When We Don't Know What We Owe

JOSHUA STEIN*

ABSTRACT

Uncertainty complicates analysis of practical and moral decision making. This paper argues that these problems are acute and underdeveloped in Effective Altruism. The paper argues that some parts of Effective Altruism, as a theory, can be salvaged by improving application of tools in practical reasoning and rational choice (along with some minor shifts in background commitments), but some widely accepted parts of Effective Altruist thought may be beyond saving. The first half of the paper (Sections I–III) argues for applying theories of practical reasons and rational choice to Effective Altruism to address both perennial problems of uncertainty and particular layers of uncertainty unique to its theory. Those sections are ultimately optimistic that the problems can be addressed using established tools in the field. The second half of the paper (Sections IV–VI) argues that unique problems are created by other commitments of Effective Altruism which cannot be addressed without changes of substantive commitments, underlying attitudes towards uncertainty held by proponents, or both.

TABLE OF CONTENTS

* Joshua Stein is a postdoctoral fellow at the Georgetown Institute for the Study in Markets in Ethics. His work focuses on the philosophy of economics, metaethics, and political philosophy.

I. EFFECTIVE ALTRUISM UNDER CONDITIONS OF UNCERTAINTY

Effective Altruism (EA) includes a decision procedure, a way to evaluate possible courses of action in complex circumstances. This paper concerns the cases in which that decision procedure fails, on its own terms and under norms of practical reasoning.[1] Some failure can be fixed by adopting established philosophical tools; some are deeper problems for EA. This analysis includes two sets of cases. The first set is the failure to account for varieties of uncertainty; the second set is systematic failures and biases particular to EA. I am optimistic about the prospects for the former set, but I am skeptical the latter can be solved without substantial revisions and reconsiderations of Effective Altruism.

Section III discusses internal failures regarding uncertainty, the first set of cases. EA fails to adequately consider both the scope and variety of uncertainty. These failures result in incomplete decisions in cases where indecisiveness is unacceptable, on EA's own terms. The section also provides some tools for

1. Effective Altruism is a decision procedure, not a descriptive decision theory. *See* Jake Chandler, *Descriptive Decision Theory*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2017).

resolving those cases. Sections V and VI argue systematic failures in EA cause several substantive failures in policy. This is structured as an internal critique, but these failures mirror external critiques. These problems are entangled with uncertainty, but they spin out into substantive (rather than merely formal) concerns, including institutional capture, fraud, and misappropriation of resources.

I am an agnostic on the overall value of Effective Altruism. On the one hand, I try to improve the decision-making processes using tools from theories governing practical reasoning. On the other hand, these failures may be too pervasive and severe to be solved. My goal is to leverage the decision procedure issues into a better means to do good. It is secondary whether "Effective Altruism" is that means.

A. Heterogeneity, Ecumenism, and Pluralism in Effective Altruism

EA has distinguishable theory and movement sides; these are not separable, but they are distinct. This is no analysis of necessary or sufficient conditions; EA is different things to different people. This is normal for any sufficiently large and diverse thing. It is a philosophical position; it is a movement. It includes substantive moral commitments and views about expected utility. There is no singular consensus about what EA is. There are a few core claims which are useful to establish.

Effective altruists hold a few views broadly. I want to unpack these core claims to articulate the areas of agreement. What elements must a theory have to qualify as Effective Altruism?

> *Impartiality:* Welfare should be considered impartially. We might disagree about who welfare bearers are and how to evaluate them, but within the scope of welfare bearers, EA requires we be impartial.[2] Put another way: we cannot prefer the welfare of those who share our ethnicity, country, religion, etc. over those who do not.
>
> *Consequences First:* One must prioritize the practical considerations and consequences ahead of moral and political theories and commitments. If moral commitments result in action plans which produce strictly worse outcomes, the outcomes take priority.
>
> *Effectiveness and Efficiency:* There is some standard for practical reason about "effectiveness" which should be central to decision-making. Following "consequences first," a plan of action is only justified if the plan is effective and/or efficient relative to available alternatives.[3]

---

2. *See* Brian Berkey, *The Philosophical Core of Effective Altruism*, 52 J. SOC. PHIL. 93, 93–115 (2021); William MacAskill, *Effective Altruism: Introduction*, 18 ESSAYS PHIL. 1, 1–5 (2017).

3. Zuolo notes that "effectiveness" is ambiguous in EA. Effectiveness may describe either the likelihood of effectuating the desired outcome or the efficiency at which the desired outcome is achieved, which are extensionally different. *See* Federico Zuolo, *Beyond Moral Efficiency: Effective Altruism and Theorizing about Effectiveness*, 32 UTILITAS 19, 19–32 (2020).

Beyond these theses, there is intellectual and ideological diversity in EA. Berg notes the ten percent donation figure is a heuristic for making EA "ecumenical."[4] If we are going to borrow the Christian notions of ecumenism and tithing, we should note MacAskill's focus on ecumenism is evangelical in nature, about converting outsiders to EA. Not all effective altruists endorse ecumenism. The version of EA formulated by Singer includes a maximizing claim: one should do as much good as one can, subject to each individual's resources.[5] This includes an obligation to give all excess income, borderline ascetism, and other moral commitments. This condition is sometimes called "demandingness," and is subject to eponymous objections.[6] Singer's theory is not alone in facing demandingness objections, but it is a useful illustration.

The effectiveness condition pits demandingness against ecumenism because the maximizing view has limited appeal and will result in limited participation (or so the argument goes). A movement must trade demandingness against evangelicalism. Singer's demanding version is likely to struggle with adherence, which limits effectiveness (at the level of the movement). Therefore, if EA requires effectiveness in its movement elements, that counts against Singer's version.[7] If EA cannot attract adherents, then it is not effective. Maximizing every case means fewer cases. These internal disagreements are not surprising. In any large group, there will be disagreement. Disagreement between fundamentalist and ecumenical EA is not concerning; this discussion is just to lay out the points of agreement and disagreement.

B. Practical reasons as the proper ecumenical decomposition of Effective Altruism

Defenders of EA insist that it is not committed to utilitarianism, that it is neutral towards moral theories.[8] The arguments for ecumenism inform this non-neutrality. If a philosophical view is committed to a theoretical claim, then if one has strong reason to reject the theoretical claim in question, one should reject the

4. *See* Amy Berg, *Why Ten Percent?*, 21 GEO. J. L. & PUB. POL'Y 655 (2023); *see also* William MacAskill, *The Definition of Effective Altruism*, *in* EFFECTIVE ALTRUISM: PHILOSOPHICAL ISSUES 14–16 (Hilary Greaves & Theron Pummer eds., 2019).

5. *See* PETER SINGER, THE MOST GOOD YOU CAN DO 3–12 (2015); Peter Singer, *Famine, Affluence, and Mortality*, 1 PHIL. & PUB. AFFS. 229, 239 (1972).

6. *See* Ryan W. Davis, *The Moral Status of Beneficence*, 21 GEO. J. L. & PUB. POL'Y 639 (2023); Berg, *supra* note 4; Brian McElwee, *Cost and Psychological Difficulty: Two Aspects of Demandingness*, AUSTRALASIAN J. PHIL. 1 (2022); Brian Berkey, *Effectiveness and Demandingness*, 32 UTILITAS 368, 368–81 (2020).

7. One need not fully reject demandingness. Sometimes moral theories require actions which cannot be adopted broadly. Strict adherence to religious doctrines, for example, are rarely followed by all members. Singer's demanding version can hold it is right to maximize while acknowledging ecumenism as prudent; it is not incoherent, but tense. Section III.E. raises tools for addressing this tension.

8. *See* William MacAskill, *Understanding Effective Altruism and its Challenges*, *in* THE PALGRAVE HANDBOOK OF PHILOSOPHY AND PUBLIC POLICY 441–53 (David Boonin ed., 2018); Jeff McMahan, *Philosophical Critiques of Effective Altruism*, 73 PHILOSOPHERS' MAG. 92, 92 (2016).

view.[9] It is not clear that this neutrality claim is defensible; the claim does not smell right. If EA is not grounding its substantive claims and motivational force in welfare utility, then what is the grounding? The neutrality claim hinges on a distinction between theoretical commitment and practical use. Adopting consequentialist approaches in practice does not entail a commitment to consequences as moral grounds. This is where the water gets murky. Suppose an effective altruist claims Carol ought to give her money to anti-malarial charities, rather than local libraries.[10] The way an effective altruist might argue such a claim is to appeal to impartiality and effectiveness claims; if they hold these considerations have moral force, the reasoning may be grounded in welfare utilitarian commitments. I argue below this can be avoided by shifting to practical reasons, but many formulations of EA do not allow that shift.

Davis[11] notes EA tends to decompose in one of two directions. Either it holds consequences instantiate an obligation or it does not. He argues Singer's demanding version is defensible *because* it holds the former. Just as one cannot be halfway pregnant, one cannot be halfway obligated. However, if EA holds the consequentialist considerations instantiate an obligation, then they are consequentialists about moral theory. So non-utilitarians are suspicious of EA, and they should be. This smells like smuggling.

Davis's analysis illuminates another possible direction.[12] Suppose there is no moral obligation. Instead, the consequences can be a practical reason to act but do not themselves ground an obligation. One is not a consequentialist in moral theory; the moral domain is not implicated.[13] Practical reasoning is doing the work. This is not exactly moral neutrality. The modes of assessment are still consequentialist and thus lean towards consequentialism. However, this is pluralist and inclusive with regards to other theories, and therefore satisfies the ecumenical requirements. On this shift, EA can avoid committing to a consequentialist moral theory.

---

9. This is the practical framing; MacAskill draws from Parfit, and Parfit's moral consensus building project late in his career emphasizes using practical reasons as a means of reconciling disparities in major moral theories. *See* DEREK PARFIT, ON WHAT MATTERS (2011). Unfortunately, adopting Parfit's conciliatory project takes on other conceptual baggage.

10. *See generally* Jennifer C. Rubenstein, *The Lessons of Effective Altruism*, 30 ETHICS & INT'L AFFS., 511, 511–26 (2016).

11. *See* Davis, *supra* note 6.

12. This move is mine, not Davis's, though private conversations suggest he agrees. Basically, Davis's minimalist assessment voids the obligations, but does not remove independent reasons to prefer the actions. These actions are less than required, but still grounded in reason. They are supererogatory. *See* David Heyd, *Supererogation*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2019).

13. Some theorists about EA hold that practical reasoning about effectiveness and efficiency is and should be taken as moral. (e.g.) If resource scarcity means waste of resources is itself immoral (and therefore maximizing effectiveness and efficiency is obligatory), then one might reject my shift to practical reasons as such reasons are practical *and* moral.

C. The Guardrails Argument and philanthropic capture

All philanthropic projects involve a power imbalance. If a project is philanthropic, one party has resources the other party cannot easily obtain. If the recipients of aid could obtain resources without donors, the project would not be philanthropic. Philanthropy is not the only model for the redistribution of resources; however, as EA is the focus of this paper, set aside alternative models of redistribution. One concern in philanthropy is the exploitation of power imbalance. EA provides a check by requiring transparency and justification of resource allocation, to verify effectiveness and efficiency elements. Donors should only contribute to a project if it satisfies conditions of transparency and public justification. Transparency and public justification are guardrails, which traditional philanthropy lacks.

Whether these guardrails work is an empirical question. If the decision procedure is defective, then how Effective Altruists engage in public justification will likely also be defective; even if the decision procedure is perfect, some projects may still be opaque. However, establishing guardrails is useful.

II. DECISION THEORETIC ELEMENTS OF EFFECTIVE ALTRUISM

EA includes substantive moral claims. For example, impartiality is a substantive and controversial moral claim.[14] Mainstream EA often discusses a "bias towards action."[15] If an agent has a set of possible actions and is unsure about which to do, it is better to pick one rather than do nothing. There are contexts where this bias towards action is worrying, such as in medical practice and ethics, where using the term causes overtreatment.[16] In EA, bias towards action is meant to address Buridan's Ass.[17]

Consider Taurek's patients:[18]

> I have a supply of some life-saving drug. Six people will all certainly die if they are not treated with the drug. But one of the six requires all of the drug if he is to survive. Each of the other five requires only one-fifth of the drug. What ought I to do?[19]

---

14. Effective altruists sometimes overlook how controversial this claim is, but "we should fix our problems locally before we invest in problems abroad" is intuitively morally compelling to many people. Warmke's analysis of EA turns on rejection of the impartiality condition and partial preference of family, local community, etc. *See* Brandon Warmke, *Saving the World Starts at Home*, 21 GEO. J. L. & PUB. POL'Y 769 (2023).

15. I cannot belabor this point here. For illustration, see Gideon Lewis-Kraus, *The Reluctant Prophet of Effective Altruism*, NEW YORKER (Aug. 8, 2022), https://www.newyorker.com/magazine/2022/08/15/the-reluctant-prophet-of-effective-altruism [perma.cc/A7U4-GEM4].

16. *See* John Z. Avanian & Donald M. Berwick, *Do Physicians Have a Bias toward Action?: A Classic Study Revisited*, 11 MED. DECISION MAKING 154, 154–58 (1991).

17. ARISTOTLE, DE CAELO bk. II, ch. 13, at 432–33, *in* THE BASIC WORKS OF ARISTOTLE (Richard McKeon ed., 2001) (c. 350 B.C.E).

18. John M. Taurek, *Should the Numbers Count?*, 6 PHIL. & PUB. AFFS. 293, 293–94 (1977).

19. *Id.* at 294.

EA analyzes Taurek's patients in consequentialist terms; there is no complex moral question. One should do whichever of the two actions promotes the greatest welfare, giving the standard utilitarian judgment that one should save the larger group.

Taurek raises moral worries about this, which should concern Effective Altruists if they take EA to be a moral position. If EA concerns practical reasons for action, this is a non-issue. The consequentialist judgment is a strong practical reason to give the drug to the larger group. However, note what is not allowed: there is no option to delay treatment in the process of deliberation. One cannot, like Buridan's Ass, be paralyzed between the options. Effective Altruists like bias towards action because it limits time wasted in deliberation. The decision procedure limits delay. However, there are circumstances where Effective Altruists jettison the bias towards action in favor of intensive research[20] or resource accumulation.[21]

A. Uncertainty and Cluelessness

Uncertainty is the conceptual core of this paper. Uncertainty has been used as an objection to consequentialism.[22] "Cluelessness objections" recognize the future is uncertain, and therefore consequences are uncertain. Because consequences are uncertain, the moral value of an act is uncertain, if evaluated in terms of its consequences. If moral value is uncertain in a sufficiently wide range of cases, then the moral theory has a problem with uncertainty. Put another way: moral theories have to be able to generate evaluative judgments of possible actions; the cluelessness objections hold that consequentialism cannot provide forward-looking evaluation, because of uncertainty, and therefore fails systematically.

EA has two ways to respond. The first response is the standard response used by consequentialist theories: the future is uncertain, but we can still have a reasonable basis for belief. We can make reasonable inferences and use our predictive powers well enough to anticipate certain consequences. Cluelessness is limiting, not blocking.[23] This response understands EA as committed to consequentialism, or at least consequentialist analysis.

Suppose an effective altruist dislikes the commitment to consequentialism as a moral theory, and instead prefers the practical reasons analysis I provide in Section I.B. If EA is not concerned with what agents ought to do, but rather presents reasons for what they should do, then uncertainty is not a serious problem. Uncertainty is an ordinary feature of human practical reasoning. Gambling and risk analysis illustrates this. These two moves are similar; they differ in whether the consequentialist analysis is bound to a moral theory or just ordinary

20. *See* discussion *infra* Section III.B.
21. *See* discussion *infra* Section V.A.
22. James Lenman, *Consequentialism and Cluelessness*, 29 PHIL. & PUB. AFFS. 342, 342 (2000).
23. Hilary Greaves, *Cluelessness*, 2016 PROC. ARISTOTELIAN SOC'Y 311.

practical reasoning. Both acknowledge uncertainty is a challenge for decision making when outcomes matter.

Observing uncertainty in deliberation makes EA no different from any other project of practical rationality.

B. Comparativism and the Grounds of Rational Choice

Rational choice matters to EA. Our best understanding of practical rationality across a range of intellectual endeavors and domains can aid in how we think about effectiveness and efficiency. Two quick concepts: First is the concept of a reason, both familiar and controversial. There are platitudes used to give the impression that we have a deep understanding of what the reason is, "a reason is a fact that counts in favor,"[24] giving the impression that the analysis of a reason is simple. I will not dwell on the metaethics here. What follows adopts an ecumenical attitude toward theories of reason.

Second is the concept of comparison. Comparison is necessary for full consideration in actual cases. It is wrong, all else being equal, to cause harm; but that tells us very little about how to evaluate harms in practical reasoning, because in any actual case, we need to figure out whether all else is equal. Chang's "comparativist" analysis of practical rationality is important.[25] A reason to act applies in considering possible actions, consequences, social facts, and the rest of the wrinkly bed sheet of reality. When I compare drinking coffee to drinking tea, I consider the time of day, whether others are drinking too, the amount of coffee I have been drinking recently, etc. These things may be reasons to drink coffee, rather than water. Rarely are they relevant when considering whether to drink coffee or vodka; vodka choices are different than coffee choices. On the comparativist view, reasons and rational choice require comparing possibilities.

III. UNCERTAINTY AND THE DECISION PROCEDURE IN EFFECTIVE ALTRUISM

This paper started as a series of puzzles. The goal was to develop an analysis of how a theory of practical reasons, under conditions of uncertainty, could improve EA. This section argues that EA, as presently understood, is deficient in handling uncertainty. EA has oversimplified the varieties and scope of uncertainty. Unlike "cluelessness" objections, these are objections with a narrow focus. These puzzles show how EA is ill-equipped to manage uncertainty and introduce helpful tools.

A. Risk as an ordinary calculation familiar to Effective Altruism

Effective altruists like gambling. Unsurprisingly, EA is equipped to deal with uncertainty about outcomes and risk. Operating under conditions of risk is operating with uncertainty about the outcome. In a simple gambling analogy, the

24. Parfit, *supra* note 9, at 1; THOMAS M. SCANLON, BEING REALISTIC ABOUT REASONS 30 (2014).

25. Ruth Chang, *Comparativism: The Grounds of Rational Choice*, 2016 WEIGHING REASONS 213; Ruth Chang, *Are Hard Choices Cases of Incomparability?*, 22 PHIL. ISSUES 106 (2012).

probability of a possible outcome is not one or zero. Conditions of risk may have well-defined probabilities (e.g., casino games) or uncertain probabilities. Either way the outcome is uncertain.[26]

The odds of winning a spin of a roulette wheel may or may not provide reasons for betting in certain patterns. The odds of winning by betting on black are 47.37 percent (in the United States), with a payout of 1-to-1. This is a suboptimal bet, as all casino games are.[27] EA frequently considers the likelihood of possible outcomes. This makes sense, given the emphasis on effectiveness and efficiency. Some targets of EA, like malaria or poverty, focus on extant problems; because those problems already exist, their probability of occurrence is one. The likelihood of a global pandemic from a contagious upper-respiratory disease was not one, but public health scholars warned for years the likelihood was high. As a result, it was reasonable to make decisions about resource allocation based on that high likelihood.

EA uses assessment of risk as we might expect experts in social science disciplines to do. EA is positioned to address risk in the same terms as global public health pertaining to pandemic preparedness. However, there are serious deficiencies in how Effective Altruists model risk. Further, Effective Altruists tend to neglect other varieties of uncertainty, (i.e.) varieties of uncertainty other than outcome. There are wrong answers when it comes to risk management. For most real-world choices, there are usually several defensible approaches; there are also indefensible approaches. Consider the following:

> *Toy Case, Roulette v. Coin-Flip:* Suppose Sam can choose between two games. The first is roulette, but Sam can only make one-to-one bets on color; the second is a coin flip which returns one-to-one. Under these conditions, Sam would be irrational to choose roulette. The coin flipping game has strictly better expected value than roulette. Roulette has an expected return of about $0.95 for every $1 put in; the coin flip game has a $1 expected return for every $1. All else equal, choosing roulette is just choosing to make less money.

EA can leverage this point into arguments for effective and efficient giving. If Sam donates money and the donation has a probability of effectuating the goal of $0.95 for every $1, this may be defensible, but if the other available option pays out $0.05 better and all else is equal, choosing the first charity is irrational.[28] Situations this cut-and-dry are rare.

Roulette v. Coin-Flip is meant to be boring; its purpose is to illustrate minimal requirements of rational choice under comparison. Questions of how much to bet,

---

26. Sven Ove Hansson, *Risk*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2018).

27. A game can be suboptimal and still be better than alternatives. If you must choose between playing European and American roulette wheels, just betting on Black, you should generally play European roulette. The payout is the same, but the odds of winning can be 1.2% higher.

28. R.A. Briggs, *Normative Theories of Rational Choice: Expected Utility*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2019).

how to pattern bets, etc. are subject to other considerations and substantial disagreement. Some possible choices can be excluded, even if the exclusion of possible choices does not settle the issue. However, any additional wrinkles complicate choice.

> *Risk-Averse Roulette:* Suppose Sam is given a choice between the two games, but with a distinction. He can either make one $100 bet in the coin-flip game or one hundred $1 bets in the roulette game.

Maximizing expected utility on a standard model would suggest Sam ought to bet the coin-flip game, but we can make sense of how a normal, defensible risk-aversion might result in Sam choosing to play 100 games of roulette. Why? Because regression to the mean in 100 games of roulette limits the likelihood of total loss.

The odds of walking away broke in the coin-flip game is 50 percent; the odds of walking away broke after 100 $1 games of roulette are incredibly low. We can acknowledge different attitudes to risk within this account. Some versions of EA are risk-tolerant in the way *Risk-Averse Roulette* pushes against. Distributing bets (assuming independent probabilities) is a strategy to mitigate risk; that's why a diverse portfolio is typically more stable.

But EA is not concerned about stability and risk-mitigation; the strategy of EA research organizations, for example, is to identify a narrow band of high-yield organizations and funnel donor money to those organizations, rather than distributing the money more widely. If the expected outcome was a sure bet, that might not raise concerns about risk profile, but (as we will see below) these bets are uncertain. This is not to say the high risk-tolerance of Effective Altruism is wrong, but it has problems. It also results in concerns about actual choices made by EA organizations, as I will discuss in Sections V.A. and VI.

B. Uncertainty about Risks; or Second-Order Uncertainty

Casino games have well-defined risks. The real world usually doesn't. Risk creates uncertainty about outcomes, but there are cases of both uncertainty about the outcome and the risks themselves. Uncertainty about risks is second-order uncertainty. One has a second-order uncertainty when one is uncertain about uncertainty; in this case, one is uncertain about the conditions of uncertainty of outcomes.

One way to model uncertainty about risk is to use a range of probabilities for the range of possible outcomes. Another is to set error bars. Both approaches generate a range; I use the former here, though this is oversimplified, because probabilities are rarely evenly distributed. There are more complex ways of modeling imprecise probabilities.[29]

---

29. These are simple ways amenable to the purposes of this section. There are more sophisticated approaches. *See* Seamus Bradley, *Imprecise Probabilities*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2019).

Second-order uncertainty is a general problem for practical reasoning and not a problem at all. It is a problem for practical reasoning because attempts to map clean cases are difficult when numbers do not behave themselves. It is not a problem insofar as this just reflects complexity of the world. For simplification purposes, I use the ranges approach below.

> *Uncertainty over Ranges:* Suppose that Sam can choose between two action plans (x and y) that instantiate exactly one of three possible outcomes (M, N, O). Suppose we can fix the values such that M is exactly twice as good as N and O is half as good as N. M, N, and O are two, one, and 0.5 respectively. If S does x, then the probability of M is a range 0.2-0.4, N is 0.6-0.7, O is 0.2-0.4. If S does y, then the odds are M: 0.4-0.6, N: 0.1-0.2, O: 0.4-0.7. What should Sam do?

Doing y maximizes the likelihood of the best possible outcome (M); doing y represents an increase of M from anywhere between 0.2 to 0.4 over doing x. On the other hand, doing y also increases the risk of the worst outcome (O) by a range of 0-to-0.5 over x. Choosing y may not make O any more likely; it may also raise the odds of O dramatically; we are uncertain about whether doing y makes a difference to the likelihood of O.

y is riskier. It improves the probabilities of the best and worst case scenarios. Risk-tolerant people may prefer y; someone who is risk-averse may dis-prefer y. There are constraints on expected utility which indicate whether x or y is likely to maximize expected utility, but the probabilities are not fixed as in the casino game cases.

The purpose of toy cases is to illustrate, as the math is much messier in real life. They lead to two observations. First, some choices may maximize expected utility with high-risk, high-reward gambles; these sorts of high-risk cases may not be the best choice for humans.[30] In some cases, risk-seeking may be the best strategy under expectation of return but be too risk-tolerant for human agents, especially in areas where stakes are high. In real-world cases, risk-aversion looks different because downsides have incomparable dimensions. Taking a high-risk strategy where the worst-case scenario is death and harm for others is different than taking a high-risk strategy at the casino. Money and suffering are different.[31]

This is an obvious point, but it spins out for EA in a troubling way. Reconsider the Risk-Averse Roulette case. We might think it's fine to bet aggressively in that case, because the stakes are the bettor's money; but what if the bettor was gambling with the welfare of a nation? If a philanthropist takes an aggressive gambling strategy with the welfare of those in danger of falling into poverty, that's not an issue of rational choice, but moral judgment. Placing an aggressive bet

30. Strictly formal models are typically more risk-tolerant than human decision makers. This results in paradoxes of practical rationality. *See* discussion *infra* Section V.

31. I will come back to this point. *See* discussion *infra* Section III.C.

with one's child's college fund or a nation's development resources or charitable donations is not defensible; unfortunately, this is where the high risk-tolerance of EA starts to create moral problems.

The second observation regards a subtler, trickier problem.[32] We sink enormous resources into limiting second-order uncertainty. As Joyce notes, this use of trials is not efficient and (in many contexts) will not produce adequate data. Joyce illustrates other approaches (especially causal analyses) which are more cost-efficient and produce better results. In EA, efficient allocation of resources matters; sinking resources into trials on the hope of limiting second-order uncertainty may not be the best use of resources.

> *Reduction of Second Order Uncertainty:* Sam has a choice between acts x and y, but Sam is uncertain both about outcomes and risks associated with each. In the meantime, Sam adopts z, which is an action plan of research to address the uncertainty around x and y. Suppose x and y both cost $100. What would the circumstances have to be for z to be worth a $20 investment?

There are ways to bring out the details of this case, but they require layers of probabilities. First, stipulate initial probabilities for outcomes under x and y; second, stipulate secondary probabilities for x and y learned through researching z. This establishes the difference z makes. These layers pile up quickly, but the details are beside the point:: Researching differences to reduce uncertainty is defensible if and only insofar as research makes a difference to planning *and* is preferrable to arbitrarily choosing between x and y.

Put another way: If the cost of the research is high enough that it depletes more resources than it prevents losing, then we have made a Buridan's Ass of ourselves, dilly-dallying in gathering information that doesn't make a sufficient difference and (therefore) doing something worse than arbitrarily choosing. This is inconsistent with the core tenets of EA.

There are contingent (but perhaps unknowable) circumstances where a large RTC is justified, (e.g.) when failing to allocate resources effectively will result in massive waste. Some Effective Altruists argue that resource-intensive approaches to research detract from "effectiveness." Joyce's proposal is to focus on more fine-grained, less resource-intensive approaches; this is a significant improvement to uncertainty about risk. EA can go a step further by engaging in more diverse approaches to funding philanthropies, to limit potential losses to second-order uncertainty.

C. Uncertainty about Values

"Moral uncertainty" refers to a range of issues.[33] Broadly, it refers to uncertainty about moral properties. EA focuses on a consequences-oriented approach

---

32. Kathryn E. Joyce, *Assessing Evidence for Purposes of Effective Altruism*, 21 GEO. J. L. & PUB. POL'Y 757 (2023).

33. *See* WILLIAM MACASKILL ET AL., MORAL UNCERTAINTY (2020); TED LOCKHART, MORAL UNCERTAINTY AND ITS CONSEQUENCES (2000).

to evaluation, so we can simplify this analysis of moral uncertainty to values of possible outcomes. Singer insists that if one comes across a drowning child, and one can save that child with little effort, one is obligated to do so. I am inclined to agree with Singer's analysis. However, there are substantive points of disagreement about even this relatively simple case. Some argue "obligation" is too strong, that individuals cannot be required morally to do things which they have not adopted as a responsibility.[34] Those arguments fall outside the scope of this discussion. Even when we have a strong moral judgment, there can be reasonable disagreement.

Even core claims of EA are subjects of reasonable disagreement and moral uncertainty. EA holds one should act for the greatest effect, including based on the greatest number.[35] Similarly, there are concerns about relative values of human and non-human animal welfare;[36] if all lives are equally valuable, we should be far more dedicated to preserving more populous species than our own. These points of moral uncertainty multiply.

Consider the problem under comparativism: we have two (or more) options of what to do and face uncertainty about how to evaluate possible outcomes. There is uncertainty about how to compare those outcomes.

> *Comparing education and health*: There are two philanthropic projects. The first provides improved health outcomes for children in a distant country; the second provides educational infrastructure in that country. Stipulate we ought to provide health care ahead have access to education on the basis that health is a necessary condition for the pursuit of education. The education project provides education for 10 times as many children. Is the relationship between healthcare and education "discounted" at 10-to-1?

EA can hold that donating to a wide range of causes is essential, and one might be justified in any of a range of possible projects so long as those projects otherwise comport with the central values. This is an EA variation on pluralism. Moral uncertainty is a persistent, serious problem.

> *Comparing conservationism across species:* Suppose within a conservation project, conservationists are confronted with competing interests between chimpanzees and arthropods. For each member of the species, chimpanzees have greater welfare consideration; however, the arthropods dominate on sheer numbers. What is the ratio of arthropods to chimpanzees such that welfare considerations favor arthropods? Is there any such ratio?[37]

---

34. *See, e.g.*, Davis, *supra* note 6.

35. *But see* Taurek, *supra* note 18; Tyler Doggett, *What Is Wrong With Kamm and Scanlon's Arguments Against Taurek*, 3 J. ETHICS & SOC. PHIL. (2008); Alan Thomas, *Giving Each Person Her Due: Taurek Cases and Non-Comparative Justice*, 15 ETHICAL THEORY & MORAL PRAC. 661 (2012).

36. Bob Fischer, *How to Express Improvements in Animal Welfare in DALYs-averted*, 21 GEO. J. L. & PUB. POL'Y 735 (2023).

37. I am grateful to Bob Fischer and Jeff Sebo for this experiment, as a variation on Taurek-style cases.

Both cases are compatible with both moral theory and practical reason versions of EA. I argue in Section VI that EA is deficient in considering moral uncertainty about certain values. Pettigrew's response to Longtermist EA, which argues some models of LEA decision theory recommend hastening human extinction, provides a useful reductio ad absurdum. Bracket this question for now.[38] Finally, this is not a "cluelessness objection." Rather, it restricts confidence based on the degrees of uncertainty about these elements. Rather than blocking outright, this analysis is qualifying; we shouldn't overstate our confidence.

D. Uncertainty about Individuation of Act Types, Consequences, Instances

Permit me a short gallop on my hobby horse. We usually individuate acts according to types. Consider the axe-murderer cases.[39] When the axe-murderer comes to the door looking for a victim, I have the choice to lie or tell the truth. Each of these represents an act type. "Lying" refers to an act type that can be carved up in different ways; fortunately, the axe-murderer case tends to cleave to a standard understanding of what "lying" is.

In some cases, there may be no controversy about how to individuate act types, but there are cases where it matters. For example, is "talking around the truth" or omitting useful information "lying?" In those contexts, individuating "lie" as a type is complicated; there are straightforward lies and borderline lies. Because EA is concerned with the macro-level, the individuation focused on intent and individual obligations like "lying" or "promise breaking" are not usually relevant. However, these problems occur in macro-level cases.

There are some morally salient properties of act types which EA may not consider adequately. If EA treats all types of acts in public health philanthropy or education identically based on outcomes, then this can flatten out important distinctions between types of acts which respect local autonomy and customs with those which do not.

> *Religious Administration of Education Case*: Suppose that there are two education charities; call one "the non-sectarian charity" and the other "the religious charity." Both increase education access for children in the region. The religious charity provides primary education for a child for $1,000/year; the non-sectarian charity provides the same for $800/year. The local community prefers providing a religious education to a non-sectarian education.

If one regards all education as of the same type, regardless of whether it is religious, then EA should straightforwardly prefer donating to the non-sectarian charity. If one treats them as different, exclusive kinds, then one may take the value of community autonomy to be worth the extra $200 per child per year in

---

38. *See* Richard Pettigrew, *Should longtermists recommend hastening extinction rather than delaying it?,* MONIST (forthcoming).

39. *See* Mark Schroeder, *The Hypothetical Imperative?*, 83 AUSTRALASIAN J. PHIL. 357, 368 (2005).

costs (or not, depending on one's views). At risk of being too technical, this is individuating at different levels. If we individuate at the determinable (i.e., higher) level where all education is of a kind, we get one assessment; if we individuate at the determinate (i.e., lower) level where non-sectarian and religious education are different things, we may get a different assessment. We must make a choice about which individuation we're using.

My broader project wonders about conceptual engineering for effective individuation. These worries intrude here: we are uncertain about the appropriate individuation of act types (among other things), and this may cause serious problems for application of practical reason in EA.

E. Uncertainty about Coordination and Problems of Scale

Humans are social. How we act, especially publicly, influences others. Effective altruists hope their donation practices will inspire others to do the same.[40] The evangelical nature of EA is an attempt to coordinate behavior. Individuals can exert indirect pressure on others to engage in good behavior by displaying good behavior and by explaining how that behavior might improve the general social welfare. In some contexts, this may be inappropriate, imprudent, or impolite; taking a family dinner as an opportunity to lecture one's grandmother about how particular choices of charitable giving are bad is unlikely to produce anything other than aggravation and tension. But in appropriate contexts, evangelizing charity is good.

Generally, social coordination is useful and valuable.[41] It is also a source of uncertainty. Many metaethical theories center behavioral coordination.[42] Some, like the public welfare theories of law, are amenable to EA.[43] But EA does not pivot on such a theory. In some cases, individual actions may significantly promote utility, but if those action plans are adopted by a sufficiently large group or if there is some other public response, then the systematic adoption can fail to promote utility.

Put another way, sometimes the best thing an individual can do to maximize utility should not be broadly adopted or universalized. This has a familiar smell of Kantianism, for good reason. Part of the Kantian uptake in theories of law

---

40. Not all views favor explicit influencing and coordination; some prefer private or anonymous giving. *See* MISHNEH TORAH, Sefer Zeraim, Gifts to the Poor 10:7–14.

41. EA has donated enormous resources to marketing; the *only* defense of such practices is that the marketing (a) results in higher levels of charitable giving *and* (b) those rates of charitable giving are higher than marketing spending *or* (c) they are efficiently allocated such that they are comparable to the alternative. I suspect all three are false (and perhaps egregiously so); it is an empirical question that cannot prudently be addressed here.

42. *See* STEPHEN DARWALL, THE SECOND-PERSON STANDPOINT: MORALITY, RESPECT, AND ACCOUNTABILITY (2009); THOMAS SCANLON, WHAT WE OWE TO EACH OTHER (2000).

43. *See* GERALD J. POSTEMA, BENTHAM AND THE COMMON LAW TRADITION (2019); Gerald J. Postema, *Bentham on the Public Character of Law*, 1 UTILITAS, 41–61 (1989); Gerald J. Postema, *Bentham's Early Reflections on Law, Justice and Adjudication*, *in* 36 REVUE INTERNATIONALE DE PHILOSOPHIE 219, 219–241 (1982).

centers rules as tools of coordination, as followable.[44] Some criticize GiveWell's deworming program for giving high rankings to deworming charities despite questionable evidence of the efficacy of deworming,[45] based on the low cost of deworming therapies.[46] Set worries about efficacy aside for now; there is a second worry about how coordinating donations towards deworming can result in improper allocation of resources, even by EA's own modes of analysis.

> *Donation Threshold Puzzle:* There are two projects: "the road" and the "the dewormer fund." The dewormer fund prevents one severe adverse childhood medical event for every $10 donated. The road prevents no adverse events until it reaches $1M; once the road reaches $1M, it prevents two adverse childhood medical events per $10 donated (200,000 at $1M). Smith has $1,000. Should he give to the road or the dewormer fund?

A naïve effective altruist analysis will hold Smith ought to donate to the dewormer fund because of the efficiency in donation; a sophisticated EA analysis may hold Smith should consider donating to the road subject to likelihood of success. Simply, if the probability of the road reaching $1M is greater than .5, this is strong reason to donate to the road.

However, simply focusing on the expected utility (in this case, the expected prevention of severe adverse events) misses the point. Social coordination is a means to establish large investments in projects which would be infeasible as single-donor endeavors. Thinking individually about donations, rather than coordination, will systematically undervalue these projects. These projects substantially improve quality of life for larger numbers of people. For example, water purification infrastructure provides greater long-term stability and regional returns on investment than small water purifiers given to individuals and families.[47]

---

44. This point regarding the role of rules in Kant is especially important for understanding a major point of incompatibility of welfare utilitarianism used to frame EA; it also raises worries about feasible moral consensus to which MacAskill and Parfit appeal. Points of consensus used to limit moral uncertainty may not be points of consensus at all. *See* ONORA O'NEILL, CONSTRUCTING AUTHORITIES 118 (2015).

45. *See* Kelsey Piper, *The Return of the 'Worm Wars'*, VOX (Jul. 19, 2022), https://www.vox.com/future-perfect/2022/7/19/23268786/deworming-givewell-effective-altruism-michael-hobbes [https://perma.cc/XE8M-KFLL]; Hauke Hillebrandt, *Commentary: Three Ways to Falsify the Case for Mass Deworming Against Soil-Transmitted Helminths*, 45 INT'L J. EPIDEMIOLOGY 2168, 2168–2170 (2016).

46. *See* GiveWell, *Changes to Our Top Charity Criteria, and a New Giving Option*, GIVEWELL BLOG (2022), https://blog.givewell.org/2022/08/17/changes-to-top-charity-criteria/ [https://perma.cc/6DYF-23CZ].

47. For an introduction to this literature, *see* Jochen G. Raimann et al., *Public Health Benefits of Water Purification Using Recycled Hemodialyzers in Developing Countries*, 10 SCI. REPS. 1, 11 (2020) (providing a useful study of the application of methods of purification which can be done at large-scale in developing countries, focusing on membrane filtration); Mark A. Shannon et al., *Science and Technology for Water Purification in the Coming Decades*, 452 NATURE 301 (2008) (preceding substantive work on the development of methods of purification which can be done at large-scale, such as waste-water recycling systems and membrane filtration); Bart Van der Bruggen, *Sustainable Implementation of Innovative Technologies for Water Purification*, 5 NATURE REVS. CHEMISTRY 217, 217–218 (2021) (providing a recent survey).

Absent tools for assessing risk under conditions of behavioral coordination, EA runs the risk of being incoherent by presenting cases which are either underdetermined (revisiting Buridan's Ass) or by producing action plans which violate norms of practical reasoning. There are some areas where EA is actively addressing coordination problems, including creating organizations to coordinate grant distribution; however, these efforts tend to under-appreciate the role of coordination outcomes of the donation process, including governmental coordination.

F. Compounding Complexity

Above, each subject of uncertainty is treated differently, but they often co-occur. This is trivially true for uncertainty about risk and outcome. If we are uncertain about risk, then (trivially) we are uncertainty about outcome. The reverse is not true, such as in cases where we have well-defined probabilities like roulette.[48] One can be uncertain about outcomes, risks, values, individuation, and social coordination all at the same time; in fact, this is true in many intellectually interesting cases. Those should compound our uncertainty about how we model cases. This does not render us clueless but should inform our reasoning.

For illustrative purposes, consider the case of choosing investments in water purification systems at two distinct levels: individual pumps and community-level purification of water sources. This is subject to different kinds of uncertainty. First, there is the individuation and interpretation of goals (e.g., whether the goal is reducing the number of individuals with access to clean water or ensuring access to water in every home). This is uncertainty about individuation. Second, there is uncertainty about the comparative valuation of these solutions; individual pumps are less secure (as they can be stolen, damaged, etc.) but easier to deliver and lower cost per person and can be extended out more easily to people in rural areas.

Third, there are concerns about national and local governance regarding development and upkeep of infrastructure which applies to the development of aqueducts and other community-based resource solutions. We are uncertain about long term governance of many such countries (as well as levels of corruption and other related issues). The second and third are short-term and long-term forms of uncertainty about risk and outcome, tangled up with uncertainty about coordination. Fourth, there are distinctions in value between community governance and individual access, and the preference of such values. This is a form of moral uncertainty. Fifth, even if we construct a risk profile for every element, there is serious uncertainty about the accuracy of any such profile and the prudential allocation of resources to develop that profile as part of research into the issue, which reiterates uncertainty about risk.

48. Uncertainty about outcome and risk are the only two that have a conceptual relation between them. The others discussed in this paper are conceptually independent.

This is a complex issue with a lot of working parts, basically all of which are subject to uncertainty except one: We are certain people fare better if they have access to clean water.

IV. EVALUATING LOCAL FAILURES TO APPLY THE DECISION THEORY WITHIN EFFECTIVE ALTRUISM

The preceding sections were supposed to be the entirety of this paper. Then, FTX crashed. The collapse shifted how internal critiques of EA Altruism operate. There is a need for immediate critical appraisal. Rather than focusing solely on the decision procedure, it is pressing to also consider critiques of the EA movement which also implicate uncertainty. Several major problems with EA are being more closely scrutinized following this massive financial fraud. These problems result from the EA movement failing to act according to norms of practical rationality and public reason, especially regarding uncertainty. There are systematic biases in EA, but we can expect systematic biases in any such movement. No amount of formalism excises the human element.[49]

Analyses of practical reasons for action are comparative. When we consider reasons for acting in terms of attitudes towards uncertainty, we must situate those choices in comparison to other courses of action. While EA may provide some pro tanto reasons for engaging in certain projects, especially for tail risk events, this approach frequently fails to consider the context of other possible courses of action. The fact that deworming is a low-cost intervention provides a pro tanto reason to donate, but it doesn't establish what we should do all things considered.[50]

Constraints on practical rationality apply under conditions of uncertainty. Even in the ecumenical view of EA, constraints on defensible action remain. There are circumstances where EA violates constraints on practical reason. If EA cannot deal with the criticisms raised below, then it will likely collapse (and probably should).

V. PRACTICAL INCOHERENCE WORRIES

The following subsections focus on two errors specific to effective altruism. Section V.A. concerns investment patterns of Effective Altruist leaders and groups and focuses particularly on "earn to give" as a mistake in practical reasoning about efficacy. Section V.B. raises the issue of non-consequentialist social values and how the failure of Effective Altruism to take those values seriously has led to damage to the movement, even on consequentialist analyses.

---

49. The concerns about St. Petersburg Paradoxes raised below also illustrate this is a necessity; purely algorithmic approaches to decision making cause buy-in and exit problems.

50. I use the word "should" rather than "ought" deliberately: to distinguish the force of practical reasons for the force of moral theory. There may ultimately be no such distinction; philosophy often takes for granted that modal terms of requirement are all identical. I want to allow room for skepticism of that thesis because of the distinction between the moral theory understanding and practical reasons understanding of EA.

A. "Earn to Give" and Aggressive, Risk-Tolerant Investing

Following the FTX fraud, there is increased attention on EA fundraising sources. Some Effective Altruists maintain an "earn to give" approach to thinking. There are two ways to understand "earn to give." The first is adopting a high-salary, but not socially valuable, profession in order to immediately flip larger portions of that wealth to EA-compatible philanthropic projects; the second is to leverage wealth to accumulate more wealth (often through aggressive investing practices), with the expectation that the wealth is eventually donated. The first version is subject to reasonable debate within EA;[51] the second, especially following the collapse of FTX, is not. This second version of "earn to give" runs contrary to maximizing requirements articulated by Peter Singer, because (as practiced) it involves reinvestment of excess income to increase future earnings rather than immediate donation.

Consider what motivates this second version of "earn to give." Basically, an individual can donate \$N at t^1 or \$M at t^2, where t^2 is later. If N and M are equal, the individual should donate at t^1, but if M is larger and represents a significant improvement in donation, there is a pro tanto reason to forego donating at t^1 subject to the difference. If I have the choice between donating \$10 today or investing such that it turns into hundreds or thousands of dollars down the road, it may be appropriate to invest to donate significantly more, later. This way of thinking is defective with respect to how it understands donation and investment, but this illustrates the basic thought pattern.[52] The success of the investment (how much the investment grows) is relevant to both the substantive and formal properties.

Many who adopt "earn to give" do not think of the practice as an investing pattern. Rather, they think of it as a choice about lifestyle (which individuals are entitled to make) incurring certain obligations down the road. That's a valid way of thinking, but we should know that it also is an investment pattern, investing time or money proper with expectations of returns in the future. To quote a sage: "Stuff can be two things."[53] There is a further affinity (including but hardly limited to FTX) between venture capitalists and EA. In some cases, this alignment promotes EA-aligned companies, such as Effective Ventures, but in other cases it does not.

There is an auxiliary argument in favor of "earn to give." This second formulation appeals to investment-minded prospective donors who are accumulating wealth. I avoid centering this argument, because it requires some deeply

---

51. There are variations on the argument against the venture capital "earn to give" practices I articulate below which push against the high-income variation accepted by Singer *inter alia*, but such arguments require adding multiple steps. If there is interest, and if EA survives, it may be the subject of a future paper.

52. This discussion focuses on good faith arguments for "earn to give" approaches; one challenge is that "earn to give" lends itself to bad faith exploitation, as in the FTX case. As a practical matter, Effective Altruism needs to take the bad faith exploitation of this dynamic more seriously. This ties into the arguments in Section V.B.

53. *Brooklyn 99: Jake and Sophia* (NBC television broadcast Nov. 9, 2014).

controversial assumptions about the good faith of such investors, assumptions which (following the FTX fraud) should be reexamined. The auxiliary argument runs full force into the objections in Section V.B.

Many hold that forgoing the good one can do today, for people in immediate and dire need, is facially wrong.[54] The challenge with a facial objection is consequentialist arguments provide a clearer grounding than facial appearance. The defense can just hang its counterargument on the size of the return on investment. This is not the best response. The best response to the facial objection is: investing a portion of charitable funds is a well-founded feature of long-term strategy for philanthropic endeavors.

Consider a traditional philanthropic organization like Rotary International. Rotary International maintains a considerable investment portfolio in service of its philanthropic giving. These investments are in projects the organization believes are themselves good. The proceeds from those investments fund future charitable activities, as well as ensure long-term financial stability for the organization.[55]

The arguments for traditional philanthropic investment practices are different from the arguments underlying "earn to give" investing. The traditional philanthropic arguments are institutionalist, focused on long-term stability, rather than consequentialist considerations. Unless one takes the facial objection all the way, that all investing is wrong, then the facial objection loses significant force. Some investment is appropriate; the questions are logistical and practical. How much investment is appropriate? How should those investments be handled? And so on.

I want to raise a more instructive objection. Effective altruists who defend the venture capital approach to "earn to give" regard it as a means of maximizing because they overvalue return on investment in comparison to the return on donations. If Joe's fund invests $1M of funds raised in a mutual fund that returns 11.5% annually, they can put that money towards a stable fundraising base. Alternatively, they could give the $1M to provide medical care and education to children. When the money is donated, we see the moral impact; when the money is in the bank, we don't. But this raises uncertainty about value; how do we understand the value of lives transformed by giving, in comparison to the returns from investing?

Giving now improves lives in the short term, but it also enables long term changes, including access to improved quality of life, income, etc. in ways which have hidden moral and economic value. Lots of analyses have attempted to flesh

54. Singer's drowning child argument (for example) entails this. *See* Peter Singer, *Famine, Affluence, and Mortality*, 1 PHIL. & PUB. AFFS. 229, 231–33 (1972).

55. For discussion, see Rotary International's running page on their investments. *See* ROTARY INTERNATIONAL, *Rotary's Investments,* https://www.rotary.org/en/rotary-investments [https://perma.cc/HRS4-FUB2] (last visited Feb. 24, 2023).

out the economic value provided by educational access to motivate investment[56] independent of the moral considerations.

The donor makes a comparative choice: to justify putting the money into the fund, the value of the appreciation of giving to the children's medical and educational resources must be greater than 11.5% per year compounding. For investing in the fund to be acceptable, the fund must offset the moral value of giving immediately and must consider the risk of the investment portfolio.

There are two shifts EA can make to justify investing practices. The first is to shift to the standard model of philanthropic investing, holding the long-term solvency of the organizations as sufficient independent reason. The second is to suggest rates of return on the investments are so large they actually do justify these patterns; this is the de facto response of crypto-driven "earn to give" investors. But this is like Jack buying magic beans. The investors believe investments (crypto or otherwise) are a magic product which invariably yield enormous returns. This is not supported by the present behavior of the market (which, for crypto, has long-since peaked), and the recent consequences of the failure of such investments, like the FTX fraud, illustrate the consequences of high-risk investing.

Note that the first shift would result in conservative, low-risk investing practices to ensure long term solvency, while the latter requires higher-risk investing for higher returns. This is a meaningful distinction in how EA organizations are structured. Investment patterns which ensure long term solvency and stability are defensible because the underlying attitudes are risk-averse; the purpose of such investments is to ensure that if things go badly in the future, there are still resources available. Aggressive investing patterns are risk-tolerant.

If the shift is based on maximizing returns, then those returns have to be significant, because the issue is comparative benefit, not just growth. This leads to aggressive investing strategies to maximize returns. These aggressive investment patterns are higher risk. Aggressive investment patterns have a place in developing personal wealth; some people are risk-seeking. However, seeking risks won't necessarily maximize funds over time; it's a short-term strategy that requires cashing out. If you keep making high-risk choices, over time, payouts regress to the mean; winnings from big wins early eventually deplete because (in high-risk circumstances) there will inevitably be more losses.

The Saint Petersburg Paradox is a paradox of practical reason. It supposes an agent is invited to play a game. The player makes an initial payment, then the coin is flipped until it comes up heads. A player wins $\$2^n$ where n is the number

56. The work of Reynolds and Temple is instructive here. *See* Arthur J. Reynolds et al., *Age 26 Cost–Benefit Analysis of the Child-Parent Center Early Education Program*, 82 CHILD DEV. 379 (January/February 2011); Judy A. Temple & Arthur J. Reynolds, *Benefits and Costs of Investments in Preschool Education: Evidence from the Child–Parent Centers and Related Programs*, 26 ECON. EDUC. REV. 126 (2007).

of times the coin was flipped.[57] Growth is exponential, and returns are potentially infinite. As a practical matter, most people do not think this is a good game to play unless the initial payment is low. It is good to play for $3, because if the first flip shows tails, the player comes out ahead. However, playing for $200 seems irresponsible, even though the potential winnings are infinite. As a practical matter, playing the game at relatively high buy-in costs is an indicator of high (even indefensible) levels of risk tolerance.

Nicholas J.J. Smith writes that "in any given context, there must be some finite tolerance—some positive threshold such that ignoring all outcomes whose probabilities lie below this threshold counts as satisfying the norm."[58] This is a minimal requirement of practical reasoning, although how to explain, justify, and represent it formally is subject to reasonable disagreement. The first permutation of coin flips, the $2 permutation, has a .5 probability of occurring. The $4 permutation has a .25 probability and is exclusive from the first. Any instance of the game has a .75 chance of returning $4 or less. The odds significantly diminish. The odds of the five-flip permutation (payout $32) is a shade over 3 percent; the 10-flip permutation, just under 0.01 percent, or a probability of 0.00097. At what point do we ignore possibilities as too unlikely to take seriously?

I will return to this point in Section VI, regarding risk analysis in longtermist EA projects. However, for now, it is worth noting that risk-tolerant approaches are worrying; if there is an investment approach at all, then it is easier to defend with the practical reasons given by traditional philanthropic endeavors.

B. The Guardrails Argument and Coordination

In Section I.C., I laid out my "guardrails argument" for EA: requiring public justification in philanthropic donation can limit capture of philanthropic institutions by the wealthy by creating an expectation that donors explain their donations, act transparently, raise funds in an ethical manner, etc. Obviously, the FTX scandal shows these guardrails failed in some EA fundraising, but the potential value of these guardrails remains.

The guardrails argument is not about consequences but rather the importance of certain non-consequentialist values in philanthropy. The guardrails argument is about public trust and being trustworthy; promoting public trust is (when guardrails are in place) a reason to prefer EA to alternatives. One can ground public trust through a consequentialist argument (e.g., individuals are more likely to participate if they have trust) but focusing on strict consequence-driven justifications opens cases where public trust can be exploited or abused; the possibility of such cases erodes the grounds of public trust.

---

57. Martin Peterson, *The St. Petersburg Paradox*, *in* The STANFORD ENCYCLOPEDIA OF PHILOSOPHY (2022).

58. Nicholas J.J. Smith, *Is Evaluative Compositionality a Requirement of Rationality?*, 123 MIND 457, 472 (2014).

Public trust should be earned; public trust should be trust in individuals based on an accurate expectation of good character. That challenge is a general problem which applies to all forms of social power. Public trust is not exhausted by its impact on outcomes, but it can produce positive outcomes, including improved behavioral coordination, responsible conservation and use of resources, etc.

A person or group is trustworthy only if, when the person or group makes a promise, we can have a reasonable belief the person or group will follow through. However, if EA organizations are bound by their commitment to assessments of the consequences ahead of their promises, then (in cases where the two conflict) they will break their promises. Consider the reverse: Suppose an EA organization promises to send their money to a set of public health and education charities during fundraising, but (during the year) the organization shifts to prioritize research in artificial general intelligence. Should the organization shift donations raised based on the promise of donating to public health and education?

If one only considers the efficacy conditions (or, rather, the organization's understanding of those efficacy conditions), then one should break the promise. This would result in the company comporting with EA but would be untrustworthy.[59] It is necessary for being trustworthy that the organizations (except in extreme circumstances, not present here) try to satisfy promises, especially promises used to raise money.

All of this lays the groundwork for a problem with the guardrails argument: Have EA organizations conducted themselves in ways that merit trust? The answer to this question is complicated because EA is not homogeneous. Hopefully we can agree the behavior of FTX, Bankman-Fried, and other leadership within that cluster of companies does not merit trust. That should be easy. But what about public voices of EA who promoted Bankman-Fried or vouched for him?[60] Unlike FTX and the family of companies now insolvent, many of those figures will remain parts of EA. Should the public trust EA organizations or the movement collectively? If so, why? Has the ground of that public trust crumbled?

VI. LONGTERMISM, AGI, AND EXPLOITATION OF MORAL UNCERTAINTY TO PRO TOTO CONCLUSIONS

"Longtermism" is a set of positions within EA which center a particular interpretation of impartiality. EA is committed to impartiality regarding gender, race, ethnicity, religious background, etc. Longtermism holds that impartiality should also include future persons.[61] A person should not be biased for the time at which he or she is born.

---

59. As a conceptual matter, having the disposition to shift the donations makes the company untrustworthy, not the actual shifting of the donations. Shifting the donations is just the realization of the disposition. But this conceptual point requires more background work than I can prudently do here.

60. We can suppose that these people were not acting duplicitously, but rather made good faith mistakes based on failures of due diligence. However, negligence can also make people untrustworthy.

61. One could argue it holds "time of birth," but what it would mean to be impartial towards past persons is a more complicated question.

There are good reasons to embrace impartiality towards future persons. It provides practical reasons for limiting risks, including climate change and future pandemics. There are two broad swaths of longtermism. One set is unobjectionable: combatting climate change, developing public health infrastructure, and making long term investments in education, science, and technology. The second raises concerns: investing heavily in "artificial general intelligence" (AGI) research based on low-probability possibilities of utopia or dystopia. The difference between these two versions is the severity of uncertainty. Public health issues and climate change are near certain problems for future generations; AGI (panacean or malignant) is not. We are uncertain about the risks of AGI, but they are lower than the probability of natural disasters or another global pandemic.

All dimensions of uncertainty apply in the AGI case. We are uncertain about: what those outcomes would look like, the probabilities of those outcomes, the relative value of the possibilities with AGI. Uncertainty does not necessitate agnosticism. Is the uncertainty sufficiently severe that we should discount the possible cases outright?

The above discussion of the St. Petersburg Paradox establishes a useful limiting principle: there is some range of probabilities such that the possibilities should be discounted. This is necessary in practical reasoning contexts because the range of possibilities is so vast that trying to handle every unlikely possibility (even if we limit only to those with extreme positive or negative values) is overwhelming.

The low probability is a familiar feature of these discussions. However, the problem is not limited to the probability of the events occurring, but also to the relative probability of those events in comparison to other events which might be prevented. The defenses of aggressive, risk-tolerant approaches focus on explosions of value with the increase in volume of lives in the far future. Pettigrew's discussion of longtermist axiology illustrates how the open-endedness of this increase can result in wildly counter-intuitive judgments, including pushing for human extinction, as reductio ad absurdum.[62]

Decision theories can be constructed in a range of ways; if one has a high risk-tolerance, then a decision theory can justify St. Petersburg games at a high entry cost because of its infinite return. If we cut the consideration of probabilities before we get to the longshot events, as Smith suggests for the St. Petersburg games, we wind up with a more reasonable decision theory; if we do not, then we run into the reductio ad absurdum developed by Pettigrew, where tolerance of tail-risks must consider hastening extinction. I hope, as does Pettigrew, that we agree hastening extinction is an unreasonable (though not formally irrational) and unacceptable action plan,[63] so we should avoid decision theories which produce this conclusion.

---

62. Pettigrew, *supra* note 38.

63. Some people may see modus ponens where Pettigrew and I see modus tollens. That is always the hazard of reductions. I just hope human extinction is viewed as off the table by most participants.

VII. DECISION THEORIES AND PUBLIC REASON UNDER UNCERTAINTY

Making decisions under conditions of uncertainty requires subjective judgment; there is no single, clear plan produced simply by application of decision theoretic tools. We can eliminate certain unsupported actions, but as we pare away the unacceptable choices (e.g., choices which are strictly worse than alternatives), we don't come down to a single, correct choice. Usually, we get a range of acceptable options. Uncertainty about probability and value ensures there are limits to our confidence, and attitudes to risk and uncertainty held by individuals or groups matter.

Variation in personal judgment necessitates public trust. If someone is going to be responsible for allocating resources based on their own discretion, we need some foundation for trusting their judgment. The present problem for EA is the collapse of public trust. The acute, proximate cause of this loss of public trust is easy to identify: the collapse of FTX as a result of fraud and other questionable business practices in EA the collapse illuminated; the above illustrates there are broader causes which resulted in poor handling of uncertainty and which can be addressed to improve public trust going forward.

Improving the way in which EA handles uncertainty is one area where public trust can be improved. Whether improvements will be actively or effectively implemented within EA, including shifting away from risk-tolerant strategies which fall outside of the realm of practical defensibility, is one way to start rebuilding that public trust.