq L_1$. Cumulative probability, on the vertical axis, is just the probability that the lottery produces an outcome no better than an outcome with some particular value. Meanwhile, Stochastic Dominance says that one lottery is at least as good as another if its probability of producing an outcome as good or better is just as high for all outcomes - or, equivalently, if the probability of an outcome as good or worse is at least as low. So Stochastic Dominance will say that $L_2 \succeq L_1$ if and only if $L_2$’s cumulative probability is always as low or lower than that of $L_1$, as it is on this graph. And, here, $L_2$ often has strictly lower cumulative probability. So Stochastic Dominance says it is strictly better than $L_1$.\footnote{This relationship between Stochastic Dominance and cumulative probability would break down if some outcomes were incomparable to others. There would then be a difference between $O_a \succ O_b$ and the negation of $O_a \prec O_b$. But, fortunately, totalism gives us a total preorder of $O$, so we can sidestep this complexity.}

But Stochastic Dominance can’t compare the two lotteries from the definition of Fanaticism, $L_{\text{risky}}$ and $L_{\text{safe}}$. Their cumulative probabilities look like this:

![Cumulative probability graphs for $L_{\text{risky}}$ and $L_{\text{safe}}$](image)

Sometimes one is higher; sometimes the other. So Stochastic Dominance remains silent. And it’s a good thing it does - to deny Fanaticism, we must not say that $L_{\text{safe}}$ is better than $L_{\text{risky}}$.\footnote{This relationship between Stochastic Dominance and cumulative probability would break down if some outcomes were incomparable to others. There would then be a difference between $O_a \succ O_b$ and the negation of $O_a \prec O_b$. But, fortunately, totalism gives us a total preorder of $O$, so we can sidestep this complexity.}
But consider a situation in which we have background uncertainty - for each option, we have a lottery over simple payoffs, but we also have a common lottery over what happens in the rest of the universe. Our lotteries over simple payoffs will be given by \( L_{\text{risky}} \) and \( L_{\text{safe}} \). And, for the background uncertainty, take some lottery \( B \) that looks like this. (I’ll describe the required properties of \( B \) below.)

![Figure 3: An example of background uncertainty \( B \): a Cauchy distribution](image)

When we have background uncertainty, it is not enough to just compare the simple payoffs \( L_{\text{risky}} \) and \( L_{\text{safe}} \) over simple payoffs. By totalism, an outcome is better than another not because it is better than the other in one narrow region, but because it is better \textit{in total} than the other. Betterness facts are grounded by facts about the total sum of value: the sum of the simple payoff and the background payoff. An outcome is better than another only if its sum of the two payoffs is higher than the corresponding sum for the other outcome. Of course, you cannot affect the background payoff - in any state of the world, no matter what you choose in Dyson’s Wager, the background payoff will be the same either way. But while considering the decision \textit{ex ante}, a good totalist’s uncertainty over the value of the outcome cannot exclude their uncertainty about the background payoff. Effectively, as a totalist, you cannot just compare \( L_{\text{risky}} \) and \( L_{\text{safe}} \) to decide what to do; you must compare \( L_{\text{risky}} + B \) and \( L_{\text{safe}} + B \).

And this is what \( L_{\text{risky}} + B \) and \( L_{\text{safe}} + B \) look like (with the above \( B \)).
As you can see, the graph for $L_{\text{safe}} + B$ is never higher than that for $L_{\text{risky}} + B$, and sometimes it is strictly lower. So Stochastic Dominance is enough to say that $L_{\text{safe}} + B$ is strictly better.

But how does this happen?

For Stochastic Dominance to hold, we need $L_{\text{risky}} + B$ to have at least as high a probability as $L_{\text{safe}} + B$ of turning out at least as good as an outcome with value $u$, for all possible values $u$. So take any real value $u < V$. What’s each lottery’s probability of doing at least that well?

Start with $L_{\text{risky}} + B$. We know that $L_{\text{safe}}$ just gives one value ($v$) with certainty. So the probability that $L_{\text{safe}} + B$ gives value $u$ or better is just the probability that $B$ gives value $u - v$ or better. (This corresponds to the area $B_2 + B_3$ on the graph below.) And then, for $L_{\text{risky}} + B$, we’ll get value at least $u$ either if $L_{\text{risky}}$ gives value $V$ or if $B$ gives at least value $u$. Denote the probability that $B$ gives value at least $u$ by $B_3$ (corresponding to that area on the graph below). Then the probability that $L_{\text{risky}} + B$ turns out at least as good as value $u$ is $\epsilon + B_3(1 - \epsilon)$. And, with a bit of simple arithmetic\(^{32}\), we can see that this will be greater than the corresponding probability for $L_{\text{safe}} + B$ if the area $B_2$ is no greater than $\epsilon \times B_1$.

\[ B_2 \leq \epsilon B_1 \iff B_2 + B_3(1 - \epsilon) \leq \epsilon \] is a sufficient condition.
Figure 5: Probability distribution of $B$

Here we have a probability graph of one possible $B$ (importantly, not a *cumulative* probability graph), with areas $B_1, B_2,$ and $B_3$ corresponding to the areas mentioned in the last paragraph. If $B_2$ is small enough compared to $B_1,$ then $L_{\text{risky}} + B$ has at least as high a probability of $u$ or better as $L_{\text{safe}} + B$ does. And for it to be at least as high for *all* real $u,$ we just need a the interval between $u$ and $u - v$ to have a tiny enough area under it. And, for that to happen, we just need $B$ to go down slowly enough as we approach $-\infty,$ and rise and fall quickly enough as we pass the peak of the curve.

And there are some possible distributions which have this property, notably any Cauchy distribution which is sufficiently stretched out.$^{33}$ And some is enough. There is then some possible situation in which an agent has that probability distribution for their background uncertainty - their uncertainty about what happens elsewhere in the universe, including in the ancient Indus Valley.

So an agent can consider lotteries $L_{\text{risky}}$ and $L_{\text{safe}}$ over simple payoffs, with background uncertainty matching $B$ above. And Stochastic Dominance says that they must make the judgement $L_{\text{risky}} + B \succ L_{\text{safe}} + B.$ But, if Fanaticism is not true, they cannot say that $L_{\text{risky}}$ alone is better than $L_{\text{safe}}$ alone. Nor can they say that $L_{\text{safe}}$ plus an additional payoff $b$ is better than $L_{\text{risky}}$ plus the same $b.$ (Adding a constant to both lotteries cannot change the verdict because totalism implies that outcomes can be compared with cardinal values alone.)

But therein lies the rub. By definition, whichever you choose, $B$ will spit out the same actual

---

$^{33}$The result holds for any Cauchy distribution with scale factor at least $\frac{\pi}{6}.$ See Tarsney (n.d.: §5) and Pomatto *et al* (2018).
background payoff $b$. If you deny Fanaticism, you know that no matter how your background uncertainty is resolved, you will deny that $L_{\text{risky}} + b$ is better than $L_{\text{safe}} + b$. But, unless you deny Stochastic Dominance, you still must say that $L_{\text{risky}} + B$ is strictly better than $L_{\text{safe}} + B$. This seems deeply irrational.\textsuperscript{34}

To bring this problem into sharper relief, suppose that the only events in the universe about which you’re uncertain are those that happened in the ancient Indus Valley. And suppose that you have that same distribution of background uncertainty plotted above. You don’t know much classical Indology but you could always learn more. Given that the first archaeological excavations of sites in Britain, Italy, and Egypt have been going for centuries longer than those of the Indus Valley (Wright 2009), there is likely plenty left to learn! Suppose that, with many years of intensive research, you could discover every relevant detail of what happened in the Indus Valley; you could pin down the exact value of $b$. And then you could make your decision between donating to malaria prevention or to positronium research, corresponding to $L_{\text{risky}}$ and $L_{\text{safe}}$. But you already know what your judgement will be! For any value to which you might pin $b$ down, you’ll establish that donating to positronium research is no better than donating to the malaria charity. So why do the many years of research? Why not just update your judgement now? You cannot. You must accept that $L_{\text{safe}} + B$ is strictly better, until you know more precisely which value $b$ takes. You must go through the years of gruelling research and digging to be able to make that judgement that you already know you would find to be correct.\textsuperscript{35}

This inconsistency seems deeply irrational. Surely we can separate $B$ from the lotteries over simple payoffs, 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 implication - an implication which, to me, seems far more absurd than simply accepting Fanaticism. And that is the Indology Objection.\textsuperscript{36}

\textsuperscript{34}The Indology Objection is the one argument I’ve given which does not apply for the full suite of axiologies I mentioned in the Introduction. It won’t apply quite the same if we accept one of the views which fall prey to the original Egyptology Objection, e.g., averageism or egalitarianism. Averageists and egalitarians might not face the Indology Objection if they reject Fanaticism, but I think the Egyptology Objection is problem enough for those views. And they’ll still face the arguments I made above.

\textsuperscript{35}cf van Fraasen’s (1984) principle of Reflection and its application by White (2010:17-8). That form of Reflection implies that, if you know that you would set your credence in some hypothesis to $p$ whether you learned proposition $P$ or not-$P$, then your current credence should be $p$. The version I am assuming is similar but different.

\textsuperscript{36}The objection as I’ve given it does apply to expected utility theory with a bounded utility function. (I find
The case I’ve described brings up several related objections, although I find them less persuasive than this one.

The first such objection is simply that your judgements are sensitive to background uncertainty if you deny Fanaticism. As we saw above, the presence of at least some possible $B$ changes your judgement in Dyson’s Wager. And this sensitivity might by itself seem implausible, even if we put aside the problem of inconsistency discussed above. When making a moral decision, intuition suggests that it is fine to ignore all of the events in the universe that are left unaltered - even if you are uncertain about exactly what happened in those events. But any theory which rejects Fanaticism must deny this.37

The second related objection is that any theory which denies Fanaticism won’t be quite so helpful in practice. You want to compare two bets that you might make? You had better consider your uncertainty about moral value everywhere in the universe, and whether that uncertainty gives a probability distribution which meets various subtle conditions. The correct judgement depends on it! While this might be possible, it makes practical decision-making at least significantly harder for limited human agents like ourselves. It may even be the case that you have no beliefs at all about what happened or will happen in remote parts of the universe, let alone precise credences. If so, you cannot act rationally in such cases. So we may not even be able to say that we ought to do something, let alone which thing we ought to do. This would make any such anti-fanatical theory a great deal less action-guiding. And, while this isn’t a decisive objection to a theory, I find it at least somewhat compelling that the correct theory of moral decision-making gives us some guidance in practice.

A third related objection is that, even if we deny Fanaticism, there is good reason to think that we often must still make judgements which seem fanatical. As we saw in the case above, with the right background uncertainty you must judge it better to donate to speculative research than to saving a life for sure - even though this judgement may still seem fanatical. In practice,

---

37This objection resembles one which has been levelled against decision theories which violate the Sure-Thing Principle (by, e.g., Briggs 2015). It turns out that following such a decision theory in each of several consecutive, independent decisions will sometimes be inconsistent with following the theory when deciding on a strategy for all of those decisions at once. Essentially, the combination of two lotteries can be evaluated quite differently to how we evaluate those lotteries separately. And that is troubling. It is unclear whether we should evaluate such decisions together or apart, globally or individually. But the correct verdict depends on which way we do it.
will our background uncertainty be right? Tarnsey (n.d.: §6) argues that it often will, for an epistemically rational agent and for a wide range cases resembling Dyson’s Wager (with \( \epsilon \) at least 1 in \( 10^9 \), and likely also a lot lower).\(^{38}\) If that is so, then Stochastic Dominance by itself will lead us to making judgements which seem fanatical. So we would gain little from rejecting Fanaticism.

Taking stock, if we deny Fanaticism then we must accept the Indology Objection, by which agents ought sometimes to make judgements which are inconsistent with anything they might learn to resolve their uncertainty. We must also accept: 1) that judgements are sensitive to uncertainty about far-off, unrelated events; 2) that our decision theory plausibly cannot give agents much practical guidance; and 3) that we’ll likely be committed to many fanatical-seeming judgements anyway. Given these implications, it seems a lot less appealing to deny Fanaticism.

7 Conclusion

These are just some of the compelling arguments for accepting Fanaticism. There are others too - for instance, all of the arguments for the stronger claims of expected value theory, or for expected utility theory. A fortiori, these also justify Fanaticism. For several of these, I refer interested readers to Feller (1968), Ng (2012), Briggs (2015), and Thoma (2018).

But some philosophers still reject expected value theory (e.g., Buchak 2013, Smith 2014, Monton 2019, Tarsney n.d.), and presumably the arguments which imply it. But this is not enough to escape Fanaticism. As I have argued here, there are compelling arguments in its favour, arguments stronger than many of those for expected value theory.

To recap, if deny Fanaticism we must deny either Minimal Tradeoffs or transitivity, and accept the counterintuitive verdicts which follow. Likewise, we must accept either: violations

\(^{38}\)In brief, Tarsney claims that moral value will be roughly correlated with the number of inhabited planets in the universe, and then shows that our cosmological evidence and the Drake equation together motivate a probability distribution over the number of inhabited planets which is relevantly similar to the above \( B \). But even apart from that reasoning, an epistemically rational agent would still have some uncertainty over which probability distribution best characterises our cosmological evidence. So they would still put some non-zero credence in distributions of the right sort. And this is enough - when we take the probability-weighted average of many such distributions to get our overall distribution, it will then have the right properties (see ibid.: 37-9 for details).
of Scale Consistency; or absurd sensitivity to the tiniest differences in probabilities and value. So too, to deny Fanaticism, we must accept the Indology Objection: you must sometimes make judgements which you know you would reject if you learnt more, no matter what you might learn.

Given all of this, it no longer seems so attractive to reject Fanaticism. As it turns out, the cure is worse than the disease. I would suggest that it is better to just live with Fanaticism.

All of this has implications for normative decision theory more broadly, at least dialectically. Philosophers often reject expected value theory because it implies Fanaticism, or because it implies fanatical verdicts in specific cases like Dyson’s Wager (e.g., Monton 2019, Tarsney 2020). The arguments I have given here suggest that this rejection is a little hasty. Doing so invites greater problems than it solves. So I, for one, am thankful that expected value theory does imply Fanaticism. We have little choice but to accept Fanaticism, so we might as well accept expected value theory too.39

References


39This paper would not be what it is were it not for input of many brilliant and generous colleagues. I am indebted to Alan Hájek, Christian Tarsney, and Jasper Hedges for their extensive comments. Likewise to Teru Thomas, Andreas Mogensen, Philip Trammel, Anders Sandberg, Shang Long Yeo, Jennifer Munt, and Oliver Rawle for valuable discussion. And likewise to seminar audiences at the Australian National University and the Forethought Foundation for Global Priorities Research for many helpful (albeit challenging) questions and comments. I am also grateful to the Global Priorities Institute at the University of Oxford for inviting me to visit for Hilary Term 2020, during which time this paper took shape.


