The unexpected value of the future

Hayden Wilkinson (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2022

GPI Working Paper No . 17-2022

The unexpected value of the future*

Hayden Wilkinson

Last updated: May, 2022

Comments welcome: hayden.wilkinson@philosophy.ox.ac.uk

Abstract

Consider *longtermism*: the view that the morally best options available to us, in many important practical decisions, are those that provide the greatest improvements in the (*ex ante*) value of the far future. Many (but not all) who accept longtermism do so because they accept an impartial, aggregative theory of moral betterness in conjunction with expected value theory. But such a combination of views results in absurdity if the (impartial, aggregated) value of humanity's future is *undefined*—if, e.g., the probability distribution over possible values of the future resembles the Pasadena game, or a Cauchy distribution. In this paper, I argue that our evidence supports such a probability distribution—indeed, a distribution that cannot be evaluated even by extensions of expected value theory that have so far been proposed. I propose a new method of extending expected value theory, which allows us to deal with this distribution and to salvage the case for longtermism. I also consider how risk-averse decision theories might deal with such a case, and offer a surprising argument in favour of risk aversion in moral decision-making.

Keywords: *Pasadena game; expected value theory; expected utility theory; longtermism; risk aversion; relative expectation theory; principal value theory.*

*I am grateful to Adam Bales, Jacob Barrett, Harvey Lederman, Andreas Mogensen, Toby Newberry, Jeffrey Russell, Christian Tarsney, Johanna Thoma, Teru Thomas, David Thorstad, Timothy L. Williamson, and two anonymous reviewers for their generous comments on various versions of this paper. For helpful discussion, I also thank Tomi Francis and an audience at the 9th Oxford Workshop on Global Priorities Research. And, for his extensive assistance with the mathematical details in §2.2, I am grateful to Alex Barry.

1 Introduction

If an agent wishes to choose the morally best options available to them, one strategy they might take is to choose options that most improve the long-term future. This is the strategy recommended (in at least some situations) by *longtermism*: the view that the best options available to us, at least in many important practical decisions, are those that most increase the *ex ante* moral value of the far future (Greaves and MacAskill, 2021, p. 3).[1]

Is longtermism true? One reason to think so is that it seems to follow from the conjunction of several highly plausible moral claims, combined with some empirical observations. (Note that this is far from the only possible justification for longtermism—many different views of moral and instrumental betterness have similar implications.)[2] But this justification for longtermism faces a serious problem. As I show in this paper, those same claims seem to give us a *reductio ad absurdum* in practice—they do not tell us that any available option is ever better than any other. In fact, they do not imply longtermist verdicts; they imply no practical verdicts at all!

But, first, what are these plausible-sounding claims that seem to justify longtermism? The first is *Impartiality*: that the moral value of a life does not depend intrinsically on when or where it occurs; that a human life lived millions of years in the future would be no more or less valuable than an otherwise identical life lived today.[3] By an impartial view, the *total* sum of value across the future may be astronomical—if humanity survives for long enough, an astronomical number of future people may exist, each contributing a similar amount of value to the total.

The second claim is that those total sums of value determine how we should compare outcomes morally. Call this *Additivity*, the claim that: an outcome is at least as good as another if and only if the former contains at least as great a total sum of the value of individual lives.[4] And if we combine Additivity with Impartiality, then it follows that it would be far better to improve many trillions of future lives than it would be to improve far fewer present lives by the same amount.

The third claim is that, when comparing *risky* options, *expected (moral) value theory* holds. This common approach (when combined with the above) says that the morally best *prospects* over outcomes are those with the highest expected moral value—the highest probability-weighted sum of (total, moral) value.

[1] Note that this is an *axiological* thesis, rather than a *deontic* one. Greaves and MacAskill (2021) present both axiological and deontic versions of what they call *strong longtermism*. The view I will focus on throughout, defined here, is approximately equivalent to their *axiological strong longtermism*.

[2] See, for instance, Tarsney and Thomas (n.d.), Thomas (n.d.), Buchak (n.d.), and Greaves and MacAskill (2021, §6).

[3] This claim is defended by many, including Sidgwick (1907, p. 414), Ramsey (1928, p. 541), Parfit (1984, p. 486), and Cowen and Parfit (1992).

[4] For strong independent justification of Additivity, see the arguments of Broome (2004, ch. 18) and Thomas (2022, §5). Note also that Additivity, as defined here, is compatible with critical-level and prioritarian views of how valuable each individual life is.

Consider one prospect that has a certainty of improving present or near-future lives, and a second prospect that has some small probability of improving far future lives. The riskier, future-benefiting prospect will be the better of the two, so long as the number of future lives improved is large enough. So say Impartiality, Additivity, and expected value theory in conjunction. Indeed, in practice, the stakes and the probabilities in many practical decisions seem to be high enough that these claims imply that, *in fact*, it is often better to do whatever will most improve the far future (see Greaves and MacAskill, 2021, for discussion).

But these same three claims, in conjunction, also have troubling implications.

By some probability distributions over how great the total moral value of the future will be, the expected total value of the future would be *undefined*. These distributions include well-known troublemakers from decision theory: the Pasadena game (originating in Nover and Hájek, 2004) and the Agnesi game (see Poisson, 1824; Alexander, 2012). But these problematic distributions aren't merely hypothetical. As I will argue below, we have compelling reasons to adopt similarly problematic probability distributions over the total value that results from *any* practical choice.[5]

If we accept such a problematic probability distribution over the total value of the future, and we accept Impartiality, Additivity, and expected value theory (and no principle for comparing risky options stronger than that), then we face a dire *reductio*. For every option ever available to us in practice, we cannot evaluate it; we cannot compare it to any other such option; not even to options *identical* to itself. We can *never* say how our options compare morally.

This implication seems absurd. But it is not immediately clear how we might avoid it in a plausible manner, at least without abandoning Impartiality or Additivity—without admitting that the time at which a life is lived *can* matter morally, or admitting that the ranking of outcomes deviates from that of their total values, either of which undermines the case for longtermism described above. Can we hold onto both claims and extend our comparisons to lotteries without slipping into absurdity?

One way we might do so is by replacing expected value theory with an alternative theory which exhibits sensitivity to risk (e.g., expected *utility* theory with a non-linear utility function, or a version of *risk-weighted* expected utility theory). With the right profile of risk aversion and risk seeking, such theories can effectively replace prospects like the Pasadena game with better-behaved ones. Given this, we have a novel argument for risk sensitivity in the moral context: it seems we may need it to compare moral lotteries *at all*, in practice. Depending on the nature of the risk sensitivity needed, this argument may well also undermine longtermism.

In this paper, I seek to determine whether this is the only way out. If you find Impartiality,

[5]Note that the distributions I describe assign no probability to outcomes of infinite or undefined value. The problems I describe arise even if we treat infinitely-valued outcomes as a conceptual impossibility. Likewise, they arise if we recognise infinitely-valued outcomes but our decision procedure brackets them off and compares lotteries only by the portion of their distributions over finitely-valued outcomes (as is proposed by Bostrom, 2011, pp. 37-8).

Additivity, and the risk neutrality of expected value theory convincing, is there some way to salvage them? If not, we have a compelling argument against the conjunction of those principles, and a compelling objection to the above justification of longtermism.

Most promising is to *extend* expected value theory to compare troublesome lotteries. The literature already features various proposals for how to do so (e.g., Colyvan, 2008; Easwaran, 2008; Easwaran, 2014a; Meacham, 2019). But, as it turns out, no existing proposals succeed in making comparisons between the prospects that I argue we face in practice. Despite this, I propose a new theory, stronger than those already on offer, that resolves the problem. With this theory, we can avoid the *reductio* that expected value theory, Impartiality, and Additivity brought upon us, and we can do so without endorsing risk sensitivity and without undermining the above argument for longtermism. Indeed, as I will show in Section 5, this new proposal provides justification for an even stronger form of longtermism.

2 Why might the expected value of the future be undefined?

Decision theorists have long recognised prospects that lack well-defined, finite expected values. Some prospects lack such expected values because they feature outcomes with *infinite* value, such as in Pascal's Wager. But I will set aside such prospects in this paper, and assume that the future of humanity must have only finite value.[6]

But even if we exclude infinitely valuable outcomes, some prospects still lack well-defined expected values. One frequently discussed such prospect is that of the *Pasadena game*.[7]

Pasadena game: (An outcome with) value 2 with probability $1/2$;
value -2 with probability $1/4$;
value $8/3$ with probability $1/8$;
...
value $\frac{2^n}{n}(-1)^{n-1}$ with probability $1/2^n$ (for each positive integer n).

What is the game's expected value? If we try to calculate it in the order the outcomes are listed, we obtain the series $1 - 1/2 + 1/3 - 1/4 + ... + \frac{(-1)^{n-1}}{n} + ...$ This series, also known as the alternating harmonic series, fails to be absolutely convergent. If we were to naively add it up in one order

[6]My reasons for setting aside such prospects are threefold. The first: it is independently interesting if we can solve the problems raised by prospects over finitely-valued outcomes alone. The second: you might in fact think that outcomes of infinite value are metaphysically or logically impossible, and so assign them probability zero in practice (cf. Al-Kindī, 1974; Craig, 1979). The third: the problems of infinitely-valued outcomes seem solvable, but in a way that leaves intact the problems of the Pasadena game and its kin (see Wilkinson, 2021; Tarsney and Wilkinson, n.d.).

[7]This game is typically presented with payoffs in terms of dollars or (decision-theoretic) utility, in amounts matching those below (e.g., Nover and Hájek, 2004; Easwaran, 2014a; Bartha, 2016). Such versions of the game pose problems for expected dollar maximisers and expected utility maximisers. Here, the game is presented in terms of *moral value* and will pose structurally identical problems for expected *value* maximisers.

or another, we could obtain *any* total we wanted, so long as we picked the right order.[8] So, we cannot say that the game has any particular expected value at all (see Nover and Hájek, 2004)—in this sense, the Pasadena game *defies expectations* (or is *expectation-defying*). And so expected value theory cannot tell us how it compares to any outcome, to any other prospect, nor even to itself. If lotteries are to be compared by expected value theory alone, then the Pasadena game will be no better than, no worse than, nor equally good as any other prospect we might consider.

A similar prospect is the *Agnesi game*. Unlike the Pasadena game, it gives a continuous (rather than discrete) probability distribution over possible values. It can result in an outcome of *any* real value v; its probability density over value is given by the following function, also known as the *Witch of Agnesi* or (an example of) a Cauchy distribution.[9]

$$p(v) = \frac{1}{\pi(1 + v^2)}$$

On a graph, the distribution looks like this:

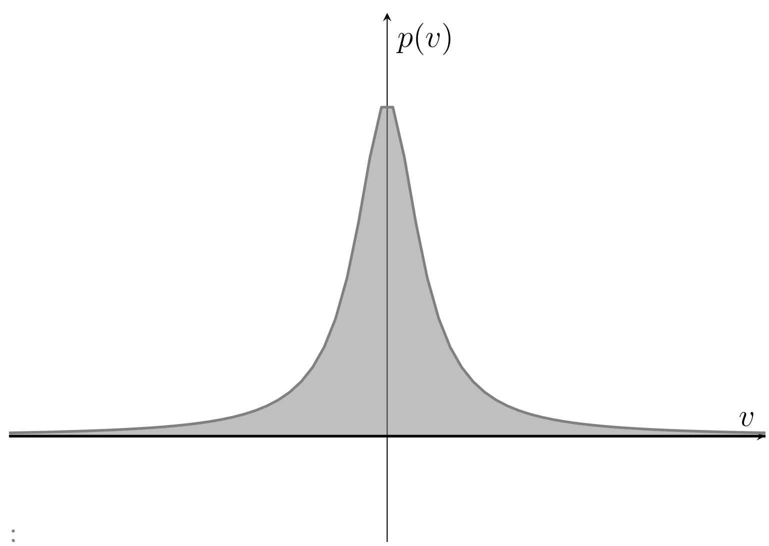

Figure 1: Probability density function $p(v)$ for the Agnesi game

Try to take the expected value of this prospect and you will find that it has none (Poisson, 1824). For continuous distributions like this, the expected value is given by the integral of $v \times p(v)$ from negative infinity to positive infinity (analogous to an expected sum: $v \times P(v)$ over all possible values v). But, for the Agnesi game, that integral between 0 and positive infinity is positively infinite! And, from 0 to negative infinity, it is negatively infinite! Sum these integrals together—equivalently, take the integral over *all* possible values of v—and the total is undefined. Much like the Pasadena

[8] Since the series is conditionally convergent, this result follows from the Riemann series theorem.

[9] This curve was first described in print by de Fermat (c. 1659) and first analysed as a probability distribution by Poisson (1824). For a discussion of this distribution in the context of decision theory, see Alexander (2012).

game's expected sum, the Agnesi game's expected integral fails to converge absolutely. It, too, defies expecations. So, expected value theory will fail to compare it to any outcome, to any other prospect, nor even to itself.

You might think that neither of these prospects are realistic—that they are merely contrived, hypothetical options that we are sure never to encounter beyond the pages of a philosophy journal. As Hájek (2014, p. 565) says of the Pasadena game, you might think that considering either prospect is "...a highly idealised thought experiment about a physically impossible game."[10] If so, you might not be troubled by the silence of expected value theory above. You might think that we should simply ignore them, and that expected value theory will still suffice for real-world decision-making.

Unfortunately, there is reason to think that we do face such prospects in practice. When we are evaluating our options morally, if we consider the prospects for humanity's long-run future and we maintain Impartiality and Additivity, then we have reason to think that every option ever available to us defies expectations. In the remainder of this section I give two discrete arguments to this effect, the second more compelling than the first.[11]

But, first, a brief note on probabilities. I assume that, for the purpose of moral comparisons, the relevant notion of the probability of an outcome must be one of two notions. The first is its *evidential* probability: how probable that outcome is to result from a given option, on the present evidence of the agent deciding between that option and others (see Williamson, 2000, p. 209). The second possible notion is the outcome's *subjective* probability: how confident the decision-making agent is that that outcome will result from a given option. If evidential probabilities are the morally relevant ones, and if our evidence prescribes expectation-defying prospects, then we will face difficulties. Or, if subjective probabilities are the relevant ones, and if we form our beliefs rationally given our evidence, we will still face difficulties.

2.1 A possibility of Pasadena

A simple argument that our prospects for the total value of humanity's future defy expectations goes like this.

It seems *possible* that a Pasadena game will be played at some point at some time in the future.

[10] Along similar lines, Jeffrey (1983, p. 154) says of the related St Petersburg game that "...anyone who offers to let the agent play [it] is a liar, for he is pretending to have an indefinitely large bank."

[11] I focus on the prospects of humanity's future rather than of the world as a whole, for three reasons. The first is simplicity. The second is that there are moral views on which the proper objects of comparison are not worlds as a whole but instead consequences—the portion of the world that it is (nomologically) possible to influence in a given decision (see, e.g., Bostrom, 2011, §3.2). And the third is that, if humanity's future prospects have undefined expected value, then so too will the prospects of the world as a whole (unless the value of events inside and outside our causal future are very strongly anti-correlated, and we have no reason to think that they are). So, it suffices to focus on the value of humanity's future.

Although perhaps physically unrealistic, we can at least *conceive* of some future agent being subjected to such a game—perhaps run by some mechanism of objective chance—and losing or gaining value in their own life with probabilities as listed above. It would be no (logical or metaphysical) impossibility for this to occur. And, given how little we know about the far future, you might think it overconfident to assign probability zero to any agent ever being subjected to such a game. So, the evidential probability of a Pasadena game someday being played, it seems, must be greater than zero.[12]

And, as has often been discussed before, *any* prospect with real, non-zero probability p of the Pasadena game, no matter what other prospects it is mixed with, inherits the problems of the game itself—like the game itself, having any such probability p of the Pasadena game brings undefined expected value (Hájek and Smithson, 2012: pp. 39-42; Bartha, 2016: pp. 802-3). So, as long as we have some probability p of such a Pasadena game over moral value being run somewhere in the future, the overall prospect for the total value of the future will be undefined.

But is there such a probability of the Pasadena game someday being played? I do not think the answer is *clearly* yes. One reason for doubt is that the correct theory of epistemic rationality may be *knowledge-based*: it may include as evidence everything the agent *knows*, and so require that evidential probabilities be assigned only after conditionalising on the agent's knowledge (see Williamson, 2000, §10.3).[13] And you might think that we *know* that no one will ever be subjected to the Pasadena game. Why? Perhaps you know that it would violate some physical law—it seems plausible that an objective chance mechanism that can produce arbitrarily large amounts of moral value would be physically impossible. Or perhaps you note that there are infinitely many different *possible* games that future people might face in their lives, but at most finitely many that anyone actually faces—from this, perhaps you can know that the Pasadena game won't be among them. Or perhaps you simply think it so implausible or subjectively improbable that the Pasadena game is ever played that you conclude that you know it will not be. Whatever the reason, you might then conditionalise on this knowledge and assign the game evidential probability zero.

Another reason to doubt that the evidential probability of the Pasadena game is non-zero is this. It's one thing to think that any possible *outcome* should be assigned non-zero probability. But it's quite another to think that any possible probability distribution *over* outcomes should be assigned non-zero probability. It may be too overconfident to assign probability zero to the future having value v or greater, for any v.[14] But it would be a strictly stronger, and so less plausible, claim to say the same of assigning probability zero to the future having any possible *probability distribution* over

[12]This line of thinking might be captured in the much-discussed principle of *Regularity*: that only logically (or perhaps metaphysically, or doxastically) impossible propositions have evidential probability zero (see Edwards et al., 1963; Easwaran, 2014b). But this principle is controversial (see, for instance, Pruss, 2013).

[13]To similar effect, you might instead think that the correct *decision* theory is knowledge-based: that, when comparing prospects, we can evaluate each prospect once we conditionalise on our knowledge (see Liu, n.d.).

[14]This claim could be treated as a weakened form of Regularity (see Footnote 12), such as: that, only for a logically (or perhaps metaphysically, or doxastically) impossible outcome O can the proposition "Outcome O occurs." have an evidential probability of zero.

values v and above. Perhaps doing the latter would not be too overconfident. Or at least, given the dire implications if you do so, perhaps epistemic rationality should not require that you entertain every such possible probability distribution (even if it *does* require you to entertain every possible *outcome*).

For either of these reasons, or perhaps others, you might be unconvinced of this argument for us facing expectation-defying moral prospects in practice. To be truly worried that expected value theory is not up to the task of comparing our moral prospects, we may need a more compelling argument that we do face such prospects.

2.2 One model of the future

Here is a more compelling argument that we face expectation-defying prospects in practice.

Consider some future time T, beyond which we have no informative empirical evidence about what will occur when. By this I mean the earliest T such that all of our specific predictions of events *after* T are merely the uniform continuation of continuous physical trends from *before* T. In effect, T is a time after all of our particular predictions for humanity's future are exhausted. Perhaps T is a billion years in the future; perhaps just 1,000 years in the future. [15]

However late T is, it is possible that humanity survives until then (or at least that *some* form of morally valuable life in our causal future survives until then). Regardless of how pessimistic you are about humanity's prospects, it seems wildly overconfident to assign probability zero to us not making it until T, or to say that we *know* that we will not survive until then. (Indeed, it seems *far more* overconfident than assigning probability zero to the Pasadena game someday being played, or claiming knowledge that it won't be.) Then, conditional on us surviving until T, what of the prospects for life *beyond* that, as time stretches out indefinitely? What is the conditional probability of a further value v arising? Since we have no empirical evidence about events beyond T, by definition, the answer is not so clear.

Consider one way we might model value after T which, I suggest, we do not know is incorrect. (There may be many plausible other models but, for my purposes, just this one model will suffice.) We might model that value as the sum of value at discrete, isolated, and reproducing *clusters* of human civilisation. At present, humanity is clustered together at one location, on a single planet. If we were to stay in this situation, it would be appropriate to assign a constant probability (or at least a minimum, non-zero probability) to us going extinct each year. But, more realistically, human

[15] Perhaps T lies after the so-called heat death of the universe. But note that even that predicted heat death is a continuation of a long-running trend of cosmological expansion—of the universe increasing in entropy which, beyond some point, it qualifies as having undergone heat death. Still, the universe will never quite reach a state of perfect entropy, so there is no genuine categorical difference between before heat death and after it (Dyson et al., 2002).

civilisation might *not* remain so clustered; perhaps we might spread through space into many such clusters. As we spread further and further, some such clusters will be more and more isolated from others. For instance, if we imagine humanity spreading to different planet-like bodies throughout space (perhaps in different galaxies, or as far from each other as we like), the maximum spatial distance between one planet and its most distant counterparts will become greater and greater. Each such planet thereby becomes more and more isolated from its most distant counterparts—its inhabitants become better and better protected from calamities that arise on the most distant planets. Indeed, given enough time, it plausibly becomes *physically impossible* for events within one such cluster of planets to affect other discrete clusters.[16] Complete isolation like this may also be achieved in other ways, such as by us perhaps even creating and inhabiting 'baby universes'.[17] But however our descendants might isolate themselves from one another, doing so makes human extinction far less likely. The extinction of humanity as a whole would then require great calamities to happen *independently* in each of many isolated clusters of civilisation—far less likely than any individual calamity.[18] The more clusters, the lower the probability of overall extinction at any given time.

Absent such calamities, in this model of the future, the number of clusters increases over time. We can assume that each existing isolated cluster has the same (independent) probability of 'reproducing' by settling a new location that will eventually be isolated from it, and thereby creating a new cluster. I will also assume, as seems at least possible, that the probability of a cluster reproducing in a given time period is at least as great as its probability of dying off.

And the more clusters, the more moral value there is. We can assume—again conservatively, as it ignores growth within each cluster—that the total value of human civilisation in a given year is roughly proportional to the number of such clusters that then exist. The total (absolute) value after T then, again assuming Impartiality and Additivity, will be roughly proportional to the sum of the lifetime of every such cluster to ever exist. But that total value may be positive or negative—there is some risk that the future of human civilisation may be one of immense misery. Or, at least, we should be uncertain about the relation between total number of cluster-years and total value—uncertain of the average value of a year of such a cluster existing. For simplicity, I will assume that there is a simple distribution over what this average value will be: probability 0.5 that it is some value v

[16] In the case of humanity being spread over planets further and further away from one another, this is made possible by cosmological expansion. With continued expansion, even star systems currently close to one another will eventually have non-overlapping causal futures (see Ord, n.d.).

[17] The possibility of doing so is somewhat supported by the prominent *inflationary view* of cosmology, under which our own universe was created by a quantum tunnelling event (see Vilenkin, 1983). It is far from settled whether inflationary cosmology would indeed allow this but, on our current understanding, it is certainly a live possibility (Farhi et al., 1990). And, independently, there is theoretical support for it being possible to create new universes via the formation of black holes, and that universes created in this way may be only temporarily accessible to their creators (Brandenberger et al., 2021; Frolov et al., 1990). Again, the science is far from settled but, based on our current evidence, it is a live possibility. (For an accessible survey of this topic, see Merali, 2017)

[18] Cf. Sandberg and Armstrong (2012).

and probability 0.5 that it is $-v$; and this is (roughly) independent of our uncertainty of how *many* clusters there are. (This distribution is unrealistic, but will be made more realistic below.)

If we combine these assumptions, the arrangement of clusters humanity's future forms a stochastic process known as a *birth-and-death* process (or, more specifically, a *Kendall process*—see Kendall, 1948). Individual clusters reproduce and die off independently, much like members of a population. And what we care about is the total number of cluster-years that are ever lived (but, by assumption, it is equiprobable that the average cluster-year is positive or negative in value). This gives us a rather complicated probability distribution over value.[19] But, fortunately, there is a prospect with a simpler distribution that shares its key properties: the Aquila game.[20] For simplicity, I will focus on the Aquila game, as given by the equation and plot below.

$$p(v) = \frac{a}{b + |v|\sqrt{|v|}} \qquad \text{for some constant } a, b > 0$$

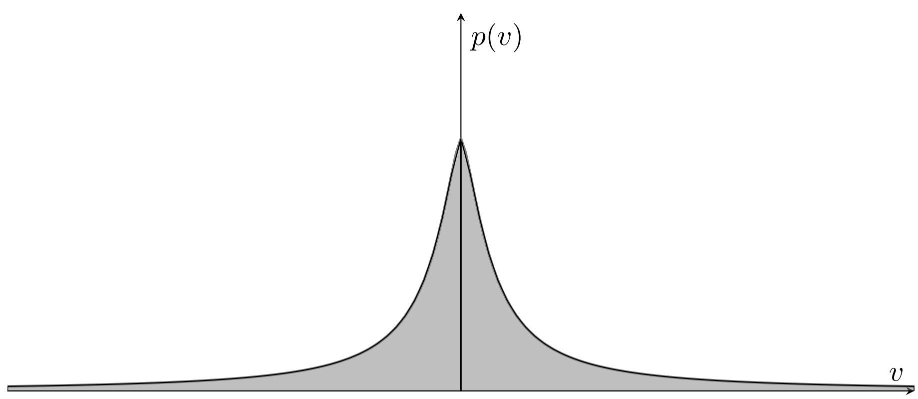

Figure 2: The probability density function over value for the Aquila game

Just like Pasadena and Agnesi, attempt to take the Aquila game's expected value and you will find that it defies expectations. Like the Agnesi game above, the probability density in its tails—as v approaches $\pm\infty$—approaches 0 sufficiently slowly that the expected value integral is undefined. And the same goes for the prospect for the total value of the world *overall*, both before and after T: like

[19]The above model, as a Kendall process with death and birth rates of μ and λ respectively, gives the following probability distribution over total population / total cluster-years (from McNeil, 1970, §5.b).

$$p(v) = \sqrt{\frac{\mu}{\lambda}} \frac{I_1(2v\sqrt{\mu\lambda})}{ve^{v(\mu+\lambda)}}$$

Here, $I_1(z)$ is the first-order modified Bessel function of the first kind, which is equivalent to $\frac{1}{\pi}\int_0^\pi e^{z\cos\theta}\cos\theta d\theta$. (I am grateful to Alex Barry for assistance with these details.)

When $\mu \leq \lambda$, that distribution has the crucial property of lacking a defined expectation. It also matches the equation for the Aquila game above in that it does not satisfy the *stability* condition discussed below in §4.2 – 4.3, and variants of it behave the same as variants of the Aquila game under the theory I propose in §4.4. For my purposes, then, it suffices to focus on the (much simpler) Aquila game.

[20]Given its connection to the St Petersburg game and its cosmic motivation, the game takes its name from the location of the Petra system in our night sky: the Aquila constellation.

Pasadena, we can mix the Aquila game with any other prospect and the overall prospect will have tails of the same shape; the overall prospect will defy expectations too.[21] Similarly, we can add the payoff of Aquila to any prospect over events before T (at least, any prospect that is not perfectly anti-correlated with the Aquila game) and the prospect over the overall payoff will defy expectations. So, if (a prospect that behaves the same as) the Aquila game is at least one minimally probable prospect for what happens after T, then expected value theory will fail to compare *every* pair of options we might ever come across in practice. And, again, that failure extends to the comparisons needed to justify longtermism.

But is the above prospect (or some mixture of it and other prospects), which is approximated by the Aquila game, the correct prospect for the value of humanity beyond time T? If there is *any* non-zero probability that it is, that is enough for my purposes—the overall prospect will then inherit its expectation-defying property. And it does seem an at least minimally plausible story, such that I suggest we *should* assign it a non-zero probability—indeed, a fairly high probability. But you might be sceptical. Here are three reasons why, and why I do not think they undermine the above model of our future prospects.

The first reason for scepticism: perhaps the number of clusters of, and value of, human civilisation could not continue growing forever. Perhaps eternal exponential growth of this sort, whether it is achieved by spreading outwards in an ever-expanding cosmos or by creating baby universes, is physically impossible. This may well be true! But we do not *know* that it is. It seems rational to assign at least *some* non-zero probability to an average such growth rate, at least in the absence of catastrophes. And if we assign *any* non-zero probability to this, and so to the above model being correct, then our prospect over the value of the future will still defy expectations.

The second reason why the above model may be unrealistic: you might think that some possible extinction scenarios would strike every cluster of civilisation at once; perhaps some exotic physical phenomenon could simultaneously remove the conditions necessary for morally valuable life everywhere. If so, the annual probability of extinction of each cluster would not be entirely independent of the others. And, given this, the annual probability of humanity's overall extinction would not be brought arbitrarily close to 0 by simply adding more and more clusters. But still this does not prevent the prospect of humanity's overall future value from resembling the Aquila game. Even if there is some annual probability of civilisation-wide extinction, whether we avoid extinction in one year (conditional on having survived until the previous year) is not independent of whether we avoid it in every other year (conditional on having survived until the year before). In some states of the world, phenomena that extinguish all of humanity at once are physically possible; in some states of the world they are not, and having arbitrarily many isolated clusters of humanity does provide

[21] As above, I assume that events within and beyond our causal future are not strongly anti-correlated (see Footnote 10).

arbitrarily much protection from extinction. We should assign at least *some* non-zero probability to such extinction-causing phenomena being physically impossible. And so we can treat the overall prospect of humanity's future value as a mixture of the prospect in which such phenomena are physically possible and the prospect in which they are not possible—in effect, a gamble between some prospect and something like the Aquila game. And so the overall prospect we obtain will still have tails resembling the Aquila game, since it offers some non-zero probability of playing such a game. And, since the Aquila game defies expectations, then the overall prospect will too. So it suffices to analyse the Aquila game in place of the more complicated overall prospect.

The third reason: it seems implausible that the average human life is just as likely to be negative in value as it is to be positive (and of equal absolute value, on whichever interval scale we use to represent value). It seems to me at least that any future civilisation will more likely aim to make its descendants happy than aim to make them miserable (or, more generally, to have valuable experiences rather than disvaluable ones), and that its probability of success in this goal is better than chance. This probability of success seems *far* better than chance once we recognise that humanity in the far future will likely have access to far more advanced technologies and greater resources than we do. Or perhaps you are pessimistic about humanity's future technological level, its available resources, or its inclination to benefit posterity. Perhaps our descendants are particularly likely to succumb to scenarios of widespread misery (for discussion of such possibilities, see Baumann, 2017). If you think so, you might think the prospect for the average human life skews towards misery rather than happiness. Either way, my earlier assumption that the average human life has probability 0.5 of having value some $k > 0$ and probability 0.5 of $-k$ seems clearly false. Rather, one of these possibilities will have higher probability than the other, and so the distribution will skew one way or the other.[22]

Given this skew, the true distribution over the value of the future of humanity will not be symmetric like the Aquila game. It will be skewed in either the positive or negative direction, as illustrated below. This more general *Skewed Aquila Game* has a probability distribution given by the following equation (for some positive $a_1 \neq a_2$, representing the relative probability of total value being positive or negative).

$$p(v) = \begin{cases} \frac{a_1}{b+|v|(\sqrt{|v|}} & \text{for } v > 0 \\ \frac{a_2}{b+|v|(\sqrt{|v|}} & \text{for } v < 0 \end{cases}$$

[22] The distribution will likely also be far more spread out than this, but I will put that complication aside, as it will simply result in an overall distribution with tails that approach 0 even more slowly than the Aquila game. The same problems as below will arise and the same solutions will hold.

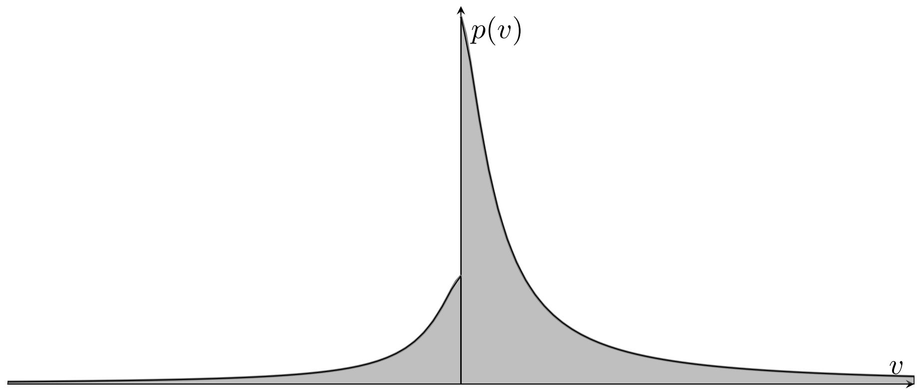

Figure 3: A probability density function over value for (a version of) the Skewed Aquila game

For simplicity, in much of what follows, I will focus on the more basic Aquila game. The problems I will describe that arise for comparing the Aquila game to alternatives will arise with equal force if we substitute in the Skewed Aquila Game.

2.3 Challenges for decision-making

If you model the value of the far future of humanity as the Aquila game, as the Skewed Aquila game, or as involving any non-zero probability of the Pasadena game or similar, then you face a serious challenge. In practice, you cannot assign an expected moral value to any of the prospects ever available to you. So, if expected value theory (and nothing stronger) is the correct theory of moral betterness under risk, then no option ever available to you will be morally better or worse than any other. But this implication is absurd.

To plausibly compare any of our available future prospects, we must replace expected value theory with some stronger alternative. In later sections, I will discuss such alternative theories. But, first, what do we want them to achieve? If they can go further than expected value theory, just how far do they need to go to suffice for practical use?

I propose five problem cases that those theories must be able to deal with—and deal with in the intuitively *correct* way—to be extensionally adequate. These cases will be crude simplifications of the options we face in practice: they exclude all sources of value in the world other than the possibility of an Aquila generated after time T; and they mostly involve symmetric such Aquila games, rather than the more realistic Skewed Aquila game from above. In practice, we face options in which many valuable events will occur before T, and in which there is perhaps only a small probability of humanity surviving until T to generate something resembling an Aquila game. Nonetheless, it will suffice to consider simplified cases like these—as noted above, our available prospects will inherit the problems of the Aquila game. If expected value theory fails in the cases below, it will fail in practice. And it turns out that it does fail, as do stronger theories designed to deal with the original Pasadena and

Agnesi games.

The first problem case, *No Change*, is (a simplification of) the decision scenario an agent faces when their available actions all produce exactly the same future prospect. For instance, an agent may choose between eating Sugar Puffs for breakfast and eating Frosties, but have no evidence for either option being more or less likely to influence the future in any particular way. (Agents with great foresight may have access to evidence supporting some story of why one cereal is more likely to produce better long-run outcomes, but suppose that the agent here lacks any such evidence.) For our purposes, the options available to her are equivalent to those below.

Scenario 1: No Change

Sugar Puffs: The Aquila game with a particular a, b.

Frosties: The Aquila game with the same a, b.

Note that both options have identical probability distributions over value. But, still, bare expected value theory cannot say how they compare—neither option has well-defined expected value, so that value cannot be equal to itself. (The same goes if we swap the Aquila game for the Skewed Aquila game.) And this is all the more troubling when, intuitively, the correct ranking of options seems clear: Sugar Puffs and Frosties are equally good. It would be desirable for our theory to say this, that the Aquila game with such and such parameters is equally as good as any other with the same parameters.[23]

The second problem case, *Improving the Present*, is that which an agent faces when they can improve some aspect of the world with certainty[24], without otherwise changing the prospect. For instance, an agent may choose whether to save the life of a child in the present day. And, regardless of whether they do so or do not, their evidence may entail an identical probability distribution over what happens in the very distant future. If so then, for our purposes, their options are equivalent to the following.

Scenario 2: Improving the Present

Do Nothing: The Aquila game (with a particular a, b).

[23]Failure to rank these options as equally good can also be characterised as a violation of *Stochastic Equivalence*.

> *Stochastic Equivalence*: For any prospects O_a and O_b, if for every possible outcome O both O_a and O_b have the same probability of resulting in an outcome equally as good as O, then O_a and O_b are equally good.

This principle is both intuitively very plausible and one that expected value theory easily satisfies for finitely-supported prospects.

[24]If that improvement is less-than-certain, we have a slightly different scenario. Fortunately, each of the proposed theories below that give the correct verdict in Improving the Present happen to give the same verdict in this different scenario, so I will not dwell on that scenario here.

Save a Life: The Aquila game (with the same a, b) with value $v' > 0$ added to every outcome.

Here, both options are identical *except* that the latter has its probability distribution shifted by some bonus value v'.[25] But, again, expected value theory cannot compare any options of this sort. (Again, the same goes if we swap the Aquila game for the Skewed Aquila game here.) And, again, this is all the more troubling given that the intuitively correct ranking is clear: that, as long as v' is positive, Save a Life is better than Do Nothing. Improving every outcome should also improve the option overall (so long as the outcomes' probabilities are otherwise held fixed).

The third problem case, *Reducing Extinction Risk*, is that which an agent faces when they can affect humanity's probability of long-term survival. If the agent does nothing, humanity will have some probability of surviving to T and beyond. If they intervene, humanity will have a *greater* probability of doing so. For my purposes, both options can be represented by some mixture of a low-value outcome (which, for simplicity, we can set to value 0) and the prospect obtained conditional on surviving the near term. For our purposes, those options are equivalent to the following.

Scenario 3: Reducing Extinction Risk

Intervene: A mixture of the (Skewed) Aquila game (with some a or a_1 and b) with probability $p > 0$ and an outcome of value 0 with probability $1 - p$.

Do Nothing: A mixture of the (Skewed) Aquila game (with the same a or a_1 and b) with probability $q < p$ and an outcome of value 0 with probability $1 - q$.

Here, both options are equivalent to having some probability of playing the Aquila game or Skewed Aquila game (with such and such parameters), with Intervene giving the higher probability. But, again, expected value theory cannot compare any two options fitting these descriptions. Worse still, consider the version of this scenario in which we have a Skewed Aquila game that is skewed heavily in the positive direction (in which humanity's future is far more likely to be better than an outcome of value 0); expected value theory cannot say that Intervene is better; it cannot say reducing the risk of extinction is an improvement. Again, this is troubling.

To satisfy intuition in this case, I would suggest that a risk-neutral theory must do two things: 1) it must be able to compare the two options, and not simply leave them as incomparable; and 2)

[25] This case is an analogue of the widely-discussed comparison of the Pasadena game to the *Altadena* game (introduced by Nover and Hájek, 2004, p. 241). In both cases, a failure to rank the latter option as better is a violation of *Weak Stochastic Dominance*.

> *Weak Stochastic Dominance*: If, for every possible outcome O, one prospect O_a has a strictly higher probability than another prospect O_b of an outcome at least as good as O, then O_a is strictly better than O_b.

Like Stochastic Equivalence, this principle is both intuitively very plausible and one that expected value theory easily satisfies for finitely-supported prospects.

it must say that Intervene is at least as good as Do Nothing if and only if it is at least as good for humanity to survive as it is to go extinct. (That is, if the corresponding Aquila or Skewed Aquila game is at least as good as an outcome of value 0.)[26]

The fourth problem case, *Improving the Future*, is one that an agent may face if they attempt to improve events occurring after T, conditional on us surviving until then. In this case, the agent's choices don't make it more or less likely that we survive but, if we do survive until then, those choices make it more or less likely that the average human life afterwards will have positive value (on whichever interval scale we represent value). It is a decision of whether to alter the skew of the Skewed Altair game one way or the other. And, to an approximation, this is the sort of case an agent might face when they can affect humanity's long-run prospects in some manner that is extremely persistent. Perhaps it applies when a political activist decides whether to campaign for a change to political institutions that would foreseeably improve decision-making. Doing so may make it ever so slightly more likely that humanity at large has better political institutions indefinitely far into the future, perhaps increasing the probability that lives or clusters of civilisation have positive value on average.

Scenario 4: Improving the Future

Campaign: The Skewed Aquila game with some $\frac{a_1}{a_2}$ and b.

Don't Campaign: The Skewed Aquila game with a *lower* $\frac{a_1}{a_2}$ (and same b).

Again, expected value theory alone cannot compare the two. Nor can it say that Campaign is better than Don't Campaign—it cannot say that it is better to make it more likely that future lives are very good than that they are very bad. Intuition demands that Campaign be ranked as better than Don't Campaign.[27]

The fifth and most challenging problem case, *Multifarious Changes*, is a combination of the previous three. Instead of merely improving/worsening the present with certainty, or changing the probability of human extinction before T, or changing the probability of a good future after T conditional on survival until then, in this case the agent can have any combination of those effects. This, I think, is a more realistic representation of our options when we have opportunities to predictably affect the long-term future. When we do so—perhaps by trying to alter political institutions or

[26] To fail to do so is to violate the *Independence* axiom of expected utility theory.

> *Independence*: For any prospects O_a, O_b, and O_c, O_a is at least as good as O_b if and only if the mixture of O_a with probability p and O_c otherwise is at least as good as the mixture of O_b with probability p and O_c otherwise.

Much like the principles given in the preceding footnotes, Independence will likely be intuitively very plausible to anyone sympathetic to expected value theory. For an extension of that theory to prospects like the Aquila game to be satisfactory, I would suggest, it must satisfy Independence (see [redacted] for related discussion).

[27] Similar to Scenario 2, a failure to rank Campaign as better than Don't Campaign is a violation of Weak Stochastic Dominance (see Footnote 24).

defend against threats of extinction—we will often have all three effects. In doing so, there is often some moral cost in the present, e.g., the opportunity cost of spending one's resources on lobbying for institutional change is that the same resources aren't directly used to help the poor. And, when attempting to reduce the risk of extinction, there is often a further effect on the well-being of future people in the event of survival—e.g., implementing some measure to reduce the incidence of deadly pandemics not only reduces the risk of extinction, but likely also causes future people to experience fewer pandemics in general, whether or not they rise to the level of threatening extinction. Likewise, attempting to make future lives more valuable conditional on survival will often affect the probability of extinction—e.g., if one succeeds in changing political institutions to better respond to the public's interests, those institutions would then likely also be better at responding to threats of extinction.

If an agent has any options have at least two of those effects in one then, for my purposes, we can model their decision as follows.

Scenario 5: Multifarious Changes

Intervene: A mixture of 1) the Skewed Aquila game with some $\frac{a_1}{a_2}$ and b, with value v' added to every outcome, with probability p, and 2) an outcome of value v' with probability $1-p$.

Do Nothing: A mixture of 1) the Skewed Aquila game with some (perhaps different) $\frac{a_1}{a_2}$ and b with probability $q \neq p$, and 2) an outcome of value 0 with probability $1-q$.

From above, *a fortiori*, we know that expected value cannot compare at least some options fitting this description. But nor, it turns out, can it compare *any* such options—any two such options will defy expectations. And again, this silence is troubling. It is not merely troubling because the correct ranking of the options is intuitively obvious; the correct ranking often won't be. But it is troubling that our normative may fall silent in a decision that we plausibly face in practice. If an agent ever has the opportunity to influence humanity's long-term future, it is plausible that they face this scenario, and they need guidance. For a decision theory to be plausible, it must offer such guidance in at least the cases we actually face in practice. But expected value theory cannot do so.

3 An argument for risk sensitivity?

Above, we saw that the prospect for the value of humanity's future defies expectations. This holds *even if* we assign only a tiny probability to humanity surviving until some indefinitely distant time T and to the model described being correct (that human civilisation will continue to grow, and that isolated clusters of humanity will face no common threats). A tiny probability of each of these results in a distribution akin to the Aquila game.

Given this, expected value theory alone cannot be the correct theory of instrumental moral betterness. If it were, no future prospect ever available to us would be better than (or even comparable to) any other. And that would be absurd. So we must replace expected value theory with something else. In its stead, we can either adopt a theory that merely extends the verdicts of expected value theory (as I consider in the next section), or adopt a theory that conflicts with expected value theory even where it already gives verdicts. Here, I will consider the second sort of replacement.

One such alternative theory is *expected utility theory* (specifically, a risk-sensitive version). It works much like expected value theory does. Where expected value theory says that the best prospects are those with the highest expected moral *value*, expected utility theory says that the best prospects are those with the highest expected *utility*.

What is utility? For my purposes, it is some representation of the betterness ranking over outcomes. But it need not be the *same* representation as the moral value function. Utility here is not the same thing as what moral theorists sometimes call utility—a cardinal measure of total welfare—but instead a purely decision-theoretic construct.[28] As von Neumann and Morgenstern (1953, p. 28) put it, utility is simply "...that thing for which the calculus of mathematical expectations is legitimate." For instance, consider three outcomes A, B, and C that have moral values 0, 1, and 2, respectively. According to at least one possible utility function, those same outcomes have *utilities* 0, 99, and 100, respectively. And consider a coin flip between A and C: it will have expected *value* 1, equal to the value of B; but it will have expected *utility* 50, much lower than 99, the utility of B. So expected utility theory is compatible with the risk-averse verdict that getting moral value 1 for sure is better than a coin-flip between 0 and 2. So too, it is compatible with the risk-inclined verdict that the coin-flip is better, if we adopt a different utility function.

In general, the utility of an outcome may be *any* real-valued function of its moral value (at least when determining instrumental moral betterness), risk-sensitive or not, so long as that function is strictly increasing. In particular, the correct utility function for use in comparing prospects morally might sometimes be *concave*: the higher the *value* of outcomes, the less their *utility* increases for each additional unit of value that is added to them. This tends to lead to risk-averse preferences. And/or the utility function may sometimes be *convex*: the higher the value of the outcomes, the *more* their utility increases for each additional unit of value. This tends to lead to risk-*inclined* preferences. One possible function, $u(v)$, that is sometimes concave and sometimes convex is plotted below.

[28]See Zhao (2021, pp. 11-2) for discussion.

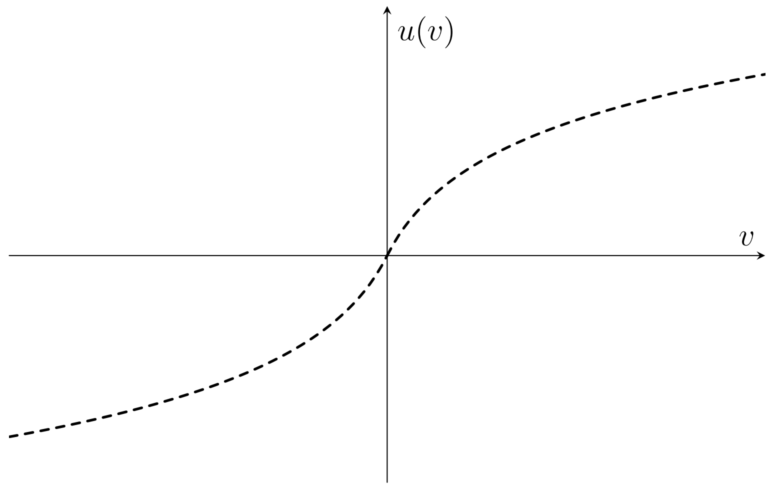

Figure 4: A utility function that is concave for $v > 0$ and convex for $v < 0$

But how does switching from expected value theory to a risk-sensitive version of expected utility theory, with a non-linear utility function, affect our comparisons of expectation-defying prospects? To see how, note that such prospects posed a problem for expected value theory only because the probability densities of outcomes didn't approach zero quickly enough as value approaches positive and negative infinity. If those extreme outcomes just had lower (perhaps *much* lower) absolute values, the prospects would no longer defy expectations, and expected value theory could evaluate them. But, in effect, they *do* have lower absolute 'values' if we switch to expected utility theory with a utility function like that plotted above—we lower the contribution that those extreme outcomes make to the expected utility calculation. Then, for the purpose of calculating expected utility, an expectation-defying prospect no longer defies expectations!

Take, for example, the Aquila game. Its troublesome distribution was given by $p(v) = \frac{a}{b+|v|\sqrt{v}}$ (for some $a, b > 0$). With a utility function that is concave enough for large positive values and convex enough for large negative values, we can turn that expectation-defying distribution *over value* into a much tamer distribution *over utility.* For instance, set $u(v) = \sqrt{|v|}$ such as that given by $p(u) = \frac{4a|u|^2}{b+|u|^3}$ (which gives an expected utility of 0).

This works in all five of the problem cases described above. In the first (No Change), where we must compare the Aquila game to an identical prospect, a utility function as above lets us say that the two prospects are equally good—each prospect gets an expected utility and, since the prospects are identical, those expected utilities are equal. So, not only can expected utility theory say something here, but it says the intuitively correct thing. In the second case (Improving the Present), expected utility theory with a utility function as above lets us say that the Aquila game sweetened by value v' is indeed better than the same Aquila game without the sweetening. In the third case (Reducing Extinction Risk), where we compare two mixtures of the (Skewed) Aquila game, expected utility theory can again provide a comparison (although what it says will depend on the exact utility function). In the fourth case (Improving the Future), it says that increasing the Skewed Aquila game's skew in the positive direction is indeed an improvement. And, in the fifth (Multifarious Changes), again, it can compare any (perhaps sweetened mixture of) one Skewed Aquila game to another. In all five cases, it satisfies the desiderata I gave above.

And we can do the same with *any* pair of expectation-defying prospects; we need only adopt a utility function that is concave (convex) *enough* for large positive (negative) values. We need only accept a certain sensitivity to risk and the problem is solved. Thus, expected utility theory can deliver verdicts in those scenarios where expected value theory was lacking.[29]

What should we make of the success of expected utility theory (with the right utility function) where expected value theory failed, and in the absence of any conservative extension of expected value theory? We might take the above as a surprising argument in favour of risk sensitivity—in favour of risk aversion for large positive values and in favour of risk inclination for large negative values.

If we accept expected utility theory (and no stronger principle) then, to be able to ever compare our moral options in practice, we must adopt a utility function like that above. Otherwise, we cannot compare the Aquila, Pasadena, or any other expectation-defying game to any other, and so we cannot compare morally any options we ever face. And we have at least some reason to accept a form of expected utility theory—it is implied by the conjunction of several appealing axioms, as shown in various celebrated representation theorems. (Note that expected *value* theory would be compatible with these axioms too, at least if weren't for expectation-defying prospects, as it is equivalent to expected utility theory with a linear utility function.) If we accept these axioms, and we require that our theory provides comparisons in practice, we have no choice but to accept expected utility theory with a utility function of the sort depicted above.

And that means accepting a certain level of risk aversion and risk seeking for large positive and negative values, respectively. This will affect everyday decision-making—in at least some circumstances, it will require agents to no longer be indifferent between, say, one additional unit of value for sure and a coin flip between two units and zero units of additional value.

But you might find this solution unsatisfying. One possible reason is that risk sensitivity of

[29] A similar result could be achieved with a modified version of Buchak's (2013) *risk-weighted expected utility* (REU) theory (even with utility linear with respect to moral value). That theory says that a prospect O_a should be evaluated by $\text{REU}(L) = u_0 + \sum_{j=1}^{n}(u_j - u_{j-1}) \cdot r\big(P(L \geq u_j)\big)$, where the utilities of possible outcomes are given in ascending order by $\{u_0, u_1, ..., u_n\}$ and $r : [0, 1] \to [0, 1]$ is some non-decreasing function describing a particular risk attitude. When applied to lotteries with continuous distributions and over outcomes with unbounded values, we might adjust the theory in two ways: 1) replace with the discrete sum with an integral; and 2) take separately the REU of the conditional lotteries i) O_a, conditional on $u \geq 0$, and ii) O_a, conditional on $u < 0$, with the latter calculated 'in reverse', using the equation $\text{REU}(L|u < 0) = u_n - \sum_{j=1}^{n}(u_j - u_{j-1}) \cdot r^*\big(P(L < u_{j-1})\big)$ (and a suitable r^* function). Doing so has an effect similar to that under expected utility of adopting the utility function illustrated above. But proponents of REU theory would likely baulk at this modification of their theory—particularly (2)—which may seem ad hoc, arbitrary, and poorly motivated.

this sort lacks independent justification. This is not the same sort of risk sensitivity that various decision theorists have recently defended as rationally permissible: risk sensitivity arising from and matching the agent's own preferences over different means to their desired ends (e.g. Buchak, 2013). Here, the risk sensitivity is imposed externally, and will often diverge from the agent's own attitudes. For instance, an agent may have be perfectly risk-neutral in their own attitudes but, to avoid the problems brought on by the Aquila game, they must instead abide by risk attitudes such as that in Figure 4. If anything, the standard motivation and standard arguments for risk sensitivity are reasons *against* accepting this solution.[30]

Another, related reason this solution may be unsatisfying is that it may be a mistake to determine a correct attitude to risk in this manner. Even if we do accept that there is a correct universal, agent-neutral attitude to risk, we might think that the correct method to set this is by considering simple, idealised cases in which our moral intuitions are particularly clear[31] and to reason from there to more complicated practical cases. If we instead determine the correct risk attitude based on the presence of prospects we *in fact* face, it may seem that we are making a methodological mistake. Yet that is what we must do if we accept this solution to the practical problem of the Aquila game and kin.

But perhaps a far more compelling reason to reject this solution is that, particularly in the moral setting, risk *neutrality* has powerful arguments in its favour. These arguments include Harsanyi's (1955) classic social aggregation theorem and many others (e.g., Tarsney, n.d.; Zhao, 2021; Thomas, n.d.; Wilkinson, 2022, n.d.a). By such arguments, if we adopt an aggregative theory of moral betterness but admit sensitivity to risk, we must violate one or another highly plausible principles. Without going into detail here, I will note simply that such defences of risk neutrality (and of such defenders) means that it is at least of interest whether we can deal with expectation-defying prospects without embracing risk sensitivity.

4 Preserving risk neutrality

Is embracing risk sensitivity our only option for dealing with prospects like the Aquila game? Or, if you find risk neutrality independently appealing, do you have any way to preserve it?

In this section, I consider possible ways we might do so—ways we might extend expected value theory to deal with expectation-defying prospects, without admitting risk sensitivity. Note that I am interested here only in whether we *can* accommodate risk-neutral verdicts—whether we can find an extensionally adequate theory—not the further question of whether we can independently motivate such a theory, which is beyond the scope of this paper.[32] I will first survey existing proposals for

[30] I am grateful to Johanna Thoma for suggesting this objection.
[31] Cf. the method of arriving at correct attitudes to risk in prudential cases in Buchak (2013, §2.3).
[32] I take up the challenge of providing independent motivation for my preferred theory in other work; see [redacted].

such a theory in the literature. As we will see, none of these proposed theories can adequately deal with a prospect as troublesome as the Aquila game. But I will propose a novel extension that can.

4.1 Relative Expectation Theory

The first such proposal is *Relative Expectation Theory*, first proposed by Colyvan (2008). Here, I will focus on the strengthened version suggested by both Colyvan and Hájek (2016, pp. 837-8) and Meacham (2019, pp. 13-7).

According to Relative Expectation Theory, we no longer attempt to assign some value to each prospect separately and compare those values. Instead, for each *pair* of prospects, we evaluate a *relative expectation* (RE): the expected *difference* in value between the two prospects; but, in calculating this difference, we match up the outcomes of each prospect by how far along the prospect's probability distribution they are. For any prospects O_a and O_b, we match up the lowest value of the possible outcomes of O_a to the lowest possible value for O_b; we match up the median values of each; we match up the best possible values of each; and likewise for every other possible value, matching each value from O_a with the value in O_b that is equally far along O_b's distribution. Put differently, we match each possible value in O_a to the value lying at the same *quantile* in O_b.

Formally, we identify the value that is fraction P of the way along the probability distribution of O with the quantile function $v_O(P)$—the function that, for each probability P, gives you the largest value v such that O has probability P (or less) of resulting in value v or less. For instance, $v_O(0.5)$ would be the median, and $v_O(0.9)$ would be the value that O has only a probability 0.1 of exceeding. (Equivalently, $v_O(P)$ is the inverse of O's cumulative probability distribution; for an illustration, see below.) With this function, Relative Expectation Theory can be stated as follows.

Relative Expectation Theory: A prospect O_a is at least as good as another prospect O_b if

$$\mathrm{RE}(O_a, O_b) = \int_0^1 \big(v_{O_a}(P) - v_{O_b}(P)\big)dP \geqslant 0$$

Relative Expectation Theory agrees with all of the verdicts given by expected value theory. But how does it fare in the cases described above? Recall, for instance, the case of No Change. Where O_a and O_b are both the Aquila game—precisely the same distribution—both will have the same quantile function v_O (matching the function labelled "Aquila game" in the figure below). So $v_{O_a}(P) - v_{O_b}(P)$ will always be 0, the integral from 0 to 1 will be 0, and they will be equally good.

Or consider Improving the Present. The option Do Nothing is simply the Aquila game, while the option Save a Life is the same Aquila game but with every outcome sweetened by value $b > 0$. These options will have quantile functions v_O as plotted below—functions that are identical, except that

Save a Life's function is shifted up by value b for all P. The difference between the functions for Save a Life and Do Nothing is always positive, so the integral of $v_{O_a}(P) - v_{O_b}(P)$ from 0 to 1 (matching the area between the two graphs below) will be positive too, and Save a Life will be better. Not only can Relative Expectation Theory compare the two, but it gives the intuitively correct verdict. For similar reasons, it also gives the intuitively correct verdict in the fourth case, Improving the Future.

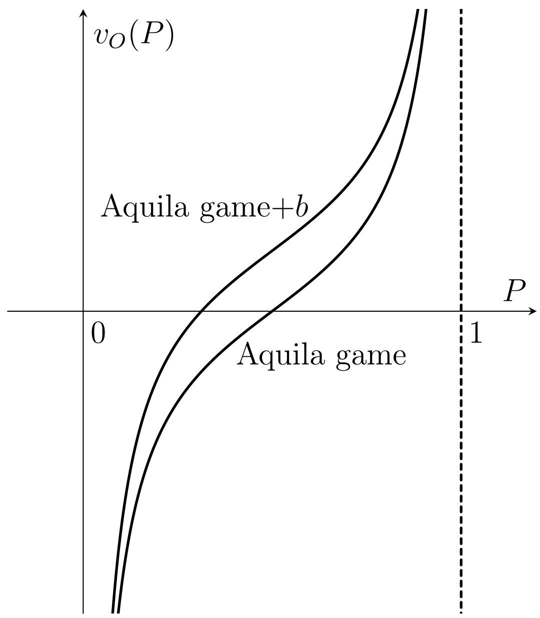

Figure 5: The quantile functions v_O of the options in Improving the Present: the Aquila game (Do Nothing); and the same Aquila game with each outcome improved by $b > 0$ (Save a Life)

But Relative Expectation Theory cannot say anything in the third and fifth scenarios (Reducing Extinction Risk and Multifarious Changes). As has been noted before, it cannot compare an expectation-defying prospect to a sure outcome of value 0 (Colyvan and Hájek, 2016; Meacham, 2019)—$\text{RE}(O_a, O_b)$ becomes the expected value of the expectation-defying prospect which, by definition, is undefined. The same goes for the Aquila and Skewed Aquila games. Where previous authors have observed this implication, they have accepted it—Pasadena and its kin are peculiar prospects, so it is not clear how we should compare them to the status quo, nor how good they are. But it cannot be a proper implication of decision theory that it falls silent in a wide range of decisions we actually face in practice, particularly moral decisions. And yet, with that verdict, Relative Expectation Theory does—that it must fall silent in Reducing Extinction Risk and, *a fortiori*, in Multifarious Changes. Whenever an agent faces a decision that affects the probability that humanity survives rather than perishes, this theory fails us. So, I suggest, it proves inadequate.

4.2 Principal Value Theory

Another alternative that extends expected value theory comes from Easwaran (2014a).[33]

By this proposal, *Principal Value Theory*, we consider *truncated* versions of each prospect: versions of each prospect with the tails of the distribution removed and that probability redistributed to value 0. More precisely, for any prospect O_a and for any positive n, define $O_a^{|v|\leqslant n}$ as the prospect that assigns the same probability as O_a to every possible value with absolute value up to n; the remaining probability mass (taken from values below $-n$ and above $+n$) is redistributed to value 0. For any such n, that truncated prospect will have some defined expected value. But we can also consider the expected value of the truncated prospect as n gets larger, tending to infinity. If, as n approaches infinity, the expected value of $O_a^{|v|\leqslant n}$ approaches some finite limit, then that limit (or *principal value*) seems an appropriate value to assign O_a. Indeed, for all prospects that have well-defined expected values, their principal values will match perfectly.

> *Principal Value Theory*: A prospect O_a is at least as good as another prospect O_b if $\text{PV}(O_a) \geq \text{PV}(O_b)$, where
>
> $$\text{PV}(O) = \lim_{n\to\infty} \mathbb{E}(O^{|v|\leqslant n})$$
>
> and the principal values of both O_a and O_b are *stable* (more on this below).

Although principal values (given by $\text{PV}(O)$) might seem a plausible way to compare prospects, they sometimes have an undesirable feature. For many prospects, the principal value is what we call *unstable*: adding a constant value v to every one of the prospect's outcomes does not increase its principal value by v.[34] For any two prospects O_a and O_b, if either has an unstable principal value, then it may hold that $\text{PV}(O_a) > \text{PV}(O_b)$ but for at least some value v we can add v to every outcome of each and the sweetened O_a will no longer have greater principal value than the sweetened O_b. And this is implausible—sweetening both prospects by the same amount shouldn't, intuitively, change their ranking.[35] So it seems correct that Principal Value Theory fall silent in these cases, where it would otherwise give us misleading verdicts.

That Principal Value Theory remains silent in such cases becomes all the more important when

[33] That proposal is itself a strengthening of an earlier one from Easwaran (2008).

[34] As Easwaran (2014a, pp. 524-5) shows, this condition is equivalent to a prospect (O) satisfying, for some positive k:

$$\lim_{n\to\infty} \Big((n-k)P\big(|O| > n-k\big) - (n+k)P\big(|O| > n+k\big) \Big) = 0$$

[35] This is especially clear in the moral setting, when we assume Additivity. Sweetening both prospects by adding a constant value to all of the outcomes of each might represent making an analogous moral decision in a world where ancient history had gone slightly differently: perhaps in ancient Egypt some number of additional happy lives were lived. It seems absurd that the correct moral verdict would be sensitive to such events, which are unaffected and held fixed across all of our options (cf. Wilkinson, 2022, n.d.c; Parfit, 1984, p. 420). I take this as a decisive objection to ranking prospects by their principal values when those are unstable.

we consider that value might be represented merely on an interval scale[36]—the relevant implication of which is that value has no natural zero. If so, the very same prospects O_a and O_b might be just as easily represented with their every outcome's numerical value increased by v, or decreased by v (much like the same level of heat can be just as easily measured in degrees Celsius, Farenheit, or Kelvin). The structure of value would entail that these different scales are equivalent; the use of some numbers rather than others is a purely cosmetic, and entirely arbitrary, choice. And yet that choice changes the prospects' principal values! And, if we apply Principal Value Theory without restricting it to prospects with stable principal values, that choice also changes our *rankings*. Thus, to give consistent, non-arbitrary rankings even if value comes only with an interval scale structure, Principal Value Theory must fall silent when comparing such prospects.

But many expectation-defying prospects of interest fall into this category. Neither the Aquila game nor Skewed Aquila game have a stable principal value (but Pasadena and Agnesi each do). As a result, in practice, Principal Value Theory cannot evaluate *any* of the prospects featured in the five cases above, nor can it compare any of those prospects to any other.

But we might strengthen Principal Value Theory further. One proposal for doing so comes from Meacham (2019, p. 1021), by which, in effect, we combine it with Relative Expectation Theory. Instead of evaluating each prospect by its principal value and comparing those values, we might take the principal value of the relative 'difference' between them. We can do so, for any prospects O_a and O_b, by considering the *relative prospect* $R(O_a, O_b)$. Roughly, this is the prospect over how much O_a and O_b would end up differing in value, if we identified their states by cumulative probability. Less roughly, this is the prospect for what value you obtain if you take the difference between $v_{O_a}(P)$ and $v_{O_b}(P)$, randomly selecting a probability P from (a uniform distribution from) 0 to 1.[37] We can then take the principal value of $R(O_a, O_b)$ rather than each of O_a and O_b. If it is greater than 0, O_a is better than O_b.

*Principal Value Theory**: A prospect O_a is at least as good as another prospect O_b if

$$\mathrm{PV}\big(R(O_a, O_b)\big) = \lim_{n\to\infty} \mathbb{E}(R(O_a, O_b)^{|v|\leqslant n}) \geq 0$$

and the principal value of $R(O_a, O_b)$ is stable.

This version of Principal Value Theory lets us say more. For instance, in the first problem case (No Change), it confirms that the Aquila game is equally as good as itself—for any prospect O, $R(O, O)$ gives a certainty of value 0. And in the second case (Improving the Present), it confirms

[36] This is advocated by, for instance, Broome (1991, p. 218). See also Adler's 2019, pp. 43-5 discussion of the *Fundamental Principle of Invariance*: roughly, that rankings of moral betterness must be invariant to certain rescalings of the measure of value.

[37] Formally, $R(O_a, O_b)$'s distribution is given by $p(v) = p\Big(v_{O_a}(P) - v_{O_b}(P) = v \quad | \quad P \propto \mathcal{U}_{[0,1]}\Big)$.

that improving every outcome by value $v' > 0$ is a strict improvement—R(Save a Life, Do Nothing) will give a certainty of value v'. Likewise, in the fourth case (Improving the Future), it confirms that shifting probability mass from a bad future to a good future is an improvement too—R(Campaign, Don't Campaign) gives a prospect in which all possible outcomes have value greater than zero.

But this version of the theory still does not let us say anything in the third case, Reducing Extinction Risk. Note that the relative prospect R generated between some prospect O_a and a sure outcome of value 0 simply *is* that original prospect O_a. So, since the Aquila game is unstable, the relative prospect between it and that outcome of value 0 will be unstable too—Principal Value Theory* cannot compare the Aquila game to the sure outcome (indeed, *any* sure outcome). This carries over to comparisons of different mixtures of the Aquila game with a sure outcome of value 0—precisely the sorts of mixtures that we must compare in Reducing Extinction Risk. So, even this strengthened version of Principal Value Theory will fall silent in Reducing Extinction Risk. And, *a fortiori*, it will fall silent in the fifth case, Multifarious Changes. So it will be inadequate for at least an important class of practical moral cases.

4.3 Invariant Value Theory

But the silence of the above proposals does not mean that there isn't *any* extension of expected value theory that can sensibly compare the Aquila game to alternatives.

I propose an alternative method: *Invariant Value Theory*. Much like Easwaran's Principal Value Theory, it uses a Cauchy principal value of a prospect's expectation, but a different one. Easwaran's proposal has us truncate prospects by the absolute *values* of their outcomes—that theory has us consider $O^{|v| \leqslant n}$, the prospect obtained from O by cutting off the tails of the distribution above n and below $-n$ (and that probability redistributed to value 0). It is little surprise that truncating the prospect in this way sometimes results in evaluations that differ if we use a different scale, with a different zero point, to represent value—after all, n and $-n$ identify different values depending on the scale. It would be much better if we could truncate in some way that is independent of the scale used for value.

Invariant Value Theory involves such a truncation. Instead of cutting off O's tails where they exceed some absolute value n, we cut them off according to probability. Take some probability ε. We cut off the right tail at whatever value the prospect has probability ε of exceeding, and the left tail at the value for which there is probability ε of falling below it.[38] In the terminology from earlier, we cut off the distribution at values $v_O(\varepsilon)$ and $v_O(1-\varepsilon)$.

Like Principal Value Theory, this proposal then takes the expectation of the truncated version

[38] Although our final proposals are very different, this method of truncation matches that used by Smith (2014).

of O_a, and evaluates O_a by the limit of this expectation (its *invariant value*) as the truncation approaches the true prospect. But this limit is taken as ε approaches 0, not simply as the value n approaches infinity. Put formally, the proposal can be expressed as follows.

> *Invariant Value Theory*: A prospect O_a is at least as good as another prospect O_b if $\mathrm{IV}(O_a) \geq \mathrm{IV}(O_b)$, where:
>
> $$\mathrm{IV}(O) = \lim_{\varepsilon \to 0^+} \int_{\varepsilon}^{1-\varepsilon} v_O(P) dP$$

And we can immediately strengthen the theory further, in line with Meacham's (2019, p. 1021) proposal described above. Instead of taking the limit of the *expectation* of each prospect as ε goes to 0, we take the limit of the *relative expectation* between the two prospects (from earlier) as ε goes to 0. Or equivalently, we simply take the invariant value of the relative prospect $R(O_a, O_b)$ (also from earlier).

> *Invariant Value Theory**: A prospect O_a is at least as good as another prospect O_b if
>
> $$\mathrm{IV}\big(R(O_a, O_b)\big) = \lim_{\varepsilon \to 0^+} \int_{\varepsilon}^{1-\varepsilon} v_{O_a}(P) - v_{O_b}(P) dP \geq 0$$

To illustrate the theory at work, consider the options in Improving the Present: Do Nothing, resulting in the Aquila game; and Save a Life, resulting in the same Aquila game but every outcome is sweetened by value $b > 0$. Their quantile functions are plotted in Figure 6 below. We consider the difference between the two functions, as we did earlier. And we take the integral of that difference (given by the shaded area below), but only between $P = \varepsilon$ and $P = 1 - \varepsilon$. This will always be well-defined and finite, for any pair of prospects. Then we let ε approach 0—and so, symmetrically, let $1 - \varepsilon$ approach 1—and see what limit that integral/area approaches. In this case, that limit is simply v', which tells us that the sweetened Aquila game is better than the unsweetened one.

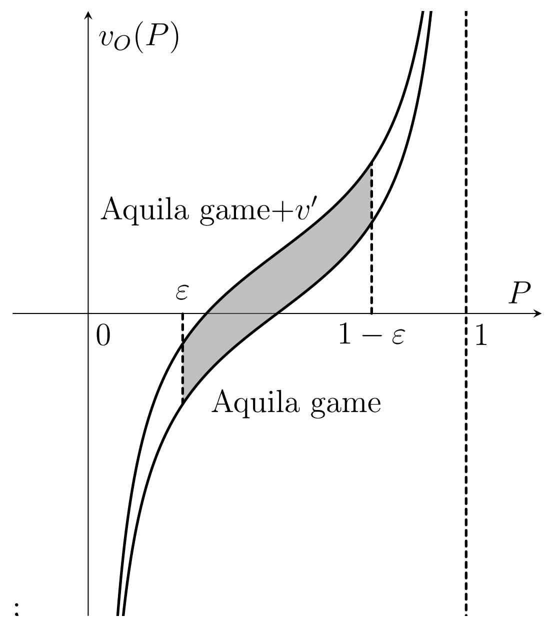

Figure 6: The quantile functions v_O of the options in Improving the Present: the Aquila game (Do Nothing); and the same Aquila game with each outcome improved by $v' > 0$ (Save a Life)

Indeed, we could take *any* prospect O_a and a modified version of O_a with v' added to the value of every outcome. Their relative difference will always have an invariant value of v'. It also follows that we can take any prospects O_a and O_b and, whatever the invariant value of their relative difference $R(O_a, O_b)$, adding value v' to every outcome of each won't change it. However Invariant Value Theory compares O_a and O_b, it will compare the modified versions each sweetened by value v' in just the same way. Thus, we do not encounter the problem faced by Principal Value Theory above—whether a prospect's principal value is stable or not, our evaluations won't be inconsistent for different sweetenings, nor for different arbitrary numerical representations of value.

This should not be surprising. The principal values used by the earlier theory involved truncating prospects at the levels of value n and $-n$; which values these are will of course depend on where our numerical representation places the zero. But the invariant value truncates prospects at whatever values (indeed, whatever outcomes) lie at given points along the probability distribution—values that are entirely determined by the probability distribution, not the numerical representation of value. So, even when our prospects have *unstable* principal values, we need not exclude them from evaluation. Under this new theory, we need no stability restriction.

How does Invariant Value Theory* fare in the problem cases from earlier? We have already seen that it compares the options in the second case, Improving the Present, and delivers the intuitively correct verdict. Likewise, it has no trouble with the first case, No Change—for any prospect O, including the Aquila game, the integral from the equation above is simply 0. So, the Aquila game will be equally as good as itself.[39] And it has no trouble in the fourth case (Improving the Future) either—since one option there turns out better than the other in *every* quantile $v_O(P)$, the invariant value is guaranteed to be positive in favour of the former. So, in both scenarios we can compare our options, indeed in the intuitively correct way.

But can Invariant Value Theory* say what the previous proposal could not? Can it deal with the remaining two problem cases? In the third problem case, Reducing Extinction Risk, we must compare two different mixtures, call them O_p and O_q, of the (Skewed) Aquila game and an outcome of value 0. O_p has probability p of resulting in the (Skewed) Aquila game, and O_q has a smaller probability q of resulting in that same game. It can be shown that $R(O_p, O_q)$ will itself be a mixture of the (Skewed) Aquila game and an outcome of value 0, with probability $p - q$ of resulting in the (Skewed) Aquila game. Since the Aquila game has a symmetric distribution, the invariant value of $R(O_p, O_q)$ will be its median value, 0 (see Wilkinson, n.d.b, §4) , the same as the Aquila game. So, the theory can compare O_p and O_q, and says that they are equally good. (The same won't hold of the Skewed Aquila game, where human survival is not equally likely to be better or worse than value 0; more on this below.)

What of the fifth case, Multifarious Changes? Again, we must compare different mixtures O_p and O_q of (perhaps Skewed) Aquila games, but they may be *different* Skewed Aquila games (with different values of $\frac{a_1}{a_2}$), and they might be sweetened different constant amounts. We are no longer comparing transformations of the same underlying prospect; we must now compare transformations of entirely *different* Aquila games. But, again, Invariant Value Theory* can do so. Any such Aquila game is symmetric about its median value, as is any mixture of it and an outcome of value 0. So, as above, it can be shown that $R(O_p, O_q)$ will itself be some mixture of some Aquila game with an outcome of value 0. This will have invariant value 0. Again, the theory can compare any two such prospects—even in this most challenging of the problem cases, this proposal succeeds in providing guidance, unlike Principal Value Theory*.

And Invariant Value Theory(*) does better than its rivals in other respects. Its basic approach is similar to Principal Value Theory(*), so seems at least as well motivated as that theory. In fact, it seems strictly *better*-motivated than that theory—unlike Principal Value Theory(*), the function of prospects by which we evaluate them (PV or IV) is not one that requires a further, ad hoc restriction on when we can use it. This theory can compare *any* pairs of prospects that have defined invariant value/s, without excluding some subset that do not also satisfy the stability condition. (For a further, axiomatic motivation of the theory, see [redacted].)

[39] We can obtain the same verdict here with the weaker version of Invariant Value Theory. The Aquila game's distribution is symmetrical about 0 and, as I show elsewhere (see Wilkinson, n.d.b, §4), the invariant value of a symmetrical distribution is *always* its median value. So the invariant value of the Aquila game is 0. Most importantly, this means that its invariant value is *defined*, so it is equal to itself. We can therefore say that the two options in No Change are equally good.

The theory has other advantages too. As I show in other work (Wilkinson, n.d.b, §5), it upholds all of the verdicts of Principal Value Theory(*) (and expected value theory, and Relative Expectation Theory) but provides a strict extension of them: wherever that theory makes a comparison, Invariant Value Theory(*) will agree with it, but it can make many more comparisons as well. Indeed, even the weaker Invariant Value Theory (without the *) can compare a *vast* range of different prospects to others, including any prospect with a symmetric distribution. And the stronger version can compare *any* pair of prospects whose quantile functions are continuous and have their second derivatives bounded above and below close to 0 and 1 (*ibid.* §5). This latter category is extremely broad, and allows us to swap out the Aquila game for Skewed Aquila in the cases above and the theory will still provide answers (more on this below). And I suspect that, even if we develop models of humanity's far future prospects more sophisticated than that described here, those models will still give probability distributions with the necessary property. Or, at the very least, the difference between such distributions available to us will give us a relative difference $R(O_a, O_b)$ with that property. This bodes well for our ability to compare our options in practice if Invariant Value Theory* is true, even if the model I have described here are not perfectly accurate.

5 The outsanding case for longtermism

We now know that, despite the presence of expectation-defying prospects like the Aquila game in practice, we can maintain risk neutrality and still make comparisons in many cases. But *what*, exactly, do these comparisons say? In particular, even if we face prospects like the (Skewed) Aquila game, do those comparisons still justify longtermism?

Recall that the Aquila game itself is unrealistic. It seems deeply implausible, given whatever evidence we might have, that the average future life is equally likely to be negative and to be positive (on any given interval scale). Far more plausible is that lives are more likely to be positive on average, or perhaps negative; either way, they will not be equal. And so too, the overall prospect of the value of the future will be skewed one way or the other, as in the Skewed Altair game.

Recall also the verdict in the third case, Reducing Extinction Risk. By Invariant Value Theory (and by intuition), reducing the probability of extinction is an improvement if and only if survival is better than extinction—if and only if the Skewed Aquila game corresponding to survival is better than the outcome of fixed value 0. But *how much* of an improvement is it? It turns out that, if we compare the Skewed Aquila game to a sure outcome of value 0 (A), the invariant value of R(Skewed Aquila, A) diverges to $+\infty$ or to $-\infty$ (if the game is skewed in the positive or negative direction, respectively). Likewise, if we compare a mixture (O_p) of Skewed Aquila with that outcome of value 0 to another, less probable mixture (O_q) then, again, $\mathrm{IV}\big(R(O_p, O_q)\big)$ is positively or negatively infinite (depending on the direction the game is skewed). So, if we face the prospect of the Skewed Aquila game skewed to the positive (or negative direction) then, not only is it better (worse) to reduce the risk of human extinction than not reduce it, it is *infinitely* better (worse).

This has practical implications in the final case, Multifarious Changes. If we compared two different mixtures of the same Skewed Aquila game but, for one, sweetened every outcome by some finite value v', then the mixture with the higher probability would be better (or worse, if the Skewed Aquila game were skewed to the negative). That holds *no matter* how large v' is, and *no matter* how small the difference between the mixture's probabilities p and q. Given a decision between reducing the risk of extinction, however slightly, and providing some guaranteed benefit, however large, it is always better (or always worse) to reduce the risk of extinction.

This is not the only sort of change that is effectively infinitely valuable. Changing the skew of the Skewed Altair game—shifting the probabilities that the average life in the far future is positive or negative—gains similar importance. Compare any version (SA_1) of the Skewed Aquila game to another version (SA_2) with a more positive skew (a higher ratio of $\frac{a_1}{a_2}$) and the invariant value of $R(SA_2,SA_1)$ will be positively infinite. We could even sweeten every outcome of SA_1 by any finite value, no matter how large, and SA_2 would still be better according to Invariant Value Theory. And the same holds *no matter* how slight the difference in skew between the two prospects.

So, if our future prospects resemble the Skewed Aquila game in the relevant respects, it would not merely be an improvement to change the probability of humanity's survival or the probability of a good future conditional on survival. It would be an *infinite* improvement to even *slightly* change these values.

These verdicts are not merely some quirk of Invariant Value Theory*. They are what a risk-neutral theory *must* say in such cases. Consider a 'Skewed Pasadena' or 'Skewed Agnesi' game (obtained from the standard Pasadena and Agnesi games in the same way, by increasing/decreasing the probability of positive/negative outcomes by a fixed proportion). For such games, Relative Expectation Theory and Principal Value Theory* each say that increasing the games' skew towards positive values is more valuable than sweetening them by *any* finite value. Indeed, the difference between the probability distributions of two such skewed games is roughly analogous to the St Petersburg game, which a genuinely risk-neutral theory must say is better than any finite value (see Hájek and Nover, 2006, p. 706). When facing the Aquila game, if a risk-neutral theory *didn't* give us the above verdicts, we should be sceptical that it was truly risk-neutral!

Given the above, Invariant Value Theory* supports longtermism, even if we do face prospects like the Aquila game in practice. It confirms that the best options available to us, in many important decisions, are those that provide the greatest increases in the invariant value of what happens after T. But given that we *do* face such prospects, it also implies a much stronger conclusion than longtermism—it doesn't just imply that it is often better to improve the far future than the present;

it implies that it is *infinitely better* to do so.

For instance, consider any option that even slightly reduces the probability of human extinction in the near future—perhaps a decision of whether to donate to advocacy efforts against nuclear weapons. If our prospects over the future resemble a Skewed Aquila game, skewed in the positive direction, then such an option will be *infinitely* better than an option that improves the world in the near term with certainty (thereby improving every outcome). Or, if those prospects are skewed in the negative direction, then options that *increase* the probability of extinction will be infinitely better than those that merely improve the near-term future.

Alternatively, consider any option that even slightly changes the probability that future human lives will, on average, have positive value—perhaps this might include a decision of whether to campaign for changes in political institutions. Such an option shifts us from one Skewed Aquila game to another one, with greater skew in the positive direction. This option will be *infinitely* better than any alternative that only improves the world in the near term, even if the latter sweetens the outcome no matter what else happens.

So, if we accept Invariant Value Theory* and we do indeed face prospects resembling the Skewed Aquila game, then our best options will often be those that most improve the far future. Longtermism holds. But not only that; those best options will be *infinitely* better than options that have no effect on the far future. No matter how slight the changes to the parameters of our far future prospects and no matter how great the benefits we could otherwise provide to the near future, our best options will *still* be those that most improve the far future. The case for longtermism thus becomes even stronger.

6 Conclusion

There is reason to think that our prospects for the total moral value of humanity defy expectations—that their expected values are undefined, even if we assume that they can only result in finite value. This is a serious problem for expected value theory as a candidate theory for comparing risky moral options.

And, so too, it may seem to be a serious problem for the moral claim of longtermism. As it is often justified by appeal to expected value theory, or to risk-neutrality more generally, those justifications might be thought to stand or fall with that theory.

One possible response to this is to abandon the verdicts of expected value theory, in favour of some alternative theory that exhibits risk *sensitivity*. By doing so we can, in effect, turn any expectation-defying prospect into a better-behaved one, but at the cost of giving up the theoretical advantages of risk neutrality. But is that the only possible solution to the problem?

It turns out that, instead, we can extend expected value theory to deal with expectation-defying prospects. We can extend it even beyond the existing proposals of Colyvan (2008), Colyvan and Hájek (2016), Easwaran (2008), Easwaran (2014a), and Meacham (2019), each of which carves off some of the remaining pairs of expectation-defying prospects for comparison. And, with Invariant Value Theory*, we can extend the theory far enough to deliver comparisons even for prospects that plausibly describe the future of humanity: (those involving some probability of) the Aquila and Skewed Aquila games.

Given these prospects, if we accept Invariant Value Theory* then the risk-neutral justification for longtermism returns in even greater force. Again, certain options that improve the long-term future will be vastly better than options that only improve the world in the near term. But, when faced with prospects like the Skewed Aquila game, such options will now be *infinitely* better than options that only improve the world in the near term—they will be better no matter how much we could otherwise improve the world in the near term. If we are to maintain risk neutrality even in the face of our real-world moral prospects, then this is the conclusion we are led to—that improving the long-term future is not just valuable; it is vastly, overwhelmingly more valuable than anything else we might ever seek to accomplish.

References

Adler, M. D., 2019. *Measuring Social Welfare: An Introduction.* Oxford University Press. (cited on page 24)

Al-Kindī, 1974. *Al-Kindī's Metaphysics: A Translation of Ya'qūb ibn Ishāq al-Kindī's Treatise 'On First Philosophy'.* State University of New York Press, Albany. (cited on page 3)

Alexander, J. M., 2012. Decision theory meets the Witch of Agnesi. *Journal of Philosophy*, 109, 12 (2012), pp. 712–27. (cited on pages 2 and 4)

Bartha, P. F. A., 2016. Making do without expectations. *Mind*, 125, 499 (2016), pp. 799–827. (cited on pages 3 and 6)

Baumann, T., 2017. S-risks: An introduction. *Center for Reducing Suffering*, available at: https://centerforreducingsuffering.org/research/intro/ (accessed March 2022). (cited on page 11)

Bostrom, N., 2011. Infinite ethics. *Analysis and Metaphysics*, 10 (2011), pp. 9–59. (cited on pages 2 and 5)

Brandenberger, R.; Heisenberg, L.; and Robnik, J., 2021. Through a black hole into a new universe. *International Journal of Modern Physics D*, 30, 14 (2021), 2142001. (cited on page 8)

Broome, J., 1991. *Weighing Goods.* Blackwell. (cited on page 24)

Broome, J., 2004. *Weighing Lives.* Blackwell. (cited on page 1)

Buchak, L., 2013. *Risk and Rationality*. Oxford University Press, Oxford. (cited on pages 19 and 20)

Buchak, L., n.d. How should risk and ambiguity affect our charitable giving? Unpublished manuscript. Available at https://globalprioritiesinstitute.org/lara-buchak-how-should-risk-and-ambiguity-affect-our-charitable-giving/. (cited on page 1)

Colyvan, M., 2008. Relative expectation theory. *Journal of Philosophy*, 105, 1 (2008), pp. 37–44. (cited on pages 3, 21, and 32)

Colyvan, M. and Hájek, A., 2016. Making ado without expectations. *Mind*, 125, 499 (2016), pp. 829–857. (cited on pages 21, 22, and 32)

Cowen, T. and Parfit, D., 1992. Against the social discount rate. In *Justice Between Age Groups and Generations (Philosophy, Politics, and Society)*, pp. 144–61. Yale University Press. (cited on page 1)

Craig, W. L., 1979. Whitrow and Popper on the impossibility of an infinite past. *British Journal for the Philosophy of Science*, 30, 2 (1979), pp. 165–70. (cited on page 3)

de Fermat, P., c. 1659. De aequationum localium transmutatione et emendatione ad multimodaum curvilineorum inter se vel cum rectilineis comparationem, cui annectitur proportionis geometricae in quadrandis infinitis parabolis et hyperbolis usus. In *Œuvres de Pierre Fermat* (Eds. P. Tannery and C. Henry), p. 216–37. Gauthier-Villars. (cited on page 4)

Dyson, L.; Kleban, M.; and Susskind, L., 2002. Disturbing implications of a cosmological constant. *Journal of High Energy Physics*, 2002, 10 (2002), 011. (cited on page 7)

Easwaran, K., 2008. Strong and weak expectations. *Mind*, 117, 467 (2008), pp. 633–41. (cited on pages 3, 23, and 32)

Easwaran, K., 2014a. Principal values and weak expectations. *Mind*, 123, 490 (2014), pp. 517–31. (cited on pages 3, 23, and 32)

Easwaran, K., 2014b. Regularity and hyperreal credences. *Philosophical Review*, 123, 1 (2014), pp. 1–41. (cited on page 6)

Edwards, W.; Lindman, H.; and Savage, L. J., 1963. Bayesian statistical inference for psychological research. *Psychological Review*, 70, 3 (1963), pp. 193–242. (cited on page 6)

Farhi, E.; Guth, A. H.; and Guven, J., 1990. Is it possible to create a universe in the laboratory by quantum tunneling? *Nuclear Physics B*, 339, 2 (1990), 417–90. (cited on page 8)

Frolov, V. P.; Markov, M. A.; and Mukhanov, V. F., 1990. Black holes as possible sources of closed and semiclosed worlds. *Physical Review D*, 41, 2 (1990), 383. (cited on page 8)

Greaves, H. and MacAskill, W., 2021. The case for strong longtermism. Global Priorities Institute Working Paper Series. (cited on pages 1 and 2)

Hájek, A., 2014. Unexpected expectations. *Mind*, 123, 490 (2014), pp. 533–67. (cited on page 5)

Hájek, A. and Nover, H., 2006. Perplexing expectations. *Mind*, 115, 459 (2006), pp. 703–20. (cited on page 30)

Hájek, A. and Smithson, M., 2012. Rationality and indeterminate probabilities. *Synthese*, 187 (2012), pp. 33–48. (cited on page 6)

Harsanyi, J. C., 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63, 4 (1955), pp. 309–21. (cited on page 20)

Jeffrey, R. C., 1983. *The Logic of Decision, 2nd edn.* University of Chicago Press, Chicago. (cited on page 5)

Kendall, D. G., 1948. On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19, 1 (1948), 1–15. (cited on page 9)

Liu, S., n.d. Don't bet the farm: Decision theory, inductive knowledge, and the St. Petersburg paradox. Unpublished manuscript. (cited on page 6)

McNeil, D. R., 1970. Integral functionals of birth and death processes and related limiting distributions. *The Annals of Mathematical Statistics*, 41, 2 (1970), p. 480–5. (cited on page 9)

Meacham, C., 2019. Difference minimizing theory. *Ergo*, 6, 35 (2019). (cited on pages 3, 21, 22, 24, 26, and 32)

Merali, Z., 2017. *A Big Bang in a Little Room: The Quest to Create New Universes.* Hachette UK. (cited on page 8)

Nover, H. and Hájek, A., 2004. Vexing expectations. *Mind*, 113, 450 (2004), pp. 237–49. (cited on pages 2, 3, 4, and 14)

Ord, T., n.d. The edges of our universe. Unpublished manuscript. Available at https://arxiv.org/abs/2104.01191. (cited on page 8)

Parfit, D., 1984. *Reasons and Persons.* Oxford University Press, Oxford. (cited on pages 1 and 23)

Poisson, S. D., 1824. Sur la probabilité des résultats moyens des observations. In *Connaissance des Temps pour l'an 1824*, p. 273–302. (cited on pages 2 and 4)

Pruss, A. R., 2013. Probability, regularity, and cardinality. *Philosophy of Science*, 80, 2 (2013), 231–40. (cited on page 6)

Ramsey, F. P., 1928. A mathematical theory of saving. *The Economic Journal*, 38, 152 (1928), pp. 543–59. (cited on page 1)

Sandberg, A. and Armstrong, S., 2012. Indefinite survival through backup copies. Future of Humanity Institute Technical Report 2012-1. Available at https://www.fhi.ox.ac.uk/reports/2012-1.pdf,. (cited on page 8)

Sidgwick, H., 1907. *The Methods of Ethics, 7th edn.* Macmillan, London. (cited on page 1)

Smith, N., 2014. Is evaluative compositionality a requirement of rationality? *Mind*, 123, 490 (2014), pp. 457–502. (cited on page 25)

Tarsney, C., n.d. Exceeding expectations: Stochastic dominance as a general decision theory. Unpublished manuscript. Available at https://globalprioritiesinstitute.org/christian-tarsney-exceeding-expectations-stochastic-dominance-as-a-general-decision-theory/. (cited on page 20)

Tarsney, C. and Thomas, T., n.d. Non-additive axiologies in large worlds. Unpublished manuscript. Available at https://globalprioritiesinstitute.org/christian-tarsney-and-teruji-thomas-non-additive-axiologies-in-large-worlds/. (cited on page 1)

Tarsney, C. J. and Wilkinson, H., n.d. Longtermism in an infinite world. In *Essays on Longtermism*. Available at https://globalprioritiesinstitute.org/longtermism-in-an-infinite-world-christian-j-tarsney-and-hayden-wilkinson/. (cited on page 3)

Thomas, T., 2022. Separability and population ethics. *The Oxford Handbook of Population Ethics*, (2022), 271–296. (cited on page 1)

Thomas, T., n.d. The asymmetry, uncertainty, and the long term. Unpublished manuscript. Available at https://globalprioritiesinstitute.org/teruji-thomas-the-asymmetry-uncertainty-and-the-long-term/. (cited on pages 1 and 20)

Vilenkin, A., 1983. Birth of inflationary universes. *Physical Review D*, 27, 12 (1983), 2848. (cited on page 8)

von Neumann, J. and Morgenstern, O., 1953. *Theory of Games and Economic Behavior, 2nd edn*. Princeton University Press, Princeton. (cited on page 17)

Wilkinson, H., 2021. *Infinite Aggregation*. Ph.D. thesis, Australian National University. (cited on page 3)

Wilkinson, H., 2022. In defence of fanaticism. *Ethics*, 132, 2 (2022), p. 445–77. (cited on pages 20 and 23)

Wilkinson, H., n.d.a. Can an evidentialist be risk averse? Unpublished manuscript. (cited on page 20)

Wilkinson, H., n.d.b. Flummoxing expectations. Unpublished manuscript. (cited on pages 28 and 29)

Wilkinson, H., n.d.c. Separating separabilities. Unpublished manuscript. (cited on page 23)

Williamson, T., 2000. *Knowledge and Its Limits*. Oxford University Press, Oxford. (cited on pages 5 and 6)

Zhao, M., 2021. Ignore risk; Maximize expected moral value. *Noûs*, (2021). (cited on pages 17 and 20)