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, PANL Perspectives, May 2023

Part 2: Effective Altruism

As noted in part 1 of this series, the charges brought in late 2022 against the crypto entrepreneur Samuel Bankman-Fried intensified existing criticisms and suspicions of Effective Altruism – the approach to philanthropy with which he is closely associated. Parts 3, 4 and 5 will examine those criticisms and derive from them broader questions and issues for the philanthropic sector to consider. But here in part 2, we examine Effective Altruism itself: its origins, ethos, analytical methods, priorities and evolution.

What're we talking about?

Effective Altruism (often shortened to EA) has two sides. 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, *contingency* 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 of this series – 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.

The next three parts of this series examine those criticisms along with their rejoinders. But they do so with the intent of holding up a mirror to the philanthropic sector as a whole – a mirror that might allow more of us across the sector to learn and possibly apply something from the downfall of Bankman-Fried and his ties to EA.