A paradox for tiny probabilities and enormous values

Nick Beckstead and Teruji Thomas

Global Priorities Institute | July 2021

GPI Working Paper No . 7-2021

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

A Paradox for Tiny Probabilities and Enormous Values

Nick Beckstead and Teruji Thomas*

Abstract

We show that every theory of the value of uncertain prospects must have one of three unpalatable properties. *Reckless* theories recommend risking arbitrarily great gains at arbitrarily long odds for the sake of enormous potential; *timid* theories permit passing up arbitrarily great gains to prevent a tiny increase in risk; *non-transitive* theories deny the principle that, if A is better than B and B is better than C, then A must be better than C. While non-transitivity has been much discussed, we draw out the costs and benefits of recklessness and timidity when it comes to axiology, decision theory, and moral uncertainty.

On your deathbed, God brings good news. Although, as you already knew, there's no afterlife in store, he'll give you a ticket that can be handed to the reaper, good for an additional year of happy life on Earth. As you celebrate, the devil appears and asks, 'Won't you accept a small risk to get something vastly better? Trade that ticket for this one: it's good for 10 years of happy life, with probability 0.999.' You accept, and the devil hands you a new ticket. But then the devil asks again, 'Won't you accept a small risk

*Working paper, June 2021. This paper was originally drafted by the first author based on his dissertation (Beckstead, 2013, chapter 6). The second author is responsible for most of the revisions and ideas newly appearing in this version. We are grateful to many people for feedback and assistance, including Amanda Askell, Andreas Mogensen, Christian Tarsney, David Thorstad, Elliott Thornley, Hayden Wilkinson, Hilary Greaves, Jacob Barrett, Larry Temkin, Petra Kosonen, Philipp Schoenegger, Stefan Riedener, Theron Pummer, Tomi Francis, and Will MacAskill.

to get something vastly better? Trade that ticket for this one: it is good for 100 years of happy life—10 times as long—with probability 0.999^2—just 0.1% lower.' An hour later, you've made 50,000 trades. (The devil is a fast talker.) You find yourself with a ticket for $10^{50,000}$ years of happy life that only works with probability $0.999^{50,000}$, less than one chance in 10^{21}. Predictably, you die that very night.

Here are the deals you could have had along the way:

Deal	0	1	2	3	⋯	n	⋯	50,000
Payoff	1	10	100	1,000	⋯	10^n	⋯	$10^{50,000}$
Probability	1	0.999	0.998	0.997	⋯	0.999^n	⋯	$< 10^{-21}$

On the one hand, each deal seems better than the one before. Accepting each deal immensely increases the payoff that's on the table (increasing the number of happy years by a factor of 10) while decreasing its probability by a mere 0.1%. It seems unreasonably timid to reject such a deal. On the other hand, it seems unreasonably reckless to take all of the deals—that would mean trading the certainty of a really valuable payoff for all but certainly no payoff at all. So even though it seems each deal is better than the one before, it does not seem that the last deal is better than the first.[1]

In this paper, we develop a general version of this paradox. In short, every theory of the value of uncertain prospects must be *timid*, *reckless*, or *non-transitive*. Timid theories permit passing up arbitrarily great gains to prevent a tiny increase in risk. Reckless theories recommend risking arbitrarily great gains at arbitrarily long odds for the sake of enormous potential. And non-transitive theories deny the principle that, if A is better than B and B is better than C, then A must be better than C.

[1] Arguments with a similar structure are often called *spectrum* or *continuum* arguments—see Temkin (2012) for a survey. The best known arguments involve tradeoffs between the quantity and the quality of a payoff (e.g. the duration of a life and its typical wellbeing at a time) rather than, as here, the quantity and the probability. Our argument has the useful feature that probability has an extremely well-established quantitative structure, with no natural threshold between 'high' and 'low' probability, whereas the structure of well-being (say) is more mysterious. Temkin discusses a probability-based spectrum argument in his chapter 8, but it works quite differently from ours.

We'll set out this general trilemma more formally in section 1. We think that timidity and recklessness are both intuitively implausible for many types of goods. Not simply resting on intuition, the remainder of the paper will draw out the troubling consequences of each. In contrast, we will not consider in depth the possibility of rejecting transitivity; for a comprehensive discussion of it, we defer to Temkin (2012) and the large associated literature. The arguments in this paper could, then, be seen as supporting non-transitivity, although we are not inclined toward that conclusion.

In section 2, we explain how the choice between timidity and recklessness fits in with some well-known approaches to the evaluation of prospects. The discussion includes expected value theory, risk-weighted expected utility theory, and a cleaned-up version of the 'Nicolausian discounting' view advocated by Smith (2014) and Monton (2019). For example, we note that the best-supported versions of utilitarianism, which specifically involve expected value theory, will lead to recklessness with respect to the creation of good lives. In general, our analysis of timidity and recklessness will be relevant to identifying the costs and benefits of many axiological theories.

In section 3, we argue that timid approaches have implications that will be implausible for many kinds of goods. For example, timidity means that events that happened in remote regions of space and time, beyond our influence, can have a dramatic effect on which actions are best in prospect. Moreover, timidity tends to require implausibly extreme risk aversion, and even timid views will tend to recommend certain kinds of long-shot bets, with counterintuitive results.

In sections 4 and 5, we describe some problems that arise from recklessness when one allows for unbounded or infinite payoffs. Versions of these problems will be familiar to many readers in the context of expected utility theory, where they are associated with the St Petersburg gamble and Pascal's wager, respectively. What's new here is that similar problems arise very generally from recklessness, regardless of whether one accepts expected utility theory. So these problems are more general and—given that the alternative is timidity—more pressing than usually thought. We also want to emphasize the importance of these problems in moral contexts, whereas discussions of St Petersburg and Pascal almost always focus on prudential value or mere preference.

Section 4 shows, more specifically, that recklessness conflicts with some very plausible dominance principles, while in section 5 we argue that reckless theories are *infinity-obsessed*: they evaluate an arbitrarily tiny probability of an infinite payoff above any guaranteed finite payoff. Although, for reasons we will explain, it's unclear how much difference this makes in ordinary situations, infinity obsession would lead to very strange decisions in some possible circumstances, and would greatly alter the grounds that we invoke to justify claims about whether one prospect is better than another.

In section 6, we show that recklessness poses a problem for creating an acceptable theory of normative or evaluative uncertainty. How should your preferences reflect your uncertainty about which theory of value is correct? Many philosophers have argued in favor of an approach to this and similar questions that relies on expected utility theory. We argue that, under such an approach, agents who have any credence at all in recklessness must themselves be reckless. (This is related to, but in some ways goes beyond, standard worries about extreme theories 'swamping' more moderate ones.) Although we will mention some approaches to normative uncertainty that avoid this problem, there is a conundrum here for philosophers broadly sympathetic to the use of expected utility theory in that context.

1 The General Trilemma

We'll now state more formally the trilemma between timidity, recklessness, and non-transitivity. In doing so, we'll generalize the initial example in two main ways. First, instead of considering extra years of happy life, we can consider other types of goods. Second, the specific numbers used in the example don't matter. The devil ramps up the size of the payoff on the table, while slowly decreasing its probability from certainty down to almost zero. As long as there's *some* way for him to do this that makes each trade look attractive, the paradox remains.

To do this properly, let's fix some terminology. By a *prospect* we mean a situation in which different outcomes can arise with different probabilities.[2] We will only consider prospects in which the possible outcomes are

[2] While we will often speak of *chance* for convenience, we are sympathetic to the view that,

adequately described in terms of a *payoff*: the quantifiable, and in principle arbitrarily large, gain or loss of some type of good relative to a fixed baseline. We will, in fact, mainly consider prospects that can be described as a matter of 'getting payoff x with probability p'—the implication being that one gets nothing (i.e. the baseline outcome) with the remaining probability.[3] Different theories of value will, of course, care about different types of goods. In a prudential context, a payoff might be some number of years of happy life, but we will be especially interested in a moral context, in which a typical payoff might be some number of people benefited in a certain way, or else some number of good lives brought into existence, or perhaps the flourishing of human civilization through some span of time. At any rate, the first two horns of the trilemma are formulated with respect to a given type of payoff and a given baseline.

The first horn, timidity, says that sometimes a slight decrease in the probability of a payoff cannot be outweighed by any increase in the payoff's size. Moreover, it says, this is true no matter how strictly we interpret 'a slight decrease'. To make this precise, let's say that a *standard of closeness* specifies when two positive numbers count as close together—in which case a decrease from the larger to the smaller would count as slight.[4]

> **Timidity:** By any standard of closeness, there's a finite payoff x, and close-together, positive probabilities p and q, such that getting x with the slightly higher probability p is no worse than getting any other finite payoff, no matter how good, with the slightly lower probability q.

strictly speaking, the relevant probabilities are epistemic. The distinction may be especially important if one thinks that chance-propositions can themselves bear value (Stefánsson & Bradley, 2015).

[3] Some views consider it crucial, when comparing prospects, to consider the *states* in which payoffs occur, not just their probabilities (e.g. Colyvan, 2008). We don't deny this. We think, however, that the idea of 'getting x with probability p' is sufficiently determinate for our discussion to make sense without the extra verbosity of explicitly mentioning states; those still worried, or who appreciate greater formalism, should refer to footnote 5.

[4] The only assumptions we will use are (i) closeness is symmetric: if p is close to q then q is close to p; (ii) the numbers that count as close to any given p include an open interval around it.

For example, by one standard, a 0.1% decrease always counts as slight. So a theory that is timid with respect to future years of happy life must say that, for some x, p, and q, living x years with probability p is no worse than living an arbitrarily large number of years with probability q, even though q is only 0.1% smaller than p.

The second horn of the trilemma is

> **Recklessness:** For any finite payoff x, no matter how good, and any positive probability p, no matter how tiny, there's a finite payoff y, such that getting y with probability p is better than getting x for sure.

So, a theory that is reckless with respect to future years of happy life will say that the prospect of living 100 more years for sure is worse than a mere one-in-a-trillion chance of living for some finite (but perhaps truly vast) amount of time—and thus, all but certainly, dying right away.[5]

The third horn, which we will not much discuss, is

> **Non-Transitivity:** There are prospects A, B, and C, such that A is better B, B is better than C, but A is not better than C.

While our focus is officially on theories of the value of prospects, we'll sometimes talk about reckless or timid *agents* as well as theories. These agents form their preferences and choose their options in line with the recommendations of a reckless or timid theory of value. We do this to make recklessness and timidity more vivid; we're not committed to the view that, in general, one ought to choose the best prospect, although we're inclined to think that this is often true.

Let's make sure it's clear why we must choose between timidity, recklessness, and non-transitivity. Start with the prospect of some finite payoff x for

[5] For those who want to track states explicitly (footnote 3), here's a good way to interpret timidity and recklessness. Consider throughout a state space S with a non-atomic probability measure (e.g. the interval $[0,1]$ with the uniform measure). For any event $E \subset S$ and finite payoff x, let x_E be the prospect of getting x on E and nothing otherwise. Timidity can be regimented as the claim that, by any standard of closeness, there's a finite payoff x and two events $E \subset F$, with close-together probabilities, such that x_F is not worse than any y_E. Recklessness can be regimented as the claim that, for any x and E, some y_E is better than x_S. Everything else in this paper can be interpreted along similar lines.

sure. Unless timidity is true, we can find a sequence of prospects, each better than the one before, and each with a slightly smaller probability of getting a payoff. Eventually, we can reduce the probability of the payoff from certainty down to any positive probability p, no matter how small.[6] By transitivity, it would be better to trade the original payoff x for the prospect of some (potentially vast) payoff y with probability p. That's recklessness.

While we'll focus on the conditions just stated, it's worth noting that one can construct an analogous trilemma using negative value. In this version, the payoffs might instead be years of miserable life, rather than years of happy life. A person who is 'timid', in the negative variant, passes up a deal that would give them a much shorter period of suffering but with slightly higher probability. More generally and precisely:

> **Negative timidity:** By any standard of closeness, there's a finite payoff x, and close-together, positive probabilities p and q, such that getting x with the slightly higher probability p is no *better* than getting any other finite payoff, no matter how *bad*, with the slightly lower probability q.

A person who is 'reckless,' in the negative variant, prefers even a very long period of misery for sure to an arbitarily small probability of a sufficiently long period of misery (e.g. preferring 1,000 years of suffering for sure to a one-in-a-trillion chance of suffering for some much longer period). More generally and precisely:

> **Negative recklessness:** For any finite payoff x, no matter how *bad*, and any positive probability p, no matter how tiny, there's a finite payoff y, such that getting y with probability p is *worse* than getting x for sure.

We think that, for many sources of value, negative recklessness and negative timidity are roughly as counterintuitive as the original positive versions, and

[6] One might worry that there is a threshold value p_0 such that a sequence of slight decreases in the probability can bring it arbitrarily close to p_0, but never below it (cf. Binmore & Voorhoeve, 2003). Our formulation of timidity avoids this worry: once the probability is close to p_0 (by the relevant standard), one can slightly reduce it to p_0 in one step, and then slightly below p_0 in another.

they will lead to analogous problems. But, if anything, our arguments will weigh especially heavily against negative timidity.

2 Examples

We now explain how the choice between timidity and recklessness looks from the point of view of some popular approaches to evaluating prospects. We organize our discussion around three ideas that one might invoke to explain timidity, or to avoid recklessness: the boundedness of value, risk aversion, and discounting small probabilities. To be clear, this discussion isn't meant to exhaust the logical space. But it will include the best-credentialed normative theories of evaluation under uncertainty, and it will aid our discussion of the costs of timidity and recklessness.

Throughout this section, we'll focus on prospects that are *simple* in that each one has only a finite number of possible outcomes, all corresponding to finite payoffs. We thereby rule out, to begin with, any difficulties that might arise from infinite payoffs or St Petersburg gambles of the sort we'll describe in sections 4 and 5.

2.1 The boundedness of value

Perhaps the most natural explanation for timidity is that the value of finite payoffs is bounded above. In our initial example, deal n would be worse than deal $n-1$ if the payoff in deal $n-1$ was already so good that it was near the upper limit of how valuable an additional number of happy years could be. Although the *number* of years available in deal n might be much greater, their *value* would not, and therefore wouldn't justify taking even a tiny additional risk. It would be strange to say (on the other hand) that the possibility of a much better outcome could not justify a slight increase in risk.

As we'll now explain, the connection between timidity and the boundedness of value is especially tight if we use *expected value theory* (EVT), the most common normative model for evaluating simple prospects.[7] According

[7] EVT is a special case of expected *utility* theory, to be discussed below. There are two different ways to understand EVT. On one understanding, there are independently given cardinal facts about value, i.e. not only facts about which outcomes are better than which

to EVT, one prospect is better than another just in case it has higher *expected value*. Here, the expected value of a prospect is computed by (i) identifying all of the possible outcomes of the prospect, (ii) multiplying the probability of each outcome by its value, and (iii) adding all those terms together. So the expected value (relative to the relevant baseline) of getting a payoff x with probability p just equals p times the value of getting x for sure.

It is easy to see that, given EVT, recklessness corresponds exactly to the value of payoffs being unbounded above, in the following precise sense:

> For any finite payoff x, and any number $n > 0$, there exists a finite payoff y whose value is at least n times greater than that of x.

Similarly, timidity means that value has an upper bound, in the precise sense that there's some x and some n such that no finite payoff is more than n times better than x. Perfectly analogous arguments show that negative timidity requires a lower bound on the value of finite payoffs, and negative recklessness requires that there is no such lower bound.

Even if EVT is not exactly right, the boundedness of value would be a natural way to avoid recklessness. It provides a clear and, at least, prima facie plausible explanation of why some of the devil's deals might not be worthwhile.

We'll discuss some of the general problems with this idea in section 3, so for now we'll stick to two preliminary points. First, while *perhaps* additional years of happy life have bounded value (we are skeptical, but see Williams (1973) for a classic statement of this view), this is less plausible for other goods. We are especially skeptical of claims that value is bounded *below*, as required for negative timidity: additional years of bad life, for example, do not seem to diminish in their badness. Second, many popular theories of value do not in fact put bounds on the value of various interesting goods. For example,

---

others, but also facts about *how much* better they are; EVT then makes sense relative to a suitable quantitative representation of those facts. But some authors (e.g. Broome, 2004, p. 90) think that the question 'how much better' is at best ambiguous, and regard EVT as providing an implicit definition, or disambiguation, of the cardinal facts. We note that the intuitive explanation of timidity in terms of the boundedness of value itself requires some cardinal facts: we want to be able to say that, at some point, increasing the payoff makes things better but *not by much*.

the most natural understanding of utilitarianism as an evaluative theory says that improving n lives by a given amount improves the world by n times as much as improving one life; so the value of improving lives is unbounded. *Total* utilitarianism and its variants likewise put unbounded value on creating good lives.[8] And utilitarianism is not special here: giving priority to the badly off, say, or introducing further dimensions, like equality or perfection, along which an outcome can be good or bad will not automatically place bounds on value. So there is ample motivation to consider other possible justifications for timidity.

2.2 Risk aversion

The very names 'timidity' and 'recklessness' suggest different attitudes towards risk. If value is not bounded, one might guess that risk aversion is the proper way to account for timidity. As we'll now argue, timidity can indeed be explained by *extreme* levels of risk aversion on *some*, but not all, ways of theorizing about risk attitudes. In the end, the risk aversion seems too extreme to be plausible; that's an issue we'll return to in section 3.

It is not straightforward to give an account of 'risk aversion' in an ordinary intuitive sense. However, according to a standard story, the characteristic feature of risk aversion is that one systematically judges that a sure payoff is better than an uncertain prospect with the same expected size: for example, it's better to save 10 lives for sure than to have a ½ chance of saving 9 lives and a ½ chance of saving 11. (This is, specifically, risk aversion with respect to the *size* of payoffs; we'll also discuss risk aversion with respect to their *value* below.) Risk seeking requires the opposite judgment; risk neutrality means that these prospects are equally good.

We'll illustrate the connection between timidity and extreme risk aversion using what we take to be the two most popular normative theories of risk attitudes.

[8] What's more, some of the best worked-out arguments for total (or critical level) utilitarianism, like that of Broome (2004), specifically rely on expected value theory. One can even argue *for* expected value theory from total utilitarianism, or for expected utility theory with an unbounded utility function, from key utilitarian principles like the Pareto principle and anonymity; see McCarthy *et al.* (2020, especially Theorem 4.10).

First, *expected utility theory* (EUT) assigns a numerical 'utility' to each outcome, and prospects are compared on the basis of their expected utility. The utility of an outcome should depend only on the outcome's value, and better outcomes should have higher utility.[9] Given the similarity between EUT and EVT (we'll explain the contrast shortly) it should come as no surprise that timidity is equivalent to the claim that the utility of finite payoffs is bounded above, and negative timidity to the claim that it is bounded below. The argument is the same as before, with 'utility' instead of 'value'. What does this have to do with risk aversion?

Risk aversion, in this framework, means that payoffs have decreasing marginal utility: the utility function increases less rapidly with each additional unit of payoff. It is, in other words, *concave*. Now, concavity does not entail that the utility function is bounded above, so risk aversion does not entail timidity. But the utility function will be bounded above if it is *very* concave, when it comes to good outcomes—that is, if the contribution to utility of each additional unit of payoff falls off sufficiently quickly. So timidity can be explained by relatively extreme levels of risk aversion when it comes to large payoffs. Risk seeking, on the other hand, means that the utility function is *convex*: it increases more rapidly with each additional unit of value. Negative timidity can be explained by relatively extreme levels of risk seeking when it comes to very negative payoffs.

How is this story different from the one about expected *value* theory in §2.1? First of all, EUT doesn't presuppose that there are any cardinal facts about value (cf. footnote 7). But even if there are such facts, the utility function need not straightforwardly reflect them. In particular, even if the value of finite payoffs has no upper bound, their utility may. This can happen if the utility of positive payoffs is very concave as a function of their value, not just as a function of their size. In that case, getting a positive payoff for sure will be better than any uncertain prospect with the same expected value. This is 'pure' risk aversion, risk aversion with respect to value itself. And similarly, even if the value of finite payoffs has no lower bound, negative timidity can arise through pure risk seeking, when it comes to negative payoffs.

[9] 'Utility' is sometimes used to refer to a person's welfare, as in 'utilitarianism' and 'total utility'; we are using it in a different sense that arises from decision theory.

The EUT account of risk aversion has often been criticized—see Buchak (2013) for an overview. We should therefore also consider the best-defended alternative, *risk-weighted expected utility theory* (REUT), developed by Buchak as a 'subjective' version of the anticipated utility theory of Quiggin (1982). According to this view, each payoff again has a numerical utility, but risk attitudes are captured by a separate *risk function* r, which is used to transform the probabilities. In the case of risk aversion, this transformation is designed to put high weight on the relatively bad outcomes of each prospect; for risk seeking it puts high weight on the relatively good outcomes.[10]

Even if this is an attractive theory of risk attitudes, risk-weighted expected utility theory will not avoid recklessness unless utility is bounded above: it is no different from expected utility theory in that respect. The reason is that the risk-weighted expected utility of getting y with probability p comes out to $r(p)u(y)$ (see footnote 10). This is different from the ordinary expected utility $p\,u(y)$, but it can nonetheless get arbitrarily large unless the utility function u is bounded above. Moreover, this time, the boundedness of the utility function cannot be explained by appeal to risk aversion: risk aversion is already supposed to be handled by the risk function r. A bound on the utility function is more naturally interpreted as a bound on value. So, according to the standard intepretation of this theory, not even extreme levels of risk aversion suffice for timidity.

It may count against timidity that it cannot be understood as risk aversion, according to this prominent theory. There is, however, a relatively natural way to make REUT more extreme, so that it does lead to timidity, even with an unbounded utility function. We'll discuss that tweak under the next heading.

[10] Suppose the risk function is r and the utility function is u. If, according to some prospect P, $P_{>x}$ is the probability of getting an outcome better than x, and $P_{\geq x}$ is the probability of getting an outcome at least as good as x, then the risk-weighted expected utility of P is defined as

$$\sum_x \left[r(P_{\geq x}) - r(P_{>x})\right] \cdot u(x).$$

In the special case where $r(x) = x$, the term in square brackets is just the probability of getting an outcome as good as x, and the whole sum is nothing but the expected utility of P. But, in general, r is only required to be an increasing, real-valued function with $r(0) = 0$ and $r(1) = 1$.

2.3 Discounting small probabilities

A third strategy for avoiding recklessness is *Nicolausian discounting*. On common formulations of this view, one simply ignores outcomes whose probabilities are smaller than some threshold.[11] So, when it comes to recklessness, one will simply end up ignoring the tiny probability of an enormous payoff. Indeed, a key motivation for Nicolausian discounting is precisely to avoid recklessness and similar phenomena. Unfortunately, the common formulation of Nicolausian discounting is fatally flawed. But since *something* in this neighbourhood has often been considered prima facie plausible, we'll take the time to formulate a principle that avoids the most obvious pitfalls.

First, the problems. Consider the (not simple) prospect of getting x years of happy life, where the real number x will be chosen at random somewhere between 0 and 1. Then each specific payoff has probability zero. Ordinary and risk-weighted expected utility theory have ways of dealing with this phenomenon (roughly, one replaces sums with integrals), but applying Nicolausian discounting in this case is completely hopeless (we can't just ignore *every* outcome). Relatedly, whether or not any outcomes fall below the probability threshold will depend on how finely we individuate them. Suppose for the sake of concreteness that the probability threshold is one-in-a-billion. If the devil offers us one year of happy life, should we evaluate the prospect as if it has exactly one payoff with probability 1, or as if it has a trillion slightly different payoffs, all below the probability threshold? Nicolausian discounting, insofar as it makes sense at all, must implausibly claim that this kind of redescription matters a lot.[12] Finally, suppose that one prospect is very similar to another, except that it will turn out much better in certain cases, which individually have probabilities below the threshold (but which may together have probability close to 1). The first prospect is clearly better, but

[11] See Smith (2014) and Monton (2019) for recent proponents of this very old view. Monton coins the name 'Nicolausian discounting' after Nicolaus Bernoulli, who, in a 1714 letter, suggested that 'cases that have a very small probability must be neglected and assumed to be zero' (in Monton's translation). The view is similar to the theory of '*de minimis* risks', for which see Peterson (2002).

[12] See Gustafsson & Torpman (2014) for an attempt to find a canonical individuation of states, in a rather different context.

Nicolausian discounting denies it.[13]

Here's a solution. Instead of giving zero weight to small-probability outcomes, give infinitesimal weight to outcomes that are unsually extreme in value.[14] We will call this type of view *tail discounting*. Here is a way to cash it out. There is some probability threshold ε. A payoff x is in the *left tail* of a given prospect if the probability of getting x-or-worse is less than ε. Similarly, x is in the *right tail* if the probability of getting x-or-better is less than ε. Say that a payoff is *extreme* with respect to the given prospect if it is in the left tail or the right tail. Otherwise, say that x is *normal*. The simplest version of tail discounting is that one prospect is better than another if its expected value, conditional on getting a normal payoff, is higher; in case of ties, use as a tie breaker the expected value conditional on getting an extreme payoff.[15] Roughly, then, the view is that one should ignore the tails of a distribution, except for breaking ties.

Tail discounting is somewhat more defensible than Nicolausian discounting as commonly formulated. It does not suffer from the problems described above. And it can be seen as an extreme form of risk-weighted expected value theory, meaning that defences of the latter can, to some extent, be adapted to defend tail discounting as well.[16] Unfortunately, tail discounting leads to an

[13] Of course, following Monton (2019, §7), one could supplement the basic Nicolausian discounting view by using some form of dominance reasoning to break ties. But this still won't give plausible results in similar cases where the first prospect almost but not quite dominates the second.

[14] Buchak (2013, pp. 73–74) makes (without endorsement) a similar suggestion in the context of St Petersburg gambles, which we'll consider in section 4; but her suggestion is to give extreme outcomes *zero* weight, which still leads to an analogue of the last problem mentioned above. At any rate, the problems we raise throughout this paper for tail discounting *also* apply to this simpler view.

[15] There is a complication about how to treat payoffs x that straddle the edge of the right tail (or similarly the left tail), in the sense that the probability of getting x-or-better is greater than ε, but the probability of getting better-than-x is less than ε. To avoid unnecessary problems, these edge cases should partly but not wholly be included in the tails; this is handled properly by the version of the view stated in footnote 16.

[16] Specifically, it can be formalized as risk-weighted expected value theory with a risk function such that $r(x)$ is infinitesimally close to 0 for $x \leq \varepsilon$, and to 1 for $x \geq 1-\varepsilon$. In detail, for an infinitesimal number ι (e.g. a surreal or hyperreal number), one could define $r(x) = x\iota$ for $x \leq \varepsilon$, $r(x) = 1-(1-x)\iota$ for $x \geq 1-\varepsilon$, and define $r(x)$ to increase linearly for x between ε

especially extreme and implausible form of timidity (§3.3), along with some more general problems to which we'll now turn.

3 The Price of Timidity

We think that timidity is troubling on its face. For example, each deal offered by the devil really does seem better than the one before. But the problems don't stop there. In this section we explain four additional problems faced by timid theories. The first two of these are completely general. The second two arise for all the particular ways of implementing timidity that we considered in section 2. Even if these latter problems may not be completely general, they illustrate the difficulty of constructing a plausible timid theory.

We'll give a brief summary at the end of the section.

3.1 Even small decreases in unlikely payoffs cannot be outweighed

In any instance of timidity, a slight decrease in the probability of a payoff cannot be outweighed by any finite increase in the size of the payoff. One might think, 'This is not *so* implausible, because, after all, the payoff may have been enormous to begin with; a slight decrease in the probability of an enormous payoff is still a weighty matter.' However, it turns out that timid theories are committed to the view that even a *small* decrease in an *unlikely* payoff sometimes cannot be outweighed by any finite increase in the size of a much more likely payoff. We can use this to argue against timidity in the following way.

Consider, for the sake of concreteness, the prospect P of saving 1,000 lives with probability 0.1. A theory that is timid with respect to lives saved might say that P is no worse than the prospect of saving even a vast number of lives with probability 0.099, only one percent smaller. We can imagine that the outcome of P depends on the outcome of a raffle: if the golden ticket is drawn, then a mechanism will be activated that will save 1,000 lives. Now consider a prospect P_1 that differs from P in two ways: first, the mechanism

and $1-\varepsilon$. Because of this connection, tail discounting satisfies the most important axioms of risk-weighted expected utility theory: dominance, comonotonic tradeoff consistency, and the related comonotonic sure thing principle, for which see Buchak (2013).

is designed to save many more lives—let's say 10,000; but second, there is a one-percent failure rate, and in the fail-state, P_1 saves only 999 lives. So P_1 almost certainly leads to ten times as many lives saved as in P, and at worst, in the highly unlikely fail-state, it will save only one life fewer. Intuitively, P_1 is much better than P. But now consider a prospect P_2. In P_2, the mechanism usually saves 100,000, but it only saves 998 in the fail-state. So, compared to P_1, there are almost certainly 10 times as many lives saved, and at worst one fewer. P_2 is clearly better than P_1, so also better than P. But if we continue this sequence along, we get to P_{1000}, in which no lives are saved in the fail-state. Thus P_{1000} leads to a vast number of lives saved (namely, 10^{1003}) with probability 0.099. By transitivity, P_{1000} is better than P. And so, the argument concludes, our theory of value must not be timid in this particular way. Nor is there anything special about this case; *mutatis mutandis*, we have an argument against timidity in general.

3.2 Strange dependence on distant space and time

Timidity requires our evaluation of prospects to depend in strange ways on what happens in distant places and times, causally isolated from the here-and-now. At least, this is true insofar as the types of goods with respect to which we are timid can be realized far away.[17]

The rough idea is easy to see if we suppose that timidity arises from the boundedness of value. How close we are to the upper bound can depend on facts about how things are in remote regions of spacetime. The value of the things we do here and now can therefore depend on those facts in dramatic ways: if we are close to the upper bound, then we can achieve only trivial improvements. But, for example, if we had some way of greatly improving the hundreds of millions of lives currently in extreme global poverty, we could not plausibly claim that this would have only trivial value just because a lot of great things had happened in remote galaxies a long time ago.

[17] Perhaps, then, the point is less telling in the context of prudential value, but we leave this open. The argument to follow is closely related to well-known arguments from 'separability' for totalist views in population ethics (see Broome, 2004; Thomas, 2020). 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.

In general, timidity leads to the same kind of problem in more complicated guise. For illustration, consider the view that the existence of additional lives (of some particular high quality) makes things better. So a payoff here is some number of lives beyond the baseline. Suppose we can create additional lives in our galaxy, but we have no contact with other galaxies, and no influence on how many lives they might contain. It seems obvious that, regardless of the situation in those other galaxies, instead of creating *some* lives with a tiny probability, it would be better in prospect to create a trillion times *more* lives with a trillion times greater probability. According to timidity, though, this is not true.

To see this, consider the following prospects A, B, and C. Relative to the baseline, A leads to the existence of n lives with probability p. B leads to the same number of lives with greater probability $p+q$, and C leads to N additional lives with the original probability p.

		Probabilities		
		p	q	$1-p-q$
A	leads to	n	0	0
B		n	n	0
C		$n+N$	0	0

According to timidity, there will be some such case in which C is not better than B, even though q is less than one trillionth the size of p and N is a trillion times the size of n: the small decrease in probability from $p+q$ to p cannot be outweighed by the enormous increase in the payoff from n to $n+N$. But now suppose that the n lives that would exist with probability p would exist in another galaxy. The choice between A, B, and C has no influence on whether they exist, or on the probability that they exist, or on anything else to do with them. So the practically relevant status quo—what we get if we create *no* lives—is prospect A. Relative to that status quo, prospect B is the prospect of creating only n lives with tiny probability q, while C is the prospect of creating N lives with a much larger probability p. In other words, compared to B, C is the prospect of creating a trillion times more lives with a trillion times greater probability. And yet timidity says that C is not better than B.

It seems strange to us that the importance of a given probability of creating a given number of lives depends in *any* significant way on what might happen, entirely beyond our influence, in far-away times and places. But the specific implications of timidity in cases like the one we just described are especially extreme.

3.3 Extreme risk aversion in very positive outcomes

The next problem is that timidity leads to implausibly extreme forms of risk aversion, at least according to all the timid theories we surveyed in section 2.

Consider first a view according to which there is an upper limit on how good finite payoffs of the relevant sort could be. For example, suppose that 10^{10} extra years of happy life would take you extremely close to the upper limit to how good extra years of life could be. Now consider two prospects, *A* and *B*, in which the outcome depends on the toss of a fair coin:

	Heads	Tails
A	10^{10} years of happy life	1 hour of misery
B	10^{80} years of happy life	2 hours of misery

Since 10^{10} years of happy life would take you extremely close to the upper limit, the additional 70 orders of magnitude of years of happy life would represent a trivial improvement. However, against the baseline of a middling-value life, there is a non-trivial difference between two hours of misery and only one. Therefore, on this approach, we should conclude that *A* is better than *B*. This implausible conclusion can be interpreted as an extreme degree of risk aversion with respect to years of happy life. *B* has a much greater expected number of years of happy life than *A*, but it is still judged worse; this is the characteristic pattern of risk aversion.

This problem seems bound to recur even if value is not bounded. To avoid recklessness, the level of risk aversion must be very high; but then it will seem unacceptably high in other cases. In the case of EUT, we can run the same example as before, supposing that 10^{10} extra years of happy life would bring you very close to the upper bound for utility. We will again conclude that *A* is better than *B*. In the case of REUT, there is a complication: one could, in

principle, avoid this conclusion by counteracting the bounded utility function with a risk function that is extremely risk seeking, so as to achieve something closer to risk neutrality in this particular case. But then the theory will be implausibly risk seeking in more ordinary cases.

Finally, what about tail discounting? Tail discounting has an advantage over the views we have just considered. Our example of extreme risk aversion seems particularly strange because the probabilities involved—those of a fair coin landing heads or tails—are middling in value, quite different from the exotically small probabilities that seem most relevant to the dilemma between timidity and recklessness. In such middling-probability cases, tail discounting will tend to agree with expected value theory, leading to generally plausible results. But it will still lead to strange results when we consider probabilities close to the threshold. Specifically, tail discounting implies the following condition that is both stronger and stranger than ordinary timidity, and which isn't implied by the other timid theories we considered.

> **Threshold timidity:** There is some positive probability threshold such that, for any finite, positive payoffs x and y, getting x with probability below the threshold is never better than getting y with probability above the threshold—no matter how much better x is than y and no matter how close together the two probabilities may be.

So, for example, the prospect of creating 10^{80} lives with some particular positive probability, one trillionth of a percent below the threshold, is no better than the prospect of creating only one life with a probability two trillionths of a percent greater. Roughly speaking, threshold timidity tells us that, close to the threshold, decreasing risk is infinitely more important than increasing expected value.

As one might guess, negative timidity is associated with extreme risk *seeking* for prospects with very bad outcomes. This makes negative timidity especially implausible. For example, the negative variant of threshold timidity says that, close to the threshold, *increasing* risk is infinitely more important than increasing expected value. This does not strike us as a tenable view.

3.4 Long-shots

Recklessness suggests that we should value some extreme long-shot bets in a way that often seems counterintuitive. However, all the timid views we considered in section 2 also recommend certain long-shot bets, so the purported advantages of timidity are undermined.

This is most obvious for tail discounting. Suppose for concreteness that the probability threshold is 0.00001. Then we avoid recklessness, but, if the value of finite payoffs is unbounded, it will still be true that any finite payoff for sure, no matter how good, is worse than a 0.000011 probability of some other finite payoff. But a longshot bet with a 0.000011 probability of a payoff is not really less objectionable than one with a 0.00001 probability. There seems to be no way to set the threshold so that tail discounting rules out all and only the objectionable longshots.[18]

Now here's an example of the type of longshots that are encouraged when value or utility is bounded above.

> **Lingering doubt:** In an alternate possible world, people live in a utopia. Life is extremely good for everyone, society is extremely just, and so on. Their historians offer a reasonably well-documented history where life was similarly good. However, the historians cannot definitively rule out various unlikely conspiracy theories. Perhaps, for instance, some past generation 'cooked the books' in order to shield future generations from knowing the horrors of a past more like the world we live in today.
>
> Against this background, let us evaluate two options: one would modestly benefit everyone alive *if* (as is all but certain) the past was a good one; the other would similarly benefit only a few people, and only if the conspiracy theories happened to be true.

In this case, the second option might yield a better prospect. Why? If the conspiracy theories are not true, then the status quo is already about as good

[18] Perhaps this problem can be mitigated by allowing the probability threshold to be vague, to match the way in which it is vague which longshots are objectionable. See Peterson (2002, p. 53) for this idea in the context of *de minimis* risk; we won't press the issue here, and Thomas (2021) for vagueness as a general response to spectrum arguments.

as it could be, and the benefits to everyone alive would make almost no difference to the value (or utility) of the world. But if the conspiracy theories are true, and the world overall is middling in value or utility, then benefiting relatively few people would make a relatively large difference.

Of course, what's strange about this case is not *only* that it involves seemingly undue attention to a highly unlikely scenario. Contrary to the timid analysis of the case, we feel that modestly benefitting everyone alive would in fact be a weighty consideration, even in the presence of some risk.

3.5 Itemized billing for the timid

To summarize, timid theories have the following features that seem problematic for many types of goods. The first three are general.

1. They pass up at least one of the devil's seemingly great low-risk trades (or similar trades for other types of goods).

2. They must claim that, in some cases, even a small decrease in one unlikely payoff cannot be outweighed by any increase in a much more likely payoff (§3.1).

3. Their ranking of prospects is sensitive in implausible ways to how things are in the far removes of space and time (§3.2).

The other features are had by all the timid theories we have identified, and it is hard to see how we could avoid them; but, for all that, they may not be completely general.

4. They rank some prospects in an extremely risk-averse way; in the case of negative timidity, they rank other prospects in an extremely risk-seeking way (§3.3).

5. Like reckless theories, they still recommend betting on certain kinds of counterintuitive longshots (§3.4).

Analogues of all of these problems arise for negative timidity as well; as we mentioned, the extreme risk seeking associated with negative timidity seems especially implausible.

All together these problems show that justifying timidity in a plausible way is a serious challenge. But recklessness will also have its costs.

4 Recklessness and Dominance

There is some hope that the basic implausibility of recklessness might be open to debunking. It is hard to comprehend payoffs that are getting arbitrarily large, so perhaps intuition is not to be trusted in these cases.[19] And our intuitions might also be confused by all kinds of confounding factors that we would expect to see in practical situations where the stakes are very high and one's evaluation of prospects is sensitive to miniscule absolute changes in probabilities: it's natural to start worrying about the reliability of one's evidence and one's cognitive facilities, one's powers of introspection, whether one is being tricked or one is simply hallucinating, and so on. Perhaps in the unusually simple and clearly specified cases that recklessness strictly speaking involves, 'reckless' behavior really would be reasonable.

Still, the case against recklessness can be pressed, and that's what we'll do in this section and the next. Here, we'll argue that recklessness leads to violations of a very compelling dominance principle. Next, we'll argue that (under some further assumptions) recklessness leads to the single-minded pursuit of infinite payoffs.

In the context of expected utility theory, we know that recklessness corresponds to the use of an unbounded utility function. And there is a large literature that focuses on the problems that arise from an unbounded utility function, centered around the *St Petersburg gamble*. It turns out that at least some of these problems arise directly from recklessness, given background assumptions that are much weaker than the assumption of expected utility theory. Given that the alternative is timidity, the problems are deeper and more general than usually supposed. That's what we'll now explain.

Recall that the St Petersburg gamble, in its modern guise, offers a ½ chance of 2 units of utility, a ¼ chance of 4 units, a ⅛ chance of 8 units, and in general a $1/2^n$ chance of 2^n units. The expected utility of this prospect is infinite.[20] Some of the problems associated with the St Petersburg gam-

[19] See Baron & Greene (1996) for enlightening examples about how pre-theoretic intuition fares very badly with respect to evaluating alternatives involving large numbers of people. See Gustafsson (2021) for a forthcoming overview in the context of population ethics.

[20] The original St Petersburg gamble had monetary payoffs, and so infinite expected *monetary* value. The observation that one can create St Petersburg gambles with infinite expected utility, as long as the utility function is unbounded, goes back at least to Menger (1934). To

ble involve interpreting this infinity in a sensible way. As but one example, the 'Petrograd' gamble (Colyvan, 2008) is like St Petersburg except that it invariably pays off one extra unit of utility. It is presumably better than St Petersburg, but it apparently has 'the same' infinite expected utility. The more fundamental problem, however, is that the St Petersburg gamble leads to violations of some extremely attractive normative principles which are often themselves assumed in axiomatic approaches to expected utility theory. Specifically, consider the following two principles:

> **Outcome-outcome dominance:** If, no matter what, the outcome of A would be at least as good as the outcome of B, then A is at least as good as B.[21]
>
> **Prospect-outcome dominance:** If prospect A is strictly better than each possible outcome of prospect B, then A is strictly better than B.

The St Petersburg gamble violates prospect-outcome dominance: it is strictly better than each of its own outcomes, but it is not strictly better than itself.[22] Some care is appropriate here. One might think that St Petersburg is better than each of its outcomes *because* it has infinite expected utility, while each outcome has only finite utility. But this may be too quick: arguably, EUT only applies to prospects insofar as they have finite expected utility. A better argument goes like this. By outcome-outcome dominance, the St Petersburg gamble is at least as good as a version that caps the payoff at 2^n units of utility.

---

emphasize: our own argument below in no way assumes expected utility theory.

[21] This principle is often called 'statewise dominance': the antecedent is that, whichever state the world is in, A results in an outcome that is at least as good as that of B. Some decision-theoretic frameworks do not formally represent prospects using states, but it should still be possible informally to specify prospects by their effects in different states, and to assert outcome-outcome dominance for prospects so specified. Alternatively, one can replace outcome-outcome dominance in the following discussion by the notionally stronger but also very popular condition known as 'stochastic dominance' that does not refer to states.

[22] See Chalmers (2002) for a version of this problem, and Hammond (1998) for the use of similar dominance principles to rule out unbounded utility functions. The strangeness of violating prospect-outcome dominance may be more vivid if we consider two independent St Petersburg gambles, A and B. No matter how B turns out, one would trade the result for A. Yet A is not strictly better than B.

This 'capped' version, to spell it out, offers a ½ chance of 2 units of utility, a ¼ chance of 4 units, a ⅛ chance of 8 units, and so on, up to a $1/2^{n-1}$ chance of 2^{n-1} units, and offers 2^n units with the remaining probability. But (one can check) this capped version has expected utility greater than n; it is better than getting n units of utility for sure. So, by transitivity, the St Petersburg gamble is strictly better than getting n units for sure, for every natural number n. Therefore, as promised, the St Petersburg gamble is strictly better than every one of its own possible outcomes.

Since the St Petersburg gamble is specified in terms of utility, the problem described so far still arguably depends on expected utility theory, or on some other theory that gives meaning to the utility scale. Unfortunately, it generalizes to all reckless views. For, assuming recklessness, we can construct a 'generalized St Petersburg gamble' in the following way. Take any payoff x_0. By recklessness, there is a payoff x_1 such that a ½ chance of x_1 is better than x_0 for sure. By recklessness, there is an payoff x_2 such that a ¼ chance of x_2 is better than x_1 for sure. And so on, obtaining a sequence of payoffs $x_1, x_2, x_3, \ldots$, such that a $1/2^n$ chance of x_n is better than x_{n-1} for sure. Now define the generalized St Petersburg gamble A to be one that offers a $1/2^n$ chance of x_n, for every $n = 1, 2, 3, \ldots$. For each n, outcome-outcome dominance entails that A is at least as good as simply getting the same $1/2^n$ chance of x_n and nothing otherwise. And by construction this is better than getting x_{n-1} for sure. So, again, A is better than every one of its possible outcomes, violating prospect-outcome dominance.

We think that prospect-outcome dominance and (especially) outcome-outcome dominance are both compelling principles. It's hard to fathom why they should be violated in the clean kinds of cases we have in mind, where the value of outcomes is apparently the only thing at stake and there are no violations of rights or objectionable unfairness or other complicating factors. Moreover, the timid theories with bounded value or utility functions that we discussed in section 2 are immune to these problems. This is a serious strike against recklessness.[23]

[23] What about views that discount small probabilities (§2.3)? While such views are immune to the recklessness-based argument we just gave, the problem still arises. To see this in the context of tail discounting, consider a prospect in which one will either, with probability below the threshold, face a St Petersburg gamble, or, with the remaining probability,

5 Recklessness and Infinity Obsession

Now we introduce a problem that arises for reckless theories when there is a possibility of an infinite payoff. Although theorizing about infinite cases in ethics is notoriously difficult (see Bostrom (2011) for an entry to the literature), the timid approaches we surveyed in section 2 don't lead to the particular problem we identify here. That speaks in favour of timidity, and against recklessness.

While in philosophy and in life we usually don't think about the possibility of achieving infinite payoffs, some philosophers, such as Pascal, claim that such infinite considerations should be decisive, at least in theory.[24] More precisely, these people claim:

> **Infinity obsession:** Any non-zero probability, no matter how small, of an infinite payoff, is better than any finite payoff for sure.[25]

Throughout this discussion, by 'an infinite payoff' we will mean specifically a *positively* infinite payoff, like an infinite number of years of happy life, or an infinite number of lives saved, or something equally good. For example, in Pascal's case, infinity obsession arises from the claim that an arbitrarily small chance of eternity in heaven is better than any finite reward. But of course the negatively reckless will face similar problems with respect to negatively infinite payoffs.

In this section, we argue that, given minimal assumptions, reckless agents *must* be infinity-obsessed (§5.1), and that, given slightly stronger assump-

get a zero payoff. This will lead to a violation of prospect-outcome dominance. On the other hand, simply ignoring the tails of the distribution (as in footnote 14) would violate outcome-outcome dominance. We do not know whether there is a reasonable way for probability discounters to thread the needle here—it's yet another challenge for advocates of these views.

[24] See Hájek (2003) for an insightful discussion of Pascal's wager from the point of view of modern decision theory.

[25] We continue with the standard assumption that probabilities are real numbers, and so exclude infinitesimal probabilities. We thus set aside the natural question of what to do about infinitesimal probabilities of infinite payoffs. Infinity obsession involves a violation of the standard 'continuity' or 'Archimedean' axiom used by expected utility theory and many other decision theories.

tions, recklessness leads to other, potentially even stranger forms of obsession (§5.2). Finally, we explain why these forms of obsession remain troubling even if it's not entirely clear to what extent they would affect our judgments about cases of immediate practical relevance (§5.3).

First, though, let us make sure the differences between recklessness and infinity obsession are quite clear. Of course, only the latter involves infinite payoffs, but there is another difference too. Though the reckless are willing to take some extreme risks for finite rewards, their willingness to take a risk for any given reward depends on the probability involved. When it comes to infinite rewards, the infinity-obsessed have no such concern. *Any* non-zero probability of an infinite reward seems better to them than any finite reward for sure. To emphasize how strange this is: in any instance of recklessness, there's some conceivable evidence you could show the reckless agent, short of definitively proving, with probability 1, that his longshot won't pay off, that will make him give up his interest in it. Not so for the infinity-obsessed; unless you definitively prove, with probability 1, that his infinite longshot won't pay off, he'll prefer to take his chances.

5.1 How recklessness leads to infinity obsession

Despite their differences, recklessness leads to infinity obsession, given only very weak assumptions. For example, suppose we have an agent who is reckless about the number of happy years of life he has. Compared to getting any finite number of years of happy life for sure, he'd prefer even a tiny chance of living for a sufficiently long but finite time. *A fortiori*, he'd prefer the same tiny chance of living for infinitely long. So he'd prefer any tiny chance of living forever to any finite number of years of happy life for sure. In other words, he'll be infinity-obsessed.

More formally, the premises we need to derive infinity obsession are recklessness, transitivity, outcome-outcome dominance, and the claim that an infinite payoff is better than a finite one.[26] The argument goes like this. Con-

[26]It's true that, in section 4, we argued that recklessness requires one to deny either outcome-outcome dominance or prospect-outcome dominance. We think it would be quite remarkable to reject the particular application of outcome-outcome dominance that follows. Be that as it may, the logical point is that if one rejects *only* the slightly less compelling

sider a prospect A that gives the agent an infinite payoff with probability $p > 0$, let B be a prospect that gives the agent some finite payoff y with that same probability p, and let C be a prospect that gives the agent some other finite payoff x for sure. We have to show that A is better than C. According to recklessness, there's some value of y that would make B better than C. But an infinite payoff is better than a finite payoff, so, by outcome-outcome dominance, A is at least as good as B. Therefore, by transitivity, A is better than C.

5.2 Other forms of obsession

We've formulated infinity obsession in a way that ties it very closely to recklessness, as the preceding argument shows. But slightly stronger assumptions lead from recklessness to other forms of obsession, which may be even more troubling. To illustrate, suppose we accept

> **The sure thing principle:** If, on the supposition that some proposition E is true, prospect A is at least as good as B, and, on the supposition that E is false, A is better than B, then A is better than B.

Although the sure thing principle in its full generality is closely associated with expected utility theory, the particular applications we will make of it are relatively uncontroversial; for example, they are compatible with the risk-weighted expected utility theory we described in section 2.[27] Here are two further types of obsession that will ensnare the reckless if they accept the sure thing principle.

First, infinity obsession compares a prospect with *some* chance of resulting in an infinite payoff to a prospect with *no* such chance: it says the former is always better. The following condition suggests, more generally, that a *higher* chance of an infinite payoff is always better.

---

condition of prospect-outcome dominance, then one still falls prey to infinity obsession.

[27] That is, we will at most use a simple case of the *comonotonic* sure thing principle discussed by Buchak (2013). We won't go into the details, since one can judge our application of the principle on its own terms.

> **Generalized infinity obsession:** Getting an infinite payoff with some probability p and nothing otherwise is better than getting the same infinite payoff with any smaller probability q and a finite payoff x otherwise—no matter how good x may be.

A bit roughly, someone who is infinity-obsessed in this generalized sense is obsessed with increasing the probability of infinite payoffs, completely heedless of finite considerations.[28]

Although generalized infinity obsession is stronger than plain old infinity obsession, one can argue for the stronger claim from the weaker. Suppose that A is the prospect of getting the infinite payoff with probability p and nothing otherwise, and B is the prospect of getting the infinite payoff with probability q and x otherwise. First, it seems harmless to assume that, p being higher than q, A results in the infinite payoff whenever B does. (If one *did* resist this step, the conclusion would be only slightly weaker.) Now suppose that A and B both result in the infinite payoff. Then they turn out equally well; on that supposition, A and B are equally good. On the contrary supposition, B results in the finite payoff x, and A either results in the infinite payoff or in nothing. By infinity obsession, A is better than B, on this supposition. Therefore, by the sure thing principle, A is better than B.

Second, here is a troubling kind of obsession that involves only finite payoffs. In section 4, we showed how to construct a 'generalized St Petersburg gamble' using recklessness and only finite payoffs. An argument very similar to the one for infinity obsession, but using the sure thing principle in place of outcome-outcome dominance, shows that reckless theories are likely to be 'St Petersburg-obsessed':

> **St Petersburg obsession:** Any non-zero probability, no matter how small, of facing a generalized St Petersburg gamble, is better than any finite payoff for sure.

We'll omit the argument for the sake of space.[29]

[28] Hájek (2003) entertains the view that, contrary to generalized infinity obsession, all prospects involving infinite payoffs are equally good. But this can't be right if one accepts outcome-outcome dominance in the relevant cases and thinks an infinite payoff is better than a finite one.

[29] In the context of expected utility theory, the point that an unbounded utility function

Of course, one can try to argue in similar ways for forms of obsession that apply in a wider range of cases, using similar or stronger dominance principles. The intention of our current arguments is simply to show that there are a range of troubling obsessions that a reckless theory will have difficulty avoiding in a plausible and principled way.

5.3 What are the implications of infinity obsession?

Let us explain why we find infinity obsession and its kin so troubling.

One might worry that, in directing all our attention to infinite payoffs, infinity obsession must have implausibly revisionary practical consequences. We find this somewhat unclear, given our actual situation. To be sure, we do not think that infinite payoffs are entirely off the table. As Bostrom (2011) and others point out, many astrophysicists now believe that the universe is infinite, and this would make it likely that there are infinitely many valuable lives. In turn, it is hard to be absolutely certain that one will not happen to achieve an infinite payoff in terms of lives improved, or perhaps in terms of good lives created; in the context of generalized infinity obsession, it seems hard to claim that all of our options have *exactly* the same probability of generating an infinite payoff.[30] So infinite payoffs may well be practically relevant, but there are at least two complicating factors.

First, one shouldn't forget that our formulation of infinity obsession only involves a class of very simple prospects (each involving at most two outcomes), and realistic cases will obviously be much more complicated. We haven't, moreover, considered how to weigh chances of positive infinities against chances of negative infinities; how infinite payoffs compare with generalized St Petersburg gambles; and so on. Infinity obsession answers *one*

leads to violations of the continuity axiom through St Petersburg gambles goes back at least to Arrow (1965).

[30] See Greaves (2016) for an attempt to defend an indifference principle that might lead to such a conclusion in some, but certainly not all, cases. A different thought is that the epistemic probability of achieving an infinite payoff will usually be 'imprecise'. Then the theory of evaluation and decision-making with imprecise probabilities will come to the fore. Here a serious worry is that imprecision about the probability of an infinite payoff will tend to mean that almost any option is permissible—see Mogensen (2020) for a version of this worry for high-stakes finitary cases.

theoretical question about infinite payoffs, but it leaves open many important questions of this kind.

Second, we are somewhat attracted to what Bostrom calls an an *empirical stabilizing assumption*—an empirical assumption which, if true, means that what's best conditional on achieving a finite payoff and what's best overall are broadly equivalent. Specifically, one might argue that whatever best promotes the long-term survival and flourishing of our descendants in the distant future will generally be what's best by either standard. It will do well with respect to finite payoffs because of the enormous (even if finite) number of potentially good lives in the future.[31] And it will do well overall because flourishing future civilisations will be relatively well-placed to achieve infinite payoffs, if it's possible at all. Indeed, it's unclear how we could realistically promote infinite payoffs without enabling our descendants in more general ways.[32] While there is much more to think about here, it does seem that, even if recklessness is true, there might be significant overlap between what's best with respect to finitary considerations and what's best overall, in our current situation.

However, even if some empirical stabilizing assumption meant that taking infinite payoffs into account would not greatly change our evaluation of the prospects open to us, it *would* significantly change the reasons behind these evaluations, and in strange ways. For someone who suffers from generalized infinity obsession, the main reason that it would be good to prevent climate change might be that it would increase the chance that future people find some way of achieving an infinite amount of value, rather than the fact that lots of future generations would be better off in some more likely, merely finite way. Prudentially, the main reason that it would be important to avoid smoking might be that it increases your odds of living long enough that someone discovers a way to give you infinite value; 'lung cancer is an awful experience' would be a relatively minor consideration. Infinity obsession would remain revisionary in this way.

Finally, there are situations that might arise in the distant future in which it is clear that no empirical stabilizing assumption would apply. In some of

[31] Parfit (2011, chapter 36), Bostrom (2003, 2013) and Beckstead (2019) all argue for this conclusion.

[32] For a similar line of thought, more fully developed, see Williams (2013).

those cases, infinity obsession still seems deeply troubling. Here is a stylized illustration.

> **Infinite research vs. utopia:** Our descendants reach the limits of technological progress and become very confident, but not certain, that it's impossible to achieve an infinite payoff. (And, indeed, it is impossible). They must decide how some vast amount of resources should be allocated between two projects: creating an extremely good (but only finite) utopia, or researching possible methods of achieving an infinite payoff.

It is likely that if these people were infinity-obsessed, they would spend nearly all of the resources on the infinite research; they would keep becoming more and more certain that their research would bear no fruit; and they would keep going at it as long as they didn't become *completely* certain that achieving an infinitely good outcome was impossible—which perhaps they never would.

6 Recklessness and Moral Uncertainty

Some philosophers have argued that we should treat uncertainty about moral or evaluative matters in essentially the same way that we should treat empirical uncertainty, and specifically that we should use expected utility theory in both cases.[33] After all, the axiomatic basis of expected utility theory is *prima facie* plausible regardless of what type of uncertainty is at stake. In this section, we raise an objection to this type of view, based on the possibility of reckless and infinity-obsessed theories of value. We then comment on how this fits into the general challenge of constructing an adequate theory of evaluative uncertainty.

Let us step back for a moment to make sure the topic is clear. In this context, a natural way to understand the question is: how should your preferences reflect your credences about value? Here is an example.

> Suppose that you have some credence in utilitarianism and some credence in a pluralistic form of egalitarianism, as theories of

[33] For early examples, see Lockhart (2000), Ross (2006), Sepielli (2010); see MacAskill *et al.* (2020) for a recent book-length treatment, and Riedener (2020) for a version of the axiomatic approach mentioned in the next sentence.

> value. Should you then prefer a more equal outcome to a less equal one when egalitarianism says that it is better but utilitarianism says that they are equally good?

For our purposes, one can understand the 'should' as a matter of coherence between your beliefs and your desires. The question is whether you could coherently have the credences described while also (for example) being entirely indifferent about equality. A theory of evaluative uncertainty would answer this question and others like it.

Here is the problem. Suppose that each of the theories T_i in which you have some credence ranks simple prospects by expected utility—each one with respect to a different utility function u_i. Suppose your preferences also rank simple prospects by expected utility, with respect to yet another utility function u. Finally, suppose that your preferences over simple prospects satisfy the following

> **Pareto assumption:** If every theory T_i in which you have some credence says that A is at least as good as B, then you weakly prefer A to B (that is, you either strictly prefer A to B, or you are indifferent between them).
>
> If, in addition, one of the T_i says that A is better than B, then you strictly prefer A to B.

(So, in the preceding example, you *do* prefer the more equal outcome to the less equal one.) Then, one can show that, if any of the T_i are reckless, your preferences must be reckless as well.[34] Indeed, it follows from Harsanyi's

[34]Two subtleties. (1) The Pareto assumption is most plausible when empirical and evaluative uncertainty are independent, so that the probabilities encoded in each prospect are the same conditional on the truth of each T_i. Thanks to Tomi Francis for pointing this out. (2) We're continuing to assume that all the theories under consideration recognize that payoffs of the relevant kind are good to some extent (or at least limited in how bad they are). This avoids some strange and exceptional cases where different theories perfectly cancel one another out. For example, if total utilitarianism and the opposite of total utilitarianism were the only theories in which you had credence, and you gave them exactly the same weight, then your preferences would not be reckless; you would be indifferent about everything. But even without this assumption, it would take a remarkable coincidence for different theories to cancel out exactly in a way that avoids recklessness.

aggregation theorem that u is a weighted sum of the u_i:[35]

$$u = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n \tag{1}$$

for some positive real numbers α_i. If any one of the T_i is reckless, or equivalently if any one of the u_i is unbounded, then u must be unbounded as well.

Despite the drawbacks of recklessness, it is hard to deny that we should have *some* credence in reckless theories of value, given the challenges faced by the alternatives, and given that, on many standard theories, the value of finite payoffs is unbounded. It is therefore hard to deny that, if empirical and evaluative uncertainty are both governed by expected utility theory, then one's preferences should be reckless. Moreover, since advocates of expected utility theory typically endorse outcome-outcome dominance and the sure thing principle as core commitments, we seem bound to fall into generalized infinity obsession and St Petersburg obsession as well.

This could reasonably be seen as an objection to the use of expected utility theory in this context. It is true, after all, that some quite different approaches to evaluative uncertainty will avoid the problem altogether. Gustafsson & Torpman (2014) defend 'My Favourite Theory', which in this context would be the view that your preferences should match the judgments of the evaluative theory in which you have highest credence; they reject both expected utility theory and the Pareto condition. And normative externalists, like Weatherson (2019), might claim that there simply aren't interesting norms in this area (let alone ones that would require reckless preferences). So our argument could be taken to lend support to such alternative approaches, or simply as raising an important problem for those attracted to expected utility theory and similar views.

This argument is closely related to, but importantly different from, a common observation in the literature. Suppose there were a universal notion of 'moral value' such that each theory T_i assigned a degree $v_i(x)$ of moral value

[35] See Harsanyi (1955, 1977, §4.8). In the traditional interpretation of Harsanyi's framework, each u_i is a utility function representing the preferences of some person, and u is a utility function representing the preferences of a social planner. The result is that the social planner's utility function should be a weighted sum of those of individuals. The idea of using Harsanyi's theorem in the context of moral uncertainty is found in Beckstead (2013) and independently in Riedener (2015, 2020).

to each outcome x. We could then make intertheoretic comparisons of moral value, i.e. we could ask whether T_i gives greater moral value to x than T_j does.[36] And suppose you formed your preferences in line with expected moral value, i.e.

$$u = \mathrm{Cr}(T_1)v_1 + \mathrm{Cr}(T_2)v_2 + \cdots + \mathrm{Cr}(T_n)v_n \qquad (2)$$

where $\mathrm{Cr}(T_i)$ is your credence in T_i; the similarity between (1) and (2) should be clear. Then the common observation is that your preferences will tend to be determined by high-stakes theories, i.e. ones that claim there are very large differences in moral value between relevant outcomes; moreover, this can be true even if you have only tiny credence in those high-stakes theories. The low-stakes theories are 'swamped' by the high-stakes ones. Again, this may seem objectionable.[37]

Because our argument focusses on the specific phenomenon of recklessness, it is able to rely on a much more minimal view about evaluative uncertainty. First, we don't need to say anything about any cardinal notion of 'moral value' or 'high stakes'; our use of utility functions is compatible with the view that utility is just a formal device based on the structural axioms of expected utility theory, such as the sure thing principle. Second, relatedly, we don't need to say anything about the meaningfulness of any sort of intertheoretic comparisons, a deep problem for moral uncertainty that goes back at least to Hudson (1989).[38] And, finally, far from *assuming* that the overall

[36] Strictly speaking, in this discussion we only need to make intertheoretic comparisons between *differences* in moral value, but we set this aside to simplify the exposition.

[37] See e.g. Ross (2006), Greaves & Ord (2017), MacAskill *et al.* (2020, ch. 6) for discussions of this problem. We note that the recent literature on normative uncertainty is highly sophisticated, and the comments in the next paragraph are not intended as objections.

[38] It might seem that the agent's preferences must end up reflecting an implicit view about intertheoretic comparisons; what can this view amount to if such comparisons are meaningless? Perhaps we should understand it in purely pragmatic terms—that is, not as an implicit view *about* intertheoretic comparisons, but a practical stance towards reconciling competing demands. See Riedener (2020) for more on this issue. At any rate, the logical point is that, insofar as you do, on whatever basis, satisfy the stated conditions, your preferences will be reckless. One could also try to generalize the argument to allow for *incomparability* in the agent's preferences, contrary to the most orthodox version of expected utility theory. We note that some generalizations of Harsanyi's theorem allow for incomparability (see McCarthy *et al.*, 2019), but if incomparability is sufficiently widespread (as one might guess if

utility function u is a weighted sum of the u_i, in analogy to (2), we derive it, via Harsanyi's theorem, from a small number of more basic principles.

7 Conclusion

In summary, as far as the evaluation of prospects goes, we must be willing to pass up finite but arbitrarily great gains to prevent a small increase in risk (timidity), be willing to risk arbitrarily great gains at arbitrarily long odds for the sake of enormous potential (recklessness), or be willing to rank prospects in a non-transitive way. All options seem deeply unpalatable, so we are left with a paradox. Of course, as we argued in section 6, the paradox may effectively resolve itself: on some views about normative uncertainty, we are bound to have reckless preferences, even if our credence in reckless theories is vanishingly small. But this result is unpalatable in itself.

If we accept timidity, we must think that sometimes even a small reduction in an unlikely payoff cannot be outweighed by a sufficiently large increase in a much more likely one. We must care about seemingly irrelevant details involving the distant reaches of space and time. We are likely to endorse extreme risk aversion in some cases and, when it comes to negative timidity, extreme risk seeking in others. And, for all that, it is not clear we can avoid favouring objectionably long-shot bets.

If we accept recklessness, then we must deny some compelling dominance principles, like prospect-outcome dominance. And we are likely to be infinity-obsessed, pursuing any chance of an infinite payoff, or even any chance of a generalized St Petersburg gamble, at any finite expense.

Some may see all this as another argument for ranking prospects non-transitively. We are not inclined toward this resolution, but we think that increasing one's confidence in this position is a fair reaction to the arguments, given that new challenges have been presented for other approaches, but not for this one.

---

intertheoretic comparisons make *no sense at all*) then the argument for recklessess will fail; see MacAskill (2013) for a discussion of similar issues.

References

Arrow, Kenneth. 1965. *Aspects of the Theory of Risk-Bearing*. Helsinki: Yrjö Jahnssonin Säätiö.

Baron, Jonathan, & Greene, Joshua. 1996. Determinants of insensitivity to quantity in valuation of public goods: Contribution, warm glow, budget constraints, availability, and prominence. *Journal of Experimental Psychology: Applied*, **2**(2), 107–125.

Beckstead, Nicholas. 2013. *On the overwhelming importance of shaping the far future*. Ph.D. thesis, Rutgers.

Beckstead, Nick. 2019. A brief argument for the overwhelming importance of shaping the far future. *Pages 80–99 of:* Greaves, Hilary, & Pummer, Theron (eds), *Effective Altruism: Philosophical Issues*. Oxford University Press.

Binmore, Ken, & Voorhoeve, Alex. 2003. Defending transitivity against Zeno's paradox. *Philosophy & Public Affairs*, **31**(3), 272–279.

Bostrom, Nick. 2003. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, **15**(3), 308–314.

Bostrom, Nick. 2011. Infinite ethics. *Analysis and Metaphysics*, **10**, 9–59.

Bostrom, Nick. 2013. Existential risk prevention as global priority. *Global Policy*, **4**(1), 15–31.

Broome, John. 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), 155–157.

Colyvan, Mark. 2008. Relative expectation theory. *Journal of Philosophy*, **105**(1), 37–44.

Greaves, Hilary. 2016. Cluelessness. *Proceedings of the Aristotelian Society*, **116**(3), 311–339.

Greaves, Hilary, & Ord, Toby. 2017. Moral uncertainty about population axiology. *Journal of Ethics and Social Philosophy*, **12**(2), 135–67.

Gustafsson, Johan E. 2021. Our intuitive grasp of the Repugnant Conclusion. *In: Oxford Handbook of Population Ethics*. Oxford University Press. Forthcoming. [MS available at johanegustafsson.net].

Gustafsson, Johan E., & Torpman, Olle. 2014. In defence of my favourite theory. *Pacific Philosophical Quarterly*, **95**(2), 159–174.

Hammond, Peter J. 1998. Objective expected utility: A consequentialist perspective. *Pages 143–211 of:* Baberà, Salvador, Hammond, Peter J., & Seidl, Christian (eds), *Handbook of Utility Theory, Volume 1*. Kluwer.

Harsanyi, John C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, **63**(4), 309–321.

Harsanyi, John C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press.

Hudson, James L. 1989. Subjectivization in ethics. *American Philosophical Quarterly*, **26**(3), 221–229.

Hájek, Alan. 2003. Waging war on Pascal's wager. *The Philosophical Review*, **112**(1), 27–56.

Lockhart, Ted. 2000. *Moral Uncertainty and Its Consequences*. Oxford University Press.

MacAskill, William. 2013. The infectiousness of nihilism. *Ethics*, **123**(3), 508–520.

MacAskill, William, Bykvist, Krister, & Ord, Toby. 2020. *Moral Uncertainty*. Oxford University Press.

McCarthy, David, Mikkola, Kalle, & Thomas, Teruji. 2019. *Aggregation for potentially infinite populations without continuity or completeness*. Manuscript.

McCarthy, David, Mikkola, Kalle, & Thomas, Teruji. 2020. Utilitarianism with and without expected utility. *Journal of Mathematical Economics*, **87**, 77–113.

Menger, Karl. 1934. Das Unisherheitsmoment in der Wertlehre. *Z. Nationalökonomie*, **51**, 459–485. Translated as 'The role of uncertainty in economics', in Shubik, M. (ed.), *Essays in Mathematical Economics in Honor of Oskar Morgenstern*, Princeton University Press, 1967.

Mogensen, Andreas. 2020. Maximal cluelessness. *The Philosophical Quarterly*, **71**(1), 141–162.

Monton, Bradley. 2019. How to avoid maximizing expected utility. *Philosophers' Imprint*, **19**(18), 1–25.

Parfit, Derek. 2011. *On What Matters: Volume Two*. Oxford: Oxford University Press.

Peterson, Martin. 2002. What is a *de Minimis* risk? *Risk Management*, **2**(4), 47–55.

Quiggin, John. 1982. A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 323–343.

Riedener, Stefan. 2015. *Maximising expected value under axiological uncertainty: An axiomatic approach*. Ph.D. thesis, Oxford University.

Riedener, Stefan. 2020. An axiomatic approach to axiological uncertainty. *Philosophical Studies*, **177**, 483–504.

Ross, Jacob. 2006. Rejecting ethical deflationism. *Ethics*, **116**(4), 742–768.

Sepielli, Andrew. 2010. *Along an Imperfectly-Lighted Path: Practical Rationality and Normative Uncertainty*. PhD Thesis, Rutgers University.

Smith, Nicholas J. J. 2014. Is evaluative compositionality a requirement of rationality? *Mind*, **123**(490), 457–502.

Stefánsson, H. Orri, & Bradley, Richard. 2015. How valuable are chances? *Philosophy of Science*, **82**(4), 602–625.

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

Thomas, Teruji. 2020. Separability in population ethics. *In: The Oxford Handbook of Population Ethics*. Oxford University Press. Forthcoming.

Thomas, Teruji. 2021. Are spectrum arguments defused by vagueness? *Australasian Journal of Philosophy*. doi: 10.1080/00048402.2021.1920622.

Weatherson, Brian. 2019. *Normative Externalism*. Oxford University Press.

Williams, Bernard. 1973. The Makropulos Case: Reflections on the tedium of immortality. *Pages 82–100 of: Problems of the Self: Philosophical Papers 1956-1972*. Cambridge University Press.

Williams, Evan G. 2013. Promoting value as such. *Philosophy and Phenomenological Research*, **87**(2), 392–416.