What can the philanthropic sector take from the downfall of Samuel Bankman-Fried and his ties to Effective Altruism?

A five-part series

Calum Carmichael

School of Public Policy and Administration, Carleton University

Ottawa, Ontario Canada

April-September 2023

Updated November 2023

PANL
Perspectives
PHILANTHROPY AND NONPROFIT LEADERSHIP

Prologue

In September 2022, the prescient but pseudonymous Sven Rone anticipated the fallout that Effective Altruism (EA) would experience before year's end:

"By relying heavily on ultra-wealthy individuals like Sam Bankman-Fried for funding, … the Effective Altruism community does not appear to recognize that this creates potential conflicts with its stated mission of doing the most good by adhering to high standards of rationality and critical thought…. [A]ttacks on the image of SBF, FTX and even crypto as a whole carry the risk of tarnishing EA's reputation. Were SBF to be involved in an ethical or legal scandal (whether in his personal or professional life), the EA ecosystem would inevitably be damaged as well."

* * *

In November 2022, following the bankruptcy of FTX International and the criminal charges against Samuel Bankman-Fried (SBF), The Economist referred to that fallout:

"The downfall of Mr Bankman-Fried, who has been apparently dedicated to the [EA] cause since his time at university, has led to a reckoning. Not only has effective altruism lost its wealthiest backer; its reputation has been tarnished by association. Many inside and outside the community are questioning its values, as well as the movement's failure to scrutinise its biggest funder—something particularly painful for a group that prides itself on logically assessing risk."

* * *

In the same month, Kate Arnoff described EA's reckoning as the one bright spot in the downfall of SBF:

"This rethinking of effective altruism may be the one bright spot in an otherwise depressing crash …. It's good that FTX's collapse is finally making people rethink Bankman-Fried and effective altruism."

* * *

Also in the same month, Erik Hoel speculated on whether such rethinking would lead to the demise of EA:

"Sam Bankman-Fried, affectionately known as SBF, was until recently effective altruism's biggest funder…. If in a decade barely anyone uses the term 'effective altruism' anymore, it will be because of him…."

Part 1: What's going on here?

Prior to November 2022, Samuel Bankman-Fried (popularly known as SBF) was seen to be a successful and philanthropic entrepreneur. In both his career and giving choices he aligned himself with the ethos, methods and priorities of Effective Altruism (often shortened to EA). In his early 20s, for example, he reportedly decided to work not for a charity but in finance, believing greater good would come from his earning more to give more. In 2018 at the age of 26, he founded the cryptocurrency trading company Alameda Research to profit from crypto prices being higher in Asia than America. The following year he founded the FTX Group of cryptocurrency exchanges to service the trades of Alameda and other clients, promoting his companies as workplaces of choice for effective altruists. He based the largest of the exchanges, FTX International, first in Hong Kong and then in the Bahamas to handle popular products not legal in the US.

For both the 2020 presidential and 2022 mid-term elections, SBF was among the biggest contributors to Democratic candidates. In February 2021 he established the FTX Foundation for charitable giving, donating $50 million US that year. In June 2022, with a net worth having peaked above $25 billion US, SBF signed the Giving Pledge to donate most of his wealth during his lifetime or by will, with $16.5 billion US from FTX sources having already been earmarked for EA causes and organizations.

But in November 2022, all of this changed. Alameda and FTX declared bankruptcy. In December and into 2023, SBF was indicted on multiple charges of fraud, conspiracy, money laundering, foreign bribery and campaign finance violations. In November 2023, he was convicted on seven charges of fraud and conspiracy, with the additional charges of bribery and illegal political donations coming to trial in March 2024.

The mismanagement and criminal behaviour of SBF and his associates have had wide effects. They've caused great financial harm to the customers and creditors of FTX, as well as reputational harm to cryptocurrency markets. But they've also intensified and refocused pre-existing criticisms and suspicions directed to not only the ethos and methods of EA with which SBF had associated himself, but also the organizations of EA and their leaders who had associated themselves with him. These include Giving What We Can, 80,000 Hours, the Centre for Effective Altruism, and their co-founder William MacAskill. It was he who advised the young SBF that greater good could come not from his working directly for a charity but from pursuing a high-earning career in finance that would allow him to give more to EA causes.

Responding to the charges and aware of the intensified criticism, in November 2022 MacAskill expressed his personal dismay, shame and need to reflect:

> If FTX misused customer funds, then I personally will have much to reflect on. Sam and FTX had a lot of goodwill – and some of that goodwill was the result of

> association with ideas I have spent my career promoting. If that goodwill laundered fraud, I am ashamed. As a community, too, we will need to reflect on what has happened, and how we could reduce the chance of anything like this from happening again…. I know that others from inside and outside of the community have worried about the misuse of EA ideas in ways that could cause harm. I used to think these worries, though worth taking seriously, seemed speculative and unlikely. I was probably wrong.

To be sure, the criticisms amplified since the downfall of SBF have been directed specifically toward Effective Altruism. But to my mind, they have ramifications that extend to the philanthropic sector as a whole. They deserve to be heard, reflected upon and heeded more widely.

What're we to take from this?

To support that wider attention and reflection, I've prepared this five-part series for PANL Perspectives. To structure it, I've condensed the major criticisms of EA into seven points, arranging them under three headings and sequencing them so they could be read as one continuous argument:

The philosophical foundations of Effective Altruism

1. *The ethical bases of EA rely on a narrow version of utilitarianism to the exclusion of other ethical theories or considerations, such that they encourage its adherents – through their philanthropy – to pursue purportedly good ends using potentially harmful or corrupting means.*
2. *Those excluded considerations include human emotion or loyalty as guides to philanthropic choice, such that EA undercuts philanthropists' agency and overlooks or opposes key aspects of human motivation.*

The analytical approaches of Effective Altruism

3. *By relying on impartial reason to identify the philanthropic interventions that will do the most good, EA idealizes a methodology that quantifies and compares the value and probabilities of alternative and highly-speculative outcomes – thereby mistaking mathematical precision for truth and ignoring important qualities of human life and flourishing that are not readily quantified.*
4. *This methodology – bolstered by its ethical assumptions and claims of impartiality – cultivates hubris, a condescension toward and dismissal of contending priorities or sources of information, and the impulse to define and control philanthropic interventions on one's own terms.*

The ultimate effects of Effective Altruism

5. *Moreover, this methodology – by focusing on separate, numerically-evaluated interventions – overlooks the wider behavioural, institutional or systemic conditions that might not only limit the effectiveness of the interventions themselves but also cause or perpetuate the societal ills they seek to address.*
6. *By not addressing those systemic conditions, EA takes on a conservative agenda: one that distracts from and thereby perpetuates the political, social and economic status quo and the inequalities and deprivations therein.*
7. *Accordingly, in its formation, methods, application and effect, EA is elitist: it risks becoming an intellectual, do-gooder playground for the privileged who ultimately benefit from – and through their philanthropy avoid substantially changing or challenging – the inequalities of the world around them.*

Part 2 of this series provides an overview of EA: its origins, ethos, analytical methods, priorities and evolution. Parts 3, 4 and 5 present the criticisms and their rejoinders that apply, respectively, to its philosophical foundations, analytical methods and ultimate effects. In presenting these, I've tried to use the words and preserve the voices of critics and defenders alike. Throughout, however, my goal is not simply to summarize the contending views of EA, but to help us derive from them implications and questions for the philanthropic sector as a whole. I provide several examples of these under each heading. But undoubtedly, readers will be able to identify others, so that regardless of our different connections to the sector, each of us will be able to take and possibly apply something from the downfall of Samuel Bankman-Fried and his ties to EA.

Part 2: Effective Altruism

What're we talking about?

Effective Altruism (EA) consists of two things. First, it's a research area that uses evidence and reasoning to determine ways to do the most good with a given amount of philanthropic resources. Second, it's a community of practice comprising individuals and organizations that use the findings of that research to direct their resources – whether time or income – toward the ways found to do the most good. In some respects, it resembles the secular philanthropy clubs founded in Paris and London during the late Enlightenment – the Société Philanthropique of 1780 or the London Philanthropic Society of 1788 – in disparaging indiscriminate forms of charity, promoting reason as a means to identify the most effective ways to improve well-being particularly of the poor, and then privately funding and managing those ways.

Origins

In its current form, EA emerged less than two decades ago out of initiatives taken by young adults in the UK and US both to measure differences in the cost-effectiveness of charities and to encourage giving on that basis.

In 2009, Toby Ord and William MacAskill, then a research fellow and graduate student with the Faculty of Philosophy at Oxford University, collaborated in measuring those differences across charities addressing poverty and its effects in low-income countries. Within months, they created the organization Giving What We Can that encourages members to donate at least 10% of their income in perpetuity to the charities found to be most effective. Its membership has grown from the original 23 to more than 8,200. In 2011, MacAskill and a fellow student Ben Todd co-founded the organization 80,000 Hours (what an average person works over a lifetime) to provide advice on careers that could combine a personal fit with high social impact. By the end of that year, MacAskill, Ord, Todd and 14 others involved with the two organizations decided to incorporate them as a charity, finally agreeing on the name Centre for Effective Altruism (CEA). The term "Effective Altruism" was established.

Meanwhile in the US, Holden Karnofsky and Elie Hassenfeld, both in their 20s, decided in 2007 to leave their jobs at an investment firm to found and staff GiveWell, a research organization that uses performance metrics to identify the most cost-effective charities in a variety of fields. In 2012, GiveWell partnered with Good Ventures, the charitable foundation established by the Facebook co-founder Dustin Moskovitz and his wife Cari Tuna. That partnership led to the creation of Open Philanthropy, a research and granting foundation that in 2017 became a separate organization with Karnofsky as CEO, leaving Hassenfeld as CEO of GiveWell. These people and organizations formed linkages with their counterparts in the UK, and adopted the term Effective Altruism to describe their own research agendas and giving practices.

Ethos

Given its origins, one can understand why EA interprets its research and giving in explicitly ethical terms. To a large degree, its philosophical roots draw from utilitarianism – the premise that actions are moral to the extent they promote total well-being. More pointedly, they draw from the work of the philosopher Peter Singer, starting with the four-fold "basic argument" that he laid out in 1972 during the famine in Bangladesh. Given:

- that the human misery from lack of food, shelter or medical care is bad;
- that if we as individuals can prevent something bad from happening, then it's wrong for us not to do so;
- that through our philanthropy, we as individuals can relieve the misery of others without sacrificing anything as important; then
- if we as individuals don't give to relieve that misery, we're doing something wrong.

Utilitarianism is a form of consequentialism which defines actions as moral based on the goodness of their outcomes. It's distinct from other ethical theories such as deontology which defines actions as moral based on their compliance with certain rules or duties like respecting human rights. And it's distinct from virtue ethics that locates morality not in the attributes of actions, but rather in the underlying character traits of individuals – whether those are virtuous or not. Accordingly, the philanthropy endorsed by EA may overlap with but is distinct from "justice philanthropy" that would have individuals fulfill the duty to compensate for the human rights violations they have collectively caused and benefited from (Cordelli; Pogge); or from "community philanthropy" that would shift and share decision-making power with those in need or who've had their rights violated (Hodgson and Pond; Villanueva); or from approaches to philanthropy that would assist givers in leading fulfilled and virtuous lives (Martin; Anderson).

There are many ways in which effective altruists could configure utilitarianism or consequentialism to assess the morality of alternative philanthropic causes or interventions. For example, such ways could differ according to:

- the benefits and costs that compose total well-being (whether these affect pleasure, pain or longevity, or include less-tangible things such as freedom, knowledge, capability; whether the relevant consequences are actual or expected ones; whether intrinsically bad actions – like torture or lying – can be considered instrumentally beneficial if they lead to intrinsically good outcomes like saved lives);
- who tallies the benefits and costs (whether an objective observer or the parties directly affected – "nothing about us without us");
- the population whose well-being is of concern (whether all sentient beings everywhere and for all time, or a smaller group limited by species, location or time period); and
- the relative importance of the entities making up that population (whether they're of equal merit such that the benefits and costs that compose well-being are weighted the

same regardless of who or what experiences them, or whether they're of different merit such that the benefits and costs are weighted differently if they affect, say, humans rather than animals, people living in Canada rather than elsewhere, or people living in the present rather than the future).

To be sure, in pursuing "the most good" EA doesn't rely exclusively on utilitarianism or on a single or simplistic way either to define total well-being, to anticipate the consequences of an intervention or to measure such things. It recognizes such tasks as being subject to ethical ambiguities as well as side constraints like "do no harm." That said, EA relies on impartiality. This involves being neutral at the outset across alternative causes and the means of advancing them, and only narrowing one's options on evidence of what would do the most good with a given amount of resources. Impartiality also involves not limiting one's circle of concern by geographical location, species or time period: what matters is the well-being of and consequences for all sentient life, regardless of where, how or when it exists. Accordingly, EA favours treating all life as having comparable if not equal merit. It emphasizes tangible benefits and costs that affect pleasure, pain or longevity. It focuses on expected outcomes. And it idealizes their assessment as if by an objective observer.

Analytical methods

EA applies evidence and reasoning in the context of an analytical framework. For the past decade, that framework has distinguished alternative causes and interventions according to their importance, neglectedness and tractability. For a given outlay of philanthropic resources, *importance* refers to such things as how many would benefit in terms of lives extended or improved, by what extent would they benefit and for how long. *Neglectedness* refers to how much an intervention would add to or alter the amounts and allocations of resources already or potentially dedicated to a problem. And *tractability* refers to how easy it would be to make and detect progress in alleviating a problem.

The INT framework was laid out in 2014 by Karnofsky at GiveWell. Within it, the outcomes of concern take place in the near future and can't be measured directly. Accordingly, decisions are based on expected values. Constructing these requires estimating for a set of possible values for each outcome – say, the number of lives affected by a certain intervention – as well as their probabilities summing to one. The expected value of the outcome is then the sum of its possible values multiplied by their probabilities. Understandably, the INT framework lends itself to indicators or outcomes that can be readily quantified. Where data are sufficient, the ability to estimate the outcome of a particular intervention can be improved by using various statistical methods whether these are experimental (e.g., randomized control trials), quasi-experimental (e.g., propensity score matching), or non-experimental (e.g., regression discontinuity designs or difference-in-difference models). As suggested by the top charities selected by GiveWell, such methods favour health interventions in low-income countries for which data are available and where philanthropic resources go further.

Although the INT framework readily accommodates quantitative methods, it doesn't require them. There are other grounds on which to build clear and rigorous arguments. These could include qualitative empirical methods that use, say, interview transcripts or survey responses. Or they could include theoretical research, whether tied to techniques such as game theory, decision theory or computer simulation, or related to disciplines such as ethics, analytical philosophy, psychology or international relations. Theoretical research could be relied upon, for example, to assess the cost-effectiveness of interventions intended to reduce threats to the future survival of sentient life for which quantitative or qualitative data are either incomplete or don't exist.

To handle such threats and accommodate theoretical approaches, EA recently introduced in conceptual form a second analytical framework that distinguishes the significance, persistence and contingency (SPC) of events or interventions. These three concepts determine the capacity of an intervention to affect the total long-term value that a future state of the world would generate for its population. There are different versions of the SPC framework, but all assume that an intervention could not only increase or decrease that total value, but also increase or decrease the duration of a future state by causing it to begin earlier or later. The *significance* of an intervention refers to the change in total value divided by the change in duration it would bring about. *Persistence* refers to that duration whether extended or shortened. Finally, c*ontingency* or non-inevitability deals with the counterfactual. It refers to how much additional time it would have taken for the future state to begin (if ever) had the intervention not taken place, expressed as a fraction of the future state's duration.

By estimating and then multiplying together the significance, persistence and contingency of an intervention, one estimates the change in total long-term value it would bring about – a concept related to *importance* from the INT framework. Because these three terms multiply, once estimated they allow for comparisons of long-term effects. For example, an intervention being 8 times as persistent as another means that it would generate more long-term value – have a greater importance – even if the alternative is 7 times as significant.

Priorities

Particularly over the past decade the causes promoted by EA have come to focus less on improving the well-being of humans and animals in the near term and more on identifying and averting possible threats to sentient life in the far future. Admittedly, such causes were there from the beginning. In Oxford, Toby Ord's position as a research fellow was with the Future of Humanity Institute, a multidisciplinary unit established in 2005 under its founding and current Director Nick Bostrom. Then as now, the Institute's research deals with "existential risk" or "x-risk" which Bostrom attributes to events that "would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential", thereby either exterminating humankind altogether or having "major adverse consequences for the course of human civilization for all time to come." Such threats include nuclear war and climate change, but now feature pandemics whether natural or bio-engineered as well as artificial superintelligence (ASI) and its potential for being used maliciously by either human agents or itself.

Ord and other early players in EA convinced MacAskill that the opportunities to avert x-risk have a cost-effectiveness which far outstrips that of most if not all other causes. Using the nomenclature of the INT framework, what such opportunities lack in tractability, they more than make up for in importance and neglectedness. As the thinking goes, because so many more intelligent beings could live in the future than have ever lived to date – trillions rather than billions – the most important thing one can do now is increase the chance that their future will actually happen.

Two books written by MacAskill for popular audiences mark his own shift from so-called "neartermist" priorities to "longtermist" ones: the book published in 2015 argues for using philanthropy primarily to relieve current human and animal suffering, and for identifying congenial careers that could offer high social impact; whereas that published in 2022 argues for using philanthropy to increase by even an infinitesimal amount the probability that many trillions of human and potentially digital entities with "Earth-originating" intelligence will exist in the far future, whether on Earth or beyond. Reflecting this shift, in 2022 the original EA organization – Giving What We Can – added to its top-rated funds one dedicated to the Long-Term Future which "aims to positively influence the long-term trajectory of civilisation by making grants that address global catastrophic risks, especially potential risks from advanced artificial intelligence and pandemics".

The pivot toward longtermist causes – particularly avoiding rogue ASI – has also taken place in the US, fueled there by the attention and philanthropy of Silicon Valley and the crypto industry. Such causes compose a key focus area for Open Philanthropy, and according to its CEO Holden Karnofsky make this century the "most important ever for humanity." The major grantees of Open Philanthropy include the Machine Intelligence Research Institute (MIRI) that was founded in 2005 by the effective altruist Eliezer Yudkowsky with the mission to make AI systems "safer and more reliable when they are developed". MIRI has also attracted large donations from: Peter Thiel, co-founder of PayPal; Jaan Tallin, co-founder of Skype; and Vitalik Buterin, co-founder of Ethereum.

Samuel Bankman-Fried had positioned himself to be the major supporter of longtermist causes. In February 2022, a year after establishing the FTX Foundation, he and his colleagues created within it a Future Fund with MacAskill on the advisory board. In its first year, this Fund was to have disbursed between $100 million and $1 billion US "to improve humanity's long-term prospects" through "the safe development of artificial intelligence, reducing catastrophic biorisk, improving institutions, economic growth, great power relations, effective altruism, and more." But late in 2022 – following the allegations, bankruptcies and criminal charges noted in part 1 – the Foundation, the Fund and those plans came to an end.

Evolution

In under two decades, EA has moved far beyond the initiatives of a few young adults working independently in the UK and US. It's now a well-connected international movement comprising thousands of supporters and affiliates, including 10 signatories of the Giving Pledge. The causes that first focused on alleviating the immediate effects of poverty in low-income countries have broadened to include and emphasize averting the extermination of Earth-originating intelligence or at least delaying it far into the future. Annual donations have gone from thousands to millions, with commitments measured in billions. Admittedly, those donations and commitments come disproportionately from the tech and crypto sectors, and thus are subject to their booms and busts. Bankman-Fried provides an extreme case: across 2022 his net worth peaked at an estimated $25.9 billion US in March, fell to $8.1 billion in June, rose to $12.8 billion in August and evaporated in December. Despite such volatility, the ongoing and potential resources now available for EA operations are huge. In the US, GiveWell, the research organization founded by its two initial employees, has developed into multiple research and granting organizations that employ hundreds. In the UK, the Centre of Effective Altruism that in 2013 worked out of the basement of a real estate agent in Oxford had by 2022 been able to purchase Wytham Abbey, a Grade I listed manor house built in 1480 that has hosted, among others, Queen Elizabeth I, Oliver Cromwell, and Queen Victoria.

Such shifts in scale, priorities and resources have come quickly. And they've required EA to adapt in terms of its *needs* and *operations*, as well as its norms and *culture*.

Needs no longer centre around funding, but rather around developing a wider set of relatively cost-effective interventions across new causes. This introduces the need to recruit, keep and accommodate more staff – people that combine the skills to identify, manage and evaluate those interventions, along with an expertise in not only, say, tropical diseases or animal welfare, but also artificial intelligence, virology, emergency preparedness or advocacy and lobbying. In other words, there is a need – expressed by its leadership – for EA to invest in itself: its personnel, skill-set, facilities and membership.

In terms of *operations*, frugality had once required great selectivity: using evidence and reasoning to locate the most reliable ways to do the most good with a given amount of resources. But with frugality no longer paramount, the standards for cost-effectiveness can now be lowered under a revised goal of doing more good with more resources. Greater risks can be taken. Deciding which intervention could do more good or the most good can be based on trial and error. Robustness of evidence – something highlighted by MacAskill in 2015 – is less relevant: expected values can be used, for example, even if what lies behind them are very speculative but highly beneficial outcomes tied to probabilities that are microscopic if even measurable. Alternatively, calculations of cost-effectiveness can be ignored altogether for the time being. Founders Pledge is a British charity to which EA entrepreneurs commit a portion of the proceeds when selling their businesses. In 2021 it opened a so-called Patient Philanthropy

Fund – also available to smaller donors through Giving What We Can – that will only make disbursements "when the highest-impact opportunities are available", perhaps centuries from now.

With respect to *culture*, that of EA has been described as "nerdy, earnest, and moral", "overly intellectual", subject to "relentless, sometimes navel-gazing self-criticism and questioning of assumptions" as well as "hero worship" whether directed to Singer, MacAskill or until recently Bankman-Fried. There's a shared confidence in the ability to identify the causes and interventions best able to promote well-being. And there's a shared esteem for those – including oneself – who undergo personal thrift and a degree of self-sacrifice in order to give more to those causes. This culture, while enduring, has had to deal with differences and tensions over the shift in priorities and surge in funding. Such tensions surface in questions and opposing viewpoints around mission drift that appear on the EA online Forum. How should the unwritten norms of deference and personal thrift apply to billionaires who associate with EA? Is EA's capacity to promote total well-being strengthened or weakened by spending on itself rather than its causes, or by spending on causes that seek to avert the unknown but imagined misery of the distant future rather than those that seek to alleviate the known and very tangible misery of the here and now?

To ease such tensions, MacAskill has called for a spirit of "judicious ambition" within the culture of EA. By his description, this spirit would be judicious in not doing potentially destructive things: "avoiding unnecessary extravagance, and conveying the moral seriousness of distributing funding", "emphasising that our total potential funding is still tiny compared to the problems in the world" and "being willing to shut down lower-performing projects". But it would also be ambitious in doing new and potentially constructive things: creating "more projects that are scalable with respect to funding", investing in additional staff "time and increased productivity", and being "willing to use money to gain information, by just trying something out".

Others, such as Carla Zoe Cremer, seek to address the tensions not by naming and encouraging a new spirit but rather by implementing structural reforms within EA organizations – reforms designed not only to foster greater transparency and accountability in decision-making, but also to counter tendencies toward group-think both among leaders and across the membership. The former would recognize and respond to different positions held around the changes in priorities and funding or around other issues. The latter would better reflect the very terrain in which EA works – one where there exist different and defensible concepts of well-being or of the good, where causes may move beyond the bounds of manageable risk and into the realm of incalculable uncertainty, and where wisdom and judiciousness don't necessarily coincide with the status awarded those who demonstrate greater wit, eloquence or intellectual profile.

Summary and prelude

Effective Altruism is a young but not unprecedented approach to philanthropy – one that aspires to promote total well-being by giving to the causes and interventions that would do the most good with a given amount of resources. It interprets and defends those aspirations on ethical grounds, and seeks to fulfill them both by relying on evidence and reason to identify the best causes and interventions, and by cultivating the ability and willingness of its members to give. Recently, the causes seen as doing the most good have moved beyond alleviating the effects of poverty or improving animal welfare in the near term. They now include and emphasize averting the extinction of Earth-originating intelligence far into the future. Associated with that pivot in priorities has been an influx of funding, particularly from the tech and crypto sectors. Such changes have required adjustments in the analyses and standards for selecting and supporting causes and interventions. And they've introduced tensions within the EA community and encouraged thinking about possible ways by which such tensions could be eased or handled.

For the past decade, Samuel Bankman-Fried has been a prominent figure within EA. As a university student he subscribed to utilitarianism. As a young professional, he acted on the advice of MacAskill that he could do more good by working not in the charitable sector, but rather in finance so as to earn more and thus give more to EA causes. Earn more he did, through the crypto companies he founded, accumulating in four years a net worth measured in billions. And give more he started to do – in part to political campaigns, but largely to causes intended to avert future threats to sentient life. He became a well-known advocate for and representative of EA. He encouraged likeminded people to join his companies. And he was upheld by EA leaders as an example of the dedication and personal thrift (e.g., maintaining a vegan diet, driving a Corolla) that would allow others to do more good by being able to give and commit more to EA causes and interventions.

Confirming this prominence, the revelations and criminal charges brought against Bankman-Fried in late 2022 not only triggered reproach and anger toward him, but also re-focused and amplified existing criticisms and suspicions of EA, its philosophical foundations, analytical approaches and ultimate effects – as outlined in parts 3, 4 and 5.

Part 3: Questioning the philosophical foundations of Effective Altruism

Criticism 1: The ethical bases of EA rely on a narrow version of utilitarianism to the exclusion of other ethical theories or considerations, such that it encourages its adherents – through their philanthropy – to pursue purportedly good ends using potentially harmful or corrupting means.

"The question is: was the FTX implosion a consequence of the moral philosophy of EA brought to its logical conclusion?" Hoel November 2022

* * *

"The problem for effective altruists is not just that one of their own behaved unethically. There is reason to believe that the ethos of effective altruism … enabled and even encouraged the disaster at every step along the way…. [I]t is little more than a fancy way of saying 'the ends justify the means'." Morris November 2022

* * *

"One key feature of utilitarianism is that it doesn't rule out any kinds of actions unilaterally. Lying, stealing and even murder could, in certain situations, yield the overall best consequences…. That doesn't mean that an effective altruist has to say that stealing is okay if it leads to the best consequences. But it does mean that the effective altruist is engaged in the same style of argument." Dunn November 2022

* * *

"If there's a lesson to be learned from the collapse of FTX, it's this: ethics is not the result of calculated consequences. If there's any good to emerge from the rubble, it's this: the demise of utilitarianism as a spiritual guide." Cook November 2022

* * *

This first line of criticism against the philosophical foundations of Effective Altruism (EA) focuses on their connections with utilitarianism and its premise that actions are moral to the extent their consequences promote total well-being. Sure enough, utilitarianism informs EA, whether through the writings of thought leaders such as Peter Singer or William MacAskill, the outlooks of the majority of effective altruists as surveyed in 2017, or the analytical methods used to identify the philanthropic causes or interventions capable of doing "the most good." And sure enough, Samuel Bankman-Fried (SBF) aligned himself with utilitarianism early on. At the age of 20 – perhaps influenced by his parents, both of whom are professors at Stanford Law School – he described himself as "a total, act, hedonistic/one level (as opposed to high and low pleasure), classical (as opposed to negative) utilitarian; in short, I'm a Benthamite." Both parentheses are original.

According to some critics, the presence of utilitarianism has "poisoned" or "corrupted" EA, in part by inviting "'ends justify the means' reasoning, … [and a] maniacal fetishization of 'expected value' calculations, which can then be used to justify virtually anything", ranging from business fraud all the way to such things as bestiality and murder. Even within the EA community there are some thought leaders – Holden Karnofsky being one – who have voiced milder concerns that utilitarianism could weaken the trustworthiness of effective altruists: "Does utilitarianism recommend that we communicate honestly … [or] say whatever it takes … stick to promises we made … [or] go ahead and break them when this would free us up to pursue our current best-guess actions? …. My view is that – for the most part – people who identify as EAs tend to have unusually high integrity. But my guess is that this is more *despite* utilitarianism than *because* of it."

Among external critics of EA, the unease around utilitarianism often focuses on the "earning to give" strategy – the idea promoted by 80,000 Hours that for some effective altruists a career with social impact might involve their working not in positions tackling major problems directly, but rather in high-paying jobs that allow them to donate more to organizations tackling those problems effectively. As noted in parts 1 and 2, it was this strategy that MacAskill proposed to the undergraduate SBF, and of which SBF came to be the most prominent and praised exemplar. Some argue, however, that "earning to give adds a darker possibility of rationalizing unethical means in service of virtuous ends."

This could take several forms. First, the strategy could place well-intentioned people in work environments likely to erode those intentions. For example, "the idea that getting rich is good (or even obligatory) so long as you're giving enough of it away, can become a justification for embracing a soul-corroding competitiveness while telling yourself you're just doing it for the greater good." Alternatively, "the Spartan tastes and glittering ideals of dogooder college students rarely survive a long marinade in the values and pressures and possibilities of expansive wealth." Second, it could encourage people to accept careers that are high-paying but socially harmful, or to undertake business practices that are profitable but shady: "[i]t's easy to see how this could translate to: Go work in crypto, which is bad for the planet, because with all that crypto money you can do so much good." Finally, the strategy might attract duplicitous or at least susceptible characters from the get-go: "[the experience of SBF] is also a warning that if you are the type to try and make billions, you should worry that your ethics are vulnerable along the way." Italics are original. According to one commentator, the italicized warning could also have applied to types who try to receive billions: "It is possible that MacAskill and his peers recognized that running a crypto exchange was inherently unethical, but concluded that it was nevertheless justifiable given the scale of the good that SBF's fortune would do."

There are rejoinders to these criticisms of the role and effects of utilitarianism and the earning-to-give strategy. First, as an ethical theory, utilitarianism offers not a cut-and-dry how-to manual for day-to-day use, but rather a general framework for thinking about what makes actions moral. Like environmentalism or feminism, it provides sufficient latitude for people holding different moral outlooks or priorities to partake.

Second, EA isn't simply utilitarianism – despite SBF labeling it "practical utilitarianism." EA makes no claim that one must sacrifice one's own interests or those of another to serve the "greater good," nor does it specify or insist upon what the "greater good" comprises. Sure enough, as noted in part 2 the EA organization Giving What We Can encourages members to donate at least 10% of their income in perpetuity to the charities found to be most effective. But such a standard isn't unique to EA: it's present in Judaism and Christianity. Moreover, in 1996 philosopher Peter Unger developed arguments akin to those of Singer from 1972 – ones that could have similarly inspired EA – but unlike Singer, he did so disavowing any particular ethical theory, including utilitarianism.

Third, effective altruists aren't all utilitarian: although MacAskill is thought to be, his co-founder of Giving What We Can, Toby Ord, isn't; and although the majority surveyed in 2017 said they were, a sizable minority said they weren't, affiliating instead with another ethical theory (e.g., deontology or virtue ethics) or none.

Fourth, the underlying principles of any ethical theory, if carried to the extreme, could be used to justify abhorrent behaviour. Sure enough, as some critics of EA argue, fanatical utilitarianism could be used to justify the murder of one to save the lives of two. But then again, fanatical deontology could be used to justify not telling a lie even to save the life of an innocent victim. And fanatical virtue ethics could be used to justify sectarian indoctrination or bloodshed. Thus, using extreme extrapolations to declare utilitarianism – or deontology, or virtue ethics – a "flawed philosophy" that has "corrupted" EA isn't only a logical fallacy, but also – if carried to the extreme – a line of reasoning that would dismantle Western ethical thought.

Fifth – turning to the dangers of recommending the "earning-to-give" strategy – such recommendations are infrequent, made perhaps to 15% of effective altruists. For most, careers combining social impact with a better personal fit would come from working directly on important problems – whether through nonprofits, charities, social enterprises, universities, think tanks, government or political organizations.

And sixth, when recommended, earning-to-give comes with guidelines: for example, don't pursue a career that violates the rights of others or that entails fraud, such things being bad both in themselves and in their likely consequences; don't enter or stay in a job where "there is a large gap between your daily conduct and your core commitment"; more generally, "avoid doing anything that seems seriously wrong from a commonsense perspective"; and "in the vast majority of cases" don't pursue "a career in which the direct effects of the work are seriously harmful, even if the overall benefits of that work seem greater than the harms." To be sure, by estimating the donations that would compensate for the harmful aspects of a career or by inserting phrases like "a large gap" or "a commonsense perspective" or "in the vast majority of cases," the guidelines could set up slippery slopes toward profitable but bad behaviour or lucrative but harmful careers. And admittedly such warnings may not have penetrated the thinking of SBF who claimed "I would never read a book. I'm very skeptical of books." Nevertheless, the insertion of such "fudge factors" within the guidelines provides the agency that effective altruists would need to make their own moral decisions around actions that might be bad in themselves but good in their side effects: actions akin to spanking a child to discourage cruel behaviour, or telling a lie to protect an innocent life. Such trade-offs exist in all walks of life, and credible, moral decisions regarding them aren't necessarily categorical.

Criticism 2: EA excludes human emotion or relationship as guides to philanthropic choice, such that it undercuts philanthropists' agency and overlooks or opposes key aspects of human motivation.

"Many EA folks come from tech; many also consider themselves 'rationalists,' interested in applying Bayesian reasoning to every possible situation. EA has a culture, and that culture is nerdy, earnest, and moral. It is also, at least in my many dealings with EA folks, overly intellectual, performative, even onanistic." Lowrey November 2022

* * *

"What the 'effective altruism' types believe in is that they can replace the inferior, subjective standards of the plebs with the superior, objective standards of the ruling class.... Armed with these tools, … [they] feel empowered to do an unhinged collection of immoral things because, frankly, they are saving the world." Skiviers November 2022

* * *

This second line of criticism against the philosophical foundations of EA focuses on their undercutting donors' agency by discouraging them from choosing philanthropic causes freely in response to their own unfiltered emotions, interests or relationships. Instead, EA uses impartial and impersonal criteria to pre-select causes and interventions that are cost effective in saving or improving lives and then asks donors to choose from these. As explained by Singer: most charitable donations are "given on the basis of emotional responses to images of the people, animals, or forests that the charity is helping. Effective altruism seeks to change that by providing incentives for charities to demonstrate their effectiveness." Or as put more bluntly by the effective altruist Eliezer Yudkowsky: "This isn't about your feelings. A human life, with all its joys and all its pains, adding up over the course of decades, is worth far more than your brain's feelings of comfort or discomfort with a plan. Does computing the expected utility feel too coldblooded for your taste? Well, that feeling isn't even a feather in the scales, when a life is at stake. Just shut up and multiply." Indeed, SBF endorsed such reasoning even in choosing among the cost effective causes pre-selected by EA – eschewing those he considered "more emotionally driven" such as global poverty and health that threaten millions of lives at present, preferring instead those he considered more intellectually driven such as runaway artificial superintelligence that could conceivably exterminate trillions in the distant future.

EA's use of impartial and impersonal criteria to pre-select causes has been criticized for both what it overlooks in the world and denies in the individual. In terms of what it overlooks, the criteria used by EA focus on concepts "of individual needs and welfare, rather than power, inequality, injustice, exploitation, and oppression". By omitting the latter set of concepts, EA gives short shrift to conditions that are inherently important to our quality of life.

In terms of what it denies, EA's reliance on utilitarianism and impartiality requires individuals to forgo "the things that constitute us as humans: our personal attachments, loyalties and identifications" along with "the complex structure of commitments, affinities and understandings that comprise social life." Moreover, imposing a "point-of-viewless" impartiality "deprives us of the resources we need to recognise what matters morally." The social world is "irreducibly," "irretrievably" and "ineluctably" normative such that acting morally does not require "acting with an eye to others' well-being" but rather acting with a "just sensitivity to the worldly circumstances in question." As a result, EA's "image of the moral enterprise is bankrupt and … [the] moral assessments grounded in this image lack authority." Such concerns echo those of the philosopher Bernard Williams who argued in 1973 that the impartiality prescribed by utilitarianism is neither possible nor desirable: it's not possible given that individuals cannot step outside their own skin; and it's not desirable if, like Williams, one assumes that our individual well-being depends upon our ability to decide and act freely in accord with our own concerns, purposes or deepest convictions and not become a conduit for the initiatives or claims of others – including the claim that we should replace our own convictions with the "impartial point of view" needed to maximize total utility.

There are rejoinders to the criticisms of what EA overlooks and denies. First, when it comes to overlooking justice or equality or freedom, Singer admits that effective altruists "tend to view values … [like these] not as good in themselves but good because of the positive effects they have on social welfare." And yet, within EA there's no "party line" on that front. Indeed, given EA's commitment to cause neutrality and means neutrality, MacAskill claims in principle that if it can be demonstrated that advancing such values directly is a "course of action that will do the most good … then it's the best course of action by effective altruism's lights." That said, putting this principle into practice is difficult: it would require agreement at the outset on what justice or equality or freedom entails, for whom, and how it and its effects can be measured. To date, such difficulties have limited EA initiatives to ones that advance equality or justice indirectly: say, countering inequality by alleviating the effects of poverty; or addressing injustice by promoting election reform, criminal justice reform or international labour mobility. Second, directly pursuing justice or equality or freedom internationally could introduce forms of cultural domination and colonization by imposing Western concepts on non-Western societies and exercising philanthropic spending power that could silence or subvert local priorities as well as challenge the sovereignty of host-nations.

Third – turning to the denial of donor agency and the suppression of emotion – MacAskill argues that EA seeks to harness such things, not eliminate them. Some effective altruists may choose to adopt an impartial perspective if given evidence that this would allow their philanthropy to do more good for more people. Moreover, as proposed by economist Tyler Cowen, "an inescapable feature of human psychology means at the normative level, there's just no way we can fully avoid partiality of some kind." In order to recognize and respond to a cause or need, we need first to identify with it and with the people or entities involved. The direction and degree of such identification differ across donors, and to honour these differences EA presents a menu of alternative cause areas and interventions deemed cost effective. And fourth, the critics of EA who echo Williams' insistence that morality is essentially first-personal rather than impersonal risk undermining the responsive regard for others that is the very basis for, and indeed the original meaning of, philanthropy: according to philosopher Jeff McMahan, "the importance to oneself of one's own projects and attachments limits the extent to which morality can demand that one provide assistance to others."

What can we take from the downfall of Samuel Bankman-Fried with regard to the philosophical foundations of Effective Altruism?

Is SBF the exception that proves the general rule that the philosophical foundations of EA are sound? Or is he the example that demonstrates they aren't? Or is he neither? How did the utilitarianism he professed as a student in 2012 apply in his professional life a decade later? Did he use it as an ethical theory to guide and justify his actions, or as a smoke screen to obscure them? Did he consider his own ethical protestations sincere, whereas those of his competitors a marketing ploy? Or was he just like the others? Such questions weren't answered definitively in his infamous Twitter exchange with journalist Kelsey Piper, soon after he came under investigation in November 2022:

Piper: *"So the ethics stuff - mostly a front? People will like you if you win and hate you if you lose and that's how it all really works?"*

SBF: *"Yeah. I mean that's not *all* of it. But it's a lot…."*

Piper: *"You were really good at talking about ethics for someone who kind of saw it all as a game with winners and losers."*

SBF: *"Ya. Hehe. I had to be. It's what reputations are made of, to some extent. I feel bad for those who get fucked by it. By this dumb game we woke westerners play where we say all the right shiboleths (sic) and so everyone likes us."*

By Ord's account: "I don't think anyone fully understands what motivated Sam (or anyone else who was involved). I don't know how much of it was greed, vanity, pride, shame, or genuinely trying to do good…. [If he remained a utilitarian, then] it increasingly seems he was that most dangerous of things – a naive utilitarian – making the kind of mistakes that philosophers (including the leading utilitarians) have warned of for centuries…. [T]he sophistications that he thought were just a sop to conventional values were actually essential parts of the only consistent form of the theory he said he endorsed."

To my mind, it's unclear what role the philosophical foundations of EA played in the professional decisions of SBF. Hence, to judge those foundations by those decisions would be misleading. Nevertheless, his downfall revived two lines of criticism that raise issues and questions relevant to not only EA but also the philanthropic sector as a whole. I select three.

i. What are or what should be our ethical anchors?

As noted above, EA has been criticized for its ties to utilitarianism and the premise that actions are moral to the extent their consequences promote total well-being.

But what gives meaning or moral worth to our engagement with the philanthropic sector – whether as donors, volunteers, workers, advisors, collaborators or beneficiaries? Has it to do with the outcomes of our actions and whether they're good, or the duties and rules fulfilled by our actions and whether they're right, or the personal qualities underlying our actions and whether they're virtuous? How do we assess, perhaps question and possibly improve that goodness, rightness or virtue? Are there limitations or dangers in the standards we use? How do we work with others or in contexts that value standards different from or contradictory to our own? To what extent can we temper or change our own standards without losing our way?

If these questions seem irrelevant to how and why you engage with the philanthropic sector, why is that? Would you feel challenged by someone who sees them as fundamentally important?

ii. How do we decide upon actions that on the one hand could be harmful or problematic in themselves, but on the other hand could allow us to do more and better things?

As noted above, EA has been criticized for tolerating actions that might be intrinsically bad but instrumentally good: say, accepting donations from crypto, or recommending – albeit with cautionary guidelines – that some effective altruists pursue high-paying but perhaps corrupting or socially-harmful careers that would nevertheless enable them to donate more.

But how do or should we manage similar situations? For example, when and why should a charity refuse or return a donation? Or when and why should a charity refuse or terminate a partnership with a for-profit corporation? Should we share the outlook associated with William Booth, who co-founded the Salvation Army in 1865, that "the trouble with tainted money is t'aint enough of it"? If not, then where do we draw the line? By what criteria does "tainted" become "unacceptable" – apart from being criminal? What sources of donations would violate your own values, or either oppose the mission of a charity you deal with or trigger irreparable reputational harm in the eyes of the public or key stakeholders: tobacco, alcohol, cannabis, extractive industries, nuclear power, social media, airlines, crypto, the pharmaceutical industry, a religious foundation? Would the size or purpose of the donation make a difference to your decision?

Consider the following timeline for SBF. By 2013 he had affiliated with EA. In 2014 he took up the earning-to-give strategy, working at Jane Street Capital and donating half his salary. He started to build his crypto empire in 2017. Although crypto may be of disputed social value, it's not illegal. And although Bankman-Fried's promotional strategies may have been questionable (e.g., placing ads during the Super Bowl or in *The New Yorker* and *Vogue* magazines), they're not unprecedented. Sure enough, starting in 2018 EA leaders received personal reports that he was duplicitous, refused to implement standard business practices, and had inappropriate sexual relations with subordinates. But these reports weren't circulating publicly, didn't allege any criminal activity and could simply have been rumours spread by disgruntled associates. Few if any foresaw the devastating events of November 2022. Certainly investors like the Ontario Teachers Pension Plan didn't see them coming.

At what point during that timeline, would you or a charity you deal with have refused or returned, say, a $1 million donation from SBF?

iii. What ways should or should not be used to influence donors' decisions on how much and where to give?

As noted above, EA has been criticized for constraining the agency of donors in deciding the amounts and destinations of their giving. It recommends 10% of one's income, discourages acting on personal relationships and emotive appeal, and encourages a reliance on impersonal indicators of cost-effectiveness. As a result, some claim it both denies individuals the ability to decide and act on their own concerns, purposes or deepest convictions, and it overlooks normative but hard-to-pin-down goals such as liberty or justice.

But if the charge against EA is that it tries to sway donors – in other words, alter their conception of their own interests in ways that would have them act in a contrary manner – then could the same charge by leveled against other if not all fundraisers or fundraising campaigns in the sense of their doing the same thing albeit on different terms? Such campaigns might employ communication and relationship-building techniques designed to persuade. Such techniques might work on emotive rather than cognitive grounds, providing only selective information and relying on narratives or verbal or visual images that evoke rather than document. They might adjust the goalposts of "impact" to match what can be evoked emotively, and encourage compliant donors to think of themselves as "generous" or "visionary" and their gifts as "transformative" or "inspired."

Could such campaigns be faulted for tampering with donor agency?

Perhaps you know of campaigns that have been truly "donor-centric" in the sense of not resorting to practices that could sway or nudge their prospects into acting against their interests or priorities. If so, then – as suggested by the taxonomy constructed by MacQuillin – could such campaigns be at the expense of important considerations apart from donor agency, including what EA emphasizes: the well-being of actual or potential beneficiaries? Consider, for example, the decision of Leona Helmsley to establish in her will a $12 million trust fund for her Maltese dog, Trouble. Or consider the reassurance offered by Bronfman and Solomon that "[i]n philanthropy, there are no wrong answers…. You might want to fund an antigravity machine or a museum for dust mites. There may be more constructive uses for your money, and these objectives may sound crazy, but there is nothing wrong with them. In philanthropy, the choices are not between right and wrong, but between right and right."

Part 4: Questioning the analytical methods of Effective Altruism

Criticism 3: By relying on impartial reason to identify the philanthropic interventions that will do the most good, EA idealizes a methodology that quantifies and compares the value and probabilities of alternative and highly-speculative outcomes – thereby mistaking mathematical precision for truth and ignoring important qualities of human life and flourishing that are not readily quantified.

"Effective altruism, perhaps because it comes out of the hothouse of the Oxford philosophy department, is a bit too taken with thought experiments and toy models of the future. Bankman-Fried was of that ilk, famously saying that he would repeatedly play a double-or-nothing game with the earth's entire population at 51-to-49 odds." Klein Dec 2022

* * *

"The question of how to do good cannot be divorced from questions of what is just and where does power reside. This is a matter of morality: people concerned with doing good should be thinking about themselves not just as individual investors but as citizen-participants of systems that distribute suffering in the world unequally for reasons that are not natural but largely man-made." Aleem December 2022

* * *

As described in part 2, the analytical methods of EA rely on frameworks that distinguish and rank alternative causes and interventions. Both frameworks emphasize the quantification of outcomes and their probabilities. The first line of criticism against the analytical methods focuses on how the emphasis on quantification introduces types of methodological blindness that could either sideline certain matters relevant to well-being, accommodate subjectivity particularly in risk assessment or downplay the uncertainties and debates around "doing the most good" – the stated objective of EA.

Using common and quantitative units of account to compare the cost effectiveness of alternative causes and interventions automatically favours projects where data can be collected and causality tested: hence, matters of health in controlled environments get attention, whereas matters such as justice or self-determination are overlooked, as noted in part 2. Even for matters of health, preliminary studies based on, say, randomized controlled trials, provide imperfect guidance. Their results are specific to the scale and context of the trials and don't readily generalize and transfer to other contexts. Moreover, the results don't capture the experiments' long-term effects that could counteract any positive ones observed early on.

More generally, "[t]rying to put numbers on everything causes information loss and triggers anchoring and certainty biases…. Thinking in numbers, especially when those numbers are subjective 'rough estimates', allows one to justify anything comparatively easily, and can lead to wasteful and immoral decisions." Expected value calculations are particularly prone to this by accommodating personal levels of risk tolerance as well as value judgements about outcomes and their probabilities. Hence, they could give cover for reckless but pet decisions if upsides are emphasized and downsides disregarded – something all the more likely in the hands of someone like SBF who welcomed risk:

> "[T]he way I saw it was like, 'Let's maximize EV: whatever is the highest net expected value thing is what we should do'…. I think there are really compelling reasons to think that the 'optimal strategy' to follow is one that probably fails – but if it doesn't fail, it's great. But as a community, what that would imply is this weird thing where you almost celebrate cases where someone completely craps out – where things end up nowhere close to what they could have been – because that's what the majority of well-played strategies should end with."

Indeed, leaders of the EA community could claim similar cover for their decision to tie their fortunes and reputation to SBF in the first place: someone known as "an aggressive businessman in a lawless industry".

The focus on quantification contributes to a methodology susceptible to not only narrow and reckless decisions, but also decisions that are misdirected or conflictual because of the confusion around what "doing the most good" actually entails. On that score, EA has boxed itself into a corner. If it remains exacting in how to define and measure "the most good," then it increases the chances of repelling most donors and simply being wrong. Alternatively, if it offers greater latitude – say, encouraging donors "to be more effective when we try to help others" or to "maximize the good you want to see in the world" – then it becomes vapid. "I mean, who, precisely, doesn't want to do good? Who can say no to identifying cost-effective charities? And with this general agreeableness comes a toothlessness, transforming effective altruism into merely a successful means by which to tithe secular rich people…."

Even within the EA community there are thought leaders – Holden Karnofsky being one, Toby Ord another – who have reservations about ethical theories and analytical techniques that downplay the uncertainties and disputes around "the most good." As Karnofsky explains: "EA is about maximizing how much good we do. What does that mean? None of us really knows. EA is about maximizing a property of the world that we're conceptually confused about, can't reliably define or measure, and have massive disagreements about even within EA. By default, …. I think it's a bad idea to embrace the core ideas of EA without limits or reservations; we as EAs need to constantly inject pluralism and moderation." As Ord adds: "[E]ven if you were dead certain … it would be a problem if you are trying to work together in a community with other people who also want to do good, but have different conceptions of what that means – it is more cooperative and more robust to not go all the way."

Rejoinders to criticism #3

There are rejoinders to the criticisms of what the analytical methods of EA sideline, accommodate or downplay. With respect to their sidelining broader conditions like justice or freedom, as noted in part 3 EA supports such things indirectly where there are ties to measurable indicators: say, countering inequality by alleviating the effects of poverty; or promoting justice through criminal justice reform. That said, its methods steer clear of initiatives that wouldn't improve well-being on terms and at levels greater than the alternatives at hand. This is a strength, not a weakness. Although not perfect, EA's methods allow one to "sift through the detritus and decide what moral quandaries deserve our attention. Its answers won't always be right, and they will always be contestable. But even asking the questions EA asks – How many people does this affect? Is it at least millions if not billions? Is this a life-or-death matter? A wealth or destitution matter? How far can a dollar actually go in solving this problem? – is to take many steps beyond where most of our moral discourse goes." By promoting that discourse, EA provides a service to the philanthropic sector by "forcing us all to rethink what philanthropy should be."

For donors, by what standards do you gauge the extent to which your contribution improves the lives of others? Are these "presentable, articulable, reproducible"? For charitable organizations, what if anything makes you more deserving of donations than other organizations? How can that be demonstrated apart from story telling, image promoting and heart-string tugging? To be sure, in recent years many in the charitable sector have pushed toward greater consistency in measuring impact. But this is usually only within a cause or at an organizational level. EA insists that "people think about how we decide on the causes themselves…. That type of thinking about charitable giving is becoming more public, and that's something an effective altruist can take some credit for." But taking credit doesn't necessarily mean receiving thanks. Indeed, some of the criticism toward EA's analytical methods may simply come from donors or entities put on the defensive: say, "donors that respond to causes that move them, regardless of their cost-effectiveness" or "activists committed to the cause of social justice" who feel offended by being asked or expected "to demonstrate that their work is effective." Those on the defensive may also include the causes and organizations that EA leaders have used as examples of ineffectiveness or excess: cultural and arts organizations, Make-A-Wish Foundation, guide dogs, well-endowed universities or emergency relief for widely-reported disasters.

With respect to tolerating the value judgements that could skew decisions, EA is at least relatively transparent in what lies behind its decisions, thereby allowing others to question or challenge them and their associated risks. At the end of the day, however, judgements – ideally, defensible ones – need to be made. Sure enough, most effective altruists would agree there are instances where it's worth risking failure and perhaps ending up with nothing. But in taking that position, they're being consistent with recommendations made to the philanthropic sector as a whole by those who see greater risk-taking as necessary if the sector is to learn and make greater change, whether in Canada, the UK, the US or elsewhere. As for SBF's bravado over extreme risk-taking, it would be wrong and unfair to attribute such recklessness to EA more broadly: "if SBF went to [William] MacAskill, or any of his largesse's other beneficiaries, and asked, 'Do you think I should make incredibly risky financial bets over and over again until I'm liquidated or become a trillionaire?,' they would have said, 'No, please do not bankrupt our institutions.'"

And finally, with respect to downplaying the uncertainties and debates around "the most good," in fact EA recognizes and responds to such things. Admittedly, the focus remains on the needs of the beneficiaries and the cost effectiveness of alternative interventions to address them. But within those confines, the original EA organization Giving What We Can offers a range of funds that allows donors to support the high-impact causes and organizations that best correspond to their individual views on what constitutes the most good – whether these involve, for example, improving human well-being or animal welfare, alleviating climate change and its effects or averting catastrophic global risks in the future. And the EA foundation Open Philanthropy applies what it calls "worldview diversification": where "worldview" refers to "a set of highly debatable (and perhaps impossible to evaluate) beliefs that favor a certain kind of giving" or cause area; and "diversification" means "putting significant resources behind *each* worldview that we find highly plausible". Parentheses and italics are original. In other words, the foundation deliberately puts its eggs in multiple baskets – both to avoid rapidly diminishing returns from supporting only one or a few causes and to avoid estranging segments of the EA community that favour different worldviews.

Criticism #3 as it applies to longtermism

Critics see extreme forms of methodological blindness affecting if not motivating EA's pursuit of so-called longtermist causes and interventions – ones that, as described in part 2, seek to reduce the "existential risk" or "x-risk" posed by events or developments that "would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." Such threats include nuclear war and climate change, but now feature pandemics whether natural or bio-engineered as well as malicious artificial superintelligence (ASI) – a not-yet-realized state of AI that exceeds on all fronts the level of intelligence of which humans are capable.

For longtermist causes, the particular forms of methodological blindness emerge from the seemingly benign three-fold premise that: "Future people count. There could be a lot of them. And we can make their lives better." First, consider the implications of "future people count." Longtermists envisage people in the future as being both human and digital. In keeping with utilitarianism, they seek to increase the total well-being of future populations – a perspective known as the "total view". Toward that end, they favour not simply protecting those populations but increasing them as long as the average well-being per person, whether human or digital, doesn't fall so quickly as to reduce the total. For that reason, MacAskill argues that "[i]f future civilization will be good enough, then we should not merely try to avoid near term extinction. We should also hope that future civilization will be big…. The practical upshot of this is a moral case for space settlement." Thus, from a longtermist perspective, a population of 10 billion where each member flourishes with a high individual well-being of 100 is only half as well off as a population of 1000 billion where each member barely survives with a low individual well-being of 2. Heavily populated dystopias that are "good enough" are better than less populated utopias: a ranking known as the "repugnant conclusion" in population ethics, but accepted as a matter of logic by longtermists.

Now consider "there could be a lot of them." Nick Bostrom – introduced in part 2 as the founding and current Director of the Future of Humanity Institute – provides a range of estimates. For biological "neuronal wetware" humans, his projections for future life years range from 10^{16} if biological humans remain Earth-bound, up to 10^{34} if they colonize the "accessible universe". Alternatively, if one assumes that such colonization takes place and that future minds can be "implemented in computational hardware", then there could be at least 10^{52} additional life years that include digital ones where "human whole brain emulations … live rich and happy lives while interacting with one another in virtual environments". Bostrom, as well as Hilary Greaves and MacAskill, assume that such digital minds will have "at least comparable moral status [as] we may have."

Finally, consider "we can make their lives better" and the acceptable cost of doing this. Turning again to Bostrom and his projections, applying his lowest estimate of 10^{16} biological human life years left to come on Earth, he concludes that "the expected value of reducing existential risk by a mere *one millionth of one percentage point* is at least a hundred times the value of a million human lives" today. Alternatively, assigning "the more technologically comprehensive estimate of 10^{52} human brain-emulation subjective life-years" a mere 1% chance of being correct, he concludes that "the expected value of reducing existential risk by a mere *one billionth of one billionth of one percentage point* is worth a hundred billion times as much as a billion human lives" today. Italics are original.

Thus by assuming future populations will be and should be massive, widening the notion of what constitutes a person in the future, claiming people who might exist in the future should be counted equally to people who definitely exist at present, and believing there's no fundamental moral difference between saving actual people today and bringing new people into existence – longtermists argue that the loss of present lives is an acceptable cost of increasing by even a miniscule amount the probability of protecting future lives.

Greaves and MacAskill provide a specific example of this trade-off. Working with a projection of only 10^{14} future life years at stake and acknowledging that "[t]here is no hard quantitative evidence to guide cost-effectiveness estimates for AI safety", they nevertheless propose that "$1 billion of carefully targeted spending [on ASI safety] would suffice to avoid catastrophic outcomes in (at the very least) 1% of the scenarios where they would otherwise occur…. That would mean that every $100 spent had, on average, an impact as valuable as saving one trillion lives … far more than the near-future benefits of bed net distribution" that would prevent deaths from malaria. Parentheses are original. As they see it, such calculations "make it better in expectation … to fund AI safety rather than developing world poverty reduction." Or as put by Bostrom, because "increasing existential safety achieves expected good on a scale many orders of magnitude greater than that of alternative contributions, we would do well to focus on this most efficient philanthropy" rather than "fritter it away on a plethora of feel-good projects of suboptimal efficacy" such as providing bed nets.

Such implications – and the assumptions and methods used to support them – have attracted criticism from both within the EA community and outside it. For example, how and to what extent future people should count are topics of debate in population ethics. There's no agreement on whether digital "brain emulations" would or could be morally comparable to human or other biological forms of sentient life – and hence whether their numbers or well-being should be included in population projections. And although longtermists endorse the "total view", others don't. Opposing it, for example, are those who endorse the principle of "neutrality" made popular by Canadian philosopher Jan Narveson (who also coined the phrase "total view"). According to that principle: "We are in favour of making people happy, but neutral about making happy people." Neutrality can be linked with the principle of "procreation asymmetry" whereby there's no moral imperative to bring into existence people with lives worth living (i.e., neutrality), but there's moral imperative not to bring into existence people with lives not worth living. To be sure, these principles and their implications are themselves topics of debate. But together, they provide a credible philosophical rationale for concluding that "the longtermist's mathematics rest on a mistake: extra lives don't make the world a better place, all by themselves…. We should care about making the lives of those who will exist better, or about the fate of those who will be worse off, not about increasing the number of [sufficiently] good lives there will be."

Such a non-longtermist conclusion supports those who tie human survival not to burgeoning populations somehow maintained by extraordinary levels of economic growth, but rather to making things work sustainably on Earth. According to the 2020 open letter signed by 11,258 scientists from 153 countries, this would require that the world's population be "stabilized – and, ideally, gradually reduced" and that public policies "shift from GDP growth and the pursuit of affluence toward sustaining ecosystems and improving human well-being by prioritizing basic needs and reducing inequality."

When it comes to the numbers of people that could exist in the future and our capacity to make their lives better, the tactic longtermists use to justify their position "is always the same: let's run the numbers. And if there aren't any numbers, let's invent some." Bostrom's projections of many trillions over the next billion years rest on assumptions about space settlement, extraterrestrial energy sources and digital storage capacity that he gathers from a range of literatures, including science fiction. These assumptions are questionable. Moreover, the time horizon is itself questionable – given the imminent threats posed by nuclear war and climate change. As put by one commentator, himself an effective altruist, "once you think like the world as we know it has a likely time horizon shorter than one thousand years, this notion of, well, what we will do in thirty thousand years … just doesn't seem very likely. The chance of it is not zero. But the whole problem …starts looking … less like a probability that should actually influence … [our] decision making."

The ways by which longtermists plan to make future lives better are themselves dubious. Part of the problem lies with our "cluelessness" both about what the distant future will be like and about what differences we could possibly make to that future by our actions today. "In truth, we cannot know what will happen 100 years into the future [let alone 30,000 years] and what would be the impact of any particular technology. Even if our actions will have drastic consequences for future generations, the dependence of the impact on our choices is likely to be chaotic and unpredictable. To put things in perspective, many of the risks we are worried about today, including nuclear war, climate change, and AI safety, only emerged in the last century or decades."

Even when it comes to mitigating specific risks in the nearer future, it's not clear what are the best actions in terms of their feasibility and effectiveness, let alone how the philanthropy of EA can uniquely advance those actions. Consider the threat of nuclear war. As reasoned by the commentator mentioned above: "I'm very keen on everyone doing more work in that area, but I don't think the answers are very legible … [and] I don't think effective altruism really has anything in particular to add to those debates." With preventing malicious ASI, "I don't think it's going to work. I think the problem is if AI is powerful enough to destroy everything, it's the least safe system you have to worry about. So you might succeed with AI alignment on say 97 percent of cases, … [b]ut if you failed on 3 percent of cases, that 3 percent of very evil, effective enough AIs can still reproduce and take things over…." Moreover, "artificial intelligence issues from a national perspective, not a global perspective. So I think if you could wave a magic wand and stop all the progress of artificial intelligence, …. you never have the magic wand at the global level."

Moreover, the willingness of longtermists to sacrifice large payoffs with probabilities arbitrarily close to one in order to pursue enormously larger payoffs with probabilities arbitrarily close to zero amounts to "fanaticism" – a willingness that could possibly be justified using expected value calculations for which the numbers were reliable. But the numbers aren't: they give "a false sense of statistical precision by slapping probability values on beliefs. But those probability values are literally just made up. Maybe giving $1,000 to the Machine Intelligence Research Institute will reduce the probability of AI killing us all by 0.00000000000000001. Or maybe it'll make it only cut the odds by

0.00000000000000000000000000000000000000000000000000000000000000001. If the latter's true, it's not a smart donation; if you multiply the odds by 10^{52}, you've saved an expected 0.0000000000001 lives, which is pretty miserable. But if the former's true, it's a brilliant donation, and you've saved an expected 100,000,000,000,000,000,000,000,000,000,000,000 lives."

Trying to reason with such small probabilities is itself nonsensical. "Physicists know that there is no point in writing a measurement up to 3 significant digits if your measurement device has only one-digit accuracy. Our ability to reason about events that are decades or more into the future is severely limited…. To the extent we can quantify existential risks in the far future, we can only say something like 'extremely likely,' 'possible,' or 'can't be ruled out.' Assigning numbers to such qualitative assessments is an exercise in futility…. [I]f you are genuinely worried about long-term risk, I suggest you spend most of your time in the present. Try to think of short-term problems whose solutions can be verified, which might advance the long-term goal…. [T]o make actual progress on solving existential risk, the topic needs to move from philosophy books and blog discussions into empirical experiments and concrete measures."

Rejoinders to criticism #3 as it applies to longtermism

To a large extent, the replies to such criticisms are calls not to toss out the baby with the more extreme implications of the longtermist bathwater. The premise that future people count rests on the basic message that "the long-term future matters more than we give currently give it credit," whether in our philanthropy or public policy. That message isn't unique to EA. It ties in with the Indigenous practice of seventh generation thinking whereby one assesses current decisions on how they will affect persons born seven generations from now. And it underlies the 2024 UN Summit of the Future.

Moreover, one doesn't need to "rely on the moral math [of longtermists] in order to think that human extinction is bad or that we are at a pivotal time in which technologies if left unregulated or unchanged could destroy us…." In terms of the technologies that pose such threats, EA was a harbinger of protecting humanity from natural or bio-engineered pandemics or from malevolent ASI at a time when, apart from experts in public health and computer science, such things were seen primarily as the stuff of movies (e.g.: *2001: A Space Odyssey* 1968; *The Andromeda Strain*, 1971; *Virus: The End*, 1980; *The Terminator*, 1984) or fixations alluded to by critics as evidence of EA hand wringing. However, given the global experience of COVID-19 and the recent attentions and efforts in the US, at the United Nations and among the G7 about keeping AI safe, the concerns publicly raised by EA now seem more prescient than far-fetched.

In terms of deciding between the use of philanthropic resources to address tangible needs in the present as opposed to hypothesized needs in the future: such decisions and their dilemmas are not unique to EA. They operate, albeit with different intensities, across the philanthropic sector – going back to at least the 18th century as noted in part 1, and continuing now, for example, in debates around so-called strategic philanthropy. Indeed, versions of these debates exist within the EA community and among its longtermists.

Criticism 4: This methodology – bolstered by its ethical assumptions and claims of impartiality – cultivates hubris, condescension toward and dismissal of contending priorities or sources of information, and the impulse to define and control philanthropic interventions on one's own terms.

"Any honest reckoning over effective altruism now will need to recognize that the movement has been overconfident. The right antidote to that is not more math or more fancy philosophy. It's deeper intellectual humility." Samuel November 2022

* * *

"… there is still a strong element of elitist hubris, and technocratic fervor, in [EA's] universalistic and cocksure pronouncements…. [T]hey could benefit from integrating much more systemic humility, uncertainty, and democratic participation into their models of the world." Lehto January 2023

* * *

"This is a movement that encourages quant-focused intellectual snobbery and a distaste for people who are skeptical of suspending moral intuition and considerations of the real world…. This is a movement whose adherents … view rich people who individually donate money to the right portfolio of places as the saviors of the world. It's almost like a professional-managerial class interpretation of Batman…." Aleem December 2022

* * *

"Getting some of the world's richest white guys to care about the global poor? Fantastic. Convincing those same guys that they know best how to care for all of humanity? Lord help us." Lowrey November 2022

* * *

This second line of criticism against the analytical methods of EA focuses on the overconfidence they cultivate and how that can reinforce the methodological blindness outlined above. Such hubris takes on multiple but interconnected forms, all of which contribute to a strong if informal hierarchy within EA organizations and the community in general.

The hubris can be intellectual. In part, this draws from the academic backgrounds of effective altruists of whom "11.8% have attended top 10 universities, 18.4% have attended top 25 ranked universities and 38% have attended top 100 ranked universities globally." And in part, it stems from their "quantitative culture" and "[o]verly-numerical thinking [that] lends itself to homogeneity and hierarchy. This encourages undue deference and opaque/unaccountable power structures. EAs assume they are smarter/more rational than non-EAs, which allows … [them] to dismiss opposing views from outsiders even when they know far more than … [EAs] do. This generates more homogeneity, hierarchy, and insularity" from the broader academic and practitioner communities. Such insularity involves "prioritising non-peer-reviewed publications by prominent EAs with little to no relevant expertise…. [T]hese works commonly don't engage with major areas of scholarship on the topics that they focus on, ignore work attempting to answer similar questions, nor consult with relevant experts, and in many instances use methods and/or come to conclusions that would be considered fringe within the relevant fields."

As a consequence, EA risks becoming "a closed validation loop" that perpetuates an "EA orthodoxy" – one that privileges "utilitarianism, Rationalist-derived epistemics, liberal-technocratic philanthropy, Whig historiography, the ITN framework [see part 2], and the Techno-Utopian Approach to existential risk. Moreover, "contradicting orthodox positions outright gets … [one] labelled as a 'non-value-aligned' individual with 'poor epistemics,' so … [one needs] to pretend to be extremely deferential and/or stupid and ask questions in such a way that critiques are raised without actually being stated."

Such intellectual hubris reinforces forms of donor hubris. The rhetoric used by EA leaders encourages those who support EA causes to think of themselves as "the hero, the savvy consumer, and the virtuous self-improver." Such self-congratulatory images lead EA donors to respect and relate to each other, but not connect with or consult the objects of their philanthropy – particularly those affected by extreme poverty and the groups representing them. There's little "effort to put the EA community in contact with activists, civil society groups, or NGOs based in poor countries," thereby cutting off the community from the insights and resources of those with lived experience, and curtailing the types of consultation and power sharing with grantees that could foster local buy-in and increase the chances of change lasting beyond the immediate philanthropic interventions. By not engaging with and applying grassroots knowledge and forgoing the strategies taken, for example, by Solidaire, Thousand Currents or WIEGO – EA denies itself a possible means of doing more of "the most good."

And finally, the hubris and hierarchy of EA takes on managerial and governance forms. As a movement, EA "is deeply immature and myopic, … and … desperately needs to grow up. That means emulating the kinds of practices that more mature philanthropic institutions and movements have used for centuries, and becoming much more risk averse. EA needs much stronger guardrails to prevent another figure like Bankman-Fried from emerging…." Indeed, the readiness of EA to align itself with SBF demonstrates the need for EA decision making to "be more decentralized," and points to a "lack of effective governance" that currently looks "so top-down and so gullible." As it stands, "[o]ne has to wonder why so many people missed the warning signs" – particularly when, as noted in part 3, those signs came as explicit warnings sent by multiple parties to MacAskill and other "[l]eaders of the Effective Altruism movement … beginning in 2018 that Sam Bankman-Fried was unethical, duplicitous, and negligent in his role as CEO…." The decisions made by EA leaders to affiliate so closely with SBF underscore the need for structural reforms within EA organizations along the lines drafted in early 2022 by Zoe Carla Cremer (see part 2) or drafted in early 2023 by the pseudonymous Concerned EAs.

In part, EA's managerial immaturity can be attributed to its rapid transition in little over a decade from comprising a few student-founded and student-run organizations that relied on the camaraderie and confidence of a small homogeneous group to becoming a range of diverse and well-funded philanthropic organizations that still have many of the original students at the helm (see part 2). Given that transition, EA as a movement could be subject to the limiting or destructive effects of "founder's syndrome" – a condition identified in both for-profit and nonprofit organizations exhibiting traits that have been observed by members of the EA community.

Rejoinders to criticism #4

There are responses to the allegations concerning the intellectual, donor and governance hubris of EA. With respect to the intellectual forms – the ability of EA to attract smart and talented young people is a strength and to its credit, not a flaw or weakness. The criticisms of intellectual hubris could have equally been directed to activist organizations which "typically present themselves as more thorough-going, and more principled, champions of justice for the global poor that are their effective altruist opponents." Intellectual insularity and organizational orthodoxies affect many philanthropic or mission-based organizations and movements, whether on terms that are religious, political, ethnic or methodological. Such organizations can't be all things to all people: their stands or actions have to be consistent with their identity and purpose. That said, within each there's usually room for intellectual diversity. At least that's the case for EA: witness the debates on the EA online Forum. Besides, as noted by Karnofsky, "most EAs are reasonable, non-fanatical human beings, with a broad and mixed set of values like other human beings, who apply a broad sense of pluralism and moderation to much of what they do. My sense is that many EAs' *writings and statements* are much more one-dimensional and "maximizy" than their *actions.*" Italics are original.

With respect to forms of donor hubris – EA is not alone in cultivating these. "Many argue that the traditional models and approaches we have for philanthropy are ones which put too much emphasis on the donor's wishes and ability to choose, and give little or no recognition to the voices of recipients." More pointedly, "a lot of charitable giving is about the hubris of the donor, rather than the needs of the recipient." If anything, EA is an exception by focusing on the needs not of donors but of recipients and by addressing only those needs over which it has competence. For GiveWell, these relate to global health and nutrition that, once addressed, will "empower people to make locally-driven progress on other fronts." GiveWell selects across alternative interventions by assigning quantitative "moral weights" to their good outcomes, where those weights reflect the priorities reported in a 2019 survey of persons living in extreme poverty in Kenya and Ghana. The weights, for example, place a higher value on saving lives as opposed to reducing poverty, and on averting deaths of children under 5 years old as opposed to older ones. Although seeking to act on the preferences of those with lived experience, GiveWell stops short of "letting locals drive philanthropic projects", reasoning that local elites "who least need help will be best positioned to get involved with making the key decisions".

Finally, in terms of managerial hubris – the broad-brush criticisms, whether valid or not, are cast as if EA comprises one organization with a single founder, organizational chart or set of governance procedures. It doesn't. Instead it comprises multiple organizations – those under the auspices of Effective Ventures (e.g., the Centre for Effective Altruism, 80,000 Hours, and Giving What We Can as introduced in part 2, in addition to others such as the Centre for Governance of AI and the Forethought Foundation for Global Priorities Research), as well as a range of research organizations (e.g., the Future of Humanity Institute, Machine Intelligence Research Institute, Global Priorities Institute, GiveWell) and foundations (e.g., Open Philanthropy). Sure enough, certain individuals have longstanding and multiple ties with these organizations, and presumably have informal influence across several. Ord, for example, co-founded Giving What We Can in 2009, and is a research fellow at the Future of Humanity Institute a trustee of both 80,000 Hours and the Centre for Effective Altruism. MacAskill – described as "a co-founder the effective altruism movement" as well as its "prophet" – co-founded Giving What We Can in 2009, 80,000 Hours in 2011, the Centre for Effective Altruism in 2012 as well as the Global Priorities Institute and the Forethought Foundation in 2017 of which he is the Director.

But simply on the basis of these personal ties, it would be wrong to conclude that all of these diverse organizations exhibit the same management styles or governance structures, let alone that those styles and structures are somehow dysfunctional. Moreover, it would be wrong to infer such dysfunctionality from the decisions of MacAskill and others to affiliate with or tolerate SBF. As noted in part 3, his alleged malfeasance went unrecognized by many investors and associates. And any rumours of his bad behaviour were not evidence of criminal behaviour.

Criticism #4 as it applies to longtermism

Critics see acute forms of hubris in EA's formulation and defence of longtermist causes. They attribute the intellectual forms to the prevalence of academic philosophers among EA thought leaders (e.g., Bostrom, Greaves, MacAskill, Ord, Singer). Philosophy is known for its "tendency to slip from sense into seeming absurdity." No doubt "[t]hese are all smart people, but they are philosophers, which means their entire job is to test out theories and frameworks for understanding the world, and try to sort through what those theories and frameworks imply. There are professional incentives to defend surprising or counterintuitive positions, to poke at widely held pieties and components of 'common sense morality,' and to develop thought experiments that are memorable and powerful (and because of that, pretty weird)." Parentheses are original. In other words, "[t]he philosophy-based contrarian culture [of EA] means participants are incentivized to produce [what at least some would consider] 'fucking insane and bad' ideas…." The types of reasoning and rhetoric that set out to be provocative may be the stuff of creative seminar discussions or fun dorm-room debates. But in their raw form, they hold little credibility beyond the inner clique of academics and the autodidacts who want to be among them.

This intellectual hubris shaping and shaped by longtermism leads to multiple problems. First, it leads to inconsistencies if not misrepresentations in communication. In packaging their thinking and conclusions for a general audience, thought leaders deliberately tone down the "more fanatical versions" in order to widen the "appeal and credibility" of longtermism. Second, intellectual insularity becomes "especially egregious" when applied to a "domain of high complexity and deep uncertainty, dealing with poorly-defined low-probability high-impact phenomena, sometimes covering extremely long timescales, with a huge amount of disagreement among both experts and stakeholders along theoretical, empirical, and normative lines. Ask any risk analyst, disaster researcher, foresight practitioner, or policy strategist: this is … where you maintain epistemic humility and cover all your bases" by consulting and learning from research areas that EA typically ignores (e.g., studies in "Vulnerability and Resilience, Complex Adaptive Systems, Futures and Foresight, Decision-Making under Deep Uncertainty/Robust Decision-Making, Psychology and Neuroscience, Science and Technology Studies, and the Humanities and Social Sciences in general").

Third, the "philosophers' increasing attempts to apply these kinds of thought experiments to real life – aided and abetted by the sudden burst of billions into EA, due in large part to figures like Bankman-Fried – has eroded the boundary between this kind of philosophizing and real-world decision-making…. EA made the mistake of trying to turn philosophers into the actual legislators of the future." Rephrasing this problem more pointedly: "[t]ying the study of a topic that fundamentally affects the whole of humanity to a niche belief system championed mainly by an unrepresentative, powerful minority of the world is undemocratic and philosophically tenuous." Or more mildly: "it does seem convenient that a group of moral philosophers and computer scientists happened to conclude that the people most likely to safeguard humanity's future are moral philosophers and computer scientists." Perhaps convenient for them but fool-hardy for us if we rely on those philosophers and computer scientists to sort out the ways to safeguard the future in accord with humanity's preferences, let alone frame the national policies or international agreements purportedly capable of doing this.

With respect to donor hubris, critics see longtermism as catering to this at the expense of methodological rigour. "As much as the effective altruist community prides itself on evidence, reason and morality, there's more than a whiff of selective rigor here. The turn to longtermism appears to be a projection of a hubris common to those in tech and finance, based on an unwarranted confidence in its adherents' ability to predict the future and shape it to their liking." Hence, it's not a coincidence that "the areas EA focuses on most intensely (the long-term future and existential risk, and especially AI risk within that) align remarkably well with the sorts of things tech billionaires are most concerned about: longtermism is the closest thing to 'doing sci-fi in real life', existential catastrophes are one of the few ways in which wealthy people could come to harm, and AI is the threat most interesting to people who made their fortunes in computing." Parentheses are original.

Rejoinders to criticism #4 as it applies to longtermism

Not surprisingly, there are replies to the allegations of intellectual and donor hubris tied to longtermism. In terms of the intellectual forms, first note that "[i]t is appropriate for philosophers to speculate on hypothetical scenarios centuries into the future and wonder whether actions we take today could influence them." Second, "even longtermists don't wake up every morning thinking about how to reduce the chance that something terrible happens in the year 1,000,000 AD by 0.001%. Instead, many longtermists care about particular risks because they believe these risks are likely in the near-term future…." Indeed, MacAskill makes this point in arguing that the costs of protecting the future are "very small, or even nonexistent" since most of the things – disaster preparedness, climate-change mitigation, scientific research – we want to do for ourselves for the near future. Nevertheless, "[t]his does not mean that thinking and preparing for longer-term risks is pointless. Maintaining seed banks, monitoring asteroids, researching pathogens, designing vaccine platforms, and working toward nuclear disarmament, are all essential activities that society should take. Whenever a new technology emerges, artificial intelligence included, it is crucial to consider how it can be misused or lead to unintended consequences."

Third, as noted above, one should not judge longtermism by the extreme positions found within the rhetoric or reasoning of a "philosophy-based contrarian culture." Indeed, so called "weak longtermism" – the position that the "long-term future matters more than we're currently giving it credit for, and we should do more to help it", climate change being a case in point – may indeed be its strongest, most persuasive and powerful form. And fourth, by focusing on the bravado, some criticisms of longtermism verge on ad hominem attacks, and most overlook signs of humility. Consider, for example, MacAskill admitting that he doesn't know the answer to the question "How much should we in the present be willing to sacrifice for future generations?" Or his acknowledging that "[m]y No. 1 worry is: what if we're focussed on entirely the wrong things? What if we're just wrong? What if A.I. is just a distraction? … It's very, very easy to be totally mistaken."

With respect to donor hubris, billionaires from Silicon Valley – regardless of whether or where they practice philanthropy – aren't known for their modesty. Moreover, longtermist forms of donor hubris don't altogether dominate EA. Recent estimates of the funding going to "Global Health" are twice those going to "Biosecurity" and "Potential Risks of AI" combined. What is more, "[m]any 'longtermists' have given generously to improve people's lives worldwide, particularly in developing countries. For example, none of the top charities of GiveWell (an organization … in which many prominent longtermists are members) focus on hypothetical future risks. Instead, they all deal with current pressing issues, including malaria, childhood vaccinations, and extreme poverty. Overall, the effective altruism movement has done much to benefit currently living people." And it still does.

What can we take from the downfall of Samuel Bankman-Fried with regard to the analytical methods of Effective Altruism?

SBF expressed great confidence in financing selective interventions to reduce the existential risk posed by pandemics and ASI, and in ranking them solely on the basis of expected value calculations, regardless of the odds. And although he supported political campaigns, he did so to promote not institutional change but rather EA and an unregulated crypto industry. Are he and the adulation he received the exceptions that prove the general rule that the analytical methods of EA are sound and unencumbered by the quantification they require, the hubris they encourage or the deflection from systemic conditions they justify? Or is he the example that demonstrates those methods are defective on those terms? Or is he neither?

Regardless of how one answers such questions, the bankruptcy of FTX International and the criminal charges brought against SBF in late 2022 enlivened the existing criticisms of EA's analytical methods. Many of us affiliated with the philanthropic sector might see the criticisms as being relevant only to EA and its approach to philanthropy and having little to do with the sector as a whole. But is that the case? Here I select six areas in which the criticisms might have implications beyond EA.

i. What data would allow the sector, your organization or you as a donor to become better at recognizing societal needs and addressing them? Do we have the skills – or even the willingness to acquire the skills – needed to interpret and apply those data?

EA has been faulted for its "quantitative culture" and "overly-numerical thinking". But could such a culture and thinking empower the philanthropic and charitable sector by strengthening its abilities to recognize societal needs and address them more effectively? Sector leaders in Canada think so – placing among their top priorities the need to acquire more data for and about the sector, and calling for the sector to "grow up" in terms of investing in the technology, the talent, and the delivery and evaluative processes that would allow it to learn from and apply those data. Rather than disparage EA's quantitative methods and talents, could we learn from and make use of them?

ii. Can catering to the needs of donors – or, indeed, your own needs as a donor – impede the effectiveness of philanthropy? If so, then how can this be avoided or overcome?

Understandably, donors need to recognize their own priorities in the mission and accomplishments of the organizations to which they give. And undoubtedly, EA's track record in managing donor relations has been far from perfect. However, its starting point has been the needs of beneficiaries: first identifying the particular causes and interventions that would do the most good for them with a given amount of resources, and then recruiting the donors who wish to support this venture. As noted in part 2, those causes and interventions are ranked according to their cost effectiveness – not their donor appeal *per se*. Many of us work in or with charitable organizations with given missions and sets of beneficiaries. But are their ways either to manoeuvre within a mission or amend it that would increase the cost effectiveness of the work you do? If not, are their other organizations with missions and beneficiaries that would better match your priorities?

iii. What is the risk tolerance of your organization or for your charitable giving? Do you have the resources and opportunities to take greater risks – ones that that could open up ways to make greater change or at least learn how to do so? If not, then what other things could enable you to make your philanthropy more effective?

As noted above, the philanthropic and charitable sector has been faulted for being too staid and risk adverse, preferring safe but modest ventures that limit what the sector can achieve. How can we learn from the EA community about tolerating greater risk without falling into the recklessness evinced by SBF? What opportunities would open up if we moved in that direction? What would be the personal or organizational costs of doing so? How could those costs be overcome?

iv. Are all charitable purposes equal in their potential to create social benefit where it's most needed? If so, why is that the case? If not, then what changes could allow the sector and the donations it receives to become more beneficial?

As noted, EA uses quantitative methods to rank the cost effectiveness of not only alternative interventions within a cause, but also the causes themselves, regardless of where or when the beneficiaries are located – relegating those that are less cost effective to the category of luxury spending. A given donor might prefer giving to an opera company over a local food bank, a local food bank over the Ice Bucket Challenge, and the Ice Bucket Challenge over a fund encouraging childhood vaccination against malaria in sub-Saharan Africa. In Canada, only the latter doesn't constitute charitable giving because the organization Malaria Consortium isn't registered here. According to GiveWell, however, only the latter doesn't constitute luxury spending based on cost effectiveness. Some jurisdictions provide higher or lower tax credits or deductions according to those causes believed to generate higher or lower social benefit. Some, although sharing Canada's common law tradition (e.g., Australia, India, Singapore), deny any tax incentives for giving to places of worship. What's your take on this?

v. How can a young but maturing charitable or nonprofit organizations remain inspired by the vision and dedication of their founders but nevertheless be able to adapt, plan and decide in part by drawing upon the expertise and input of others whose views differ from or challenge those of the founders?

To endure and adapt, organizations need to become more than extensions of their founders' original vision and initiative. Such growth is not an easy or straightforward process either for the founders whose identity may tied to the work of the organizations, or for the organizations and their stakeholders that have become accustomed to simply trusting and following the personal decisions or priorities of the founders. Whether or not EA organizations are subject to so-called "founder's syndrome" – there is still a value in those organizations and others establishing decision-making procedures that are transparent and consultative, and avoiding founder or funder burnout. Have the organizations you have worked in or with been able to mature on those terms? If so, has anything been lost in the process? If not, what or who is impeding this?

vi. Do the needs of future generations explicitly and regularly fit into the mission and work of your organization or the priorities that guide the directions and amounts of your charitable giving? If not, then how do you justify not taking those needs into account – beyond claiming there are already too many needs in the present day?

How should the philanthropic sector divide its resources across the needs of current and future generations? How does or should this differ from the responsibilities of government?

Part 5: Questioning the ultimate effects of Effective Altruism

Criticism 5: Moreover, the methodology of EA – by focusing on separate, numerically-evaluated interventions – overlooks the wider behavioural, institutional or systemic conditions that might not only limit the effectiveness of the interventions but also cause or perpetuate the societal ills they seek to address.

"On some level, maybe it makes sense to ensure that your actions have the greatest possible positive impact – that your money is donated effectively to causes that improve people's lives to the greatest degree possible…. But it's not clear why this top-down, from-first-principles approach is the right one…." Ongweso November 2022

* * *

"…[I]n the hands of Bankman-Fried (commonly known as SBF), effective altruism was neither effective nor altruistic…. It's an outlook that breeds a bizarre blend of elitism, insularity and apathy to root causes of problems…. This crowd seems clueless about the reality that funding research into protecting against dangerous artificial intelligence will be impotent unless we structure our society and economy to prize public safety over capital's incentive to innovate for profit. If longtermists want to mitigate climate change, they should probably be radically reappraising an economic system that incentivizes short-sighted hyper-extractionism and perpetual growth." Aleem December 2022

* * *

The first line of criticism against the ultimate effects of EA points to its focus on separate stand-alone interventions instead of the systemic conditions that could generate or compound the problems the interventions seek to address. Critics attribute this focus to two things. The first is the bias built into the analytical methods described in part 3. This includes "observational bias" that requires collected or collectable data, "quantification bias" that expects the data and indicators to be numerical, and "instrumental bias" that privileges initiatives that are controllable. Such preconditions lead EA toward interventions that work within or apart from the social, economic and political conditions at hand.

In other words, because "EA's metrics are best suited to detect the short-term impact of particular actions, … its tendency to discount the impact of coordinated actions can be seen as reflecting 'measurability bias'…. [T]his bias is politically dangerous because it obscures the structural, political roots of global misery, thereby contributing to its reproduction by weakening existing political mechanisms for positive social change, rather than seek to reform them" – a point picked up by criticism 6. Moreover, those metrics divert attention from economic development being perhaps the most comprehensive anti-poverty strategy – a longer-term process that could be advanced more effectively by supporting not selective interventions in the fields of global health and nutrition, but rather watch-dog charities that promote strong institutions and human rights, or advocacy organizations that push for changes to the international trade deals and protectionist practices that harm low-income countries.

The second explanation for why EA overlooks systemic conditions is its assumption "that the individual, not the community, class or state, is the proper object of moral theorising. There are benefits to thinking this way. If everything comes down to the marginal individual, then our ethical ambitions can be safely circumscribed; the philosopher is freed from the burden of trying to understand the mess we're in, or of proposing an alternative vision of how things could be…. [EA doesn't] address the deep sources of global misery … or the forces that ensure its reproduction."

Put differently, "[w]hat matters for a good human life, in which basic needs are met and individuals have some autonomy, is that institutions and practices function to the advantage of every person, now and in the future. But most existing institutions are defective in this respect. They serve small elites, exploit the environment, and keep large numbers of people in poverty and inequality…." Hence, "the most dangerous underlying assumption … of effective altruists … [is that] they take the current institutional order as given, implicitly denying that it is open to change." The assumption that the institutional status quo is unchangeable leaves effective altruists with few alternatives beyond the strategy of supporting separate interventions that are small, brief and in the end largely ineffectual.

Rejoinders to criticism #5

As a response to the charge that EA overlooks the systemic conditions that generate human deprivation – the critics themselves often provide defensible rationales: "efforts to restructure the normative organisation of society, … far from obeying merely causal laws, are at home in the unpredictable realm of politics", or "institutions are very difficult to change, not only because human beings are creatures of habit, but also because there are powerful vested interests that want to keep the current order in place." If there's little evidence that EA can bring about institutional change, let alone change that would be beneficial, then EA will put its philanthropic resources elsewhere: a matter of triage imposed by resources being finite and more reliable interventions being available.

Moreover, the critics' depiction of the EA community as "monolithic or focusing only on making charitable donations" subject to an "individualistic bias" is a straw characterization. "[E]ffective altruists … view charitable donations as simply one way of maximizing the amount of good that anyone can do in the world, not as the full extent of our moral obligation to others or as a 'silver bullet' for the problems of humanity. Philosophers, activists, entrepreneurs and others in the movement are actively seeking changes to the institutions and systemic forces that constrain human potential….[S]ocial change is about improving life as much as we can, rather than furthering any specific ideology or strategy. Moreover, not all social problems can or should be solved through collective action either (although many, of course, might be)…. [O]penness to different methods of pursuing social change … [is] crucial …. Unlike many other social movements, … [EA] aspires to be 'cause-neutral,' identifying what to work on according to how we can have the greatest possible impact rather than what we're most passionate about or closest to."

With respect to EA assuming individuals will act only as individuals – that's simply because that's what they are. To be sure, individuals can join with others in a common cause: in fact, that's what effective altruists do, both through their donations, their jobs and the social movements they join. But "I am neither a community nor a state. I can determine only what I will do, not what my community or state will do. I can, of course, decide to concentrate my individual efforts on changing my state's institutions, or indeed on trying to change global economic institutions, though the probability of my making a difference to the lives of badly off individuals may be substantially lower if I adopt this course than if I undertake more direct action, unmediated by the state…. To suppose that the only acceptable option is to work to reform global economic institutions and that it is self-indulgent to make incremental contributions to the amelioration of poverty through individual action is rather like condemning a doctor who treats the victims of a war for failing to devote his efforts instead to eliminating the root causes of war…." Indeed, such arguments can backfire: "if others … become persuaded the appropriate agents for addressing problems of global poverty are communities, classes, and states, they are likely to be quite content to leave the problems to those entities and not bother with them themselves."

Criticism 6: By not addressing systemic conditions, EA takes on a conservative agenda that distracts from and thereby perpetuates the political, social and economic status quo and the inequalities and deprivations therein.

"[EA] isn't premised on a strong critique of the way that money has been made. And elements of it were construed as understanding capitalism more generally as a positive force, and through a kind of consequentialist calculus. To some extent, it's a safer landing spot for folks who want to sequester their philanthropic decisions from a broader political debate about the legitimacy of certain industries or ways of making money." Benjamin Soskis quoted by Conroy November 2022

* * *

"It's good that FTX's collapse is finally making people rethink Bankman-Fried and effective altruism. But the problem with effective altruism isn't that it's populated by insufferable dweebs. The problem is these dweebs' alliance with a profoundly anti-democratic project: to let rich people continue to make as much money as possible, whatever the cost to people and the planet." Aronoff November 2022

* * *

"…[I]n the hands of Bankman-Fried (commonly known as SBF), effective altruism was neither effective nor altruistic. Instead, he illustrated how the do-gooder ideology can serve as a … natural vehicle for bad behavior … [because] its cardinal demands do not require adherents to shun systems of exploitation or to change them…. Mainstream effective altruism displays no understanding of how modern capitalism – the system that it eagerly chooses to participate in – can explain extreme destitution in the Global South or the vulnerability of our society to pandemics. This crowd seems clueless about the reality that funding research into protecting against dangerous artificial intelligence will be impotent unless we structure our society and economy to prize public safety over capital's incentive to innovate for profit…." Aleem December 2022

* * *

The second line of criticism against the ultimate effects of EA is an extension of criticism 2 from part 3 and criticism 5 above. That's to say, the philosophical foundations if EA lead it to focus on individual needs and welfare rather than things such as freedom and equality, and its analytical methods lead it to focus on separate stand-alone interventions rather than the systemic conditions. As a consequence, "effective altruism today is a conservative project" – one that "doesn't try to understand how power works, except to better align itself with it. In this sense it leaves everything just as it is."

Critics account for this conservative bent in three ways. The first stems from the charge that EA is "in the business of selling philanthropic indulgences for the original sin of privilege." It therefore caters to the incentives and outlooks of the privileged folk it targets. Leaving everything just as it is "is no doubt comforting to those who enjoy the status quo – and may in part account for the movement's success … [in attracting] privileged, ambitious millennials [who] don't want to hear about the iniquities of the system that has shaped their worldview." More specifically, the emphasis that many EA thought leaders place on "longtermism" appeals to wealthy funders from the tech sector who "don't have to dirty their hands by dealing with actual living humans in need, or implicate themselves by critiquing the morally questionable systems that have allowed them to thrive. A not-yet-extant population can't complain or criticize or interfere, which makes the future a much more pleasant sandbox in which to pursue your interests – be they AI or bioengineering – than an existing community that might push back or try to steer things for itself."

Even between longtermist causes, one can "[c]ontrast the AI situation to climate change [which is] routinely dismissed in EA, where the problems are messy, often mundane, predominantly political, and put the very concept of economic growth under debate, and where the greatest risk is posed to poor people from the Global South. Compare also with issues like global poverty, which very few people within EA are directly affected by (and which the funders are not by definition!) and which has come to be deemed 'lower impact' within some of EA." Parentheses are original. The movement is "not above motivated reasoning."

The close ties between EA and SBF prior to November 2022 demonstrate such reasoning. Before then, a "major part of the funding pledged to organizations belonging to the Effective Altruism (EA) ecosystem … [came] from Sam Bankman-Fried alone…. [L]ittle to no effort … [was] made to recognize and minimize the potential for conflicts of interests in the EA community, in particular those linked to the funds pledged by SBF…. Those conflicts of interests are real and painfully obvious: EA … [was] incentivized to please SBF, to adopt his views and beliefs, to work on projects that he believes are important, and to shed a positive light on his activities (EA-related or not) and his person, within the EA community and to the general public." Parentheses are original.

The second explanation for EA's conservative bent is that by relying on separate interventions – say, in providing health care – it reduces the capacity or willingness of governments to expand and improve public services. Separate privately-funded interventions distract "from the urgent but thorny process of institution building. And investing in these interventions may even work to undermine the consolidation of functioning institutions. The availability of free health services reduces pressure on the state to finance and provide public goods on its own. This hinders the development of effective public administration and a sustainable tax system. It lures competent professionals away from public agencies and discourages the civic participation necessary for holding the state accountable." "The result is a disengagement of the most mobilized, discerning poor citizens from the state. These are the citizens most likely to have played a previous role in monitoring the quality of state services and advocating for improvements. Once they exit, the pressure on the government to maintain and improve services eases, and the quality of government provision is likely to fall…."

Finally, the third explanation for why EA "leaves everything just as it is" points to its failure "to confront capitalism directly" despite there being "no principled reason why effective altruists should endorse the worldview of the benevolent capitalist. Since effective altruism is committed to whatever would maximise the social good, it might for example turn out to support anti-capitalist revolution…. [However, EA thought leader William] MacAskill does not address the deep sources of global misery – international trade and finance, debt, nationalism, imperialism, racial and gender-based subordination, war, environmental degradation, corruption, exploitation of labour – or the forces that ensure its reproduction…. [C]apitalism, as always, produces the means of its own correction, and effective altruism is just the latest instance."

Rejoinders to criticism #6

There are responses to the charges that EA serves and preserves the status quo. These include the rejoinders to criticisms 2 and 5. But in addition, with respect to the claim that EA caters to the incentives of the privileged – "[i]n many ways this reflects a wider critique of philanthropy: namely that it is inherently a reflection of existing inequalities and power structures, and therefore will always be part of the problem rather than part of the solution when it comes to addressing such issues at a fundamental level. There are those within the world of philanthropy who are trying to overcome these challenges, by finding models and approaches that allow for genuine structural reform rather than simply addressing the symptoms of structural issues (perhaps through supporting social movements or embracing participatory methods which empower recipients)." Parentheses are original. Sure enough, the power of philanthropy – whether wielded by EA or not – originates from those with the means and incentives to give. Nevertheless, many fundraisers would understand and sympathize with MacAskill's outlook that "[i]f I can help encourage people who do have enormous resources to not buy yachts and instead put that money toward pandemic preparedness and AI safety and bed nets and animal welfare that's just like a really good thing to do."

In terms of undercutting the abilities and incentives for governments to step up their game – again, this is a criticism not of EA but of philanthropy in general. Sure enough, greater government funding in, say, health, education or housing may "crowd out" philanthropic spending in those areas because private donations become less needed to preserve the quality and quantity of services. But "crowding out" could occur in the opposite direction – greater private funding could reduce the pressures for more government funding. Consider, for example, the so-called "Mother Theresa effect" – the reference being to the work of the Missionaries of Charity allegedly reducing the pressures on public authorities to provide better care for the Dalits and lepers of Calcutta. Or closer to home – consider the growth since the 1980s of Food Banks in Canada and their allegedly reducing the pressure on governments to ensure the food security of their citizens. Whether affiliated with EA or not, philanthropic organizations that try to compensate for inadequate public services could end up perpetuating that inadequacy.

And in terms of kowtowing to capitalism – living and working and giving philanthropically within capitalist economies doesn't prevent individuals from seeking ways to reduce capitalism's imperfections or harmful effects. This could be by joining or supporting political organizations and social movements – as some effective altruists do. Or it could be by giving to the types of charitable causes endorsed by EA – as all effective altruists do. Those who criticize EA for emphasizing the latter route typically "combine ambitious accounts of how our societies and/or the world at large must be changed in order to become just, with moderate accounts of what individuals are obligated to do in response to the overwhelming injustice and suffering that continues to plague our world." "[P]erhaps some day the world will be receptive to rational reforms of the global economic system. But until this Utopian condition prevails, there is much that a single individual can and should do."

Sure enough, effective altruists through their giving might treat the symptoms of poverty rather than the root causes, particularly if they either don't know those causes or aren't able to change them. But "we should not forget that … [treating even symptoms] will mean saving lives, alleviating hunger or chronic malnutrition, eliminating parasites, providing education, helping women to control their fertility, and preserving sight." And we shouldn't forget that the moral significance of our efforts to improve the well-being of a finite number of people will not be diminished by there being many others that our efforts can't reach. Besides, in the more accusatory words of an anti-capitalist effective altruist, "the harm I might do in regularly buying a chai latte rather than letting that money be used to feed someone is only different in degree from the harm a capitalist does in not directly releasing grain to the hungry. The power is of the same kind."

Criticism 7: In its formation, methods, application and effect, EA is elitist: it risks becoming an intellectual, do-gooder playground for the privileged who ultimately benefit from – and through their philanthropy avoid substantively changing or challenging – the inequalities of the world around them.

"…the movement is insular. Its demographics skew very young, very male, very white, very educated, and very socioeconomically privileged." Lowrey November 2022

* * *

"Effective altruism posits that making money by (almost) any means necessary is OK because you, Elon and Zuck and SBF are so brilliant that you absolutely should have all the power implied by billions of dollars in the bank." Morris November 2022

* * *

"During Bankman-Fried's ascent, … his proximity to EA's brand of self-sacrificing overthinkers often helped deflect the kind of scrutiny that might otherwise greet an executive who got rich quick in an unregulated offshore industry." Tiku November 2022

* * *

"Philanthropy has a simple power structure: the haves give, the have-nots receive. If organizations veer too far from the wishes of their benefactors, funders can withhold their money. Few meaningful guardrails exist to stop the rich from dictating what happens to the money hoarded in philanthropic organizations. Effective altruism was supposed to be one such protection: discouraging wasteful, suboptimal spending. But they developed that culture in a way that fails to constrain funders' control over philanthropy and hands them new tools to organize the world around their cynical aspirations and unhinged preoccupations." O Táíwò November 2022

* * *

The third line of criticism against the ultimate effects of EA – and the seventh criticism overall – is a culmination of the preceding six. Effective Altruism – as informed by its philosophical foundations that require impartiality and abjure emotion or relationship, as guided by its analytical methods that focus on quantification and cultivate hubris, and as defined by its selective interventions that leave intact and unquestioned the entrenched inequalities of power and resources – is a version of philanthropy that's the product and practice and ultimately the protector of an intellectual and monied elite. It's fundamentally undemocratic.

Such concerns are voiced both outside the EA community and within it. From outside, one critic observes that "[p]aradigmatic effective altruists … are relatively well-off individuals who donate large amounts of money to organizations that aid impoverished strangers. In contrast, a poor person who devotes all her time and resources to effectively alleviating her family's or community's poverty is not an altruist and so cannot be a member…. Effective altruism is a movement that excludes poor people. This exclusion is compounded by the effective altruism movement's primary strategy for attracting new members: showing how easy it is to save lives cheaply…. This analogy may be stirring, but it encourages donors to think of themselves as heroes or saviors. This orientation overlooks poor people's central role in alleviating their own poverty and rich people's role in contributing to and benefiting from it…. By excluding poor people and encouraging a savior complex and insularity among its members, the effective altruism movement fails to meet normative criteria of democracy and equality."

Critics from within the community expand upon such observations. "EA is very white, very male and dominated by tech industry workers. And it is increasingly obsessed with ideas and data that reflect the class position and interests of the movement's members rather than a desire to help actual people. In the beginning, EA was mostly about fighting global poverty. Now it's becoming more and more about funding computer science research to forestall an artificial intelligence-provoked apocalypse. At the risk of overgeneralizing, the computer science majors have convinced each other that the best way to save the world is to do computer science research. Compared to that … global poverty is a 'rounding error'."

Moreover, across both neartermist and longtermist causes "[t]he issues of transparency and accountability become especially problematic when dealing with tasks as huge as eradicating poverty or preventing human extinction: these are communal projects, with stakeholders numbering in the billions. We cannot be so arrogant as to assume that we, the 'epistemically superior' elite of wealthy white dudes, should simply impose our preferred solutions from the top down. Projects with the aim of doing the most good should be embarked upon in cooperation and consultation with the people affected."

To account for why consultation doesn't happen, one can't ignore that "EA is largely reliant on the goodwill of a small number of tech billionaires, and as a result fails to question the practice of elite philanthropy as well as the ways by which these billionaires acquired their wealth…. Relying on a small number of ultra-wealthy members of the tech sector incentivises us to accept or even promote their political, philosophical, and cultural beliefs, at the expense of the rigorous critical examination EA prides itself on."

Such "motivated reasoning" is no more evident than in the current promotion and pursuit of longtermist causes – the ones favoured by SBF. "The things … [EA leaders] push tend to be things that Silicon Valley likes, and they almost always focus on technological fixes rather than political or social ones…. These big decisions about the future of humanity should be decided by humanity. Not by just a couple of white male philosophers at Oxford funded by billionaires. It is literally the most powerful, and least representative, strata of society imposing a particular vision of the future which suits them…. I don't think EAs – or at least the EA leadership – care very much about democracy."

Rejoinders to criticism #7

Just as this criticism can be derived from the preceding six, so can its rejoinders.

But in addition, the charges of EA being elitist and undemocratic aren't unique to EA. They are tailored versions of the more sweeping charges made against "big philanthropy" (i.e., large private charitable foundations) particularly in the US. As described by Robert Reich, for example, (non-EA) big philanthropy as "a form of power that is unaccountable, low on transparency, donor directed, and by default perpetual [when exercised by a private foundation]. Big philanthropy is a plutocratic element in democratic society." According to David Callahan, it now finances an "influence sector" comprising "think tanks, advocacy groups, litigation outfits, voter mobilization organizations and media outlets – on both the left and right – that amplify the preferences of the wealthy in public life." And as seen by Anand Giridharadas, the "world-changing initiatives funded by the winners of market capitalism do heal the sick, enrich the poor and save lives. But even as they give back, American elites generally seek to maintain the system that causes many of the problems they try to fix – and their helpfulness is part of how they pull it off…." "For when elites assume leadership of social change, they are able to reshape what social change is – above all, … the idea that people must be helped, but only in market-friendly ways that do not upset fundamental power equations … [or] the underlying economic system that has allowed the winners to win and fostered many of the problems they seek to solve."

To the extent that the charges against EA's elitism are versions of those against big philanthropy, their rejoinders can be fashioned along the lines compiled by Beth Breeze. That's to say, philanthropy has never operated on strictly democratic principles, has never tackled inequality exclusively and has always allowed donors to get something out of their giving whether in tangible or intangible forms. Admittedly, some big philanthropists may be corrupt and duplicitous. But the charges against big philanthropy are exaggerated, fail to recognize the unique role and limitations of philanthropy, and attribute the foibles and flaws of a few philanthropists or philanthropic interventions to most or all.

For example, countering the claims of big philanthropy being undemocratic Breeze points out that: civil society is key to democracy, and is funded in part by philanthropy, big and small; philanthropic interventions are episodic and relatively small-scale, in no way rivals to government programs; and the decisions and narrow constituencies of journalists, trade unions, community activists are also "undemocratic" but receive less criticism, perhaps because it's less fun to criticize those who aren't so wealthy.

In addition, Breeze argues that many critics – Giridharadas in particular – ignore the big philanthropists, including several from the tech sector, who, despite having benefited from the capitalist system, want to reform it. For example: the Omidyar Network (established by Pierre Morad Omidyar, the founder of eBay) seeks ways to change the current "broken" form of capitalism and alleviate the inequalities caused by unrestrained free-market ideologies; similar goals are pursued by the MacKenzie Scott Foundation (former wife of Amazon founder Jeff Bezos), the Economic Security Project and the Anti-Monopoly Fund (established by Chris Hughes, co-founder of Facebook); and the Hewlett Foundation (established by William Hewlett, co-founder of Hewlett-Packard) seeks "a new way of thinking about policy, law and the proper role of government to shift the underlying terms of debate and open up space for solutions that neoliberalism is currently choking off." Sure enough, these initiatives could be brushed off as capitalism producing "the means of its own correction." But they also indicate that there are members of the elite who, although having personally benefited from the status quo, want to change it.

The perspective I'm presenting as a rejoinder – that the criticisms of EA's elitism are for the most part criticisms not of EA in particular but rather of big philanthropy and philanthropy in general – may not convince or assuage the critics from within the EA community who want and expect their brand of philanthropy to be uniquely capable of doing "the most good." Neither may it convince commentators from outside the community who, like Callahan, want or consider EA to be "an overdue, much-needed counterweight to the typical practice of elite philanthropy, which has been very inefficient."

Nevertheless, the perspective is consistent with the overall argument of this series: that although the bankruptcy of FTX and criminal charges against Samuel Bankman-Fried focused and amplified existing criticisms of Effective Altruism, those criticisms have implications and pose questions for philanthropy and the philanthropic sector as a whole. They deserve to be heard, reflected upon and heeded by not simply EA, but the broader sector of which it's a part.

What can we take from the downfall of Samuel Bankman-Fried with regard to the ultimate effects of Effective Altruism?

As noted in part 2, Samuel Bankman-Fried eschewed the "more emotionally driven" neartermist causes of EA such as global poverty and health and animal welfare, and instead positioned himself to be the major supporter of its longtermist causes. In February 2022, he created a Future Fund within the FTX Foundation that in its first year was to have disbursed between $100 million and $1 billion US "to improve humanity's long-term prospects" through funding separate interventions that among other things would promote "the safe development of artificial intelligence" and reduce "catastrophic biorisk." But late in 2022 – following the allegations, bankruptcies and criminal charges – the Foundation, the Fund and those plans came to an end. Are he and the reputation he built as the "world's most generous billionaire" the exceptions that prove the general rule that the ultimate effects of EA are indeed altruistic, effective and enduring? Or is he the example that demonstrates those effects are defective on those terms? Or is he neither?

Regardless of how one answers such questions, the downfall of SBF re-activated existing criticisms and suspicions of EA's ultimate effects. Many across the philanthropic sector might see those criticisms as being relevant only to EA with little to say about the effects of the sector as a whole. I don't see it that way. Here, I select five areas in which the criticisms might have implications and raise questions beyond EA. Readers will no doubt be able to recognize many others.

i. Critics of EA attribute its focus on separate, short-term interventions in part to its "bias" toward having quantitative evidence of cost effectiveness. Does the philanthropic sector's parallel quest for having measurable "impact" strengthen or weaken what it can do and do well?

What role does measurable "impact" play in the charities or nonprofits you work with or in? As a means of appeasing donors or other funders – as suggested in part 3? Or as a basis for evaluation and learning about what works and what doesn't – at least in terms of the indicators used to define effectiveness or impact? Something else? Are there activities or opportunities – potentially beneficial ones – that are being overlooked or avoided because of the role it plays?

ii. Some critics fault EA for not pursuing social transformation through institutional change. Should social transformation be the goal of philanthropy, or at least a goal? Can it be a genuine goal given that philanthropy comes primarily from those who have prospered under the status quo? If social transformation should and can be a goal, then by what initiatives or interventions could the sector advance it?

Some commentators and scholars assess the achievements of philanthropy in terms of "the transformation of society, rather than increased access to socially-beneficial goods and services

– a noble goal for sure, but insufficient to lever deeper changes in the distribution of power and resources across the world." Such transformation would fundamentally alter the institutions, laws or societal norms that are seen to perpetuate inequality. EA doesn't pursue such systemic change because it sees little evidence that it could bring it about, or bring it about to beneficial effect. Where do you stand? Is social transformation within the reach and purview of philanthropy, or does it lie within the realm of politics? How does your stand square with your views on the expanded abilities of Canadian charities to engage in "public policy dialogue and development activities?"

iii. Certain norms, expectations, routines and allocations of power and responsibility internal to a charity or nonprofit make the delivery of its services possible. But norms and protocols might also limit the effectiveness of those services and the populations they reach. How can such unintended barriers be identified and changed – and by whom?

Again, think of the charities or nonprofits you work with or in – whether they deal with the arts, education, health, housing, sports, religious observance or something else. Who is missing in the populations that those organizations serve or privilege? Who is missing in their leadership? Are there barriers that limit the organization's reach and impede its mission? As case studies of such internal change, consider these arts organizations that designed or revised their programming to widen their audience and advance the well-being of their communities.

iv. Critics of EA claim that it doesn't consult or work in solidarity with its beneficiaries, particularly when alleviating the effects of poverty – a point also raised in part 4. What are the barriers to such consultation, collaboration and accountability? Should all charitable organizations overcome them, and if so, how?

Between grantors and grantees, or between a charity and its beneficiaries, it's not easy or straightforward to maintain an open and frank exchange of information, let alone create systems for shared decision-making and mutual accountability. With its longtermist causes, EA claims to be aware of and responsive the needs of its beneficiaries – the silent, unborn generations that are otherwise "utterly disenfranchised." But in terms of sharing decision-making power with present-day grantees, or soliciting advice from ultimate beneficiaries of its neartermist causes, its efforts are modest. Once more, think of the charities or nonprofits you work with or in. Are there opportunities for, and would there be benefits from, greater consultation and collaboration? If so – how could these be initiated, and by whom? To get you thinking, here are some case studies on sharing and shifting power and foundations' efforts toward Indigenous reconciliation.

v. Critics of EA fault it for providing social services that allow governments to dodge their responsibilities to provide those services more comprehensively – often to the disadvantage of the worst off who can't access EA's services. Is such criticism fair?

The welfare state developed in Canada in the 1960s, encouraged by the federal government's expansion during and after WWII, and by the inabilities of philanthropic efforts and piecemeal policies to handle the greater social needs that surfaced in the 1930s and then accompanied the post-war baby boom. However in the 1990s, given the pressures for government retrenchment and the popularity of "new public management," governments scaled back their spending and increasingly transferred the delivery of public services to nonprofits and charities, funding them through contribution agreements. But now, after demonstrating their competence and in light of the growing demand for social services, have nonprofits and charities made it easier for governments – and businesses, for that matter – to avoid designing and funding more comprehensive programs that would better meet the health, housing, integration and reconciliation needs of their populations? Is it aways better for the philanthropic sector to take on greater responsibilities?

Epilogue

Effective Altruism (EA) was the approach to philanthropy that Samuel Bankman-Fried (SBF) promoted and practiced. And prior to the bankruptcies and criminal charges of late 2022, the EA community had upheld him as an exemplary practitioner. In the aftermath of those events and given its ties to SBF, EA came under intense scrutiny and criticism for having possibly encouraged his actions or given them cover.

Sure enough, the criticisms were directed against EA – its philosophical foundations, analytical methods and ultimate effects. But as I see it, they have implications and raise questions that can be generalized and applied to the philanthropic sector as a whole and to our individual involvements with it – whether as donors, volunteers, employees, advisors, regulators, or beneficiaries. For that reason the criticisms and their rejoinders deserve to be heard and reflected upon more widely. And it's for that reason I've prepared this five-part series.

As noted in part 1, soon after the downfall of SBF William MacAskill underlined the need for both he and the EA community to reflect on what should change:

> Sam and FTX had a lot of goodwill – and some of that goodwill was the result of association with ideas I have spent my career promoting. If that goodwill laundered fraud, I am ashamed. As a community, too, we will need to reflect on what has happened, and how we could reduce the chance of anything like this from happening again…. I will be reflecting on this in the days and months to come, and thinking through what should change.

I hope this series provides a better understanding of EA – one that is both critical but admiring, that recognizes not only its distinctiveness in the world of philanthropy, but also the many priorities and concerns that it holds in common with others involved in that world.

And I hope that we, mindful of what we all hold in common, will, along with the EA community, be able to reflect on how and what we should change.