A Systematic Response to Criticisms of Effective Altruism in the Wake of the FTX Scandal

Summary

Effective altruism (EA) has been in the news recently following the crash of a cryptocurrency exchange and trading firm, the head of which was publicly connected to EA. The highly-publicized event resulted in several articles arguing that EA is incorrect or morally problematic because EA increases the probability of a similar scandal, or that EA implies the ends justify the means, or that EA is inherently utilitarian, or that EA can be used to justify anything. In this post, I will demonstrate the failures of these arguments and others that have been amassed. Instead, there is not much we can conclude about EA as an intellectual project or a moral framework because of this cryptocurrency scandal. EA remains a defensible and powerful tool for good and framework for assessing charitable donations and career choices.

Introduction

Recently, there has been a serious scandal primarily involving Sam Bankman-Fried (SBF) and his cryptocurrency exchange FTX, precipitating a crash of billions of dollars into bankruptcy. I am talking about this because SBF has been publicly connected to the effective altruism movement, including being upheld as a good example of "earning to give," which is where people purposely take lucrative jobs in order to donate even more money to effective charities. For example, Oliver Yeung took a job at Google and is able to donate 85% of his six-figure income to charities while living in New York City; for four years, he lived in a van to push this up to 90-95% of his income.

SBF met William MacAskill, one of the leaders and founders of the effective altruism (EA) movement, in undergrad, and MacAskill convinced him to go into finance to "earn to give." SBF did very well, working at a top quantitative trading firm, Jane Street, and he decided to work with some other effective altruists (EAs) to start a trading firm Alameda Research and eventually a cryptocurrency exchange FTX that was intimately connected with Alameda. FTX and Alameda were doing really well, ballooning in the past several years. At his peak, right before the downfall, SBF had a net worth of $26 billion.

Like many other cryptocurrency exchanges, FTX produced its own altcoin, FTT, which gives some discounts and rewards to customers and acts as stock, and SBF had some of his company's own assets in FTT. Trouble started in early November when CoinDesk published an article expressing concern over Alameda's balance sheet, revealing an unhealthy amount of assets invested in FTT, which is essentially its own made-up currency. FTT-related assets amounted to over $6 billion assets of Alameda's $14 billion assets, leaving Alameda extremely vulnerable to sudden drops in investment due to their limited ability to liquidate enough assets to pay the sellers.

Unfortunately for SBF, the Binance CEO decided to sell all of Binance's FTT tokens, collectively worth $529 million. The CEO also publicly announced the sale, triggering a bank-run where many other customers decided to sell their FTT and withdraw their funds from FTX entirely. As a result of the run, $6 billion was withdrawn from FTX within 72 hours. FTX did not have the liquid assets to cover all of this and rapidly collapsed, declaring bankruptcy.

It became apparent that the investments of Alameda were extremely risky, even though they repeatedly told customers they have loans with "no downside" and high returns with "no risk." It was revealed that Alameda's risky bets were made with customer deposits, which is apparently a big "no-no". As far as I can tell, it is not clear whether SBF actually committed fraud, but he clearly mishandled funds and misled customers about their funds, possibly in a way that violated the business's terms and conditions.

In the fallout of this disaster, which included the closing of over 100 other organizations and the loss of many employees' life savings, etc., effective altruism came under fire for their connection to SBF. SBF, was, after all, following suggestions given by EA organizations when he decided to "earn to give." Further, he has explicitly advocated for EA-adjacent reasoning in maximizing expected value, though he also champions a more risk-tolerant approach than EAs tend to prefer.

The question everyone is asking (and most are poorly answering) is: "Is effective altruism to be blamed for SBF's behavior?"

Many articles in popular media have denounced effective altruism in the wake of the crash, characterizing the philanthropic approach as "morally bankrupt," "ineffective altruism," and "defective altruism." They say the FTX scandal "is more than a black eye for EA," "killed EA," or "casts a pall on [EA]." Articles linking the scandal and EA, most of them critical of EA, have been published in the New York Times, the Guardian, the Washington Post, New York Magazine, the Economist, MIT Technology Review, Philanthropy Daily, Slate, the New Republic, and many other sites.

In this post, I am going to subject these articles and their arguments to scrutiny to see what exactly we can conclude about EA's framework of evaluating the effectiveness of charities and careers and how they advocate for why and how we should do so in the first place. **In short, my answer is: not much. There is not much we can conclude about EA from the FTX scandal.**

I am only going to be investigating in search of critiques and assessing the articles as critiques of effective altruism. Some of these articles might have additional or entirely different purposes but sound sufficiently negative toward EA that I will nonetheless assess whether we can construct an argument against EA as a result.

Furthermore, I want EA to be criticized in the same sense that, for any given position, I want the best arguments and evidence for and against each side to be raised and assessed in the most rigorous way. Of course, that doesn't mean every argument is equally good. I have spent much time looking at academic critiques of effective altruism, which I (normally) find more compelling, as they are more rigorous. However, most recent online criticisms are just not good.

In this post, I will 1) give a precise characterization of effective altruism, 2) mention possibly relevant background information that informs my perspective in evaluating EA, 3) address what seems to be the most frequent concern, yet to my mind remains the most perplexing concern, that SBF's association with EA reveals that EA has an incorrect framework, 4) respond to arguments against EA that rely on the utilitarian origins of EA or its leadership, 5) clarify "ends justify the means" reasoning in recent discourse and normative ethics more broadly, 6) introduce six differences between EA and utilitarianism, showing that is EA independent of any commitments to consequentialism, and, finally, 7) respond to the concern that EA or consequentialism or longtermism can be used to justify anything and is therefore incorrect. With each argument, I try to reconstruct what is the best version of the critique against EA, since much of the argumentative work in these articles is left implicit or neglected entirely.

I welcome responses, better reconstructed arguments, corrections, challenges, counter-arguments, etc. Let's dive in.

Effective Altruism Revealed

In "The Definition of Effective Altruism,"[1] William MacAskill characterizes effective altruism with two parts, an intellectual project (or research field) and a practical project (or social movement). Effective altruism is:

> (i) the use of evidence and careful reasoning to work out how to maximize the good with a given unit of resources…
>
> (ii) the use of the findings from (i) to try to improve the world.

We could perhaps summarize this to say that **someone is an effective altruist only if they try to maximize the good with their resources, particularly with respect to charitable donations and career choice**, since that is EA's emphases. A few features of this definition that MacAskill emphasizes are that it is: non-normative, maximizing, science-aligned, and tentatively impartial and welfarist.

We can further distinguish between different kinds of effective altruists[2]: *normative EAs* think that charitable donations that maximize good are morally obligatory, and *radical EAs* think that one is morally obligated to donate a substantial portion of one's surplus income to charity. *Normative, radical EAs* combine these two together, and I independently argue for normative, radical EA in a draft paper (see n. 2). It is helpful to distinguish these kinds of EAs (minimal, normative, radical, or normative radical), where the summary of MacAskill's definition is considered the *minimal* definition that constitutes the *core* of effective altruism, while the normative and radical commitments are *auxiliary hypotheses* of effective altruism. I will revisit this in the Effective Altruism is Not Inherently Utilitarian section.

Based on the characterization above, we can quickly dispel two key errors that articles repeatedly made. One error is that "effective altruism requires utilitarianism" (then "utilitarianism is false", concluding "EA is incorrect"). The truth is that utilitarianism (trivially) implies effective altruism, but effective altruism does not imply utilitarianism. In fact, I would put effective altruism at the center of the Venn diagram of the three moral theories (see Figure 1). There are strong deontological and virtue ethical arguments to be made for effective altruism. See Effective Altruism is Not Inherently Utilitarian section for more on this, including one theory-independent and two virtue ethical arguments for EA. Also, see this 80,000 Hours Podcast episode on deontological motivations for EA.

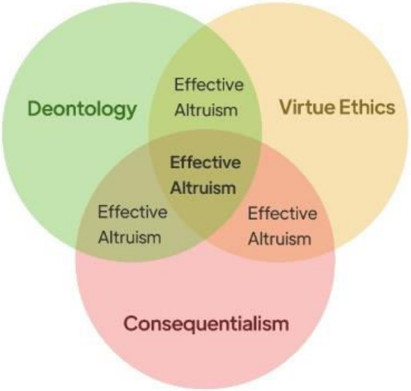


Figure 1: A Venn diagram showing what moral theories imply effective altruism

The second important flawed criticism is that longtermism is an essential part of effective altruism. The *core commitments* of effective altruism do not imply longtermism, and longtermism does not require effective altruism. Instead, longtermism is an *auxiliary hypothesis* of EA. Longtermism could be false while EA is correct, and EA could be false while longtermism is correct. To get from EA to longtermism, you need an additional premise that "the best use of one's resources should be put towards affecting the far future," which longtermists defend, but EAs can reasonably reject. EA is committed to *cause neutrality*, so it is open to those who think non-longtermist causes should be prioritized.

As we will see, many people writing articles with criticisms of effective altruism could really stand to read the FAQ page on effectivealtruism.org, as many of the objections have been replied to at-length (not to mention academic level pieces), including the difference between EA and utilitarianism or neglecting systematic change. Another, slightly more advanced, but more precise discussion on characterizing effective altruism is in the chapter "The Definition of Effective Altruism" by MacAskill. The very first topic MacAskill covers in the "Misunderstandings of effective altruism" section is "Effective altruism is just utilitarianism."

My Background

I call myself an effective altruist. I think that effective altruism is obviously correct with solid arguments in its favor. It follows from very simple assumptions, such as i) it is always permissible to do the morally best thing,[3] ii) acting on strong evidence is better than acting on weak evidence, iii) if you can help someone in great need without sacrificing anything of moral significance, you should do so, etc. If you care about helping people, you are spending money on things you don't need, and you don't have infinite money, then you might as well give to where it helps the most. This just makes sense. On the other hand, I wouldn't call myself a longtermist[4] (regarding either *weak longtermism* that says affecting the longterm future is a key moral priority or *strong longtermism* that says it is the most important moral priority), as I am skeptical about many of their claims. I simultaneously think most critiques I have heard of longtermism (I have not read much, if any, academic work on this) are lacking.

I have known about effective altruism since early 2021 and took the Giving What We Can pledge in March 2021. However, I was convinced of its way of thinking for several years, since early in undergrad. I have mostly been a part of Effective Altruism for Christians (EACH) more than the broader EA movement. I have not worked for an EA organization directly and do not have a local EA group to be a part of. I had never even heard of Sam Bankman-Fried until this whole scandal happened, though I heard other people talking about the FTX Future Fund (but I didn't know what FTX was).

The closest to an "insider look" I have gotten into EA as an institutional structure is conversations with some people at an EACH retreat in San Francisco, one of which worked for an EA startup and started an EA city chapter. The other has been involved in the EA Berkeley community. Some of the things they said suggested that there are ways that various EA suborganizations could be further optimized in their use of funding, but nothing super concerning.

I will be mostly looking at recent pieces insofar as they contribute to the debate about the intellectual project and moral framework of EA, as I find that to be the most interesting, important, and fundamental questions at hand. The end result of this inquiry has direct bearing on whether we should give to EA-recommended charities like GiveWell, rather than asking e.g., whether the Center for Effective Altruism should spend less on advertising EA books, which is a different question entirely and not central to the EA project. Additionally, I have engaged with enough material on the moral frameworks in question (and normative ethics more broadly) to hopefully have something to contribute to evaluating the EA moral framework .

SBF Association Argument Against Effective Altruism

A lot of recent critiques of EA appeared to have the general outline of the form:

1. Sam Bankman-Fried (SBF) engaged in extremely problematic practices.
2. SBF was an EA/was intimately connected to EA/was a leader of EA.
3. Therefore, EA is a bad or incorrect framework.

(1) is uncontroversial. On (2), SBF was clearly connected in a very public way to EA. The extent to which he was following or internalized EA principles can be challenged, and I will also question in inference from (1) and (2) to (3). What exactly is the argument from SBF's actions and connection to EA to concluding that EA is either inherently or practically problematic?

Was SBF Acting in Alignment with EA?

The most relevant question in this whole debacle is that whether the EA framework implies that SBF acted in a morally permissible manner. The answer is this: **it is extremely unlikely that, given the EA framework, what SBF did was morally permissible.**

EA leaders have repeatedly repudiated the general type of scenario that SBF engaged in numerous times. In fact, William MacAskill and Benjamin Todd give financial fraud as a go-to example of what would be an impermissible career choice on an EA framework. Eric Levitz in the Intelligencer acknowledges this by saying that "MacAskill and Todd's go-to example of an impermissible career is 'a banker who commits fraud.'" Eric says that MacAskill and Todd specifically argue that "engaging in harmful economic activity to generate funds for charity probably is [wrong]." Additionally, "they suggest that performing a socially destructive job for the sake of bankrolling effective altruism is liable to fail on its own terms."

It is very difficult to see how a virtually guaranteed bankruptcy, when thousands of people are depending on you for their lifesavings, jobs, and altruistic projects, is actually the best moral choice. Fraud is just a bad idea and is completely independent of effective altruism. The disagreement here may merely be on the empirical question rather than the moral question (it is notoriously difficult, at times, to separate empirical from moral disagreement, as empirical disagreement is often disguised as moral disagreement).

MacAskill calls out SBF's behavior as not aligned with EA: "For years, the EA community has emphasised the importance of integrity, honesty, and the respect of common-sense moral constraints. If customer funds were misused, then Sam did not listen; he must have thought he was above such considerations." Furthermore, "if he lied and misused customer funds he betrayed me, just as he betrayed his customers, his employees, his investors, & the communities he was a part of."

Additionally, his practices were just clearly horrible financially. He misplaced $8 billion dollars. John J. Ray III, who oversaw the restructuring of Enron and is now overseeing FTX, said about the FTX financial situation, "Never in my career have I seen such a complete failure of corporate controls and such a complete absence of trustworthy financial information as occurred here. From compromised systems integrity and faulty regulatory oversight abroad, to the concentration of control in the hands of a very small group of inexperienced, unsophisticated and potentially compromised individuals, this situation is unprecedented." These practices obviously do not give a maximum expected value on any plausible view.

SBF Denies Adhering to EA?

In addition, Sam Bankman-Fried himself appeared to deny that he was actually attempting to implement an EA framework, though he later clarified his comments were about crypto regulation rather than EA. As Nitasha Tiku in The Washington Post (non-paywalled) puts it, "[SBF] denied he was ever truly an adherent [of EA] and suggested that his much-discussed ethical persona was essentially a scam." Tiku is referring to an interview between SBF and Kelsey Piper in Vox. Piper interviewed SBF sometime in the summer, where SBF said that doing bad for the greater good does not work because of the risk of doing more harm than good as well as the 2nd order effects. Piper asked if he still thought that, to which he replied, “Man all the dumb sh*t I said. It’s not true, not really.”

When asked if that was just a front, as a PR answer rather than reality, to which he said, “everyone goes around pretending that perception reflects reality. It doesn’t.” He also said that most of the ethics stuff was a front, not all of it, but a lot of it, since it’s just about winners and losers on the balance sheet in the end. When asked about him being good at frequently talking about ethics, he said, “I had to be. It’s what reputations are made of, to some extent…I feel bad for those who get f---ed by it, by this dumb game we woke Westerners play where we say all the right shiboleths [sic] and so everyone likes us.” He said later, though, that the reference to the “dumb game we woke Westerners play” is to social responsibility and environmental, social, and governance (ESG) criteria for crypto investment rather than effective altruism.

Perhaps the most pessimistic and antagonistic of people would say, perhaps as Tiku did, that SBF only said what he did to protect EA. The idea is that he actually was an effective altruist, believed it, but lied about it just being a front in order to help save face for EA. Tiku says that EA’s brand “helped deflect the kind of scrutiny that might otherwise greet an executive who got rich quick in an unregulated offshore industry,” also reflected in the title of the article, “The do-gooder movement that shielded Sam Bankman-Fried from scrutiny.” Since we do not have access to SBF’s mental states, I do not care to speculate much about his reasoning for saying what he said. Armchair psychoanalysis is not exactly a reliable methodology.

People argue about whether or not SBF was being truthful here or not. He appeared to believe he was speaking off the air, suggesting honesty. If so, then he did not believe he was actively trying to implement the EA framework (unless SBF’s answers about his ethics in the Vox interview were intended to be disconnected from the EA framework and solely about regulations, which to me is not clear either way but didn’t seem entirely disconnected). Ultimately, I do not think much hinges on whether SBF believed he was implementing the EA framework, since it is more important whether or not SBF’s actions are a reflection of what is inherent in the EA framework, which they are not.

Now, I have little interest in attempting to disown SBF because he is now a black sheep. There is no doubt that EA painted SBF as a paradigm case of an actor doing great moral good by using his money to invest in and donate to charity. We EAs have to own that, and EAs got it incorrect due to our lack of knowledge about what was happening behind the scenes. Could there have been more to be done to prevent this from happening? Probably, and EAs are taking this very seriously, doing a lot of soul searching. It is likely there will be more safeguards put into place. These are reasonable questions, but they have little to do with the moral framework of EA itself, since the EA framework still ends up rendering SBF’s gamble as impermissible.

Next, I will investigate whether or not the mere connection between SBF and EA, rather than an alignment between EA’s framework and SBF’s actions, is sufficient to challenge EA’s framework.

EA is Not Tainted by SBF

Now that we know SBF's actions do *not* coincide with EA principles, we can investigate how the connection between SBF and EA could be used as an argument against EA. Recent articles mostly seem to just toss the two names next to each other in an obscure way without making any clear argument, hoping that one will be tainted by the other.

An Irrelevant "Peculiar" Connection

For example, Jonathan Hannah in Philanthropy Daily says, "MacAskill claims to be an ethicist concerned with the most disadvantaged in the world, and so it seems peculiar that he was inextricably linked to Bankman-Fried and FTX given that FTX claimed to make money by trading cryptocurrencies, an activity that carries serious negative environmental consequences and may play a role in human trafficking." The environmental consequences have to do with crypto mining that uses a lot of electricity (more than some countries as a whole), and the role in human trafficking is that virtual currencies are harder to track, so they are frequently used in black market activities.

It is hard to understate how much of a stretch this argument is. Here is an equivalent argument against myself (relevant background is that I studied chemical engineering at Texas A&M, which also has a strong petroleum engineering program). I say I care about the disadvantaged, yet I have many friends that went into the oil and gas industry (and some of them listened to my suggestions about charitable donations). Oil and gas bad. Curious! Further, I have many more friends that love, watch, and/or attend football and other public sporting events, and yet these events are associated with an increase in human trafficking.[5] Therefore…I don't care about the disadvantaged? And therefore my thoughts (or knowledge of evidence like randomized control trials) about helping others are wrong? Looks not much better than Figure 2.

WE SHOULD CARE FOR THE DISADVANTAGED
YET YOUR FRIEND HAS A JOB THAT DOESN'T HELP THE DISADVANTAGED. CURIOUS!
9-13-16
DIST. BY UNIVERSAL UCLICK
BORS
imgflip.com

Figure 2: I am very intelligent.

Of course, effective altruists have spent a great deal of time working on the issue of weighing the moral costs and benefits of working in plausibly harmful industries vs working for charities. This isn't exactly their first rodeo. See *80,000 Hours: Find a Fulfilling Career That Does Good* and *Doing Good Better: Effective Altruism and How You Can Make a Difference* (you can get a free copy of either of these at 80,000 Hours). We can also quickly consider SBF's scenario (I am only considering my first-glance personal thoughts, and not attempting to use the 80,000 Hours framework). In SBF's case, he has earned enough money from cryptocurrency to carbon offset all the cryptocurrency greenhouse emissions in all of the U.S. many times over.[6] Additionally, it is hard to see why employees (or employers) of cryptocurrency can be blamed for human trafficking purchases with crypto, especially no more than the U.S. treasury can be blamed for human trafficking purchases done with cash (which seems negligible at best). Plus, many other things he can do with the remaining sum not spent on carbon offsetting, resulting in a net good (especially compared to what other job opportunities he could take, many of which have comparable negative effects).

Skills in Charity Evaluation ≠ Skills in Fraud Detection in Friends

The same author also asks, "If these 'experts' failed to see what appears to be outright fraud committed by someone they were close to, why should we look to these utilitarians to learn how to be effective with our philanthropy?" This is again a strange conditional. Admittedly, I have not had many friends that committed billions of dollars' worth of fraud (perhaps the author has more experience), but I would not expect them to go to their close friends and say, "Hey I'm committing fraud with billions of dollars, what do you think?" Acts like those done by SBF are done in desperation with a sinking ship, like a mouse backed into a corner, or someone with a gambling habit (especially apropos for the given situation). You get deeper into debt, take more risks, assuming and desperately hoping that it will work out in the next round. Repeat until bankruptcy. This is not something you go telling all your friends about (instead, you lie and try to siphon money from them, as was recently done by a Twitch scammer).

In addition, the skills and techniques it takes to assess the effectiveness of charities are quite different from the skills it takes to discover that your friend is committing massive fraud with his business. So, the reason we should look to EAs to be effective in philanthropy is because they have good evidence for charity effectiveness. Randomized control trials (or other comparable methods) are not exactly the tools optimized for detecting fraud in friends' businesses.

Now, was there nothing suspicious about SBF prior to this point? No. There was some reason for suspicion. And of course, hindsight is 20-20. They evidently attempted to evaluate SBF and his ethical approach in 2018. I'm unsure the details of this, and I don't know how much changed in SBF's behavior in 4 years. As I mentioned earlier, like the desperation of a gambler, the risks and bad behavior likely exponentially increased over time leading to the present failure. Thus, we would expect most of the negative behavior to be heavily weighted towards 2022 rather than 2018 when he was reviewed. This debacle will likely increase scrutiny into this type of behavior (as much as possible across organizational lines), and with good reason. I won't say EA as an organization or community is blameless here. But that doesn't change the EA framework has being the best (and correct) framework for evaluation of charity effectiveness.

Without making this connection more explicit, this looks like a fallacious argument; however, like all informal fallacies, there is likely a reasonable argument form in the vicinity. Let us try to consider some of these possibilities.

EA Does Not Problematically Increase the Risk of Wrongdoing

Here is one way of putting the key inference for this argument: *if something increases the probability of believing or doing something wrong, then it is bad or incorrect* (and EA does this, so EA is incorrect). Of course, this is implausible, as then we couldn't do anything (re: MacAskill's paralysis argument). If we always had to minimize the probability of engaging in wrongdoing (through violating constraints) or false beliefs, then we should do (or believe) nothing.[7] This is one standard argument for global skepticism. If the only epistemic value is minimizing false beliefs, then having zero beliefs would ensure you have the minimum number of false beliefs, which is zero. This approach is clearly incorrect, since we do have knowledge and it is permissible to get out of bed in the morning.

Here's another reductio: becoming a deontologist increases the probability that you will believe that we have a deontological requirement to punch every stranger we see in the face, since consequentialism does not include deontological requirements while deontology does, so deontologists need to put higher credence in variants of deontology. However, this is an implausible view that no one defends, so this mild increase in probability is uninteresting at best.

A second, more plausible version of the inference for this argument is: *if something <u>substantially</u> increases the probability of believing or doing something wrong, then it is bad or incorrect* (and EA does this, so EA is incorrect). Random on Twitter seems to suggest something like this in response to Peter Singer's (too) brief article when he identified the criticism as being that EA is "a philosophy that tends to lead practitioners to believe the ends justify the means when that's not the case." In any case, this is an extremely difficult and unwieldy claim to deal with at all, as this empirical premise is quite difficult to substantiate. First of all, increases the probability compared to what? What is the base rate for how frequently someone does the relevant wrong in question? And what is the probability given one is an EA? Do we only compare billionaires? Do we compare millionaires and beyond? Do we only compare SBF to other crypto businessmen?

In the absence of a more clear and substantiated argument, it is hard to see how this argument can be successful. Maybe we can ask, of the people that we know made incorrect assessments of ends vs means and thought the ends sometimes justifies the mean, what percent of them accept the EA framework? Good luck with that investigation. Plus, we are inevitably going to end up doing armchair psychoanalysis, a notoriously unreliable method.

Furthermore, there is another response. Plausibly, *a framework can substantially increase the probability of people doing something wrong, and yet the framework entails that we should not do that thing.* In such a case, it is hard to see why the framework goes in the trash if it gives the correct results even if in practice people's attempted implementation end up doing the wrong thing.

To see this, consider is the difference between a criterion of rightness, which is how we evaluate and conclude if an action is morally right or wrong (as a 3rd party), and a decision-making procedure, which is the method an agent consciously implements when deciding what to do. This is a standard distinction in normative ethics that diffuses various kinds of objections, especially having to do with improper motivations for action. It may be that the decision procedure that was implemented is wrong, but this does not show that the normative or radical EA's criterion of rightness is incorrect. I suspect that Richard Chappell's meme about this distinction is actually a reference to this (or a closely related) mistake, since his other tweets and blog posts around the same time are referring to similar errors in commentary on EA and the FTX scandal (such as this thread on a possible connection between guilt-by-association arguments and inability to distinguish criterion of rightness and decision procedure).

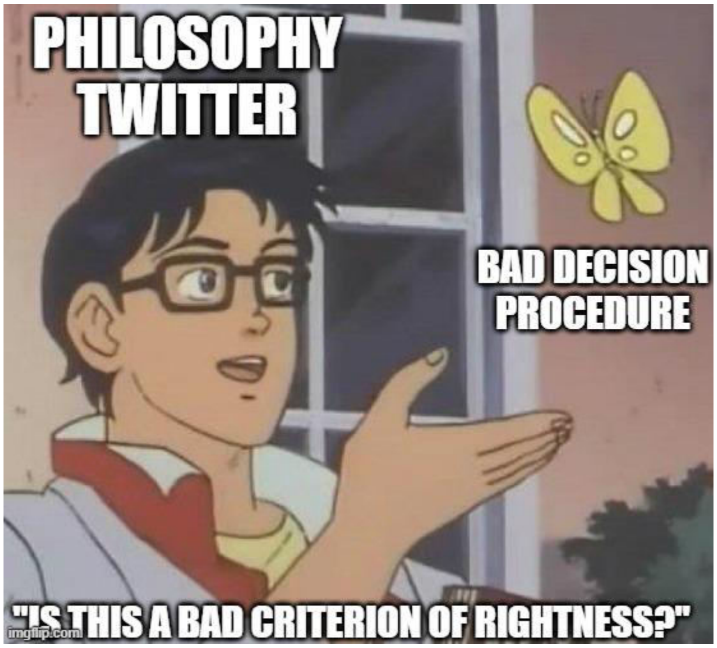


Figure 3: Richard Chappell's meme on bad EA criticism, referring to philosophers on Twitter that confuse the two

In summary, to answer Eric Levitz's question "Did Sam Bankman-Fried corrupt effective altruism, or did effective altruism corrupt Sam Bankman-Fried?", the answer is "Neither." SBF did not act in a way aligned with EA, whether he thought he was or not. **Until a better argument is forthcoming that SBF's incorrect approach implies that EA's framework is flawed, I conclude very little about the EA framework.**

The EA framework is well-motivated, even on non-consequentialist grounds (as we will see later), and EA is an excellent way to help others through your charitable donations and career. To the extent that the FTX scandal makes EA look bad, it is only because of improper reasoning. There are likely additional institutional enhancements that can be implemented as protections against these kinds of disasters, but my intent here was to investigate the EA framework more than the EA practice in all of its institutional details, to which I am not privy. **Therefore, I can conclude that the EA framework is correct and unmoved by the SBF and FTX scandal.**

Genetic Utilitarian Arguments Against Effective Altruism

There is another set of claims I will assess in these critical articles related to effective altruism's connection to utilitarianism in the form of historical and intellectual origins. Inevitably, especially from opponents of utilitarianism, any connection to utilitarianism is deemed hazardous and not to be touched with a ten-foot pole. For example, I have had several Christian friends be terrified of effective altruism because they hear that Peter Singer is connected to it.[8]

Genetic Personal Argument Against EA

I can briefly consider this genetic personal argument against EA. The best version of the principle in question to make an inference against EA is probably something like, "if a person is wrong about the majority of claims you have heard from that person, then the prior probability of the person being right about a new claim is fairly low." The principle should likely be restricted to the claims that you have heard from that person that you got from a source including many more of that person's beliefs and even arguments for said position. Otherwise, you risk making inferences from an exaggerated source, and the principle would be false. Even then, the principle would only tell you the prior probability. You need to update your background knowledge with further evidence to get the posterior probability of any given claim, so it remains important to actually investigate the person's reasons for believing the new claim before making a definitive judgment on the new claim. Therefore, EA cannot be dismissed on a personal basis without assessing the arguments for EA, such as those referenced in the independent motivation section.

Genetic Precursor Argument Against EA

There may be another genetic argument raised against EA, which is that "the historical and intellectual precursors to EA involved utilitarian commitments, and so EA is inextricably linked to utilitarianism. Further, utilitarianism is false, and therefore EA is false." I will examine each part of this argument in turn.

First, we need to examine the factual basis of the historical and intellectual connection between EA and utilitarianism in the first place. A number of recent critical articles point out the genetics of the EA tradition. I think facts about this connection are worth pointing out; yet it is important to clarify the contingent nature of this linkage, especially given how despised utilitarianism is to the average person. If this clarification was neglected as a kind of "poisoning the well" or "guilt by association", shame on the author, though I do not make that assumption.

The Economist (non-paywalled) writes, "The [EA] movement…took inspiration from the utilitarian ethics of Peter Singer." It would be more accurate to say that "the movement took inspiration from arguments using common sense intuitions from Peter Singer, and Peter Singer is a utilitarian." Of course, that's much less zingy to acknowledge that the arguments from Singer inspiring EA were not utilitarian in nature (from his "Famine, Affluence, and Morality"), as we discuss with more detail in the utilitarian-independent motivation subsection of Effective Altruism is Not Inherently Utilitarian section.

Rebecca Ackermann in Slate writes, "The [EA] concept stemmed from applied ethics and utilitarianism, and was supported by tech entrepreneurs like Moskovitz." This is just a strangely worded sentence. It would make more sense to say it stemmed from *arguments in* applied ethics, but applied ethics is merely a field of inquiry. Moreover, utilitarianism is a moral theory. So, you could say it is an *implication* of utilitarianism, but proposing that EA *stemmed* from a moral theory is a bit weird. That's mostly nit-picking, and I also have absolutely no idea what the support from tech entrepreneurs has to do with anything. I guess the “technology” audience cares? Other articles appear to poison the well against EA merely by saying rich tech billionaires support EA, as though everything tech billionaires support is automatically incorrect, though this article may not be attempting to make such a faulty ‘argument’.

Rebecca Ackermann in MIT Technology Review writes “EA’s philosophical genes came from Peter Singer’s brand of utilitarianism and Oxford philosopher Nick Bostrom’s investigations into potential threats to humanity.” Similar to above, the ‘genes’ of utilitarianism are connected in the *person* of Peter Singer but not in the *arguments* of Peter Singer, which is an incredibly important distinction. EA does not rely on his brand of utilitarianism, and it is important to clarify this non-reliance to the public that wants to throw up anytime the word “utilitarianism” is mentioned. Also, Bostrom’s existential risks aren’t even a core part of EA; they are a more recent development. From my perspective, this development is much less of the genes of EA (though Bostrom was writing about longtermism- and extinction-related topics before EA) and more of a grafting into EA, at least as far as how much weight or significance the existential risks have.

Now, it is quite possible that the authors of these articles were merely noting the historical roots of the movement, which is of perfectly legitimate interest to note. Given that the average person finds utilitarianism detestable, however, suggests that it would be important for neutrality’s sake to clarify that effective altruism is not, in fact, wedded to the exact beliefs of the originators or even the current leaders.

If this connection was made to critique EA, this amounts to a kind of genetic argument against effective altruism. Whether these authors were attempting this approach (implicitly) is not my primary concern, and I will not comment either way, but since this is a fairly popular type of argument to make, I will investigate it. In fact, it does seem like the general structure of recent critiques of EA due to SBF and FTX are a guilt by association argument, which I explored in the SBF Association Argument Against Effective Altruism. My best attempted reconstruction of the genetic utilitarian argument is of the form:

1. If the originators and/or leaders of a movement espouse a view, then the movement ineliminably is committed to that view
2. The originators and/or leaders of the EA movement espouse utilitarianism
3. Therefore, the EA movement is ineliminably committed to utilitarianism
4. If a movement is ineliminably committed to a false view, then the movement has an incorrect framework
5. Utilitarianism is false
6. The EA movement has an incorrect framework

A Movement’s Commitments are not Dictated by the Belief Set of Its Leaders

One problem with this argument is that premise (1) is obviously false. Regarding the originators, movements can change. Additionally, leaders have many beliefs 1) unrelated to the movement, and 2) even related beliefs may not imply nor be implications of the framework. This can be true even if the originators and leaders all share some set of views $P_1 = \{p_1, p_2, p_3 \dots p_7\}$, as the movement may be characterized by a subset of those views $P_2 = \{p_1, p_2\}$ where P_2 does not imply $\{p_3 \ldots p_7\}$. This is likely the case in the effective altruism movement, as P_2 does not encapsulate an entire global moral structure and so does not imply the entirety of the leader's related views. Further, there can be a common cause of the beliefs of the leaders that are non-identical to the common cause of the beliefs of the core of the movement.

Another way to remit the concern above is to consider the core of the theory vs auxiliary hypotheses, as discussed in philosophy of science. If P_2 is the core of effective altruism, it can be true that beliefs in P_1, that are not in P_2, are auxiliary hypotheses but can still be freely rejected by those in the movement and remain true to EA.

There is a parallel in Christianity as well. There is substantial diversity in the movement that is Christianity, yet there is a common core of essential commitments of Christianity, called "essential doctrine". These commitments constitute the core of the theory of Christian theism. Beyond that, we can have reasonable disagreements as brothers and sisters in Christ. As 7th century theologian Rupertus Meldenius said, "In Essentials Unity, In Non-Essentials Liberty, In All Things Charity."

This disagreement extends from laymen to pastors and "leaders" of the faith as well. I think this should be fairly obvious for people that have spent much time in Christian bubbles. Laymen can and do disagree with pastors of their own denomination, pastors of other denominations, the early church fathers, etc., and they remain Christian without rejecting essential doctrine. (Of course, some church leaders and laymen are better than others at not calling everyone else heretics).

EA Leaders are Not All Utilitarians

The second point of contention with this argument is that premise (2) is also false. William MacAskill can rightly be called both an originator and a leader of EA, and he does not espouse utilitarianism. He thinks that sometimes it is better to not do what results in the overall greatest moral good. He builds in side-constraints (though sophisticated forms of utilitarianism can do a limited version of this, and consequentialism can do precisely this in effect). Furthermore, he builds in uncertainty in the form of a risk-averse expected utility function with distributed credences between (at least) utilitarianism and deontology, which motivates side-constraints.

In this section, we examined two arguments against effective altruism in view of its connection to utilitarianism, finding both arguments substantially lacking. In conclusion from the previous two sections, we do not see a successful argument against effective altruism due to its theoretical or historical connection to utilitarianism. EA remains a highly defensible intellectual project.

Do the Ends Justify the Means?

There is a need for clarity around "ends-justifying-means" reasoning and claims like "the end doesn't justify the means." Many recent criticisms make this claim in response to the FTX scandal. They connect effective altruism to what they see as "ends-justifying-means" reasoning in Sam Bankman-Fried (SBF) and use that as a reductio against effective altruism.

This argument fails on virtually every point.

First, let's see what people have said about it. Eric Levitz in the Intelligencer says that "the SBF saga spotlights the philosophy's greatest liabilities. Effective altruism invites 'ends justify the means' reasoning, no matter how loudly EAs disavow such logic." Eric also writes, "Effective altruists' insistence on the supreme importance of consequences invites the impression that they would countenance any means for achieving a righteous end. But EAs have long disavowed that position." Rebecca Ackermann in Slate mentions, "EA needs a clear story that rejects ends-justifying-means approaches," referencing Dustin Moskovitz's Tweets.

As the authors above mention, EA thinkers typically, on paper at least, disavow "ends justify the means" reasoning. More recently, MacAskill in a recent Twitter thread says, "A clear-thinking EA should strongly oppose 'ends justify the means' reasoning." Holden Karnofsky, co-founder of Open Philanthropy and GiveWell, in a recent forum post says, "I dislike 'end justify the means'-type reasoning." This explicit rejection is not solely in the wake of the downfall of FTX; MacAskill 2019 in "The Definition of Effective Altruism" says, "as suggested in the guiding principles, there is a strong community norm against 'ends justify the means' reasoning."[9] I talk more substantively about the use of side constraints in EA in the 4th difference between EA and utilitarianism below.

Of course, critics of EA readily acknowledge that EA, on paper, disavows ends-means reasoning. The problem, they think, is that EA "invites" ends-means reasoning, or that EA "invites the impression that they would countenance any means for achieving a righteous end" over and against EA's claims.

All of the above discussion fails to acknowledge two very key points, which is due to the ambiguity in what "ends justify the means," in fact, means. These two points become obvious once we adequately explore ends-means reasoning[10]; they are: (1) some ends justify some means, and (2) "ends justify the means" is a problem for every plausible moral theory.

Some Ends Justify Some Means

Obviously, some ends justify some means. Let's say I strongly desire an ice cream cone and consuming it would make me very happy for the rest of the day with no negative results. Call me crazy, but I submit to you that this end (i.e., Ice Cream) justifies the means of giving $1 to the cashier. If this is correct, then some ends justify[11] some means. Therefore, it is false that "the end never justifies the means."

Various ethicists have pointed this out. Joseph Fletcher says that people "take an action for a purpose, to bring about some end or ends. Indeed, to act aimlessly is aberrant and evidence of either mental or emotional illness."[12] Though, it may be that this description in line with the "Standard Story" of Action in action theory entails a teleological conception of reasons that has distorted debates in normative ethics in favor of consequentialism, as Paul Hurley has argued.[13]

Nonetheless, Fletcher is right that even this commonsense thinking on everyday justification for any action "leads one to wonder how so many people may say so piously, 'The end cannot justify the means.' Such a result stems from a misinterpretation of the fundamental question concerning the relationship between ends and means. The proper question is – 'Will *any* end justify *any* means?' – and the necessary reply is negative."[14] It is obviously false that any end justifies any means, and everyone in the debate accepts that, including the hardcore utilitarian.

What happens when we raise the stakes of either the end or the means?

Some Ends Justify Trivially Negative Means

We can consider raising the moral significance of the end in question. Let us consider the end of preventing the U.S. from launching nuclear missiles at every other country on the globe (i.e., Nuclear Strike). Although lying is generally not morally good, I submit that it is morally permissible to fill in your birthday incorrectly on your Facebook account if it prevents Nuclear Strike. An end of great moral magnitude like Nuclear Strike justifies a mildly negative means like a single instance of deception on a relatively unimportant issue. Therefore, a very good moral end justifies a mildly negative means.

Similarly, when James Sterba considers the Pauline Principle that we should not do evil so that good may come of it, he acknowledges it is "rejected as an absolute principle…because there clearly seem to be exceptions to it." Sterba gives two seemingly obvious cases where doing evil so that good may come "is justified when the resulting evil or harm is: (1) trivial (e.g., as in the case of stepping on someone's foot to get out of a crowded subway) or (2) easily reparable (e.g., as in the case of lying to a temporarily depressed friend to keep him from committing suicide)."[15]

No End Can Justify Any Means

Further, there is no end that can justify any means. For any given end, we can consider means that are way worse. For example, consider the end of saving 1 million people from death. Is any means justified to save them? Of course not. For example, killing 1 billion people would not be justified as a means to save 1 million people from death. For any end, we can consider means that are 10x as bad as the end, and the result is that the means is not justified. From one perspective, in the scenario of killing 1 to save 1 million, the absolutist deontologist justifies terrible means (i.e., letting 1 million people die) to the end of saving 1; of course, they would not word it this way, but it amounts to the same thing. Ultimately, for a particular end, no matter how bad, it is false that we can use *any* means possible to achieve that end and doing so would be morally permissible.

As Joseph Fletcher (a consequentialist) said, "'Does a worthy end justify *any* means? Can an action, no matter what, be justified by saying it was done for a worthy aim?' The answer is, of course, a loud and resounding NO!" Instead, "ends and means should be in balance."[16]

A Sufficiently Positive End Can Justify a Negative Means

Let us investigate further just how negative of means can be justified. Let us reconsider Ice Cream with a more negative means. Clearly, Ice Cream does *not* justify shooting someone non-fatally in the leg to get the ice cream cone. For an end to even possibly justify non-fatal shooting, it would require something much more significant. Is there any scenario that would make a non-fatal shooting morally permissible? I think there is. Consider a scenario that is rigged such that if you non-fatally shoot a person, one billion people will be saved from a painful death. It should be obvious that preventing the death of a billion people does justify shooting someone non-fatally in the leg. Therefore, it is possible for a massively positive end to justify a negative means.

Uh oh! Did I just admit I am a horrible person? I think it is okay to shoot someone (non-fatally) if the circumstances justify it, after all. Of course, most people think it is permissible to kill in some cases, such as self-defense or limited instances of just war.[17] After explaining the typical EA stance on deferring to constraints including a document by MacAskill and Todd, and how MacAskill said that SBF violated them, Eric Levitz in the Intelligencer complains that “yet, that same document suggests that, in some extraordinary circumstances, profoundly good ends can justify odious means.” My response is, “Yes, and that is trivially correct.” If I could prevent 100,000,000 people from being tortured and killed by slapping someone in the face, I would and should do it. And that shouldn’t be controversial.

As MacAskill and Todd note (which the author also quotes), “Almost all ethicists agree that these rights and rules are not absolute. If you had to kill one person to save *100,000* others, most would agree that it would be the right thing to do.” If you will sacrifice a million people to save one person, you are the one that needs to have your moral faculties reexamined. Killing a person, while more evil than letting a person die, is not 999,999 times more evil than letting one person die. Probably, the value difference between killing a person and letting a person die is much less than the value of a person, i.e., the disvalue of letting a person die. Therefore, letting two people die is already worse than killing one person, but it even more obvious that letting 1,000,000 people die is worse than killing one person.

I do not believe I have said much that is particularly controversial when looking at these manufactured scenarios.[18] We are stipulating in these tradeoff considerations that the tradeoff is actually a known tradeoff and there is no other way, etc.

In sum, the ends don't justify the means...except, of course, when they do. Ends don't *never* justify the means and don't *always* justify the means, and virtually no one in this debate thinks otherwise. Almost everyone thinks ends *sometimes* justify the means (depending on the means). What we have to do is assess the ends and assess the means to discern when exactly what means are justified for what ends.

Absolutism is the Problem

This whole question has very little to do with consequentialism or deontology, contrary to popular belief, and everything to do with absolute vs relative ethics (not individually or culturally relative, but situationally relative).[19] There is a debate internal to non-consequentialist traditions about this question of when the ends justify the means. For example, with deontology there is what is called threshold or moderate deontology, and in natural law theory there is a view called proportionalism. Neither of these are absolutist views, and both views include the results of actions as justification for some means. Internal to these non-consequentialist families of theories typically characterized as absolutist remains the exact same debate about ends-means reasoning. In fact, the most plausible theories in all moral families allow extreme (implausible but possible) cases to violate absolute rules.

For example, it is uncommon to find a true absolutist deontologist among contemporary ethicists. As Aboodi, Borer, and Enoch point out, "hardly any (secular) contemporary deontologist is an absolutist. Contemporary deontologists are typically 'moderate deontologists,' deontologists who believe that deontological constraints come with thresholds, so that sometimes it is impermissible to violate a constraint in order to promote the good, but if enough good (or bad) is at stake, a constraint may justifiably be infringed."[20] In other words, almost all (secular) deontologists also think the ends sometimes justify the means. Absolutism is subject to numerous paradoxes and counterexamples discussed previously and in the next subsection (see Figure 4)

Figure 4: Absolutism in a nutshell

Paradoxes of Absolute Deontology

Why is it that even deontologists think there are exceptions to constraints? Because absolute deontology is subject to substantial paradoxes and implausible implications that render it unpalatable, even worse than the alternatives. One example is the problem of risk, which is that any action raises the probability of violating absolute constraints, and no action gives 100% certainty of violating constraints. Therefore, it looks like the absolutist needs to say either that any action that produces a risk of violation is wrong, leading to moral paralysis since you would be prohibited from taking any action, or pick an (arbitrary) risk threshold, which implies that, in fact, two wrongs do make a right, and two rights make a wrong (in certain cases).[21] There have been responses, but what is perhaps the best response, stochastic dominance to motivate a risk threshold, is still subject to a sorites paradox that again appears to render absolutism false.[22] MacAskill offers a distinct but related argument from cluelessness that deontology implies moral paralysis.[23]

Alternatively, we can merely consider cases of extreme circumstances just like the one I gave earlier. A standard example is lying to a visitor to your house in order to prevent someone from being murdered, which Kant famously and psychopathically rejected. Michael Huemer considers a case where aliens will kill all 8 billion people on earth unless you kill one innocent person. Should you do so? The answer, as Huemer and any sane person agrees, is obviously yes.[24] (If the reader still thinks the answer is no, add another 3 zeros to the number of people you are letting die and ask yourself again. Repeat until you reject absolutism). These types of cases show quite quickly and simply that absolutism is not a plausible position in the slightest, and it is justified to do something morally bad if it results in something good enough (or, alternatively, prevents something way worse). There are other problems for absolutist deontology I neglect here.[25]

Of course, in a trivial sense, consequentialists are absolutist: it is *always* wrong to do something that does not result in the most good. However, that is not what anyone means when they call theories absolutist, which refers to theories that render specific classes of actions (e.g., intentional killing, lying, torture, etc.) as always impermissible.[26]

In summary, any plausible moral theory or framework has to reckon with the fact that something negative is permissible if it prevents something orders of magnitude worse. When people say "the end doesn't justify the means" when condemning an action, they, in practice, more frequently mean *those* ends don't justify *those* means. Equivalently, they mean that the ends don't justify the means *in this circumstance*, rather than *never*, as the latter results in a completely implausible view.

Application to the FTX Scandal

So, where does that leave us in the FTX scandal? Everyone in the debate can say that, in this case, the ends did not justify the means. Although criticizing EA, Eric Levitz in the Intelligencer appears to challenge this, saying perhaps SBF may reasonably be considered justified if there are exceptions to absolute rules, "In 'exceptional circumstances,' the EAs allow, consequentialism may trump other considerations. And Sam Bankman-Fried might reasonably have considered his own circumstances exceptional," describing the uniqueness of SBF's case. Levitz asks, "If killing one person to save 100,000 is morally permissible, then couldn't one say the same of scamming crypto investors for the sake of feeding the poor (and/or, preventing the robot apocalypse)?" If I were to put this into an argument, it may be: (1) if ends justify the means sometimes, then SBF's actions are justified, (2) if EA, then ends justify the means sometimes, (3) if EA, then SBF's actions are justified (or reasonably considered so).

There are several problems here (found in premise 1). First, it is not *consequentialism* that may trump other considerations, but *consequences*.[27] The significance of the difference is that any moral theory can say (and the most plausible ones *do* say) that consequences can, in the extreme, trump other considerations, as we saw earlier. Second, SBF's circumstances may be exceptional in the generic sense of being rare and unique, but the question is "are they exceptional *in the relevant sense*," which is that his circumstances are such that violating the constraint of committing illegal actions or fraud would result in a sufficient overall good to warrant breaking the constraint. It is a general rule that fraud is not good in the long run for your finances or moral evaluation.

Third, it is much too low a bar to say that it is *reasonable* for SBF to think that his circumstances were exceptional in the relevant sense, but we are (or should be) much more interested in whether SBF *was correct* in thinking his circumstances were exceptional in the relevant sense. An assessment of irrationality requires us to know his belief structure and evidence base for this primary claim as well as many background beliefs that informed his evidence and belief structure of the primary claim (and possibly knowing the correct view of decision theory, which is highly controversial).

Fourth, one *can say* anything one wants (see next section Can EA/Consequentialism/Longtermism be Used to Justify Anything?). We are and should be interested in what one can *accurately* say about such a comparison between killing one person to save 100,000 and 'scamming crypto investors for the sake of feeding the poor (and/or, preventing the robot apocalypse).' Fifth, it is unlikely that one can accurately say that these are comparably similar, such that it is incredibly unlikely that SBF was correct in his assessment. This rhetorical question about comparing saving 100,000 lives vs scamming crypto investors does very little to demonstrate otherwise.

SBF's approach, which approved of continuing double-or-nothing bets for eternity, evidently did not consider the fallout associated with nearly inevitable bankruptcy and how that would set the movement back, as that would render each gamble less than net zero. Secondly, almost everyone agrees his approach was far too risk-loving. Nothing about EA or utilitarianism or decision theory, etc. suggests that we should take this risk-loving approach. As MacAskill and other EA leaders argue, we should be risk averse, especially with the types of scenarios SBF was dealing with (relevant EA forum post). Plus, there is the disvalue associated with breaking the law and chance of further lawsuits.

Levitz appears to accept the above points and concedes that it would be unfair to attribute SBF's "bizarre financial philosophy" to effective altruism, and that EA leaders would likely have strongly disagreed with implementing this approach with his investments. Given Levitz's acceptance of this, it is unclear what the critique is supposed to be from the above points. Levitz does move to another critique though, which is that EAs have fetishized expected value calculations, which I will address in the next section.

In summary, the ends sometimes justify the means, but violating constraints almost never actually produces the best result, as EA leaders are well-aware. Just because SBF made a horrible call does not mean that the EA framework is incorrect, as the typical EA framework makes very different predictions that would not include such risk-loving actions.

Effective Altruism is Not Inherently Utilitarian

There was a lot of confusion in these critiques about the connection between utilitarianism and effective altruism. Many of these articles assume that effective altruism implies or requires utilitarianism, such as (not including the quotes below) Erik Hoel, Elizabeth Weil in the Intelligencer, Rebecca Ackermann in MIT Technology Review (see a point-by-point response here), Giles Fraser in the Guardian, James W. Lenman in IAI News, and many more. I will survey and briefly respond to some individual quotations to this effect, showcase the differences between effective altruism and utilitarianism. Throughout, I will extensively refer to MacAskill's 2019 characterization of effective altruism in "The Definition of Effective Altruism."

As a first example, Linda Kinstler in the Economist (non-paywalled) writes "[MacAskill] taught an introductory lecture course on utilitarianism, the ethical theory that underwrites effective altruism." Nitasha Tiku in The Washington Post (non-paywalled) writes, "[EA's] underlying philosophy marries 18th-century utilitarianism with the more modern argument that people in rich nations should donate disposable income to help the global poor." It is curious to call it 18th century utilitarianism when the version of utilitarianism EA is closest to (yet still quite distinct from) is "rule utilitarianism", only hints of which were found in the 19th century with its primary development in the 20th century. Furthermore, while it may be a modern development that one can easily transfer money and goods across continents, it is certainly no modern argument that the wealthy should give disposable income to the poor, including across national lines. The Parable of the Good Samaritan advocates for helping explicitly across national lines, the Old Testament commanded concern for the poor by those with resources (for a fuller treatment, see *Christians in an Age of Wealth: A Biblical Theology of Stewardship*), and "the early Church Fathers took luxury to be a sign of idolatry and of neglect of the poor."[28] The fourth century St. Ambrose condemns rich neglect of the poor, "You give coverings to walls and bring men to nakedness. The naked cries out before your house unheeded; your fellow-man is there, naked and crying, while you are perplexed by the choice of marble to clothe your floor."[29]

Timothy Noah in The New Republic writes, "E.A. tries to distinguish itself from routine philanthropy by applying utilitarian reasoning with academic rigor and a youthful sense of urgency," and also "Hard-core utilitarians tend not to concern themselves very much with the problem of economic inequality, so perhaps I shouldn't be surprised to find little discussion of the topic within the E.A. sphere." It is blatantly false that economic inequality is of little concern to utilitarians (as explained in the link that the author provided himself), including "hard-core" ones, as the state of economic inequality in the world leads to great suffering and death as a result. Now, it is correct that utilitarians do not see inequality as an *intrinsic* good, but merely an *instrumental* good. Yet, I do not see the problem with rejecting inequality's *intrinsic* value rather than its *instrumental* value; it would be surprising that, on a perhaps extreme version of egalitarianism, there being two equally unhappy people is better than one slightly happy person and one extremely happy person. Alternatively, we should be much more concerned that people's basic needs are met, so they are not dying of starvation and preventable disease, than we should that, if everyone already had their needs met, the rich have equal amounts of frivolous luxuries, as sufficientarianism well-accommodates. Finally, as MacAskill 2019 notes, EA is actually compatible with utilitarianism, prioritarianism, sufficientarianism, and egalitarianism (see next section).

Eric Levitz in the Intelligencer states, “Many people think of effective altruism as a ruthlessly utilitarian philosophy. Like utilitarians, EAs strive to do the greatest good for the greatest number. And they seek to subordinate common-sense moral intuitions to that aim.” EAs are not committed to doing the greatest good for the greatest number (see the next section for clarification), and they do not think any EA commitments subvert commonsense intuitions. In fact, EAs attempt to take common sense intuitions seriously along with their implications. The starting point for EA was originally that, if we can fairly easily save a drowning child, we should.[30] This is hardly a counterintuitive claim. Then, upon investigating the relevant similarities between this situation and charitable giving, we get effective altruism.

Jonathan Hannah in Philanthropy Daily asks, “why should we look to these utilitarians to learn how to be effective with our philanthropy?” First, we should look to EAs because EAs have evidence backing up claims of effectiveness. Secondly, again, EAs are not committed to utilitarianism, though many EAs are, in fact, utilitarians.

Theo Hobson in the Spectator claims, “Effective altruism is reheated utilitarianism… Even without the ‘longtermist’ aspect, this new utilitarianism is a thin and chilling philosophy.” Beyond the false utilitarianism claim, the accusation of thinness is surprising, since there are substantial and life-changing implications of taking EA seriously. These are profound implications that have resulted in protecting 70 million people from malaria, giving $100 million directly to those in extreme poverty, giving out hundreds of millions of deworming treatments, setting 100 million hens free from a caged existence, and much more. Collectively, GiveWell estimates the $1 billion donations through them will save 150,000 lives.

The aforementioned claims are misguided, as not everything that is an attempt to do the morally best thing is utilitarianism (see Figure 5).

Figure 5: Utilitarianism is a specific moral theory (or, rather, a family of specific theories), actually

Now, I seek to make good on my claim that effective altruism and utilitarianism are distinct. There are six things that distinguish EA from a reliance on utilitarianism, and I will examine each in turn:

1. [Minimal] EA does not make normative claims
2. EA is independently motivated
3. EA does not have a global scope

4. EA incorporates side constraints
5. EA is not committed to the same "value theory"
6. EA incorporates moral uncertainty

1. [Minimal] EA Does Not Make Normative Claims

Effective altruism is defined most precisely in MacAskill 2019, who clarifies explicitly that EA is non-normative. MacAskill says, "Effective altruism consists of two projects [an intellectual and a practical], rather than a set of normative claims."[31] The idea is that EA is committed to trying to do the best with one's resources, but not necessarily that it is morally obligatory to do so. Part of the reason for this definition is to be in alignment with the preferences and beliefs of those in the movement. There were surveys both to leaders and members of the movement in 2015 and 2017, respectively, which suggested a non-normative definition may be more representative to current EA adherents. Furthermore, it is more ecumenical, which is a desirable trait for a social movement as it expands.

Of course, a restriction to non-normative claims is limited, and Singer's original argument that prompted many towards EA was explicitly normative in nature. His premises included talk of moral obligation. Many people in EA do think it is morally obligatory to be an EA. Thus, I think it is helpful to distinguish between different types or levels of EA, including minimal EA, normative EA, radical EA, and radical, normative EA.

Minimal EA makes no normative claims, while normative EA includes *conditional obligations*.[32] Normative EA claims that *if* one decides to donate, one is morally obligated to donate to the most effective charities, but it does not indicate *how much* one should donate. This could be claimed to be absolute, a general rule of thumb, or somewhere in between. Radical EA, on the other hand, includes unconditional obligations, but no conditional obligations. Brian Berkey, for example, argues that effective altruism is committed to unconditional obligations of beneficence.[33] Radical EA, as I characterize it, says one is morally obligated to donate a substantial portion of one's surplus income to charities. Finally, radical, normative EA (RNEA) combines conditional and unconditional obligations of beneficence, claiming one is morally obligated to donate a substantial portion of one's surplus income to effective charities. I expand on and defend these further elsewhere.[34]

Thus, while minimal EA does not include normative claims, there are expanded versions of EA that include conditional and/or unconditional obligations of beneficence. Minimal EA, then, constitutes the *core* of the EA theory, while these claims of obligations constitute *auxiliary hypotheses* of the EA theory. Since the core of EA does not include normative claims, it cannot be identical to (any version of) utilitarianism, whose core includes a normative claim to maximize impartial welfare.

2. EA is Independently Motivated

Effective altruism is distinct from utilitarianism in that EA can be motivated on non-consequentialist grounds. In fact, even Peter Singer's original argument, inspiring much of EA, was non-consequentialist in nature. Singer's original "drowning child" thought experiment relied only on a simple, specific thought experiment, proposing midlevel principles (principles that stand in between specific cases and moral theories) to explain the intuition from the thought experiment, and deriving a further conclusion by comparing relevant similarities in the thought experiment to a real world situation, all of which is a standard procedure in applied ethics. Of course, this article has been critically responded to in the philosophy community many, many times, some more revolting[35] than others,[36] but many (such as I) still find it a compelling and sound argument that also demonstrates EA's independence from utilitarianism.

Theory-Independent Motivation: The Drowning Child

Singer's original thought experiment is: "if I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing."[37]

Singer proposes two variants[38] of a midlevel principle that would explain this obvious result:

(1) If it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, to do it.

He also proposed a weaker principle,

(2) If it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it.

These principles are extremely plausible, are quite intuitive, and would explain why we have the intuitions we do in various rescue cases comparable to the above. Next, Singer defended why this principle can be extended to the case of charitable giving by examining the relevant similarities. The reasoning is that, given the existence of charities, we are in a position to prevent something bad from happening, e.g., starvation and preventable disease. We can do something about it by 'sacrificing' our daily Starbucks, monthly Netflix subscription, yearly luxury vacations, or even more clearly unnecessary purchases, for example additional sports cars or boats that are not vocationally necessary, etc. None of these things are (obviously) morally significant, and they are certainly not of comparable moral importance of the lives of other human beings. Therefore, we have a moral obligation to take action in donating to effective charities, particularly from the income that we are using for surplus items.

Notice that we did not appeal to any kind of utilitarian reasoning in the above argument, and one can accept either of Singer's midlevel principles without accepting utilitarianism. This example shows how effective altruism can be independently motivated apart from utilitarianism. This fact was pointed out previously by Jeff McMahan when he noticed that even philosophical critiques of EA make this false assumption of reliance on utilitarianism. McMahan, writing in 2016, said, "It is therefore insufficient to refute the claims of effective altruism simply to haul out [Bernard] Williams's much debated objections to utilitarianism. To justify their disdain, critics must demonstrate that the positive arguments presented by Singer, Unger, and others, which are independent of any theoretical commitments, are mistaken."[39]

Martin Luther's Drowning Person

Interestingly, the Christian has a surprising connection to Singer's *Drowning Child* thought experiment, as a nearly identical thought experiment and comparison was made by Martin Luther in the 16th century.[40] In his commentary on the 5th commandment "Thou shalt not kill" in *The Large Catechism*, Luther connects the commandment to Jesus' words in Matthew 25, "For I was hungry and you gave me nothing to eat, I was thirsty and you gave me nothing to drink, I was a stranger and you did not invite me in, I needed clothes and you did not clothe me, I was sick and in prison and you did not look after me." Luther then gives a drowning person comparison: "It is just as if I saw some one navigating and laboring in deep water [and struggling against adverse winds] or one fallen into fire, and could extend to him the hand to pull him out and save him, and yet refused to do it. What else would I appear, even in the eyes of the world, than as a murderer and a criminal?"

Luther condemns in the strongest words those could "defend and save [his neighbor], so that no bodily harm or hurt happen to him and yet does not do it." He says, "If…you see one suffer hunger and do not give him food, you have caused him to starve. So also, if you see any one innocently sentenced to death or in like distress, and do not save him, although you know ways and means to do so, you have killed him." Finally, he says, "Therefore God also rightly calls all those murderers who do not afford counsel and help in distress and danger of body and life, and will pass a most terrible sentence upon them in the last day."

Virtue Theoretic Motivation: Generosity and Others-Centeredness

Beyond a theory-independent approach to motivate EA, we can also employ a non-consequentialist theory, virtue ethics, to motivate EA. Some limited connections between effective altruism and virtue ethics have been previously explored,[41] but I will briefly give two arguments for effective altruism from virtue ethics. Specifically, I will argue from the virtues of generosity and others-centeredness for normative EA and radical EA, respectively. Thus, if both arguments go through, the result is radical, normative EA.

First, I assume the qualified-agent account of the criterion of right action[42] of virtue ethics given by Rosalind Hursthouse.[43] Second, I employ T. Ryan Byerly's accounts of both generosity and others-centeredness.[44] Both of these, especially from the Christian perspective, are virtues. The argument from generosity is:

1. An action is right only if it is what a virtuous agent would characteristically do
2. A virtuous agent would characteristically be generous
3. To be generous is to be skillful in gift-giving (i.e., giving the right gifts in right amounts to the right people)
4. A charitable donation is right only if it is skillful in gift-giving
5. A charitable donation is skillful in gift-giving only if it results in maximal good
6. A charitable donation is right only if it results in maximal good (NEA)

The argument from others-centeredness is:

7. An action is right only if it is what a virtuous agent would characteristically do
8. A virtuous agent would characteristically be others-centered
9. To be others-centered includes treating others' interests as more important than your own
10. Satisfying one's interests in luxuries before trying to satisfy others' interests in basic needs is not others-centered
11. An action is right only if it prioritizes others' basic needs before your luxuries
12. A substantial portion of one's surplus income typically goes to luxuries
13. Therefore, a person is morally obligated to donate a substantial portion of one's surplus income to charity (REA)

I don't have time to go into an in-depth defense of these arguments (though see my draft paper [pdf] for a characterization and assessment of *luxuries* as in the above argument, as well as independent arguments for 11-13), but it at least shows how one can reasonably motivate effective altruism from virtue ethical principles.

3. EA Does Not Have a Global Scope

Unlike utilitarianism, effective altruism is not a global moral theory in that it cannot, in principle, give deontic outcomes (i.e., right, wrong, obligatory, permissible, etc.) to any given option set (a set of actions that can be done by an agent at some time *t*). Utilitarianism is a claim about what explains why any given action is right, wrong, obligatory, etc., as well as the truth conditions for the same. In other words, utilitarianism makes a claim of the form, *an action is right if and only if xyz*, which are the truth conditions of deontic claims, and a claim of the form, *an action is right because abc*, which is the explanatory claim corresponding to the structure of reasons of the theory (that explains why actions are right/wrong).

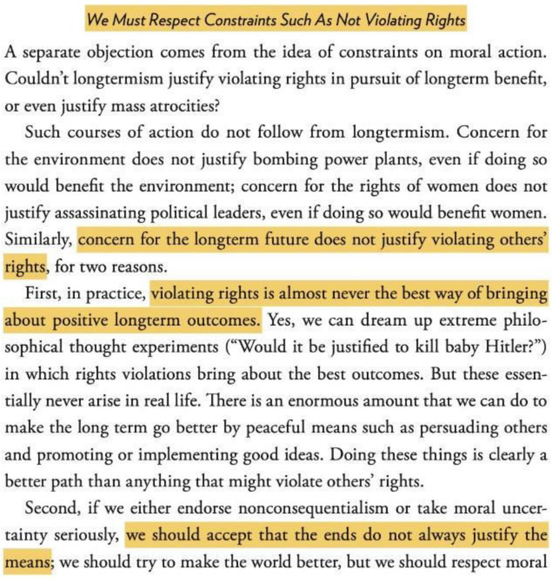
While minimal EA trivially does not match utilitarianism in making global normative claims, even radical, normative EA does not govern every possible action set, and it does not propose to. At the most, EA makes claims about actions related to (1) charitable donations and (2) career choice, including RNEA. As MacAskill 2019 says, "Effective altruism is not claiming to be a complete account of the moral life." There are many actions, such as, say, those governing social interactions, that are out of scope of EA and yet within the scope of utilitarianism.

Therefore, utilitarianism and effective altruism differ in their scopes, as EA is not a comprehensive moral theory, so EA does not require utilitarianism.

4. EA Incorporates Side Constraints

In "The Definition of Effective Altruism," MacAskill 2019 is clear that EA includes constraints, and not any means can be justified for the greater good. MacAskill says that the best course of action, according to EA, is an action "that will do the most good (in expectation, without violating any side constraints)."[45] He only considers a value maximization where "whatever action will maximize the good, subject to not violating any side constraints."[46] He says that EA is "open in principle to using any (non-side-constraint violating) means to addressing that problem."[47]

In *What We Owe the Future*, MacAskill says that “naïve calculations that justify some harmful action because it has good consequences are, in practice, almost never correct” and that “plausibly it’s wrong to do harm even when doing so will bring about the best outcome.”[48] On Twitter, MacAskill shared relevant portions of his book on side constraints when responding to the FTX scandal, including the page shown below. He states that “concern for the longterm future does not justify violating others’ rights,” and “we should accept that the ends do not always justify the means…we should respect moral side-constraints, such as against harming others. So even on those rare occasions when some rights violation *would* bring about better longterm consequences, doing so would not be morally acceptable.”[49]



We Must Respect Constraints Such As Not Violating Rights

A separate objection comes from the idea of constraints on moral action. Couldn’t longtermism justify violating rights in pursuit of longterm benefit, or even justify mass atrocities?

Such courses of action do not follow from longtermism. Concern for the environment does not justify bombing power plants, even if doing so would benefit the environment; concern for the rights of women does not justify assassinating political leaders, even if doing so would benefit women. Similarly, concern for the longterm future does not justify violating others’ rights, for two reasons.

First, in practice, violating rights is almost never the best way of bringing about positive longterm outcomes. Yes, we can dream up extreme philosophical thought experiments (“Would it be justified to kill baby Hitler?”) in which rights violations bring about the best outcomes. But these essentially never arise in real life. There is an enormous amount that we can do to make the long term go better by peaceful means such as persuading others and promoting or implementing good ideas. Doing these things is clearly a better path than anything that might violate others’ rights.

Second, if we either endorse nonconsequentialism or take moral uncertainty seriously, we should accept that the ends do not always justify the means; we should try to make the world better, but we should respect moral

Figure 6: Excerpt from *What We Owe the Future*

Utilitarianism, on the other hand, does not have side constraints, or at least, not easily. Act utilitarianism (which is normally the implied view if the modifier is neglected) certainly does not. However, rule utilitarianism can function as a kind of constrained utilitarianism in two ways; one way is strong rule utilitarianism that has no exceptions, which is absolutist. Another is with weak rule utilitarianism that still allows some exceptions. MacAskill’s wording above makes it sound like there would not be any exceptions, “even when some rights violation *would* bring about better longterm consequences.”[50]

However, elsewhere, he makes it sound as though there can be exceptions. He (with Benjamin Todd) says, “Almost all ethicists agree that these rights and rules are not absolute. If you had to kill one person to save *100,000* others, most would agree that it would be the right thing to do.” I am in perfect agreement there. I think, as I discuss below in the Do the Ends Justify the Means? section, absolute rules are trivially false. In fact, MacAskill has an entire paper (with Andreas Mogensen) arguing that absolute constraints lead to moral paralysis because, to minimize your chance of violating any constraints, you should do nothing.[51] It is likely that MacAskill thinks there are extreme exceptions, though these would never happen in real life.

Finally, there remains a distinction between constrained effective altruism and rule utilitarianism, and that distinction is the same difference as between a consequentialized deontological theory and a standard deontological theory. The difference is that even rule utilitarianism explains the wrongness of all wrong actions ultimately by appeal to consequences (we should follow rules whose acceptance or teaching or following would lead to the best consequences), while constrained effective altruism explains the wrongness of constraint violations by appeal to constraints and to rights without a further justification in terms of the overall better outcomes.

In conclusion, EA incorporates side constraints, though with exceptions (as any plausible ethical theory would allow), while act utilitarianism does not. In addition, while EA has some structural similarities as rule utilitarianism, EA has different explanations of the wrongness of actions as utilitarianism, which turns out to be the key difference between (families of) moral theories,[52] and thus the two are quite distinct.

5. EA is Not Committed to the Same Value Theory

The fifth reason effective altruism is not utilitarian is because the value theory is not identical between the two. One reason they are not identical is because EA is not, strictly speaking, committed to a value theory. However, that does not mean the value theory is a free-for-all. EA is compatible with other theories in the vicinity of utilitarianism, such as prioritarianism, sufficientarianism, and egalitarianism.

Utilitarianism is committed to impartial welfarism in its value theory. There are a range of views within welfarism about what makes something well-off. Welfarism includes a range of views about well-being, including hedonism, desire or preference satisfactionism, or objective list theories. Hence, we can have hedonistic utilitarianism, preference utilitarianism, or objective list utilitarianism. Further, utilitarianism is committed to a simple aggregation function that makes good equal to the sum total of wellbeing, as opposed to a variously weighted aggregation function, such as in prioritarianism that gives additional weight to the wellbeing of those worse off.

The value theory that MacAskill 2019 describes in the definition of EA is "tentative impartial welfarism,"[53] where the 'tentative' implies this is a first approximation or working assumption. MacAskill expresses the difficulty here that arises from intra-EA disagreement: we do not want the scope of value maximization to be too large so that it can include maximizing *whatever* the individual wants, but we do not want the scope of maximization too small to exclude a substantial portion of the (current or future) movement.

MacAskill seems to do some hand-waving on this point. When defending EA as distinct from utilitarianism, he says, "it does not claim that wellbeing is the only thing of value," so EA is compatible "with views on which non-welfarist goods are of value."[54] However, two pages previously, his "preferred solution" of "tentative impartial welfarism…excludes non-welfarist views on which, for example, biodiversity or art has intrinsic value." On the same page, he suggests that if the EA movement became convinced that "the best way to do good might well involve promoting non-welfarist goods, then we would revise the definition to simply talk about 'doing good' rather than 'benefiting others.'"[55]

Perhaps one way of reconciling these is to say that, while "tentative impartial welfarism…excludes non-welfarist views," there is instead a tentative commitment to 'impartial welfarism', as opposed to a commitment to 'tentative impartial welfarism', and it is the impartial welfarism (ignoring the tentative here) that excludes non-welfarist views. When Amy Berg considers the same problem of "how big should the tent be?", she concludes that EA needs to commit to promote the impartial good in order to ensure that the effectiveness can be objectively measured.[56]

I suggest that the best way to combine these is to say that EA is committed to maximizing the impartial good that can be approximated by welfarism. If a view cannot even be approximated by welfarism, then it would be fighting a different battle than EA is fighting. This approach combines the tentative nature of the commitment with ensuring it can be objectively measured and in line with the current EA movement, while remaining open to including some non-welfarist goods that remain similar enough to the movement as it currently stands.

Finally, MacAskill says that EA can work with "different views of population ethics and different views of how to weight the wellbeing of different creatures,"[57] which is why EA is compatible with prioritarianism, sufficientarianism, and egalitarianism, in addition to utilitarianism.

Therefore, EA is distinct from utilitarianism by having a different commitment in both what is valuable as well as the aggregation principle.

6. EA Incorporates Moral Uncertainty

The final reason I will discuss on why EA is not utilitarianism is that EA incorporates moral uncertainty, which is an inherently *metatheoretical* consideration, while utilitarianism does not. *Utilitarians* do, just as everyone else has to deal with moral uncertainty, but *utilitarianism* does not automatically include this. Since EA includes inherently metatheoretical considerations, then it cannot be the same as a theory, which does *not* inherently include metatheoretical considerations, by definition.

The first way EA includes moral uncertainty was above in the characterization of "tentative impartial welfarism." EA is open to multiple different normative views; at the very least, it is open to hedonistic, preference, or objective list utilitarianism, while no single theory of utilitarianism can be open to multiple theories of utilitarianism, by definition. Further, this value theory does not rule out non-consequentialist views, and, if my virtue theoretic arguments above (or others) are successful, then virtue ethicists can be EAs. Therefore, EAs can reasonably distributed their credences across many different normative views, both utilitarian and non-utilitarian.

EA does not endorse a specific approach to moral uncertainty, which would likely be considered an auxiliary hypothesis of EA, though EA leaders do seem to clearly favor one particular approach, which is maximum expected choiceworthiness. Furthermore, MacAskill, who has done much work in moral uncertainty, reasons quite explicitly using uncertainty to distribute non-negligible credence in both utilitarianism and deontology, combining that with a risk-averse expected utility theory to motivate incorporating side constraints (aka agent-centered or deontic restrictions). I personally tentatively support the My Favorite Theory[58] approach to moral uncertainty, though EA does not require one or the other.

Objections

Savannah Pearlman argues that even though EA and utilitarianism are distinct moral frameworks, they share core philosophical commitments, and therefore EA is still dependent on utilitarianism. As I argue above, the exact differences between the two are such that EA is *not* dependent on utilitarianism. It is perfectly sufficient that EA and utilitarianism are (1) distinct frameworks and (2) independently motivated to conclude that EA is not inherently utilitarian. I showed the independent motivation (in the form of theory-independent midlevel principles as well as virtue ethical motivation) in section 2 above.

Pearlman evidently was not convinced that the theory-independent motivation was, in fact, theory-independent because there are shared commitments between EA and utilitarianism. Of course, we would expect that plausible moral theories will share some commitments. For example, that wellbeing is morally significant, and so are the consequences of one's actions, is true on any plausible moral theory. Shared commitments, unless they are the totality of the theories' commitments, do not show dependence. In the case of EA and utilitarianism, utilitarianism is sufficient for EA, but not necessary, since we can use virtue ethical arguments (or deontological, but I do not discuss that here).

Pearlman, however, misidentified the shared commitments. She says, "Rather clearly, Effective Altruism and Utilitarianism share the core philosophical commitments to Consequentialism, Impartiality, and Hedonism (repackaged by Effective Altruists into Welfarism)." A few noteworthy items on this. First, utilitarianism is not committed to hedonism; hedonistic utilitarianism is committed to hedonism, while preference utilitarianism is committed to preference satisfactionism, etc. In other words, utilitarianism is committed to some version of welfarism, which can be cashed out in various ways, which is the same as EA's welfarism. There are no commitments to the family of utilitarian theories nor EA to a specific account of well-being.

Secondly, Pearlman includes consequentialism as part of the core commitments of EA, which she does without argument. It is unclear why she does so. There are a non-negligible number of non-consequentialist EAs. I would guess Pearlman thinks that *maximizing* only makes sense given consequentialism. I have more faith in other moral theories than Pearlman does (since maximizing is the morally correct option), apparently, since I think that deontology and virtue ethics can make sense of maximizing welfare with a given unit of resources particularly in the restricted domains of concern to EA, such as charitable donations and career choice. Maximizing in this restricted domain can also be understood as an implication of the theory-independent principles that Singer proposed in the drowning child case.

Pearlman appears to take issue with some deontic outcomes in question, namely, in comparing two charities, that one *should* donate to a charity that is 100x more effective than another. Although minimal EA does not even commit to any obligation, we can consider the auxiliary commitment of normative EA (though this would still mean EA is not *inherently* utilitarian). Pearlman takes this moral obligation to imply that EA must be committed to a more general utilitarian principle. However, ignoring any moral theorizing, it just makes sense that you should not intentionally do an action that is much less good than another when it does not affect you much to do so. Normative EAs do not need to say more than this, while utilitarians do. As Richard Chappell points out in the comments, normative EA is only committed to efficient benevolence, but not constraint-less benevolence or unlimited beneficence that requires actions at great personal cost.

All things considered, from the clarification above we can see that Pearlman is incorrect that EA is inherently utilitarian and that criticisms of utilitarianism fairly apply to EA, as well.

Conclusion

In summary, effective altruism incorporates moral uncertainty in such a way that distinguishes itself from being inherently utilitarian in any interesting sense of the term. Of course, even an absolutist deontologist should have nonzero credence in some form of consequentialism to avoid being irrational, but that hardly makes them a consequentialist. So, EA is not inherently utilitarian.

All together, we saw six reasons that effective altruism is not reliant on utilitarianism. One is that minimal EA does not make normative claims. Furthermore, we saw that EA is also motivated by non-consequentialist reasoning, both theory-independent and virtue ethical in nature. More generally, EA, unlike utilitarianism, has a restricted scope, incorporates side constraints, has a different value theory, and includes moral uncertainty.

Can EA/Consequentialism/Longtermism be Used to Justify Anything?

Multiple authors express worries suggesting that EA or consequentialism or longtermism can be used to justify anything. In this section, I will show that this claim is either false or uninteresting, depending on how the claim is interpreted.

Émile P. Torres, a big fan of "scare quotes," wrote in a *salon* article titled "What the Sam Bankman-Fried debacle can teach us about 'longtermism'" that "For years, I have been warning that longtermism could 'justify' actions much worse than fraud, which Bankman-Fried appears to have committed in his effort to 'get filthy rich, for charity's sake'." Eric Levitz in the Intelligencer says that effective altruism "lends itself to maniacal fetishization of 'expected-value' calculations, which can then be used to justify virtually anything." I have also heard this claim made about consequentialism and utilitarianism maybe 400 times, so I will address the issue broadly here.

Drawing from my own manuscript titled "Worst Objections to Consequentialism," I will show why these attempted points are silly. We can generalize the concept of moral theory to a moral framework that would include effective altruism and longtermism as their own moral frameworks, and then moral theories would also be included as their own moral framework. I will focus on moral theories because this is more well-defined and discussed among ethicists.

All Families of Moral Theories Can Justify Anything

First, **any family of moral theories (e.g., consequentialism, deontology, virtue ethics) can justify any action as morally permissible**. If this is correct, then it amounts to an entirely uninteresting claim that e.g., consequentialism can justify anything, as any family of theories can justify anything until you flesh out the details of the specific theory you actually want to compare. The reason these are called families, and not theories, is because there are a bunch of different versions of each of these theories combined in a family resemblance between them. Moral theories have a global scope, meaning they apply to all actions, and a deontic predicate, meaning they say whether an action is permissible, impermissible, obligatory, etc.

For any given family of theories, we can construct a theory in that family that renders any given action permissible by manipulating what we find valuable, dutiful**,** or virtuous. For example, we can construct a consequentialist theory that says that only harmful intentions have value. We can have a deontological theory that says that our only obligation is to punch as many people in the face as possible every day. We can invent a virtue ethical theory where an action is virtuous if and only if it has the worst moral consequences. All of these theories are part of the respective family of theories (consequentialism, deontology, and virtue ethics). Now, none of these are particularly plausible versions of these theories, but adhering to these views would justify some pretty terrible actions. Thus, it is uninteresting to make the point that these kinds of moral theory families (including utilitarianism, which is a family subset of consequentialism) can justify immoral actions (see Figure 7).

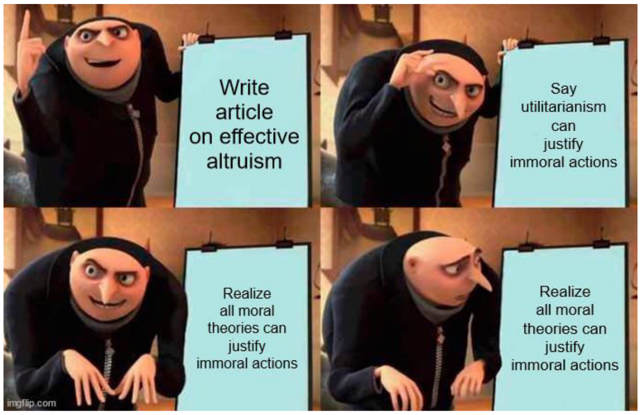

Write article on effective altruism
Say utilitarianism can justify immoral actions
Realize all moral theories can justify immoral actions
Realize all moral theories can justify immoral actions
imgflip.com

Figure 7: As it turns out, it is not helpful to point out that [inset moral theory or theory family here] "can justify" [insert immoral action here], and this is especially true since EA is not inherently utilitarian.

Another way to see why any family of theories can justify any action as permissible is because these families are interchangeable in terms of their deontic predicates. In other words, for any deontological theory, we can construct a consequentialist theory that has all the same moral outcomes for all the same actions (deontic predicates like permissible, obligatory), and vice versa. This construction is called consequentializing.[59] In the same way, we construct a deontological theory for any consequentialist theory, using a method called deontologizing.[60] There is debate over the significance of this, but the key conclusion here is that for any specific action that a deontologist can say is wrong, a consequentialist can say is wrong, and vice versa.

The takeaway from our exploration so far is that any objection to some theory for making some actions permissible needs to reference a specific version of the theory rather than the whole family of theories. For example, it is no objection to *consequentialism* that hedonistic utilitarianism makes it morally obligatory to go through the experience machine, since hedonistic utilitarianism is a subset of the family of theories, but it is a legitimate objection to hedonistic utilitarianism. Therefore, the claim that *consequentialism* can justify anything is true but uninteresting, since the same exact claim can be made of deontology, virtue ethics, or any other theory or anti-theory.

Specific Moral Theories Do Not Justify Any Action

Second, **while any specific theory "can" justify any action, any specific theory *does not* justify any action**. A significant chunk of applied ethics, and one of the primary methods of applied ethics, is taking a moral theory (or framework) and plugging in the relevant descriptive and evaluative information in order to ascertain the moral outcome of various actions. In other words, a large goal in ethics is to figure out what a moral theory actually implies for any given situation. People write many papers for and against various views, including when working from the same starting points, including the same specific theory at times. All of these contradictory implications cannot be correct. However, there is a fact of the matter about the proper implication of the theory for the specific actions, and so therefore a specific theory does not, though it can, justify any action.

Part of the issue here is obscured by the lack of definition of the word "can" in this claim. The word "can" (or "could") is doing all the work in this claim. It is never specified how this is supposed to be translated. It is common in philosophical circles to distinguish different types of possibility (or claims about what can but not necessarily will happen): physical (aka nomological), metaphysical, epistemic, and broad logical possibility. Most common (depending on the context), especially for ethics circles, is metaphysical possibility, which is typically cashed out in terms of possible worlds as implemented in modal logic.

In other words, my best guess is that to say a theory "can justify" an action means that the theory implies that some action is permissible in a possible world (aka a way that the world could have been). Presumably, the worry here is about classes of actions, like lying, running, boxing, stealing, killing, etc. So, a theory can justify any action is that for any class of actions, there is a possible world where it is permissible to do that class of action. If conceivability is at least a good guide to possibility, then any thought experiment will do to show that a class of actions can be permissible in other possible worlds.

Furthermore, as we discussed earlier, on any plausible theory (including versions of consequentialism, deontology, and virtue ethics), there is some point where contextual considerations render the results so significant that it must be permissible. To deny this is to accept absolutism with all of its many problems discussed earlier. Therefore, all plausible moral theories will have members of all classes of actions that are permissible in *some* possible world, however fantastical. Therefore, all specific moral theories "can" justify in action in the sense that there are possible worlds where some action type is permitted.

However, **any given specific theory does not justify any action**. The reason for this is simple: the actual world is a subset of cardinality 1 of the set of all possible worlds, which is infinite. So, while a theory "can" justify any action, it does not justify any action or it faces incoherence. While a theory can justify an action in a world very different from our own, different physics, people, circumstances, laws (physical and political), etc., it does not justify any action in the actual world.

Since the much more interesting concern is about what is permissible or impermissible in the actual world, we care much more about whether theories *do* in fact justify various actions rather than that they *can* justify various actions.

Specific EA and Longtermism Frameworks Do Not Justify Any Action

The same applies to moral frameworks like effective altruism and longtermism, not just theories. EA and longtermism can also be understood as having a family resemblance of models. There is a correct way of filling in the details, but since we are not certain what that is at this time, and we have substantial disagreement, EA is committed to cause neutrality. So, because there is substantial disagreement on filing in these details, they "can" justify a wide range of actions. Yet, just like all moral theories, there *is* a *correct* way of working out the details. Thus, we need to investigate this question seriously to know what the exact implications of their commitments are.

In addition, Levitz has a suspicion that 'expected-value' calculations can be used to justify anything. Well, if all you have is an equation for expected value, and you ignore the rest of a moral framework, then yes. But that's why you have the rest of the moral framework. If you only have agent-centered restrictions without filling in the details of what they are, you can say that it's obligatory to punch every stranger in the face as soon as you see them. Therefore, deontology can justify virtually anything right? Not really. Obviously, you have to fill in the details, and the details need to be remotely plausible to be worth consideration. If I defend a version of virtue ethics where the only virtue is being self-centered, I will justify many terrible actions. You obviously have to compare the actual theories themselves, and you need to compare plausible theories. See the helpful discussions on this general point by Richard Yetter Chappell here and here.

Therefore, the phrase considered at the beginning is either false or uninteresting, depending on how it is interpreted. I will reemphasize Fletcher's comments, "'Does a worthy end justify *any* means? Can an action, no matter what, be justified by saying it was done for a worthy aim?' The answer is, of course, a loud and resounding NO!"[61] At least, not in any interesting way.

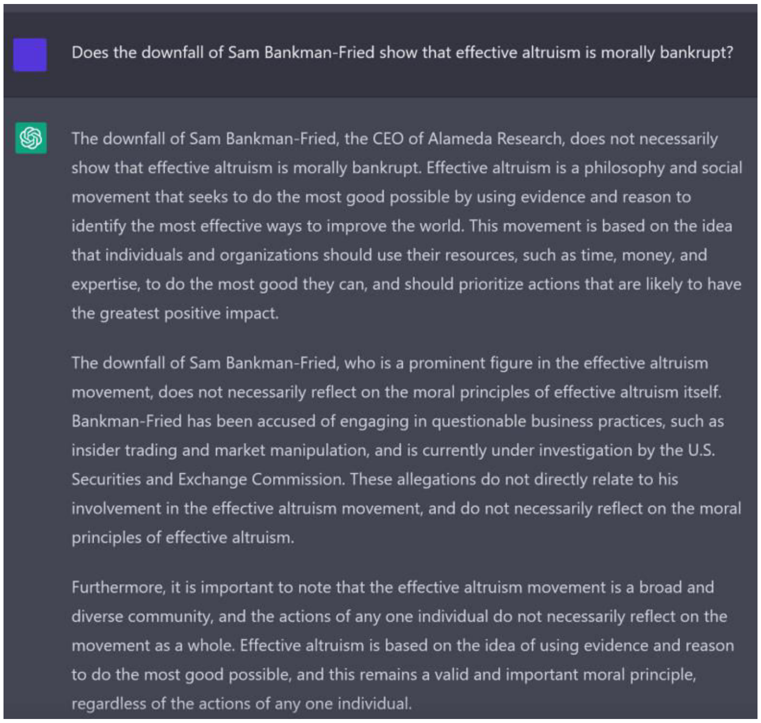
Takeaways and Conclusion

The FTX scandal is very sad for effective altruism, cryptocurrency, and beyond, since a lot of money which was, or that would be going to, saving (or sustaining) people's lives no longer will. Lots of people were hurt and will be worse-off as a result, to say the least. But as far as presenting an argument against effective altruism goes, I think there are, fortunately, no takeaways whatsoever here. The people that used SBF as an opportunity to critique a commitment to "doing the most good with one's donations and career" failed to present a decent argument.

From a Christian perspective, this debacle is similar to many scandals in Christendom that have occurred, where important or powerful leaders have committed vicious actions or formed cults of personality that have completely wrecked many people's lives and entire churches and communities. Examples include Mark Driscoll, Ravi Zacharius, and many others. These are tragedies and the actions of these leaders must be viciously condemned. Yet, from the very beginning, we know people go horribly astray. They make terrible mistakes. The only person we can have perfect faith in, and always strive to exemplify, is Jesus. Leaders do not always (and in fact rarely do always) reflect the core of their commitments. We've all heard this point 50,000 times, and yet somehow people keep thinking that leaders' mistakes are a direct result of following the teachings that they supposedly espouse. This is not always (perhaps even rarely) the case.

For someone interested in purely assessing how effective altruism's framework and approach fares, and whether EA should change its key commitments, the scandal remains entirely uninteresting and uneventful. Another day, another round of horrid critiques of effective altruism. It remains a very good thing to prevent people from dying of starvation and preventable disease, and if we can save more people's lives by donating to effective charities, I am going to keep donating to effective charities.

If you have not yet been convinced of my arguments, listen to what ChatGPT (an artificial intelligence chatbot recently launched by OpenAI) had to say about the implications of SBF for EA in Figure 8, which is that the scandal does not necessarily reflect the moral principles of EA, and this same conclusion is true for any given individual. ChatGPT also agreed that EA is not inherently utilitarian.



Does the downfall of Sam Bankman-Fried show that effective altruism is morally bankrupt?

The downfall of Sam Bankman-Fried, the CEO of Alameda Research, does not necessarily show that effective altruism is morally bankrupt. Effective altruism is a philosophy and social movement that seeks to do the most good possible by using evidence and reason to identify the most effective ways to improve the world. This movement is based on the idea that individuals and organizations should use their resources, such as time, money, and expertise, to do the most good they can, and should prioritize actions that are likely to have the greatest positive impact.

The downfall of Sam Bankman-Fried, who is a prominent figure in the effective altruism movement, does not necessarily reflect on the moral principles of effective altruism itself. Bankman-Fried has been accused of engaging in questionable business practices, such as insider trading and market manipulation, and is currently under investigation by the U.S. Securities and Exchange Commission. These allegations do not directly relate to his involvement in the effective altruism movement, and do not necessarily reflect on the moral principles of effective altruism.

Furthermore, it is important to note that the effective altruism movement is a broad and diverse community, and the actions of any one individual do not necessarily reflect on the movement as a whole. Effective altruism is based on the idea of using evidence and reason to do the most good possible, and this remains a valid and important moral principle, regardless of the actions of any one individual.

Figure 8: ChatGPT knows what’s up regarding SBF and the implications for EA (i.e., not much). Note: I only include this on a lighthearted note, not as a particularly substantive argument (though I 100% agree with ChatGPT).

Post-Script

If I have time and energy (and there appears to remain a need or interest), I will write a part 2 to this in perhaps early January. Part 2 would include criticisms I found even less interesting or plausible, those that relate to the connection between longtermism and EA, the danger of maximizing, the homogeneity of EA, concerns about community norms, and more point-by-point responses to various critical pieces published online recently. Perhaps there also will be more relevant information revealed or more poignant responses between now and then; one very recent piece has more thoroughly suggested that EA leaders should have known about SBF’s dealings, and I may investigate that more carefully. Let me know what else, if anything, I should include, and if you would be interested in a follow-up.[62]

Endnotes

[1] MacAskill, William. "The Definition of Effective Altruism." in *Effective Altruism: Philosophical Issues* (2019), p. 14.

[2] See Strasser, Alex. “Consequentialism, Effective Altruism, and God’s Glory.” (2022). Manuscript. [pdf] for more discussion about these distinctions and their motivation.

[3] This is the so-called *Compelling Idea* of consequentialism, which trivially entails normative EA and non-trivially entails radical, normative EA.

[4] Although, I realized when writing this that I might actually be a strong longtermist for Christian reasons. Namely, I probably think evangelism is the most important moral priority of our time, and the concern for the longterm future (e.g., afterlife) is sufficient to make evangelism the most important moral priority of our time. It looks like this makes me a strong longtermist after all. I need to consider this further.

[5] Boecking, Benedikt, et al. "Quantifying the relationship between large public events and escort advertising behavior." *Journal of Human Trafficking* 5.3 (2019): 220-237.

[6] Cryptocurrency emissions estimated as the maximum of the range given by the White House, which is 50 million metric tons of carbon dioxide per year. Cost per metric ton of offset is $14.62 by Cool Effect (accessed 11.27.22). This amounts to $731 million to carbon offset the entirety, which is 1/35 of SBF's net worth before the scandal. Of course, SBF and FTX's contribution to the U.S. crypto emissions is a small fraction of that, so he could even more easily offset his own carbon emissions. Another difficulty is that it is unlikely that Cool Effect could easily (or at all) implement projects at the scale required to offset this amount of emissions, which is more than some countries in their entirety.

[7] This may assume that we have more negative duties than positive duties. It is frequently defended (or assumed) that we have stronger reasons to prevent harm than to promote beneficence, in which case the argument would go through.

[8] This distaste is mostly because of his views on abortion and infanticide. While I vehemently disagree with Singer on these specific issues, Singer's thoughts on these issues do not affect his thoughts on poverty alleviation or the EA framework in general. It is also true that Singer's views on these issues are sometimes distorted, which is why Eric Sampson (a Christian) wrote a clarifying piece on Singer's views in context of backlash to Singer's visit to Eric's campus.

[9] MacAskill, William. "The Definition of Effective Altruism." in *Effective Altruism: Philosophical Issues* (2019), p. 20.

[10] For an insightful and conversational introduction to this debate, including on whether the ends justify the means or there are intrinsically evil acts that cannot ever be done, and more, see Fletcher, Joseph and Wassmer, Thomas. Edited by May, William E. *Hello, Lovers! An Introduction to Situation Ethics*. Cleveland: Corpus, 1970.

[11] I will assume "justify" means something like "renders morally permissible," whether as a truth condition or an explanation of its moral permissibility.

[12] Fletcher, Joseph. "Situation Ethics, Law and Watergate." *Cumb. L. Rev.* 6 (1975): 35-60, p. 52.

[13] Hurley, Paul. "Consequentialism and the standard story of action." *The Journal of Ethics* 22.1 (2018): 25-44.

[14] Fletcher, "Situation Ethics, Law and Watergate," p. 52.

[15] Sterba, James P. "The Pauline Principle and the Just Political State." *Is a Good God Logically Possible?* Palgrave Macmillan, Cham, 2019. 49-69, p. 49.

[16] Fletcher, "Situation Ethics, Law and Watergate," p. 51, emphasis in original.

[17] Ironically, I am quite skeptical that killing in 1-to-1 (or more generally m attackers vs n victims where $m \geq n$) self-defense scenarios or war are ever justified *in real-world scenarios*. We can construct scenarios where it obviously would be, but I am skeptical we have sufficient reason before starting a large-scale war to think the foreseeable consequences of the war would result in fewer deaths (or other goods) long term than we would have without the war. I still need to investigate further though. It is ironic that I am less likely to think killing is ever permissible in the real world than those who frequently verbalize their opposition to ends-means reasoning.

[18] Of course, some natural law theorists and some Kantians may disagree, but I am more concerned about those with plausible moral theories.

[19] It is possible that this phrase is intended to claim what the moral explanation of any deontic outcome is or its structure of reasons. Namely, that why actions are right or wrong are never its consequences, which is the distinguishing aspect of consequentialism, that the rightness/wrongness of actions are ultimately explained by consequences instead of e.g., duty. As such, it would merely be a restatement of the claim "Consequentialism is false," and then it could not even be used in the debate, since it begs the question against the consequentialist. I do not think the principle is intended to make a technical point about the proper structure of normative theories and normative explanation, but if so, it remains impotent as a moral principle.

Also, for threshold deontology, it may be the case that the explanation for why post-threshold actions are right is by appeal the consequences, so then this understanding of the phrase would be more clearly neutral between theories.

[20] Aboodi, Ron, Adi Borer, and David Enoch. "Deontology, individualism, and uncertainty: A reply to Jackson and Smith." *The Journal of Philosophy* 105.5 (2008): 259-272, p. 261 n. 5.

[21] Huemer, Michael. "Lexical priority and the problem of risk." *Pacific Philosophical Quarterly* 91.3 (2010): 332-351.

[22] Tarsney, Christian. "Moral uncertainty for deontologists." *Ethical Theory and Moral Practice* 21.3 (2018): 505-520.

[23] Mogensen, Andreas, and William MacAskill. "The Paralysis Argument." *Philosophers' Imprint* 21.15 (2021).

[24] Huemer, Michael. *Knowledge, Reality and Value*. Independently published (2021), p. 297 of pdf.

[25] For example, there is the paradox of deontology as well as the related problem of inconsistent temporal discounting. The paradox of deontology is that deontology implies violating constraints is impermissible even when doing so means that you (and/or others) will violate the constraint many fewer times in the future, which is quite counterintuitive. The second problem occurs because modelling absolute constraints requires infinite disvalue for the immediate action but a discounted, finite disvalue for the same action in comparable circumstances in the future. The circumstances are only finitely different yet there is an infinite difference in the disvalue of the same action, which appears inconsistent.

[26] See related and helpful discussion in Fletcher, Joseph and Wassmer, Thomas. Edited by May, William E. *Hello, Lovers! An Introduction to Situation Ethics*. Cleveland: Corpus, 1970, pp. 6-7. Fletcher, who identifies situation ethics as necessarily consequentialist or teleological, also says that for the single principle of situation ethics, he is deontological in a twisted sense.

[27] Consequences as understood in moral theory encompasses more than the term is used in common parlance. Consequences, in this sense, refers to the action and everything that follows from that action. It is not merely the effects after the action. Consequentialism sums the intrinsic value of the action and everything that follows from that action for all time. Lying, for example, can have intrinsic disvalue, and so can the results of lying, such as destroying a relationship. Anything, in principle, can be assigned value in a consequentialist theory, including intentions, motivations, virtues, and any subcategory of action. Further, these categories can be assigned infinite disvalue so that there are absolute constraints, if so desired.

[28] Cloutier, David. *The Vice of Luxury: Economic Excess in a Consumer Age*. Georgetown University Press, 2015, p. 137.

[29] Ambrose, "On Naboth", cited in Phan, Peter C. *Social Thought*. Message of the Fathers of the Church series, Vol. 20, 1984, p. 175.

[30] Singer, Peter. "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1.3 (1972), pp. 229-243.

[31] MacAskill, "The Definition of Effective Altruism," p. 14.

[32] See Pummer, Theron. "Whether and Where to Give." *Philosophy & Public Affairs* 44.1 (2016): 77-95 for a defense of this view, and Sinclair, Thomas. "Are we conditionally obligated to be effective altruists?" *Philosophy and Public Affairs* 46.1 (2018) for a response.

[33] Berkey, Brian. "The Philosophical Core of Effective Altruism." *Journal of Social Philosophy* 52.1 (2021): 93-115.

[34] See Strasser, Alex. "Consequentialism, Effective Altruism, and God's Glory." (2022). Manuscript. [pdf]

[35] For example, Timmerman, Travis. "Sometimes there is nothing wrong with letting a child drown." *Analysis* 75.2 (2015): 204-212 or Kekes, John. "On the supposed obligation to relieve famine." *Philosophy* 77.4 (2002): 503-517.

[36] Haydar, Bashshar, and Gerhard Øverland. "Hypocrisy, poverty alleviation, and two types of emergencies." *The Journal of Ethics* 23.1 (2019): 3-17.

[37] Singer, "Famine, Affluence, and Morality," p. 231.

[38] He also proposed a third one in *The Life You Can Save:* (3) if it is in your power to prevent something bad from happening, without sacrificing anything nearly as important, it is wrong not to do so. See discussion in Haydar, Bashshar, and Gerhard Øverland. "Hypocrisy, poverty alleviation, and two types of emergencies." *The Journal of Ethics* 23.1 (2019): 3-17, who argue that none of these three principles are needed to retain the intuition in the drowning pond case. We only need a weaker principle: (4) if it is in your power to prevent something bad from happening, without sacrificing anything significant, it is wrong not to do so.

[39] McMahan, Jeff. "Philosophical critiques of effective altruism." *The Philosophers' Magazine* 73 (2016): 92-99.

[40] Thanks to Dominic Roser for pointing this out to me.

[41] See Miller, Ryan. "80,000 Hours for the Common Good: A Thomistic Appraisal of Effective Altruism." *Proceedings of the American Catholic Philosophical Association* (forthcoming) and Synowiec, Jakub. "Temperance and prudence as virtues of an effective altruist." *Logos i Ethos* 54 (2020): 73-93.

[42] For discussion of different criteria of right action proposed in virtue ethics, see Van Zyl, Liezl. "Virtue Ethics and Right Action." *The Cambridge Companion to Virtue Ethics* (2013): 172-196.

[43] Hursthouse, Rosalind. *On Virtue Ethics*. OUP Oxford, 1999, p. 28.

[44] Byerly, T. Ryan. *Putting Others First: The Christian Ideal of Others-Centeredness*. Routledge, 2018.

[45] MacAskill, "The Definition of Effective Altruism," p. 23

[46] MacAskill, "The Definition of Effective Altruism," p. 17

[47] MacAskill, "The Definition of Effective Altruism," p. 20

[48] MacAskill, William. *What We Owe the Future*. Basic Books, 2022, p. 241.

[49] MacAskill, *What We Owe the Future*, pp. 276-277 of my pdf, emphasis in original.

[50] Ibid.

[51] Mogensen, Andreas, and William MacAskill. "The Paralysis Argument." *Philosophers' Imprint* 21.15 (2021).
[52] Schroeder, S. Andrew. "Consequentializing and its consequences." *Philosophical Studies* 174.6 (2017): 1475-1497.
[53] MacAskill, "The Definition of Effective Altruism," p. 18
[54] MacAskill, "The Definition of Effective Altruism," p. 20
[55] MacAskill, "The Definition of Effective Altruism," p. 18
[56] Berg, Amy. "Effective altruism: How big should the tent be?" *Public Affairs Quarterly* 32.4 (2018): 269-287.
[57] MacAskill, "The Definition of Effective Altruism," p. 18
[58] One of the biggest challenges here is theory individuation, or how you distribute credences in theories with slightly varied parameters or structures. See discussion in papers with "My Favorite Theory" in the title by Gustafsson and also MacAskill's book *Moral Uncertainty*.
[59] Portmore, Douglas W. "Consequentializing." *Philosophy Compass* 4.2 (2009): 329-347. There are various challenges to the success of this project, but I won't address those here. I think the challenges can be met.
[60] Hurley, Paul. "Consequentializing and deontologizing: Clogging the consequentialist vacuum." *Oxford Studies in Normative Ethics* 3 (2013).
[61] Fletcher, "Situation Ethics, Law and Watergate," p. 51, emphasis in original.
[62] Featured image adapted from FTX Bankruptcy, common creative license, downloaded here.