Should Effective Altruists Always Maximize Expected Value?

Claire Field

University of Zürich

Expected Value Theory is often assumed to offer the comprehensive theory of rational choice. If this is true, then rationality requires that we sometimes take enormous risks – specifically, when we are faced with options that have very low probability and very high value. This implication underpins some recent ethical views about our obligations to the extreme long-term future: consequentialist commitments to impartiality and doing the most good as efficiently as possible appear to imply that effective altruists should prioritize the extreme long-term future. I argue that the reasoning that leads to this conclusion rests on an overly narrow understanding of rationality, one that leaves out substantive and structural rational requirements. I show how, once we take account of these, we see that rational effective altruists should not always maximize expected value.

Keywords: rationality; risk; coherence; epistemic attitudes; effective altruism.

Expected Value Theory is often assumed to offer the comprehensive theory of rational choice. If this is true, then rationality requires that we sometimes take enormous risks – specifically, when we are faced with options that have very low probabilities and very high values. Expected Value Theory says that rational agents maximize expected value, even when the probability of success is low. In this paper, I argue that Expected Value Theory does not give the full account of what rational agents ought to do. Specifically, it is unable to capture some important further aspects of rationality: substantive and structural rational requirements.[1] While Expected Value Theory is a useful tool for comparing options, I argue here that it does not offer the complete story of rational choice. While I am happy to accept that there is an important sense in which the option that has the most expected value is 'best', I do not think that it is always rational to choose the option that maximizes expected value, even if it is best.

Instead, what I suggest here is a two-stage process of rational decision-making. First, the agent should filter out all the options that would conflict with substantive or structural

[1] For discussion of such requirements, see Broome 1999, 2013; Kiesewetter 2017; Lord 2018; Way 2011; Wedgwood 2017.

requirements of rationality. Once this has been done, she is free to maximize expected value on the remaining options.

§1 explains how the assumption that Expected Value Theory gives the comprehensive theory of rational choice leads to surprising implications that are particularly relevant for effective altruists. §2 examines the limitations of existing strategies for dealing with these implications. §3 outlines three additional rationality requirements and shows how they prevent the surprising implications of Expected Value Theory. §4 summarizes and shows how the view I argue for improves upon Expected Value Theory as a theory of rational choice for effective altruists by avoiding 'fanatical' results.

1 Rationality as Expected Value Maximization

According to Expected Value Theory, rational agents maximize expected value, and any rational agent can be represented as an expected value maximizer. Expected value of an option is calculated by multiplying the probability of each outcome with its value, and summing all these probability-weighted values together. Assuming that according to the correct axiology each life has equal (moral) value, each of the following options has equal expected value:

(a) Save 100 people for sure.
(b) 10% probability of saving 1000 people (and 90% probability of saving none).
(c) 1% probability of saving 10,000 people (and 99% probability of saving none).
(d) 0.1% probability of saving 100,000 people (and 99.9% probability of saving none).
(e) 0.01% probability of saving 1,000,000 people (and 99.99% probability of saving none).
(f) 0.001% probability of saving 10,000,000 people (and 99.999% probability of saving none).

If Expected Value Theory offers the comprehensive theory of rational choice, it would be equally rational for someone who valued each additional life equally to choose any of these options. According to Expected Value Theory, rational agents are risk neutral: they are indifferent between options with the same expected value, but different levels of risk. So, if you care about saving as many lives as possible, rationality permits you to freely choose any of options (a) – (f) or any other option with equal expected value, including smaller probabilities of saving larger numbers of lives that would have equal expected value.

Furthermore, if we were to adjust one of these options so that even more lives are potentially saved (even just one more), then Expected Value Theory would require someone who cared about saving as many lives as possible to choose that option, even when the choice is made only once.[2] It is a well-known feature of Expected Value Theory that it sometimes requires us to choose options that have extremely low probability of success over options with very high probabilities of success (Allais 1953; Nover and Hájek 2004; Quiggin 1982; Schmeidler 1989). This feature has motivated some to propose amendments and alternatives to Expected Value Theory (Buchak 2013; Monton 2019).[3]

Here, I focus specifically on the implications of this feature of Expected Value Theory for effective altruism. Effective altruism encompasses a range of moral positions focused on doing the most good as efficiently as possible. In their encyclopedia entry on the topic, MacAskill and Pummer define effective altruism as "the project of using evidence and reason to try to find out how to do the most good, and on this basis trying to do the most good". They further clarify that effective altruism is impartial about how this good is distributed, and broadly welfarist, in that the good to be promoted is well-being, broadly construed (2020: 2). Within these constraints, various versions of effective altruism are possible.

Effective altruism can be understood as involving two claims: a moral claim about what agents ought to aim for in acting (i.e. altruism), and a practical rationality claim about how agents ought to go about achieving that aim (i.e. effectively). Regarding the practical rationality claim, MacAskill and Pummer say only that effective altruists ought to 'use evidence and reason' to figure out the most efficient way to do the most good. In other words, we might think, they should act rationally.[4] While this is consistent with many different theories of rational choice, in practice, effective altruists have tended to assume that the correct theory of rational choice is Expected Value Theory.[5] This paper scrutinizes this assumption. Specifically, I raise problems that arise from that assumption for versions of effective altruism committed to the following two claims:

Moral Claim: You should aim to do the most good possible, be impartial about when, where, and how this good arises, and act as a rational agent with these concerns would.

Rationality Claim: Expected Value Theory offers the full account of rational choice. Namely, that rational agents maximize expected value.

[2] Assuming that (a) – (f) represent the agent's only available options.

[3] Two such proposals discussed here (§2) here are Buchak's risk-weighted decision theory (2013) and Monton's discounting strategy (2019) for outcomes with very low probabilities.

[4] Elsewhere, when discussing how to act under moral uncertainty, MacAskill, Bykvist, and Ord (2020: 329) describe the correct action to choose under moral uncertainty as that which would be chosen "by a rational and morally conscientious agent", indicating further that by 'effective' something like 'rational' is meant.

[5] For example, see MacAskill (2022: 53), Wilkinson (2022). One notable exception is Pettigrew (2024).

Together, these claims say that one should act in such a way that would maximize expected value for someone who had the ideal moral concerns. That is, someone concerned to do the most good impartially. This is true regardless of whether *you* in fact have these noble concerns. Here, I scrutinize the Rationality Claim. In order to do so, I will assume here that the Moral Claim, or something close to it, is true. For the purpose of this discussion, I assume a fairly simplistic background moral theory according to which outcomes are morally better when they bring about more happy lives, and extra happy lives are valued linearly, such that extra happy lives always make for a morally better outcome. In doing this I do not mean to imply that this is the only possible view that could be called effective altruism.

Commitment to the Rationality Claim has previously led effective altruists to advocate causes aimed at saving large numbers of lives cheaply, such as preventing malaria, alleviating global poverty, or ending factory farming. For example, effective altruists have emphasized how if one cared impartially about doing as much good as possible, spending $3 on small luxuries (such as coffee) generates much less expected value than donating the same $3 to a charity that could use that $3 to, for example, buy a bed net that would prevent a child dying prematurely from malaria. Expected Value Theory shows how this is true even when we cannot be certain about that our $3 donation will have the desired effect. Even effective charities might encounter unexpected costs or setbacks that prevent one's donation doing as much good as one hoped. Suppose one donates $3 to the Anti-Malaria Foundation. However, suppose there is only a 75% probability that your donation will actually go towards this cause, rather than being wasted inefficiently. Nevertheless, Expected Value Theory still evaluates donating to this charity as more rational for the ideal effective altruist than spending the $3 on some small luxury. Suppose, using arbitrary numbers, that the value of saving a child from malaria is 7000 (set the value of one year of happy life at 100, and assume that if the bed net saves the child from malaria, they go on to live an extra 70 happy years). If there is only a 75% probability of this happening, we multiply this value by 0.75 to get 5250. For the ideal effective altruist who values doing the most good impartially, it is not plausible that the pleasure they would get from a small luxury such as coffee could be this much. This reasoning shows how it would be rational for the ideal effective altruist to choose to donate to the Anti-Malaria Foundation over buying a coffee.

So far, so plausible. However, more recent strands of effective altruism have used the same expected value reasoning to move away from these traditional causes aimed at saving many lives efficiently now, and towards causes focused on improving the extreme long-term future. For example, causes aimed at preventing or reducing the risk of premature human extinction. This is because the extreme long-term future potentially contains a vast number of people and other creatures, all of whom have the potential to live long, happy lives if things go well. Longtermists argue that if we are genuinely impartial about the good that we do, this should extend not only to where the good occurs, but also when. Since the long-term future is so vast, and potentially contains so many lives, the ideal effective altruist should regard an option in which there is a probability (even a very small probability) to ensure that there are very many happy lives in the future as a better prospect than one that brings about a smaller number of happy lives now.

For example, despite its extreme unlikelihood, an option such as the following should be regarded by the rational effective altruist as more enticing than any of (a) – (f):

> (g) 0.0001% probability of saving 100,000,000,000,000,000,000,000,000 people (and 99.9999% probability of saving none).

This option has expected value equivalent to saving 10,000,000,000,000,000,000,000 lives, much greater than options (a) - (f). This implication of effective altruism – that we are sometimes morally required to choose extremely low probabilities of very high values – is sometimes referred to as 'fanaticism'.[6] An option such as (g) can be described as 'fanatical' because it is a very tiny probability of a very large value, and it has greater expected value than any alternative available option (i.e. (a) – (f)) with greater probability of value. Bostrom characterizes the view particularly vividly when he notes that "even the tiniest reduction of existential risk has an expected value greater than that of the definite provision of any 'ordinary' good, such as the direct benefit of saving 1 billion lives" (Bostrom 2013: 18).[7] This result is indeed striking – if one cares about ensuring as many happy lives as possible, choosing (g) would mean doing something that you should, rationally, be almost certain will bring about no value – you know that there is a 99.999% probability of achieving nothing. Meanwhile, you would be giving up the opportunity to choose (a) and save 100 people for sure. Although option (g) maximizes expected value, it also carries a high risk of squandering the opportunity to save anyone at all.

This is a familiar feature of Expected Value Theory. The following section evaluates existing strategies for dealing with its implications, arguing that these strategies face common limitations. Specifically, they lack principled motivation for departing from Expected Value Theory, and they are insensitive to the epistemic significance of the different magnitudes of probability. After outlining these limitations of existing approaches, I argue that a good replacement for effective altruism's Rationality Claim would be the two stage process described above. That is, rational agents maximize expected value but only on options that have not been ruled out by substantive or structural requirements of rationality.

[6] See Balfour 2021; Bostrom 2009; Wilkinson 2022.

[7] Of course, it's not at all obvious that saving people is equivalent to preventing the continued non-existence of future people. Important worries for longtermism arise from the non-identity problem (Boonin 2014; Parfit 1982, 1984, 2017; Mulgan 2006; Woollard 2012). Since my focus is on the Rationality Claim, I leave these aside here.

2 Existing Strategies

That Expected Value Theory sometimes requires rational agents to choose options with extremely low probability of success over options with high probabilities of success is widely acknowledged. In addition to the 'fanaticism' worries raised by effective altruists, it has also been discussed in the context of Pascal's Wager, the St Petersburg Game, the Pasadena game, and the Allais Paradox. This section discusses some of the most compelling attempts to accommodate, within decision theory, the sense that maximizing expected value recommends excessively risky options. I conclude that none of these existing strategies are completely satisfying; all leave out some important features that a theory of rationality ought to be sensitive to. I am by no means aiming to rule out every possible strategy for accommodating what needs to be accommodated within decision theory – the aim is to motivate an alternative approach, one that considers substantive and structural requirements of rationality.

2.1 Representation

A common way to respond to this feature of Expected Value Theory is by disputing how the decision situation is represented. For example, if the outcomes involve finite goods, such as money, then we would not expect a rational agent to value these linearly. Most rational agents value additional finite goods only up to some limit. In other words, there is an upper bound to how much value finite goods can provide. At some point, our marginal utility for finite goods diminish fast enough so that our total utility never exceeds an upper bound. This idea can thus explain how someone could rationally reject a tiny probability of a huge amount of some finite good, like money, in favour of a higher probability of a smaller amount, even when the option with tiny probability has greater expected value. The explanation is that the agent does not, and so ought not be represented as, valuing each additional unit of the finite good linearly. Because they do not value the finite good linearly, higher probabilities of smaller amounts of it have greater expected value for them.

However, effective altruism's Moral Claim is based on what the ideal effective altruist values, and thus would be rational to do. It says that we ought to do what someone who valued doing the most good impartially would be rational to do, even if we ourselves do not share those values. With this in mind, we might take intuitive resistance to low probability gambles with high expected value as evidence of our moral failings – evidence that most of us do not have the values that we ought to have. We can tell we do not have the values we ought to have because we do not value tiny probabilities of huge numbers of happy lives as much as we value the certainty of saving of fewer lives, even when these tiny probabilities have greater expected value.[8]

[8] For some valiant attempts to push our intuitions towards accepting fanatical results, see Bostrom 2013 and MacAskill 2022. For an attempt to resist intuitions against risk neutrality in decision-theory more generally, see Nover and Hájek (2004: 241). Alternatively, we might take this as evidence of a rational failing: incompetence with probabilistic reasoning, or susceptibility

Nevertheless, resistance to prioritising tiny probabilities of high value is not easily dismissed, and indeed has motivated research programs within decision-theory aiming to avoid these counter-intuitive results (Allais 1953; Buchak 2013; Monton 2019; Quiggin 1982; Schmeidler 1989). The following subsections discuss some of these.

2.2 Discounting

Another existing strategy is to amend Expected Value Theory so that very unlikely outcomes below some minimal probabilistic threshold are discounted: that is, treated as if they had 0 value.[9] This strategy rationally prohibits choosing highly unlikely options, thus vindicating the sense that these are not rational choices. Unfortunately, discounting introduces problems of its own.

For a start, as Kosonen (2024) points out, discounting violates some central axioms of decision theory. This makes it not particularly helpful it is as a strategy to preserve Expected Value Theory against challenges arising from fanatical results. Not only this, but sometimes it can be rational to choose extremely unlikely options. For example, if we know we only have one desperate shot at a good outcome, and no alternatives. Suppose you are in a burning building and your only chance of survival is to jump from a window. You are highly unlikely to survive, but certain to die if you do not. In this case, maximizing expected utility (without discounting) seems rational, despite the slim chance of survival.[10]

Discounting approaches cannot distinguish between different outcomes of low probability if all are below the minimal threshold.[11] This generates some implausible results. Consider:

(1) 0.0001% probability of saving 100 lives

(2) 0.0001% probability of saving 1000 lives

It would clearly be rational to choose (2) over (1), but traditional discounting approaches cannot accommodate this if a threshold of 1 in a million is used, since both options are

to bias or framing effects (Kahneman, Slovic, and Tversky 1982; Gilovich, Griffin, and Kahneman 2002).

[9] See Monton's discounting strategy (2019) and Smith's Rationally Negligible Probabilities principle (2014).

[10] We might think that something like this thought is what motivates Ord (2020) to argue that we ought to prioritize the long-term because we are on a 'precipice', and this is our only opportunity to prevent extinction in the long-term. MacAskill (2022) says similar things about the opportunity to prevent undesirable value lock-in through malicious AI. However, these motivations are much less convincing as a justification for choosing an option such as (g) when we appreciate how small the probabilities are, and what we would be giving up in order to do it. Notably, most of the long-term orientated interventions that Ord regards as valuable have comparably high probabilities of success.

[11] Cibinel (2023) and Isaacs (2016), make similar points. Isaacs' case differs from this one in that it turns on the irrationality of ignoring tiny probabilities of very bad outcomes, but the central point is the same.

below the minimal threshold of 1 in a million. On a threshold-based discounting approach, both must be treated as worthless.

Monton has an answer to this. He proposes that one should choose (2), but only if it is "cost-free". (2019: 20). However, this might depend on the cost. Perhaps it would be reasonable to pay a very small price to choose (2) rather than (1). Similarly, Beckstead and Thomas suggest that discounting views could avoid this difficulty by allowing value below the threshold to act as a "tie-breaker". This would vindicate the sense that Option (2) is a more rational choice than Option (1). However, as they point out (2024: 443), this would introduce further complications when choosing between options that have low probabilities lying either side of the threshold.

Indeed, it is difficult to set the threshold in the right place, so that it says the right thing about the relevant cases. Monton (2019) sets his threshold at 1 in a million. Historically, others have sent it higher.[12] However, it is not clear that this threshold would be high enough. Option (c) involves a 1% probability, and we might doubt whether this a rational choice when presented with the opportunity to save 100 people for sure. Nor is it clear that this can be resolved merely by adjusting the level of the threshold. Rather, what this suggests is that there are important contextual features of decisions involving extremely unlikely possibilities that discounting strategies cannot capture. For instance, choosing the low-probability options seems irrational not merely because the probability of success is low, but rather because it means squandering the opportunity to save 100 people for sure. If the choice was instead between, say (d) (a 0.1% probability of saving 100,000 people (and 99.9% probability of saving none)) and something much less attractive, such as an even smaller probability of saving 10,000 people, or nothing at all, then it would not be irrational to choose (d). In other words, it does not seem right to say that unlikely options should be discounted entirely. Rather, how we should treat them depends on what else we know about the situation. For example, what we would be sacrificing in order to take that option.

2.3 Risk Sensitivity

Buchak (2013) offers an alternative 'risk-weighted' decision theory that permits agents to be more risk-averse than Expected Value Theory demands – when that aligns with their risk attitudes. Risk-weighted decision theories evaluate the rationality of decisions by reference not only to the agent's preferences and credences, but also her risk attitudes: how risk-averse or risk-seeking she is. This allows it to vindicate choices that are more risk-averse (or more risk-seeking) than Expected Value Theory requires. Risk-weighted decision theory thus offers a way of rationalizing someone who chooses a sure option such as (a) over a riskier option with more expected value. We can say that someone who chooses in this way is risk-averse, and such a choice is thus rational in virtue of aligning with their risk attitudes.

[12] For example, Buffon sets it at 1 in 10,000 (1777: §8).

However, while this strategy is able to rationalize choosing less risky options, it does not by itself allow effective altruists to avoid commitment to Fanaticism. For any particular risk attitude we might choose, it will still be possible to construct additional risky options with sufficient value such that they would be rational even for a risk-averse agent.[13] Risk aversion means that risky options are weighted less than less risky options, but they can still maximize expected value for a risk-averse agent if enough value is added to sweeten the deal. Indeed, Buchak acknowledges this, and suggests dealing with it by discounting very small probabilities down to zero (2013: 73). As argued in the previous section, this carries its own problems.

There are also more fundamental worries about the general strategy of appealing to the agent's risk attitudes to vindicate the rationality of risk-averse choices. Effective altruists are interested in what it would be *rational* for someone impartially concerned to do the most good to do on the assumption that rationality can provide guidance on the most e*ffective* means to bringing about the most good. However, risk-weighted decision theory evaluates the rationality of decisions based on whether they align with the agent's preferences, credences, and risk attitudes. Whether a choice aligns with my risk attitudes is not the same thing as whether it is the most effective means for me to get what I want. Expected Value Theory tells me that the best means to my ends is to maximize expected value risk neutrally, because this is the most effective way to get more of what I want, given my preferences and my information about the world. Risk-weighted decision theory tells me only which means to my ends cohere best with the attitudes I have towards different means of getting ends (i.e. my risk attitudes). While it may well be, as Buchak argues, reasonable to have attitudes to how we achieve our ends, and thus rational to make choices in line with our risk attitudes, this is a different sense of "best means to our ends" than the one that *effective* effective altruists should be interested in. What is the most efficient means to my ends does not depend directly on my psychological attitudes to risk, even if Buchak is ultimately right that the decision it is rational for me to take does depend on this.

The following section outlines a proposal according to which rational requirements on substantive and structural rationality act as constraints on which options agents can maximize expected value over. This account is able to vindicate the sense that choosing 'fanatical' options is irrational.

3 Substantive and Structural Requirements of Rationality

A long tradition says that rationality is a matter of having a coherent and intelligible first-person perspective.[14] As Wedgwood puts it:

> Rationality is a kind of virtue displayed in some of the mental states (like the beliefs and intentions) that agents have, and in the ways in which

[13] That is, assuming we add also the assumption that the risk function is strictly increasing and defined on the reals.

[14] See Broome 2007, 2009, 2013; Kiesewetter 2017; Kolodny 2005, 2007; Lord 2018; Way 2011; Wedgwood 2017.

> agents form and revise those mental states in response to reflection and experience. (2017: 2)

Indeed, decision-theory can be seen as aiming to preserve this idea. It aims to tell you what it would be rational for you to do, given a subset of your attitudes: your preferences, credences, and – perhaps – your risk attitudes. Beyond decision theory, this idea has been captured by *structural* requirements of rationality: requirements that govern relations between attitudes.

Not only this, but rational agents comply with *substantive* requirements of rationality: they respond appropriately to their reasons and they believe what their evidence supports.[15] Responsiveness to reasons and evidence is sometimes viewed as part and parcel of having a coherent first-person perspective (Lord 2018; Kiesewetter 2017; Wedgwood 2017) and sometimes as something that conflicts with it (Lasonen-Aarnio 2014, 2020; Weatherson 2019; Worsnip 2018). Decision theory, perhaps, aims to capture this idea by considering the agent's credences. As I argue in this section, according to this broader tradition of theorizing about rationality, agents are subject to requirements of rationality that constrain which options can be rationally chosen. In other words, rationality cannot *only* be a matter of maximizing expected value, since some options that maximize expected value, if chosen, would violate substantive and structural requirements of rationality.

At this point, one might wonder whether effective altruists should care about having substantively or structurally rational attitudes. Indeed, there is a sense in which a fully committed effective altruist should be willing to sacrifice some rationality if she believes it will bring about most good. For example, if substantive rationality requires always proportioning one's beliefs to the evidence, but an eccentric philanthropist offers to make a large charitable donation if I form a belief not supported by the evidence, then if I am committed to doing the most good I should surely endeavour to adopt a belief against the evidence, thus sacrificing a little of my rationality. However, accepting that a good effective altruist should sometimes sacrifice a little of her rationality in order to do the most good is entirely compatible with the claim I am defending here: namely, that a full answer to what it would be rational for someone to do must take into consideration requirements of substantive rationality on our epistemic attitudes. This is because there are some choices that, given the agent's rational epistemic attitudes, would not be rational for them to take. To take a very simple example, if I believe that flicking the light switch will set off a bomb, and I believe that I should not set off a bomb, then it would not be rational for me to flick the light switch. Doing so would be irrationally incoherent, given my beliefs and goals.

In this section, I outline two ways that substantive and structural requirements of rationality constrain which options can be rationally chosen. Specifically, I focus on requirements on the attitudes one should have, given particular rational credences and

[15] Or, as Hume put it, "the wise man [...] proportions his belief to the evidence" (Hume 1748/1861: 78). More recently, see Brown 2018; Clifford 1877; Williamson 2000.

requirements of enkratic coherence. These requirements, I claim, outline important aspects of rationality and rational choice, and they are important constraints on the options that a rational agent impartially concerned to do the most good can rationally choose.

First, requirements of substantive rationality. Rational agents proportion their beliefs to their evidence - they comply with requirements of substantive rationality. Different magnitudes of credence will permit different coarse-grained attitudes. Sometimes, rational credences licence coarse-grained attitudes that are rationally incompatible with choosing an option that maximizes expected value. This means that sometimes it is (substantively) irrational to maximize expected value.

Rational credences constrain the attitudes it is rationally permissible to have. For example, different magnitudes of credence have different implications for course-grained epistemic attitudes such as belief, disbelief, or suspension. Something you have a 0.9 credence in is something you should regard as more likely to be true than something you have a 0.1 credence in, and your coarse-grained epistemic attitudes should reflect this. However, decision theory has no mechanism for distinguishing between these differences in epistemic significance. An option which you have credence only 0.1 that it will result in a better outcome than some alternative can thus come out just as good as an option you have 0.9 credence in, provided it is multiplied by a big enough value.

Minimally, if the rational credence to have in P in a particular situation is 0.9, this means that the situation affords evidence that supports P to 0.9. Since beliefs should be proportioned to evidence, information about which credences are rational is also information about which beliefs are evidentially supported. For example, a 0.9 credence in P plausibly permits the belief that P is likely and prohibits the belief that P is false. A 0.1 credence in P permits and prohibits a very different set of attitudes. There are various possible views of exactly how credences and coarse-grained attitudes ought to combine. On threshold views, a 0.9 credence in P puts you above the threshold required for rational full belief. On views of knowledge that does not require certainty, a rational high credence can be sufficient for knowledge.[16] Similarly, the idea that it is rational to believe that a very unlikely event will not occur has found wide support,[17] and some have even thought that we can *know* this.[18] Some have identified particular coarse-grained attitudes with particular levels of credence. For example, some have claimed that doubt that P is equivalent to low credence that P (Bricker 2022: 38). Others have thought that all credal states below 1 are merely partial belief and are not compatible with full belief (see Maher 1993; van Frassen 1995). Taking a different approach, others have argued that credence and belief bear no relationship to each other (Buchak 2014). Nevertheless, all these different views on the relationship between belief and credence must recognize some

[16] See Hong (2024) for a detailed argument from standard fallibilist and anti-skeptical accounts of knowledge against fanaticism. See Brown (2018) for a fallibilist account of knowledge.

[17] For example, this is often taken for granted in discussions of the lottery paradox (see Kyburg 1961; Foley 1979; Nelkin 2000), though Ryan 1996 provides a rare argument against it.

[18] See Nelkin 2000.

minimal rational requirements on belief, given particular credal states. A 0.9 credence in P rationally permits at least the belief "P has probability 0.9" and "P is very likely". It prohibits beliefs inconsistent with this, such as "P is very unlikely" or "P has probability 0.1". Similarly, a 0.1 credence in P rationally permits the belief "P has probability 0.1" and "P is very unlikely" and rules out beliefs such as "P is very likely" or "P has probability 0.9".

To meet standard requirements of substantive rationality, these beliefs about likelihood need to be integrated within the wider perspective of a fully rational agent. For example, some beliefs about likelihood are incompatible with choosing options that maximize expected value. Having such beliefs rationally prohibits choosing some options that maximize expected value.

Expected Value Theory does not capture these rational constraints on coarse-grained attitudes, because it handles all credences in the same way. To calculate expected value, credences are multiplied with the value of each option. Expected Value Theory calculates the value of each option in the same way, regardless of the particular magnitudes of the credences involved. This is not a shortcoming of Expected Value Theory - there are good reasons to simplify a theory of practical decision making by using only credences, rather than attempting to also incorporate coarse-grained attitudes. For example, doing so makes calculating expected value much easier, and avoids any need to appeal to intuitive judgments of rational decision making. However, recognising this advantage of a simplified theory does not mean we need to assume that Expected Value Theory is the comprehensive theory of practical rationality. Sometimes, constraints on rational coarse-grained attitudes make it irrational to choose an option that maximizes expected value. These constraints function as restrictions on Expected Value Theory, restricting the options that it is permissible to maximize expected value over. Sometimes, options are ruled out in this way because of the conflict this would produce with rational coarse-grained attitudes. Choices that we are rational to think highly likely to be failed attempts to bring about value, that we perhaps even know will be failed attempts, are not choices that it can be rationally permissible to choose, even if they maximize expected value. Since Expected Value Theory treats all credences in the same way, this insensitivity to the rationality of coarse-grained attitudes is a predictable consequence of endorsing a theory of rationality that relies on maximizing expected value as the sole guide to what is rational.

Second, requirements of structural rationality. The idea that rationality is a matter of having an intelligible first-person perspective has been captured by various specific coherence requirements prohibiting particular combinations of attitudes. The most obvious of these is the requirement to avoid contradictory beliefs.

Non-Contradiction Requirement Do not believe both P and not-P

However, there are also more general coherence requirements prohibiting combinations of attitudes that, while not strictly logically inconsistent, exhibit some kind of tension when held by the same person at the same time. Coherence requirements prohibit combinations of attitudes that, in some sense, "don't fit together right" (Worsnip 2021: 3), though it proved difficult to give a more precise definition of this. Indeed, some philosophers who endorse structural rationality requirements explicitly endorse appeal to intuition in defining coherence requirements.[19] One example of these more general coherence requirements is enkratic requirements demanding coherence between our beliefs about what we ought to do or believe, and what we actually do (or intend to do) and believe.

Enkratic Requirement $O(BO\varphi \rightarrow \varphi)$[20]

Reading O as "rationally required" and B as belief, agents violate this principle when their normative beliefs about what they ought to do, intend, or believe are out of line with their first order attitudes. For example, I violate the Enkratic Requirement if I believe that I ought to go to the gym but fail to form the intention to go to the gym. I also violate the Enkratic Requirement if I believe that I ought, rationally, believe P but I fail to do so. Enkratic requirements aim to capture the idea that rational agents intend to do what they believe they ought to do and avoid having beliefs that they regard as irrational.

Another example of a violation of these more general coherence requirements is "Moorean" absurdity, captured in assertions such as "P, but I don't believe that P" or "P, but I don't know that P" (see Smithies 2012). Yet another is having beliefs that one regards as not supported by one's evidence (see Adler 2002; Owens 2002). These combinations of attitudes are not *logicall*y inconsistent, but they nevertheless seem incompatible with having a fully rational first-person perspective.

This more general sense of incoherence is intuitively recognizable, but difficult to define precisely. Worsnip, noting this, suggests defining incoherence in the following way:

> a set of attitudes is incoherent just if it is constitutive of these attitudes that anyone holding them is disposed to revise at least one of them, under conditions of full transparency (2021: 133).

In other words, combinations of attitudes are incoherent when we would not happily hold them together. Rational agents, intuitively, do not hold such attitudes.

[19] For example, Broome notes that in defining coherence requirements we are "forced to appeal largely to our intuition" (2013: 150).

[20] There is significant debate over exactly how to formulate the Enkratic Requirement. For example, whether it should be read wide or narrow scope, and what kind of first-order attitudes it can or should range over (Broome 1999, 2007; Kolodny 2005, 2007). Here, it is stated wide-scope and as generally as possible. Whichever way it is read, it lies outside Expected Value Theory.

Using this standard analysis of incoherence, we find instances of it if we think that rational agents must *always* choose the option that maximizes expected value. Specifically, when we imagine agents choosing a low probability, high expected value option such as (g), these agents will sometimes have incoherent combinations of attitudes. Specifically, the desire to do the most good, and the intention to do something that they know is highly unlikely to bring about any good. Someone who makes such a choice would be correct to assert and believe statements such as the following, all of which have a Moorean flavour: "I am choosing E, but E is very unlikely to pay off", "I am choosing E, but E is extremely unlikely to bring about any good", or "I am choosing E, and I know that E will achieve nothing". These assertions seem strange for someone primarily concerned with doing the most good. Nevertheless, they are correct descriptions of the agent's actions, given the beliefs about their action that they are rational to hold.

An ideal effective altruist who chooses such an option will, rationally, hold the following attitudes:[21]

1. Desire to do the most (actual) good.
2. Belief that Option X is highly unlikely to bring about any (actual) good and will most likely achieve nothing.
3. Intention to choose Option X.

I think it is plausible that a fully rational agent in conditions of full transparency would be disposed to revise one of these. There are different ways to revise these in order to maintain coherence. Not all of them are available given the constraints of the example: the ideally rational effective altruist facing a choice between a very low probability of a huge amount of expected value and a certainty of lower expected value. For instance, they could revise the desire so that they desired to maximize expected value, rather than do the most good – but this desire is given to us by effective altruism's Moral Claim. Or, they could revise the belief so that they believed Option X would be more likely to be successful in bringing about good – but this belief is underpinned by rational credences. We might also assume that the agent holds an additional theoretical belief that maximizing expected value is the only or best way to do the most good - this would make them at least coherent, but it is of course a further question whether this is true, or whether they should believe this. This leaves the option of revising the intention. So, it seems that the ideally rational effective altruist should not choose Option X. So, they should not choose a very low probability of a huge amount of expected value over a certainty of lower expected value.

[21] The incoherence can be made even more apparent if we consider knowledge rather than belief. For example, replacing (2) with something like "*knowledge* that Option X *will* bring about nothing". Hong (2024) discusses such cases in more detail, arguing too that they prevent Expected Value Theory licensing fanatical decisions.

At this point it is worth addressing some objections to my claim that it is irrationally incoherent for someone who desires to do the most good to choose an option that is very unlikely to bring about value.

First, it might be noted that there are some ways of fleshing out the details of the situation such that a combination of attitudes such as (1) – (3) could be rational. Sometimes, we might think, it could be rational to choose options that we know are highly likely to fail. Suppose you are in a burning building and you know that your only chance of survival is to jump from a window. Given the height, you are highly unlikely to survive, but if you do not jump you are certain to die. If this is the situation, then it is not irrational to jump, even though you should believe that you have only a slim chance of success. However, this is because you have no alternatives. This is thus a very different situation from the one we are considering, and we might express it by adding a fourth attitude to the set, which makes the set consistent. For example:

> (4) Belief that Option Y is likelier than Option X to be successful in bringing about good.

With the addition of (4), this becomes a very different case to the one faced by the ideal effective altruist choosing between many competing causes, many of which have good chances of success.

Second, one might worry that there cannot be structural requirements of rationality, particularly for effective altruists, because we can easily imagine cases in which it would be practically rational for someone to have incoherent attitudes. However, in appealing to standard accounts of structural rationality, I do not mean to imply that it could *never* be rational for someone concerned to do the most good to have incoherent attitudes. Of course, if an eccentric philanthropist offers to donate a huge sum to a charity of my choice if I believe a contradiction, then I should endeavor to believe a contradiction. This would be practically rational in so far as I desire to do good. However, this case of incoherent attitudes is importantly different to the one under discussion. The difference lies in what the agent knows about the situation. When I endeavor to believe a contradiction in order to secure the philanthropist's charitable donation, I know that by believing the contradiction I will bring about good. This case is importantly different from the cases of structural incoherence under discussion: namely, combinations of a desire to do good and an intention to do something known or rationally believed to be highly unlikely to being about good. The agent who believes the contradiction so that the eccentric philanthropist will donate to charity knowingly takes the best means available to her to bring about good. Thus, her desire to do good is coherent with her intention to take the best available means to doing good. Indeed, she maximizes expected value. However, an agent who desires to do good and is faced with a fanatical option cannot coherently choose the option that maximizes expected value. This is because they rationally ought to believe that choosing the fanatical option would be a failure, or at least be highly likely to be a failure. Unlike the agent who believes the contradiction, it is not at all clear that they would not be knowingly taking the best available means to bringing about good (assuming they have some other non-fanatical option available). So, agents who desire to do the most good, and are faced with fanatical options, cannot choose the option that maximizes expected value and remain structurally coherent in their desires and intentions. While a commitment to effective altruism might sometimes ask us to believe contradictions, it should not ask us to do things that, as rational agents, we should believe will fail to bring about our ends. If it did, this would inhibit its ability to offer useful guidance.

Third, one might worry that this line of reasoning overgeneralizes, making *any* choice of an option for which the balance of probabilities is against it occurring (i.e. less than 50%) would be irrational. After all, if "I am choosing E , but E is very unlikely to pay off" indicates structural irrationality, then perhaps "I am choosing E, but E is unlikely to pay off" does as well. However, if that is correct, then we end up with some highly counter-intuitive results. For example, it would be irrational to pay £1 for a bet that has a 49% probability of one million pounds and a 51% probability of nothing. Given such a bet, I know that I am choosing to accept the bet, and that the bet is unlikely to pay off.[22] However, it seems irrational to turn down this bet, even though the probabilities are not in my favor - turning it down would mean turning down a decent chance of a million in exchange for a tiny sum that I can afford to lose.

Fortunately, the reasoning does not generalize in this way. In short, this is because of the gradibility of the incoherence involved in various possible assertions that would be appropriate for different choices. For example, "I am choosing E, but E is extremely unlikely to pay off" sounds, intuitively, more incoherent than "I am choosing E, but E has only a 49% probability of success". Even assuming that both options maximize expected value, the first sounds intuitively more incoherent that the other. For the fanatical options, the probability of achieving the value is very small. Since success is so unlikely, the agent would be rational to believe that it will not come about.[23] The same is not true for the 49% at a million pounds. When the option is merely more unlikely than not, but not very unlikely, the agent would not be rational to believe that it will not occur. If the weather forecast says 49% probability of rain, I should not believe that it will *not* rain, and nor should I behave as if it will not rain. There is no comparable incoherence in choosing an option with a middling, but less than 50% probability.

What this consideration of structural requirements of rationality has revealed is that it is irrationally incoherent to believe that an option is extremely unlikely and also choose that option. Since substantive requirements of rationality *require* agents in particular

[22] That is, assuming we think that a 49% probability can be correctly described as 'unlikely', which it not obvious.

[23] She may even be rationally *required* to believe this, though I won't take a stand on this.

evidential situations to believe that an option is extremely unlikely, structural rationality rationally prohibits such agents from choosing that option.

4 Should Effective Altruists Maximize Expected Value?

What this discussion has revealed is that substantive and structural requirements of rationality provide important constraints on which options can be rationally chosen. If this is right, then the Rationality Claim is false. Expected Value Theory does not give the comprehensive account of rational choice. Rational effective altruists must also comply with substantive and structural requirements of rationality, and this will sometimes conflict with maximizing expected value across all possible options.

In particular, options that offer a very low probability of very high value are ruled out as rational choices. Recall Option (g):

(g) 0.0001% probability of saving 100,000,000,000,000,000,000,000,000 people (and 99.9999% probability of saving none).

The low probability of (g) means that a substantively rational agent - that is, an agent who has the appropriate coarse-grained epistemic attitudes required by her evidence - would be rational to believe that the option is highly unlikely to be successful. On some views, she may even *know* that the option *will not* be successful. Requirements of structural rationality then prohibit someone who has the preferences of the ideal effective altruist from choosing an option like (g). As argued in the previous section, someone who desires to do the most good impartially and in the most effective way possible would be structurally irrational to choose (g) over the certainty of saving 100 people. Such an agent would have incoherent attitudes, and having incoherent attitudes is, according to the requirements of structural rationality, incompatible with full rationality.

So, complying with substantive and structural requirements on rationality means that rational effective altruists should not maximize expected value unrestrictedly. However, the arguments given here do not prevent agents from maximizing expected value over the remaining options not ruled out by substantive and structural requirements of rationality. Indeed, it seems very plausible that this is precisely what a rational effective altruist should do, and the usual arguments given for maximizing expected value still apply for this restricted range of options.[24] In other words, effective altruists ought to choose the option that maximizes expected value, provided there is no other requirement of rationality ruling that option out.

[24] For example, money-pump arguments (Gustafsson 2022), or arguments from the long-run (Thoma 2018).

On this view, fanatical options are no longer rational, because they are ruled out by substantive and structural rationality. Acknowledging this should be good news for effective altruists, since it allows them to avoid commitment to some of the fanatical implications traditionally attributed to their view, while preserving the intuitively correct idea that rational agents do maximize expected value, albeit only over options that do not conflict with the requirements of substantive or structural rationality.

Acknowledgements

This paper was written during my time as a postdoc on the Varieties of Risk project. I am very grateful to other members of the project (in particular Philip Ebert, Martin Smith, Giada Fratantonio, Xintong Wei, and Lilith Mace) for their input and feedback as these ideas came to fruition. I am additionally grateful to audiences at the Institute for Futures Studies, Stockholm, and the University of Stirling, where I presented earlier versions of this paper.

Funding Information

Arts and Humanities Research Council (Varieties of Risk, Grant #AH/T002638/1)

ORCID

Claire Field
https://orcid.org/0000-0002-0810-0685

References

Adler, Jonathan (2002) 'Akratic Believing', *Philosophical Studies* **110**: 1-27. Doi: 10.1023/a:1019823330245.

Allais, Maurice (1953) 'Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine', *Econometrica* **21**: 503–546. doi:10.2307/1907921.

Balfour, Dylan (2021) 'Pascal's Mugger Strikes Again', *Utilitas* **33**: 118–124. doi:10.1017/s0953820820000357.

Beckstead, Nick and Teruji Thomas (2024) 'A Paradox for Tiny Probabilities and Enormous Values', *Noûs*, 58, 431–455. Doi:10.1111/nous.12462.

Boonin, David (2014) *The Non-Identity Problem and the Ethics of Future People*. Oxford University Press.

Bostrom, Nick (2009) 'Pascal's Mugging', *Analysis* **69**: 443–445. doi:10.1093/analys/anp062.

Bostrom, Nick (2013) 'Existential Risk Prevention as Global Priority', *Global Policy* **4**: 15–31. doi:**10.1111/1758-5899.12002**

Bricker, Adam (2022) 'I Hear You Feel Confident', *Philosophical Quarterly* **73**:24-43. Doi:10.1093/pq/pqac007.

Broome, John (1999) ‘Normative Requirements’, *Ratio* **12**: 398–419. doi:10.1111/1467-9329.00101.

Broome, John (2007) ‘Wide or Narrow Scope?’, *Mind* **116**: 359–370. doi:10.1093/mind/fzm359.

Broome, John (2009) ‘The Unity of Reasoning’, in Simon Robertson, ed. *Spheres of Reason*: 62-92. Oxford University Press

Broome, John (2013) *Rationality Through Reasoning*. Wiley-Blackwell.

Brown, Jessica (2018) *Fallibilism: Evidence and Knowledge*. Oxford University Press.

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

Buchak, Lara (2014) ‘Belief, credence, and norms’, *Philosophical Studies* **169**:1-27. Doi: 10.1007/s11098-013-0182-y

Buffon, Georges Louis Leclerc de (1777) ‘Essai d’Arithmétique Morale’ in his *Supplement a l’Histoire Naturelle, Volume IV.* Paris: L’imprimerie Royale.

Cibinel, Pietro (2023) ‘A Dilemma for Nicolausian Discounting’, *Analysis* **83**: 662–672.

Clifford, William (1877) *The Ethics of Belief and Other Essays*. Amherst, New York: Prometheus Books.

van Fraassen, Bas (1995) ‘Fine-Grained Opinion, Probability, and the Logic of Full Belief’, *Journal of Philosophical Logic* **24**: 349-377. Doi: 10.1007/BF01048352

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press.

Gustafsson, Johan (2022) *Money-Pump Arguments*. Cambridge University Press.

Hong, Frank (2024). “Know Your Way Out of St. Petersburg: An Exploration of ‘Knowledge-First” Decision Theory’, *Erkenntnis* **89:** 2473–2492. doi:10.1007/s10670-022-00639-2

Hume, David (1861). *An Inquiry Concerning Human Understanding*. United Kingdom: J.B. Bebbington.

Isaacs, Yoaav (2016) ‘Probabilities Cannot Be Rationally Neglected’, *Mind* **125**: 759–762. doi:10.1093/mind/fzv151.

Kahneman, Daniel, Paul Slovic, and Amos Tversky (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kiesewetter, Benjamin (2017) *The Normativity of Rationality*. Oxford University Press.

Kolodny, Niko (2005) ‘Why Be Rational?’, *Mind* **114**: 509–563. Doi:10.1093/mind/fzi509

Kolodny, Niko (2007) ‘How Does Coherence Matter?’, *Proceedings of the Aristotelian Society* **107**: 229–263. doi:10.1111/j.1467-9264.2007.00220.x

Kosonen, Petra (2024) ‘Probability Discounting and Money Pumps’, *Philosophy and Phenomenological Research*: 1-19. doi:**10.1111/phpr.13053**

Kyburg, Henry (1961) *Probability and the Logic of Rational Belief. Wesleyan University Press.*

Lasonen-Aarnio, Maria (2014) ‘Higher-Order Evidence and the Limits of

Defeat', *Philosophy and Phenomenological Research* **88**: 314–345. Doi:10.1111/phpr.12090.

Lasonen-Aarnio, Maria (2020). 'Enkrasia or Evidentialism? Learning to Love Mismatch', *Philosophical Studies* **177**: 597–632. doi:10.1007/s11098- 018-1196-2.

Lord, Errol (2018) *The Importance of Being Rational*. Oxford University Press.

MacAskill, William (2022). *What We Owe the Future: A Million Year View.* Oneword Publications.

MacAskill, William, Krister Bykvist, and Toby Ord (2020) *Moral Uncertainty*. Oxford University Press.

MacAskill, William and Theron Pummer (2020) 'Effective Altruism', in Hugh LaFollette, ed. *The International Encyclopedia of Ethics*:1-9. Wiley-Blackwell.

Maher, Patrick (1993) *Betting on Theories.* Cambridge University Press.

Monton, Bradley (2019) 'How to Avoid Maximizing Expected Utility', *Philosophers' Imprint* **19**: 1-25. doi:http://hdl.handle.net/2027/spo.3521354.0019.018

Mulgan, Tim (2006) *Future People: A Moderate Consequentialist Account of Our Obligations to Future Generations*. Oxford University Press.

Nelkin, Dana (2000) 'The Lottery Paradox, Knowledge, and Rationality', P*hilosophical Review* **109**: 373–409. Doi:10.1215/00318108-109-3-373.

Nover, Harris and Alan Hájek (2004) 'Vexing Expectations', *Mind* **113**:237–249. doi:10.1093/mind/113.450.237.

Ord, Toby (2020) *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury.

Owens, David (2002) 'Epistemic Akrasia', *The Monist* **85**: 381–397. Doi:10.5840/monist200285316

Parfit, Derek (1982) 'Future Generations: Further Problems', *Philosophy and Public Affairs* **11**: 113–172.

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

Parfit, Derek (2017) 'Future People, the Non-Identity Problem, and Person-Affecting Principles', *Philosophy and Public Affairs* **45**: 118–157. doi:10.1111/papa.12088.

Pettigrew, Richard (2024) 'Should Longtermists Recommend Hastening Extinction Rather Than Delaying It?', *The Monist* **107**:130-145. doi:10.1093/monist/onae003.

Quiggin, John (1982) 'A Theory of Anticipated Utility', *Journal of Economic Behavior and Organization* **3**: 323–343. doi:10.1016/0167-2681(82)90008-7.

Ryan, Sharon (1996) 'The Preface Paradox', *Philosophical Studies* **64**: 293–307. Doi: 10.1007/bf00365003.

Schmeidler, David (1989) 'Subjective Probability and Expected Utility Without Additivity', *Econometrica* **57**: 571–589. doi:10.2307/1911053.

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

Smithies, Declan (2012) ‘Moore’s Paradox and the Accessibility of Justification’, *Philosophy and Phenomenological Research* **85**: 273–300. Doi:10.1111/j.1933-1592.2011.00506.x.

Thoma, Johanna (2018) ‘Risk aversion and the long run’, *Ethics* **129**:230-253. Doi: 10.1086/699256.

Way, Jonathan (2011) ‘The Symmetry of Rational Requirements’, *Philosophical Studies* **155**: 227–239. doi:10.1007/s11098-010-9563-7.

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

Wedgwood, Ralph (2017) *The Value of Rationality*. Oxford University Press.

Wilkinson, Hayden (2022) ‘In Defense of Fanaticism’, *Ethics* **132**: 445–477. doi:10.1086/716869.

Williamson, Timothy (2000) *Knowledge and its Limits*. Oxford University Press.

Woollard, Fiona (2012) ‘Have We Solved the Non-Identity Problem?’, *Ethical Theory and Moral Practice* **15**: 677–690. doi:10.1007/s10677-012-9359-2.

Worsnip, Alex (2018) ‘The Conflict of Evidence and Coherence’, *Philosophy and Phenomenological Research* **96**: 3–44. Doi:10.1111/phpr.12246.

Worsnip, Alex (2021) *Fitting Things Together: Coherence and the Demands of Structural Rationality*. Oxford University Press.