PHILOSOPHICAL PERSPECTIVES

Philosophical Perspectives, 00, 2020
doi: 10.1111/phpe.12133

COUNTERPRODUCTIVE ALTRUISM: THE OTHER HEAVY TAIL

Daniel Kokotajlo
UNC Chapel Hill

Alexandra Oprea
Australian National University

Effective Altruism (henceforth EA)[1] is an influential new social movement "which applies evidence and reason to working out the most effective ways to improve the world"[2] and is dedicated to asking "How can I make the biggest difference I can?"[3]

To some, this may sound like an unhealthy obsession with maximization. Yes, research and experimentation can help us design a swimsuit that is 3% percent faster, or find a stock to invest in that yields 3% higher returns. And yes, in the context of Olympic swimming or hedge fund management, it makes sense to pursue those minor improvements. But altruism is not a game or a competition. It's not about being the best, or doing the most. The quest for the optimal intervention seems like a distraction from actually helping people. We sympathize with this criticism. However, what if the improvements are major? What if, by investing time and money in researching effectiveness, we can do not 3% more good, but 3,000%? Then EA starts to seem more reasonable—perhaps even morally obligatory.

The first goal of this paper is to expand on the point just made. We will argue that the appeal of EA depends to a large extent on an implicit claim about the distribution of opportunities to do good that we call the Heavy Tail Hypothesis (henceforth HTH). The HTH has been introduced in prominent EA publications, albeit not under that name. Roughly, it means that the probability of finding an extremely beneficial altruistic intervention declines *slowly* as the amount of benefit increases. This implies that donating to the most effective causes, charities, or interventions can often do orders of magnitude more good than donating to moderately effective ones, constituting a substantial portion of the total amount of good generated through altruistic interventions. If the HTH is true, EA methods—including, but not limited to, investing heavily in effectiveness research, and trying to make the *biggest* difference—really can bring about orders of magnitude more good than non-EA methods. If the HTH is false, then EA methods are more likely to indicate an unjustified obsession with

efficiency. Although the literature includes empirical and normative criticisms of EA,[4] none to our knowledge have focused explicitly on whether the HTH justifies EA methodology or whether the HTH is true.[5]

The second goal of this paper is to offer a constructive critique of the narrow construal of the HTH as referring only to altruistic interventions that can generate large amounts of good (i.e. interventions that are located in the right-sided tail of the heavy tailed probability distribution). After canvassing arguments given for the existence of this right heavy tail, we argue that they also support the existence of a left heavy tail where counterproductive interventions do orders of magnitude more harm than ineffective or moderately harmful ones. We explore the implications of this other heavy tail for EA methodology and for the debate surrounding the so-called "institutional critique" of EA.[6]

We proceed in eight sections. Section I briefly introduces effective altruism and its core commitment to efficiency. Section II explains why efficiency considerations matter in ethical decision-making, setting up the groundwork for discussing the probability distribution of altruistic interventions. Section III introduces the heavy tail hypothesis and the theoretical differences between heavy tailed distributions and thinner-tailed ones such as the normal distribution. Section IV explains how assuming that opportunities for doing good are distributed with a right heavy tail lends support to the EA approach concerning cause prioritization, effectiveness research, and the assessment of classes of charitable interventions. Section V then introduces the implications of adding a left heavy tail, noting the ways in which it reinforces, qualifies, or undermines the implications in section IV.

At this point the reader may be wondering whether the HTH is true (both in the case of the right heavy tail and in the case of the left). Section VI discusses three empirical arguments effective altruists have provided in support of the HTH. Section VII then introduces three novel arguments. Although a full assessment of the empirical evidence is beyond the scope of the paper, we find support for both heavy-tails and contend that the burden of proof lies with those who would argue against it. Section VIII concludes.

I. What is Effective Altruism?

Following the impetus of its primary advocates, we think of EA as a research field and a social movement, not as a first order normative theory. William MacAskill, one of the founders of EA, provides the following definition:

"Effective altruism is:

(i) the use of evidence and careful reasoning to work out how to maximize the good with a given unit of resources, tentatively understanding the good in impartial welfarist terms, and
(ii) the use of findings from (i) to try to improve the world"[7]

MacAskill characterizes his definition as intentionally non-normative. However, other characterizations of EA include a series of shared ethical commitments in addition to the shared methodological commitments. Brian Berkey, for example, identifies the following four such commitments:[8]

EA1: There are very strong moral reasons, grounded in fundamental values, for the well off to direct significant resources to efforts to address important moral issues (e.g. to alleviate the plight of the global poor).
EA2: These fundamental values include (but are not necessarily limited to) impartially promoting increases in welfare, or quality of life, for individuals, and the reasons provided by this value are at least fairly weighty.
EA3: There are strong reasons to prefer giving to efforts that will promote the relevant values most efficiently.
EA4: We should employ the best empirical research methods available in order to determine, as best we can, which efforts promote those values most efficiently.

Much of the critical engagement with EA has come from normative theorists debating the strength, limits, and validity of EA1 and EA2. Although these ethical commitments have received some criticism, they can draw support from a range of normative theories, both consequentialist and non-consequentialist.[9] Our paper tentatively accepts the plausibility of EA1 and EA2. We are primarily concerned with the methodological commitments of EA, particularly the preference for efficiency in EA3, which also features prominently in MacAskill's definition.[10] More importantly, we believe EA3 represents the most distinctive aspect of effective altruism and the aspect that has received comparatively little attention in the philosophical literature.

II. Why Efficiency Matters for Ethics

One of our goals in this paper is to show that the empirical claim about the distribution of opportunities for doing good that we call the "Heavy Tails Hypothesis" accounts for the strength of the efficiency argument as well as the success of EA as a social movement (i.e. in attracting both members and resources to its cause). By "efficiency," we understand optimizing the means for achieving one's goals (in this case, one's altuistic goals). For a fixed goal, such as reducing the incidence of death from malaria, efficiency means minimizing the amount of resources (i.e. time, effort, or money) required to achieve it. For a fixed amount of resources, efficiency means maximizing the amount of good one can do through deploying those resources.[11]

For a familiar starting point, we turn to the well-known argument for duties to assist others provided by Peter Singer in "Famine, Affluence, and Morality."[12]

As others have noted, the argument has both ethical and empirical premises and can be summarized roughly as follows:

Ethical Premise: If it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it.[13]

Empirical Premises:

Magnitude: Something very bad is happening in the world right now (e.g. poverty, famine).

Effectiveness: It is in our power to prevent it. (For example, by donating some portion of our income or time to a charitable organization helping the global poor).

Cost: Preventing it would not sacrifice anything morally significant.[14]

Conclusion: We should act to prevent the very bad thing from happening. (e.g. We should donate some portion of our income to a charitable organization helping the global poor.)

The ethical premise relies on the famous Shallow Pond thought experiment, which Singer introduces as a straightforward application of the principle above:

> If I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing.[15]

The ethical premise has the advantage of appealing to a broad range of ethical perspectives, including versions of deontology and virtue ethics, particularly once one specifies the type of morally significant cost involved.[16] Regardless of how one spells out the relevant ethical principle, however, it is clear that the argument depends on the empirical premises.

The *effectiveness* premise represents a version of the "ought implies can" principle. If our donations to charity are either impotent to change the situation or — more concerningly, would lead to a worse state of affairs than before — our obligations to assist would either disappear or at least significantly weaken. Furthermore, effectiveness may be a matter of degree — unlike the all-or-nothing example in Shallow Pond, one's charitable donation may only alleviate a fraction of the problems caused by poverty or famine. The *cost* premise is also essential for the argument to work and can be spelled out as either an objective or a subjective measure, depending on one's ethical theory. As in the case of effectiveness, one can understand cost as a matter of degree. To the extent that one is able to do the same amount of good while sacrificing something less morally costly, one should do so as a matter of instrumental rationality.

Finally, the *magnitude* premise can be understood as capturing how bad the harm is. Realistically there are many degrees of harm, and probably no precise cutoff between "bad" and "very bad." However, the strength of the argument's

conclusion is proportional to the magnitude premise. Take the following pair of cases:

Flu

An infectious disease such as the flu will lead to the death of 0.1% of individuals who catch it in a given village. You have the resources to distribute 100% effective flu shots to everyone in the village at minimal cost to yourself and those who receive the vaccine.

Ebola

An infectious disease such as ebola will lead to the estimated death of between 25% and 90% of individuals who catch it in a given village. You have the resources to distribute 100% effective ebola shots to everyone in the village at minimal cost to yourself and those who receive the vaccine.

Assuming the two diseases have a similar rate of infection, the magnitude of the harm in the Ebola case is higher than in the Flu case. Although it is morally praiseworthy to assist in both cases, you have *ceteris paribus* stronger moral reasons to assist in the second case rather than the first due to the magnitude of the harm you would be preventing.[17] We believe that this point does not depend on any specific utilitarian calculations about the amount of good one is able to do. The stronger appeal of the second case can be justified on prioritarian, egalitarian, virtue ethics, or deontological grounds just as well as it can be on purely consequentialist grounds. A number of recent papers have provided convincing arguments that, *ceteris paribus*, non-consequentialists should care about the number of people that can be helped in any given intervention and should prefer to help the many rather than the few if confronted with the choice. Tom Dougherty, for example, argues that deontologists should help the many because you are (a) morally required to want the survival of each person for their own sake, and (b) rationally required to achieve as many of these ends as possible, if you have these ends.[18]

Returning to the Singer example, one can see the individual and collective importance of the three empirical premises: magnitude, effectiveness, and cost. Without the empirical premises, Singer's ethical principles would still be plausible, but they would be largely academic and limited to the rare events of saving children drowning in shallow ponds. In order to (1) account for the moral urgency with which many view the tragedy of global poverty and (2) motivate the substantial charitable contributions that Singer recommends for those living affluent lives, one requires some combination of high magnitude, high efficacy, and low cost. Put differently, the real-world significance of Singer's argument depends on efficiency (i.e. effectively achieving the highest amount of good at the lowest cost). Moreover, we conjecture that many people find Singer's arguments compelling largely because of they believe that the empirical premises are true. If more people thought these premises false, the success of the movement

started by Singer and other effective altruists would be significantly diminshed. The first goal of our paper is to argue that the HTH plays a similar role in motivating support for the EA project. By amplifying the efficiency gains and losses at stake, belief in the HTH can explain the appeal of the movement and its growing recent popularity. In the next three sections, we will show why the HTH substantially amplifies the efficiency gains and losses at stake in effective altruism.

III. The Heavy Tail Hypothesis

The goal of this section is to illustrate the difference between a world where the probability distribution of charitable interventions has thin tails (such as the normal distribution) and one where it has one or more heavy tails. For simplicity, we use the general term "intervention" to refer to a range of possible ways EAs recommend making a difference. For example, assume that you have decided to donate $10,000 and you are trying to assess the amount of good you could do by directing your donation to one among a large number of possible charities. Effective altruists contend that the probability distribution of possible altruistic impact per such a donation has a heavy tail. Or assume that you have decided to dedicate the approximately 80,000 hours of time of your expected career time to maximizing the amount of good you can do. Then we would be considering the probability distribution of altruistic impact by career choice.

A probability distribution is a function that maps potential events onto their probabilities of occurring.[19] In the case of one roll of a fair die, the probability distribution would identify the probability of each of the six possible outcomes (in this case, all outcomes have an equal probability of 1/6 or approximately 16.6%).[20] This is called a uniform probability distribution because all outcomes are equiprobable. One of the most commonly discussed probability distributions in the social sciences is the standard normal distribution. Many physical attributes such as height, weight, blood pressure, birth weight, and shoe size appear to be normally distributed. This means that the function describing the probability of a random individual having a specific value for one of the aforementioned physical attributes has the classic bell-shape illustrated in Figure 1 below. For illustrative purposes, we can look at height and assume that it in fact follows a normal distribution.[21] According to Our World in Data,[22] the average height for men in the most recent cohort available was 178.4 cm (or approximately 5 feet and 10 inches) and that the standard deviation is 7.59 cm (or approximately 2.5 inches).[23]

Table 1 below shows the probability that a given man is taller than the average by a specific amount. Each increase represents one standard deviation. The probability that a man is taller than 185.9 cm (or 6 feet 1 inches), for example, is approximately 16%. That means that there is a 1 in 6.3 chance of

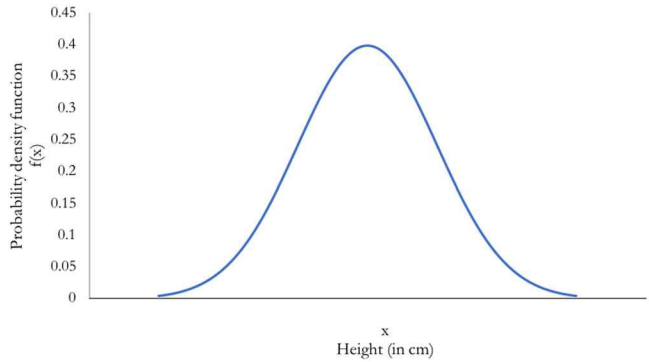

Figure 1. A Normally Distributed Variable (Height)

Table 1. Normal Distribution (Height)

Height Threshold (h)		Probability that a Man is Taller than h	Odds of Finding a Man Taller than h
in centimetres	*in feet and inches*	*approximation*	*approximation*
178.4	5' 8"	50%	1 in 2
185.9	6' 1"	15.86%	1 in 6.3
193.6	6' 4"	2.27%	1 in 44
201.2	6' 7"	0.13%	1 in 741
208.7	6' 10"	$3.17*10^{-5}$	1 in 31,574
216.3	7' 1"	$2.87*10^{-7}$	1 in 3,488,556
223.9	7' 4"	$9.87*10^{-10}$	1 in 1,013,594,635
231.5	7' 7"	$1.28*10^{-12}$	1 in 781,332,000,000
239.1	7' 10"	$6.22*10^{-16}$	1 in 1,607,470,000,000,000
246.7	8' 1"	$1.13*10^{-19}$	1 in 8,860,630,000,000,000,000

finding a man taller than 185.9 cm. As you can easily see in the Table above, the odds of finding a man taller than a given threshold declines extremely rapidly.[24]

The distribution of height illustrates two typical features of normally distributed variables. First, extreme values are highly unlikely (even if never impossible) and their likelihood declines at a quick and accelerating rate. Going from one standard deviation to two makes it a little over 3 times less likely that a given man will exceed that height. Going from nine standard deviations to ten makes it over 5,000 times less likely that a given man will exceed that height. Second, extreme values constitute a minuscule percentage of the total value. According to the Guinness Book of World Records, the tallest man ever recorded was Robert Pershing Wadlow from Illinois at 272 cm or 8 feet 11 inches tall (over 12 standard deviations above the mean). Even at almost 9 feet, Robert Wadlow's height is still the tiniest fraction of the total height of men in the world. He is not even as tall as two average-sized men! This constitutes the second typical feature. Even if extreme values do occur, their magnitude

Table 2. Power Law Distribution (Population of US Cities)

Population Threshold (pop) *In number of residents*	Probability that a US City is larger than pop *approximation*	Odds of Finding a US City larger than pop *approximation*
50,000	74.82%	1 in 1.3
100,000	30.39%	1 in 3.3
200,000	12.34%	1 in 8
500,000	3.75%	1 in 27
1,000,000	1.52%	1 in 66
2,000,000	0.62%	1 in 162
5,000,000	0.19%	1 in 532
10,000,000	$7.63*10^{-4}$	1 in 1,310
100,000,000	$3.82*10^{-5}$	1 in 26,141

has a completely negligible effect of either the total or the average of the relevant variable.[25] Together, these two features justify ignoring extreme values (or so-called outliers).

Matters appear entirely different when one considers heavy-tailed probability distributions. The most common definition of a heavy-tailed probability distribution is a probability distribution whose tail index or tail function decays more slowly than that of any exponential distribution.[26] The tail index measures the likelihood that a value larger than x (or smaller than −x) can still be found in the data as x gets larger and larger (or smaller and smaller).[27] As the definition indicates, probability distributions can have a single heavy right tail (i.e. disproportionate probability of large positive events), a single left tail (i.e. disproportionate probability of large negative events), or two heavy-tails. Such probability distributions have been argued to characterize a range of natural and economic phenomena, including earthquake magnitudes, word frequencies, deaths in wars, commercial book sales, academic citations, income, wealth, population in cities, insurance claims from floods, to name a few. As in the case of height above, let's assume that the population of cities in the US follows a specific type of heavy-tailed distribution called a power law distribution (where cities are defined as having populations larger than 40,000).[28]

Unlike the case of normally distributed data, Table 2 below shows that the probability of extreme values declines very slowly. This has two important implications for the role of extreme values in our analysis. First, extreme values are much more likely to persist as we move away from the mean, preventing us from truncating the data and ignoring so-called "outliers" (the way one could when dealing with normally distributed data). Second, the magnitude of extreme values matters for our analysis. The largest city in the US is New York City. Its population during the last census was reported as 8,398,748. While the tallest man to ever live was not even twice as tall as the average person, the largest city in the US is approximately 60 times as large as the average city above 40,000. (And much larger if we include towns and smaller cities.)

Table 3. The Significance of Extreme Values

Share of total from top…	Height	US Cities	Heavier-Tailed Distribution
1%	1.11%	19.25%	41.50%
5%	5.43%	36.98%	57.60%
10%	10.74%	47.32%	65.70%
20%	21.18%	59.91%	75.02%

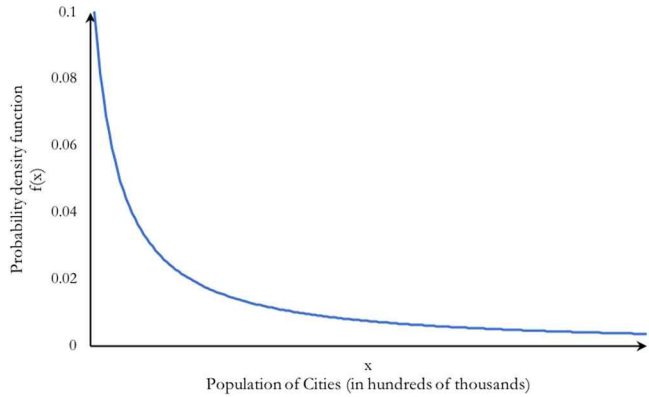

Figure 2. Power Law with $\propto = 1.1$ and $x_{min} = 1$ Probability Density Function

To get an even clearer idea of the impact of extreme values, Table 3 above summarizes the proportion of the total value that comes from the top 1%, 5%, 10%, and 20% of the data in the case of male height, US city population, and a simulated power law distribution with an even heavier tail. As you can see, the tallest 1% of men represent barely more than 1% of the total height of all men. However, the most populous 1% of US cities represent 19.25% of the total US population living in cities with a population above 40,000. Finally, the last column shows a simulated power law distribution with an even heavier tail where the top 1% represent almost 42% of the total. Figure 2 above illustrates the probability density function of the power law that generated the values in the last column.

Figure 3 below then illustrates a probability distribution with two heavy tails similar to the Student's t-distribution commonly used in statistical analysis. The distribution is centered at zero so that the left tail represents outcomes that are negative (such as counterproductive altruistic interventions) and the right tail represents outcomes that are positive (such as effective altruistic interventions).

In sections IV and V, we connect the interventions and approaches favored by effective altruists with the assumption that the distribution of opportunities for doing good has one or more heavy tails. We note that some of the interventions pursued by EA are more justified if there is also a heavy left tail, and others are less justified. After establishing the theoretical connections between

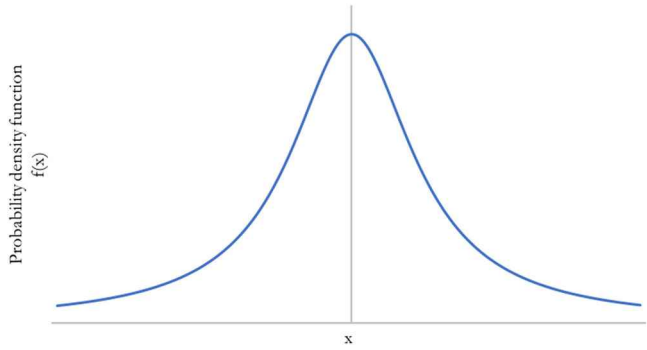

Figure 3. Two-Tailed Probability Distribution

the EA approach and the heavy tail(s) hypothesis, the following sections turn to arguments in favor of the hypothesis itself.

IV. Implications of the Heavy Right Tail for Altruism

Assume that the probability distribution of charitable interventions has a heavy-right tail (for example, like the power law described in the previous section). This means that your expectation about a possible new or unassessed charitable intervention should include the large values described above with a relatively high probability. It also means that existing charitable interventions whose effectiveness is known (or estimated with a high degree of certainty) will include interventions differing in effectiveness by orders of magnitude. We contend that this assumption justifies well-known aspects of EA practice such as (1) effectiveness research and cause prioritization, (2) "hits-based-giving," and (3) skepticism about historical averages.

First, it justifies extensive investments into effectiveness research. Suppose the distribution of charities by effectiveness is like the distribution of US cities by population. Donating $1 to a charity in the 99.9th percentile would then be *four times better* than donating *ten times as much* to a charity in the 80th-to-90th-percentile range![29] In this scenario, you should be willing to spend a large portion of your budget on research to identify the very best charities, even if you have already identified some that appear quite good. Effective altruists therefore seek to identify the best career one should pursue to maximize one's altruistic impact,[30] the best existing charity to donate money to,[31] and the best cause areas to prioritize in terms of donations of time or money.[32]

Second, it provides support for "hits-based-giving" — the official strategy of the prominent EA organization Open Philanthropy Project.[33] Hits-based giving aims to maximize the expected value of a portfolio of philanthropic "investments". In maximizing expected value, the organization does not exhibit

a preference for only low-risk interventions, but will often support higher risk, higher-upside ones. Their preference for the latter is supported by the HTH. For example, using the numbers from the scenario above, it would be *two times better* to fund 100 interventions, each of which have a 95% chance of achieving nothing whatsoever and a 5% chance of being in the 99.9th percentile, than to fund 100 interventions, each of which have a 100% chance of being in the 80th-90th percentile range. Holden Karnofsky (the CEO of the organization) compares this strategy to venture capital investments in start-ups where a few big winners more than make up for a series of failed investments.

Third, it justifies skepticism about historical averages when it comes to altruistic interventions. A small random sample from a normally distributed variable (e.g. a sample of 30 observations) is likely to provide quite a good approximation of the true mean of all possible values. This is not the case for a variable whose probability distribution has a heavy tail. A small random sample is very likely to underestimate the expected value — sometimes by orders of magnitude.[34] When aid skeptics such as Dambisa Moyo, William Easterly, and Angus Deaton point to the dismal record of foreign aid by both governments and nonprofits, supporters of EA such as William MacAskill point out that looking at the average altruistic interventions is insufficiently informative: "In response to Dambisa Moyo, I pointed out that, because the best programs are so good, they make aid very effective on average. But we don't need to fund programs of merely average effectiveness. We can deliberately choose to fund only the very best programs, which allows us to do a tremendous amount of good."[35]

V. Implications of the Heavy Left Tail for Altruism

What if the probability distribution of altruistic interventions includes both a left and a right heavy tail? In this case, we cannot assume either that (1) one's altruistic interventions are expected to have at worst a value of zero (i.e. to be bounded on the left side) or (2) that the probability that a charitable intervention is counterproductive or harmful approaches zero very rapidly. In this section, we consider the implications of a distribution with two heavy tails.[36] We believe this revised HTH reinforces the importance of cause prioritization, expands the scope of effectiveness research, introduces skepticism about some types of interventions that are becoming more prominent such as billionaire philanthropy, and provides a more nuanced approach to political or institutional change as an area of the EA project. We discuss these implications in turn.

Downside Risk Research

Many catastrophic interventions — whether altruistic or not — generate large amounts of (intentional or unintentional) harm. When someone in the world is engaging in an intervention that is likely to end up in the heavy left

tail, there is a corresponding opportunity for us to do good by preventing them. This would itself represent an altruistic intervention in the heavy right tail (i.e. one responsible for enormous benefits). The existence of the heavy-left tail therefore provides even stronger justification for the prioritization research preferred by EAs. Furthermore, it commends the emphasis on avoiding catastrophic planetary-level risks that has been central to "longtermist" EA projects which aim to prevent or mitigate the harm from terrible outcomes such as a global pandemic, nuclear war, or uncontrolled AI. The EAs who do this readily admit that it is unlikely that their individual actions will make the crucial difference between human extinction and human survival. Nevertheless, they pursue these interventions because the good done if successful is large enough to justify the investment of time and effort.[37]

The existence of the second heavy tail also supports investments into what might be called "downside risk research." Effective altruists often intentionally or unintentionally neglect the potential for counterproductive interventions in ways that may miss important opportunities for doing good. Consider the following example from the GiveWell blog, where Holden Karnofsky explains the organization's decision not to report on the questionable dealings of the organization Smile Train:

> This puts us in an odd situation: we have very little interest in bad charities, yet others are far more interested in us when we talk about bad charities. To us, credible positive stories are surprising and interesting; to others, credible negative stories are surprising and interesting. A good example is Smile Train. Nothing in our recent post is really recent at all – we did all of that investigation in 2006. We've known about the problem with Smile Train's pitch for over three years, and have never written it up because we just don't care that much.[38]

Although Karnofsky is correct that some organizations do care about charity failures and publish lists of worst charities (e.g. the Hall of Shame from Charity Watch or the list of worst charities from Charity Navigator), these lists are unfortunately relying on the same flawed methodology as the lists of best charities provided by these two organizations. By focusing only on input data and financials, the organizations may occasionally be able to identify corruption or financial mismanagement. However, the worst a (merely) financially mismanaged charity can do is waste resources without creating positive impact. In our assessment, the most relevant metric is whether the organization is actively causing harm through its donations. In this respect, a financially sound organization focused on a counterproductive intervention can do more harm than a financially unsound one, even after factoring in the opportunity cost of foregone donations.

If the assumption of a left heavy tail is correct, it may also be worth identifying interventions which are currently doing or which might in the future do large amounts of harm, and calling them out in the hopes of reforming or defunding them. This applies to a range of individual interventions, but also to discussion

of political engagement. For example, note this analysis from William MacAskill about how much good an Oxford PPE student can do by pursuing elected office:

> In coming up with this number, we made conservative assumptions at every stage, assuming no impact if she didn't become an MP or cabinet minister, and assuming that her impact as an MP would come only through government expenditure rather than through legislation. We should therefore think that the £8 million figure is an underestimate of her expected impact.[39]

While estimating the magnitude of potential impact for Laura Brown (the PPE graduate from above), MacAskill never mentions the possibility that the estimated impact may have a negative sign rather than a positive one. However, given the numerous cases of counterproductive interventions by well-intentioned politicians and policy-makers, we believe this possibility should be explicitly considered. If the evidence for the left heavy tail is convincing, this relative neglect of downside risk by EA is very dangerous, potentially leading them to support interventions that could do a substantial amount of harm.

Assessing Types of Interventions Requires Both Tails

Another conclusion we draw from the revised HTH is that the value of a class of interventions should be estimated by considering the worst as well as the best. Following such analysis, a class of interventions could turn out to be net-negative even if there are some very prominent positive examples and indeed even if almost all examples are positive. This sharply contradicts MacAskill's earlier claim that the value of a class of interventions can be approximated by the value of its best member.

One potential such area in need of further investigation is the subset of interventions called "billionaire philanthropy." Billionaire philanthropy has recently come under strong criticism from political philosophers for allowing plutocratic influences into the democratic process and providing too much discretion to wealthy donors.[40] In the area of billionaire philanthropy into K-12 education policy in the US, for example, empirical studies have uncovered large investments into what is called "disruptive philanthropy" — i.e. investments into political advocacy to introduce more elements of private markets into the governance of public schools (e.g. charter schools, merit-based pay for teachers, voucher programs, etc.). Rechkow and Snyder find that "five grantees alone received over $150 million—18% of the grant dollars distributed by the 15 largest education foundations in 2010.[41] This concentration of funds has allowed billionaire philanthropists to have a large impact in reshaping US education. Although the jury is still out on the direction and magnitude of these changes, the possibility of a two-tailed distribution leads us to urge caution and further research into the specific subset of interventions.

The Institutional Critique Reassessed

The assumption of two heavy tails also carries important implications for the institutional critique that has been recently levied at the EA movement.[42] We find that addressing interventions that are in the far left tail requires a new approach to institutional change that, depending on the specific application, may either strongly recommend or strongly condemn it.

If we are right about the existence of the left tail, certain interventions (even well-intentioned ones) are or can be expected to be extremely net-negative. Furthermore, even certain classes or subclasses of charitable interventions (e.g. foreign aid, food aid, or billionaire philanthropy) can be net-negative as a whole. In these cases, the most good an effective altruist can do may not be to launch new charitable ventures of her own or even to donate to the most effective charities. As noted above, the most efficient intervention might be to stop oneself or other people from launching massively negative interventions. A similar point is raised by Angus Deaton:

> Like Singer, I am privileged to teach at Princeton. I too see students who want to relieve suffering in the world. Should they go to Dhaka or Dakar? Focus on bed nets or worms? I tell them to go to Washington or London and to work to stop the harm that rich countries do; to oppose the arms trade, the trade deals that benefit only the pharmaceutical companies, the protectionist tariffs that undermine the livelihoods of African farmers; and to support more funding to study tropical disease and health care.[43]

However, stopping people from doing counterproductive things presents unique challenges that may not be present in the case of making a positive contribution through individual donations to charity. The options available are almost all forms of collective action. These may include political action such as government regulation or other forms of policy change. Or they may include large scale social movements, protests, or other attempts to change prevailing public opinion and social norms. Both of these options are explicitly political and institutional. Effective altruists already attempt to limit factory farming, to restrict activities conducive to climate change, or to eliminate harsh sentences for drug offences in the US. Further research may reveal other types of activities that can be included on this list that explicitly aim to eliminate activities on the left side of the tail. These might include lobbying to end counterproductive aid programs, to repeal tax exemptions for billionaire philanthropy, or to restrict the prescription of antibiotics. In these cases, the presence of the heavy left tail can serve as an additional argument for EA to focus on structural change rather than individual interventions.

Although preventing harm often requires large scale collective action, it is also the case that many of the examples of counterproductive altruism come from precisely such attempts to change the political, economic, or social institutions of an entire society.[44] Many of the most destructive interventions of the

20th century were the product of revolutionary politicians claiming to be pursuing the good of the working class, of the poorest citizens, or of developing nations. In this context, it is plausible that private charity and charitable research are less likely to cause massive negative harms compared to interventions that target governments or society as a whole. If this is true, then the assumption of the second heavy tail militates against attempting structural change. Given the complex and interconnected nature of most political and economic institutions and cultural and social norms, changes at this level would appear likely to have the largest amount of unintended consequences.

On balance, does the revised HTH make structural change more or less promising? We do not know. However, we are hopeful that systematic research into this question will illuminate this question.

VI. The Evidence for the Heavy Tail(s) Hypothesis: Existing Arguments

Our argument thus far has focused on the implications of assuming that altruistic interventions follow a heavy-tailed probability distribution with either (a) one right heavy tail or (b) a right heavy tail *and* a left heavy tail. Section IV summarized the importance of the right tail in justifying EA methodology, while Section V highlighted the implications that follow from our novel introduction of the left heavy tail. Having considered all of these implications theoretically, we now turn to the empirical plausibility of the HTH and the evidence available in support of both the left and the right heavy tails.

Identifying the probability distribution of a given variable (i.e. charitable interventions) based on real-world empirical data is a complicated and potentially intractable task. It is also largely a matter for empirical research rather than philosophical analysis. However, given the importance of the heavy-tail hypothesis, we believe it is important to undertake at least a preliminary analysis of the available evidence. Although William MacAskill and Owen Cotton-Barratt both mention heavy-tails as an assumption underlying the EA methodology,[45] there has been no systematic attempt to argue in favor of its empirical plausibility that we are aware of. In this section, we begin by reconstructing three arguments in favor of a single right heavy tail that EAs have sketched: (i) the argument from examples of extreme values, (ii) the argument from nore systematic observational studies; and (iii) the argument from inefficient markets. For each of the arguments presented, we note that they should be extended to the existence of a heavy left tail.

The Argument from Examples of Extreme Values

In *Doing Good Better*, William MacAskill identifies a number of exceptional individuals whose altruistic contributions have made a disproportionately positive impact on the lives of others. These and other similar examples are also

included in the "Introduction to Effective Altruism" on the EA blog and in the EA Handbook.[46] The list includes: Norman Borlaug, the inventor of a strand of disease resistant wheat that substantially increased crop yields and who is credited with initiating the so-called Green Revolution in agriculture; Viktor Zhdanov, a Ukrainian virologist who first proposed the eradication of smallpox at a time when no other disease had been eradicated and who thereby sped up the process of eradicating the disease in 1979; and Paul Rusesabagina, the Rwandan humanitarian who hid and protected over 1,200 Hutu and Tutsi refugees during the Rwandan genocide at great personal risk. The impact of such individuals is often many orders of magnitude higher than the impact of the average individual, even of the average individual aiming at positive global impact. Similarly, Holden Karnofsky notes that philanthropic support for the Green Revolution by the Rockefeller Foundation and the support of philanthropist Katharine McCormick (advised by Margaret Sanger) in early stage research into the development of an oral contraceptive pill have generated enormous positive impact: "there are at least a few cases in which a philanthropist took a major risk — funding something that there was no clear reason to expect to succeed — and ended up having *enormous* impact, enough to potentially make up for many failed projects."[47] Our argument is that the study of such historical examples provides similarly compelling evidence concerning the other tail of the distribution of impact.

It is noteworthy that the most devastating losses of human lives have been the product of specific interventions, and often of a small number of individuals in key policy positions. Take, for example, the UN peacekeeping mission in Haiti in the aftermath of the 2010 earthquake. The peacekeepers deployed by the UN were brought in from other countries, including areas infected with cholera such as Nepal. By participating in the peacekeeping mission, the UN personnel re-introduced cholera to Haiti — a disease that had claimed no victims during the previous century.[48] Since 2010, cholera has been responsible for an estimated 9,145 deaths and the infection of 780,000 Haitians.[49] It was only in 2016 that the UN Secretary General Ban Ki-Moon apologized for the harm done to the people of Haiti. Estimates suggest that the disaster could have been avoided by disease screenings for as little as 2,000 USD or at a cost of less than $1 per peacekeeper.[50] As it stands, eradicating the disease is likely to cost over 2.2 billion USD.

Alternatively, consider the example of Allan Savory. During the 1950s and 1960s, Savory's research on desertification suggested that elephants in Zimbabwe (at the time, Rhodesia) were responsible for the erosion of the soil. Based on his research, the government approved a culling program that resulted in the deaths of 40,000 elephants. Savory refers to this "as the saddest and greatest blunder of my life."[51] The previous are just two examples of interventions that, while well-intentioned, turned out to be highly detrimental — often to the very population one was trying to help. Unfortunately, such examples abound across a variety of policy contexts and involve a range of actors from international organizations to

individual actors. Attempts by the US Congress to limit violence due to traffic in so-called "blood diamonds" through Section 1502 of the Dodd-Frank Act backfired and resulted in increased conflict and casualties in the Democratic Republic of Congo.[52] Large donations of food from developed countries have been found to exacerbate conflict without alleviating hunger, malnutrition, and starvation in the target regions.[53] These examples suggest that extraordinarily counterproductive interventions are unfortunately plausible, justifying the assumption of a left heavy tail.

One might be tempted to dismiss the argument from extreme values as cherry-picking. Non-heavy-tailed distributions such as the normal distribution still occasionally generate very large and very small values. Instead of dealing with a heavy tailed probability distribution, what we might be observing is an unlikely outlier. There are two ways to respond to this concern. First, the presence of a number of extreme values within even a small sample should lead one to adjust their credence in the direction of heavy-tailed probability distribution.[54] Second, this is precisely why one requires more systematic observational studies similar to the one described below.

The Argument from More Systematic Observational Studies

Another argument commonly deployed in favor of the HTH comes from more systematic comparisons of different interventions aimed at addressing the same problem. According to William MacAskill, "[w]hen it comes to doing good, fat-tailed distributions seem to be everywhere."[55] In a 2013 essay, Toby Ord (the co-founder of Giving What We Can with William MacAskill) makes a very similar claim drawing on examples from public health interventions.[56] Ord, like MacAskill, compares five potential interventions focused on the prevention and treatment of HIV and AIDS and finds that "the best of these interventions is estimated to be 1,400 times as cost-effective as the least good."[57] He makes a similar case using data of cost estimates from 108 public health interventions analyzed in the compendium *Disease Control Priorities in Developing Countries*:

> In total, the interventions are spread over more than four orders of magnitude, ranging from 0.02 to 300 DALYs per $1,000, with a median of 5. Thus, moving money from the least effective intervention to the most effective would produce about 15,000 times the benefit, and even moving it from the median intervention to the most effective would produce about 60 times the benefit.[58]

The interventions surveyed are all expected to have a positive impact (even if some have very small impact altogether). However, one can construct similar cost effectiveness estimates for interventions that are expected to have a negative impact. Unfortunately, no systematic comparisons of this kind are available in this issue area. The relative absence of this information from the academic and policy literature across numerous policy areas suggests an important oversight that we pointed to in section V.

The Argument from Inefficient Markets

Another argument for the HTH comes from analogies to other domains, particularly finance and venture capital investing. Owen Cotton-Barratt argues that the presence of regular market mechanisms that incentivize individuals to identify and exploit the best opportunities in a given domain can lead to the elimination of heavy tails. As individuals abandon opportunities that generate losses and shift towards more lucrative opportunities, one expects rates of return to go down and stabilize around the normal distribution. According to Cotton-Barratt: "So, afterwards, you end up with a much more narrow distribution of the value that is being produced by people doing these different things, than we started with."[59] If Cotton-Barratt is correct, then we should expect to see fewer heavy-tails in the domains where (somewhat) efficient market mechanisms exist and more heavy-tails in the domains where they do not. Arguably, charitable donations and investments into the types of interventions pursued by effective altruists are precisely in the category where there are few of the feedback loops and market mechanisms that would push in the direction of a normal distribution.

The argument from analogy is more difficult to assess. Even if it is true, however, the argument does not uniquely identify distributions with a single heavy tail as the default. In many of the domains where a few interventions generate enormous returns, there are also interventions that generate enormous losses — often losses many orders of magnitude higher than ordinary investments, particularly in the case of leveraged investments. To the extent that analogies to other domains suggest that the distribution of altruistic interventions by degree of effectiveness has a heavy rightward tail, the same analogy would suggest a similar assumption for the leftward tail, justifying the attention to the counterproductive side as well.

VII. The Evidence for the Heavy Tail(s) Hypothesis: New Arguments

In this section, we present three new arguments. The first, the crowding out argument, shows why the existence of a right heavy tail provides reasons to suspect the existence of a left leavy tail. The second, the argument from the data generating process, notes that the underlying mechanisms determining the opportunity set for charitable interventions are conducive to heavy-tailed probability distributions with right and left heavy tails. Finally, the burden of proof argument identifies another argument for the prima facie plausibility of the HTH and briefly notes what types of arguments its detractors would have to provide in order to challenge it.

The Crowding Out Argument

Consider any big goal you wish to achieve—the sort of goal that would put your intervention far out in the right tail if you were to achieve it. There is some

chance that the goal will be reached anyway without your effort, due to the effort of someone else. There is also a chance—perhaps a smaller chance, but a chance nonetheless—that your effort will cause an effective intervention not to happen or to be less effective than would have been the case without your action. For example, perhaps if you had not chosen to work towards this goal, someone more competent would have noticed the need and taken up the project in your absence. Thus, your choice to work on the project has a chance of backfiring, and if it does, it is a failure of the same magnitude as your success would have been.

To see the plausibility of this crowding out effect, consider the role of charity evaluators such as Charity Watch (founded in 1992) and Charity Navigator (founded in 2001). These organizations emerged in order to satisfy a demand for transparency and information in the non-profit sector. They provide information to donors about individual charities and provide rankings and categories of the best and worst charities. During the two or more decades of operation, these organizations have influenced charitable giving in substantial ways. Charity Navigator notes that it had over 11 million site visits in the past year and that its recommendations were featured in every top financial publication aimed at donors both small and large. By comparison, Give Well has closer to 750,000 views despite the growing attention to the effective altruism movement.

Despite their undeniable impact, Charity Navigator and Charity Watch only focus on input metrics such as overhead ratios and other metrics of financial health. These metrics for assessing charities have been widely criticized as unreflective of the amount of good done by any particular charity.[60] Moreover, numerous academic studies have found that the focus on overhead ratios has led to a "nonprofit starvation cycle" that threatens the ability of non-profits to serve their target populations.[61] However, the first mover advantage of these organizations has substantially influenced how donors think of measuring the effectiveness of charities in ways that have crowded out the influence of output oriented metrics of the kind proposed by effective altruists. It may take decades for donors to adapt their donations to focus on output metrics rather than input metrics such as overhead ratios. Although further research is necessary, we believe that the crowding out effect is plausible and worthy of serious consideration when assessing the impact of any proposed intervention. The existence of instances of crowding out explains why, if there are heavy right tails in the distribution of charitable interventions, we should also expect to find heavy left tails.

The Data Generating Process Argument

In this second argument, we briefly turn our attention to possible processes that would generate heavy-tailed probability distributions consistent with the evidence discussed in the earlier section. Research from statistics, as well as from the natural and social sciences has identified a series of useful heuristics about when to expect different distributions to arise. Normal distributions, for

example, are typical when data points are the *sum* of many independent inputs.[62] Log-normal distributions (which have a heavy right tail)[63] are typical when data points are the *product* of many independent inputs.[64] Distributions with even heavier tails in one or both directions (such as power-law distributions) are typical when more complex, non-linear interactions are involved (e.g. combinations of exponentials, inverses, random walks, etc.).[65] Based on these heuristics, we can make an intuitive case for why our prior should be heavy-tailed. Consider a typical philanthropic intervention such as donating $1,000 to a charity which aims to limit early deaths from tropical diseases in Africa. Assume there are at least a few thousand charities working to address this problem, at least a few hundred of which distribute anti-malaria bed nets as one among their supported interventions. How many deaths from malaria do you expect to prevent through your donation? The answer involves a calculation that looks something like this:

Number of lives saved, in expectation, via my donation =

(fraction of my donation that goes to buying and delivering nets) ×
(number of nets bought and delivered per dollar) ×
(fraction of delivered nets installed and used appropriately) ×
(how many days appropriately-used nets last) ×
(how many mosquito bites prevented per day of net use) ×
(how many cases of malaria prevented per mosquito bite prevented) ×
(how many deaths prevented per case of malaria prevented)

Given that calculating the effectiveness of even a narrow range of philanthropic interventions we are considering typically involves multiplying together a large number of independent variables, we should expect the distribution of philanthropic interventions by effectiveness to be at least log-normal.[66]

Of course, the argument as currently stated is an oversimplification. Depending on the comparison class of interventions (e.g. all charities working on public health, all charities addressing global poverty, all charities in the world), the relevant calculation would become exceedingly (and likely intractably) complex. Furthermore, the variables multiplied would be likely to differ across different sub-classes of interventions. Furthermore, we should expect interventions to have morally relevant effects along more than one dimension, and the overall effect of the intervention to be the sum of these effects. (For example, distributing bed nets might have an effect on the local economy, helping people have more time for school and work, but also increase pollution in the local environment).

However, this complication, as well as potential non-linear interaction effects between the different variables only serves to increase the probability that the resulting distribution of charitable interventions will be heavy-tailed. This applies to both the positive and the negative tails as some of the relevant inputs

are likely to have negative effects similar to the ones described in the previous section.

The Burden of Proof Argument

Finally, we believe the burden of proof lies with those who posit that the probability distribution of altruistic interventions is best described by a normal probability distribution or another similarly thin-tailed distribution. In this section, we offer a brief argument for our priors in this matter. Consider how you might guess at the effectiveness of a donation to the Against Malaria Foundation, when you have barely begun to learn about the topic. Your uncertainty should range over a broad set of possible values for each relevant variable. For example, you might think the following:

> The cost to the AMF to distribute a bed net could range between 10 cents and $10 per bed net. The number of cases of malaria prevented per bed net might be between 0.001 and 10 for all I know. The number of deaths prevented per case of malaria prevented could be between 0.001 and 0.9. So for all I know, my $1000 donation could save thousands of lives, or a tiny fraction of one life. Probably the truth is somewhere in the middle, but I have very little idea where.

Something like the previous estimate will be true for most of the possible charitable interventions you would be comparing. Without specific research into effectiveness, your uncertainty about how effective they are will range over many orders of magnitude. Lining them up side-by-side in your position of ignorance, they might look something like Figure 4 below.

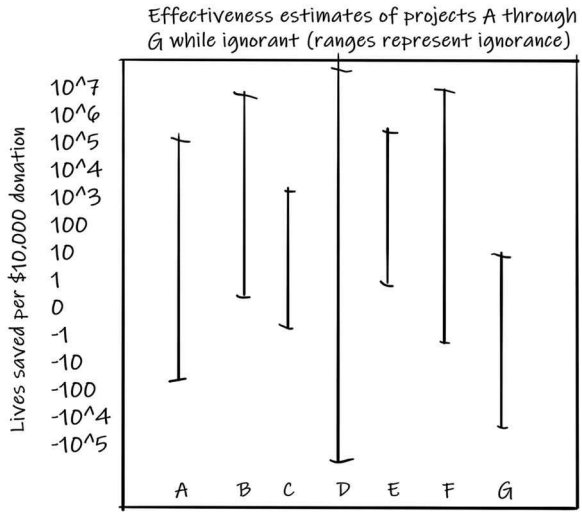

Figure 4. Uncertainty about effectiveness of interventions[67]

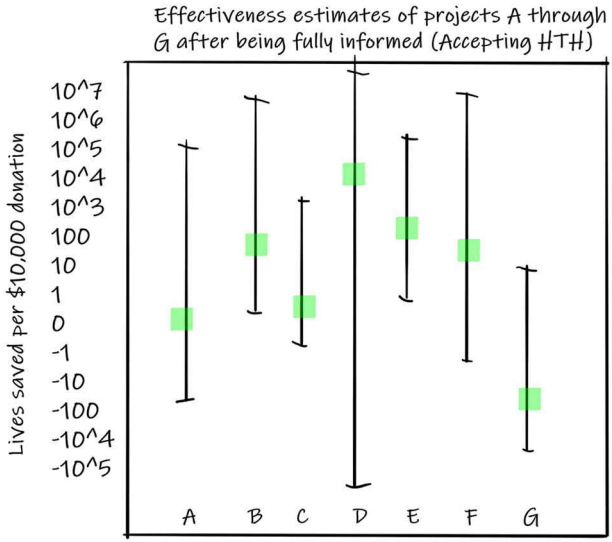

Figure 4a. Possible effectiveness estimates if the HTH is true

Each error bar represents your uncertainty about the effectiveness of the corresponding charitable interventions from A to G. If the HTH is true, then the likely effectiveness of these interventions will often differ from each other by multiple orders of magnitude. Figure 4a above shows what that might look like.

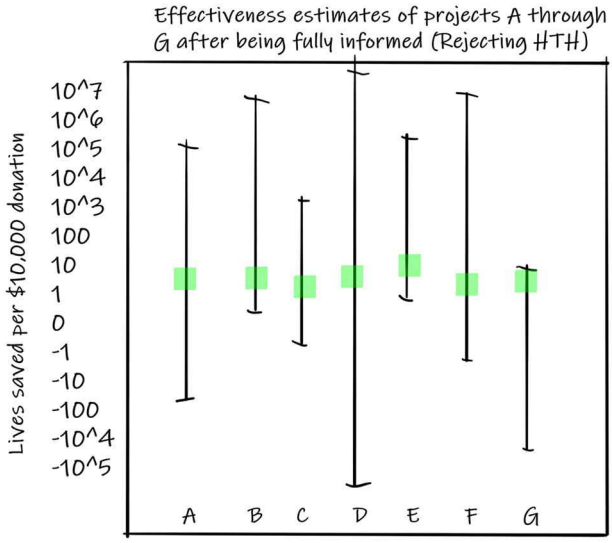

Figure 4b. Possible effectiveness estimates if the HTH is false

The green squares represent the true effectiveness of the intervention in question, i.e. what your error bars would collapse to if you became fully informed. Notice how the vast majority of the lives saved by interventions A through G are saved by intervention D, the one with the highest effectiveness. Stepping back from this specific example, we can see that pretty much any random distribution of green squares across the error bars would be consistent with the HTH.

By contrast, only a few distributions of green squares are consistent with rejecting the HTH. Figure 4b represents the central example. In order for the HTH to be false, the green squares would have to be arranged in such a way that it is not true that the very best interventions are orders of magnitude better than the typical interventions. Thus the green squares must all be clumped together on the same line, or at least a significant portion of them must be.[68] One ought therefore to think that the HTH is more probable unless we have some specific evidence or argument against it. If you have a specific argument in mind for why all of the interventions you are considering should be roughly in the same ballpark of effectiveness—perhaps an argument as simple as "If they weren't, surely it would be a big scandal and I would have heard about it by now"—then it makes sense for you to predict that the dots will line up as Figure 4b illustrates, even if you are extremely ignorant about each intervention. Our claim is merely that the HTH is a good default assumption; it is what we should believe from a position of ignorance, in the absence of an argument for expecting otherwise. The burden of proof is therefore on those who would reject the HTH to provide such an argument.

VIII. Conclusion

Effective altruism provides a promising approach to doing good. The first contribution of our paper has been to identify and systematically analyze a powerful reason for the appeal of EA methodology, namely the heavy tail hypothesis. If the HTH is true, then what is at stake in key aspects of EA practice such as effectiveness research, cause prioritization, and hits-based-giving is a large improvement in the efficiency of altruistic giving. In practice, this can translate to large differences in lives saved and lives substantially improved, accounting for the appeal of EA to individuals from a range of ethical and political perspectives. Not only is the HTH theoretically important, but sections VI and VII have also provided a number of reasons to believe that it is true — at least until further investigation.

Our second contribution has been the novel treatment of counterproductive altruism or the left tail of the probability distribution. While we commend the strides that MacAskill and others have already made in developing the ethical underpinnings and methodological guidelines for EA, we believe that there is still room for improvement. By incorporating the other heavy tail of interventions that are likely to be massively counterproductive, there are reasons

to believe effective altruists will be better prepared to avoid the errors of past attempts to do good while continuing to pursue the goal of identifying opportunities for making the largest positive contribution at the individual and systemic level. Many of these suggestions are consistent with recent developments within the EA community, including a growing attention to political and institutional interventions. A fuller examination of the tentative conclusions of this paper requires a research agenda dedicated to developing a "do no harm" toolkit in the core areas of EA. This will require interdisciplinary research and substantial investments of resources — investments that we believe to be justified in light of the possibility of counterproductive altruism.

Acknowledgements

For constructive comments and conversations, the authors are grateful to Luc Bovens, Mark Budolfson, Eric Cheng, Devin Christensen, Keith Dowding, Schmulik Nili, Alexandru Marcoci, Susan Pennings, Geoff Sayre-McCord, Katie Steele, Daniel J Stephens, Johannes Treutlein, Anil Venkatesh, and the organisers and audience of the Moral, Social, and Political Theory Workshop at the Australian National University and at the Taiwanese Political Science Association in 2019. Alexandra Oprea is also grateful to Xuejun (Kathy) Jiang for excellent research assistance.

Notes

1. For an introduction to effective altruism, see William MacAskill, *Doing Good Better* (London, 2016), Peter Singer, *The Most Good You Can Do* (London, 2015), as well as the Effective Altruism website: <https://www.effectivealtruism.org/articles/introduction-to-effective-altruism/>
2. Cited in Singer, *The Most Good You Can Do*, p. 4. See also, https://en.wikipedia.org/wiki/Effective_altruism
3. MacAskill, *Doing Good Better*, p. 11.
4. For an overview, see especially the contributions to the *Boston Review's* July 2015 forum on "The Logic of Effective Altruism" (http://bostonreview.net/forum/peter-singer-logic-effective-altruism), as well as Iason Gabriel, "Effective Altruism and its Critics", *Journal of Applied Philosophy* 34 (2016), pp. 457-73, Jeff McMahan, "Philosophical Critiques of Effective Altruism", *The Philosophers' Magazine* 73 (2016), pp. 92-99, and Jennifer Rubinstein, 'The Lessons of Effective Altruism", *Ethics & International Affairs* 30 (2016), pp. 511-526.
5. For other discussions of EA methodology see Antonin Broi, "Effective Altruism and Systemic Change", *Utilitas* 31 (2019), 262-276, Mark Budolfson and Dean Spears, "The Hidden Zero Problem: Effective Altruism and Barriers to Marginal Impact," in Effective Altruism: Philosophical Issues, ed. By Hillary Greaves and Theron Pummer (Oxford: Oxford University Press, 2019), Gabriel, "Effective Altruism and its Critics", especially the discussion of observational bias, quantification bias, and randomized control trials on pp. 6-10, Federico Zuolo,

"Beyond Moral Efficiency: Effective Altruism and Theorizing about Effectiveness", *Utilitas* (2019): 1-14 (online first); Daron Acemoglu, "Forum Response: The Logic of Effective Altruism" *Boston Review* <http://bostonreview.net/forum/logic-effective-altruism/daron-acemoglu-response-effective-altruism>

6. For an overview of the recent debate about the institutional critique, see especially Brian Berkey, "The Institutional Critique of Effective Altruism", *Utilitas* 30 (2018), pp. 143-171, with responses from Alexander Dietz, "Effective Altruism and Collective Obligations", *Utilitas* 31 (2019), pp. 106-115, and Brian Berkey, "Collective Obligations and the Institutional Critique of Effective Altruism: A Reply to Alexander Dietz" *Utilitas* 31 (2019), 326-333.
7. William MacAskill, "The Definition of Effective Altruism," in *Effective Altruism: Philosophical Issues*, ed. By Hillary Greaves and Theron Pummer (Oxford: Oxford University Press, 2019).
8. Berkey, "The Institutional Critique of Effective Altruism," p. 146-7.
9. For an example of a deontological argument in support of EA, see Theron Pummer, "Whether and When to Give," *Philosophy & Public Affairs* (2016): 77-95.
10. We believe EA4 is broadly uncontroversial.
11. For further discussion and critique, see Federico Zuolo, "Beyond Moral Efficiency."
12. Peter Singer, "Famine, Affluence, and Morality," *Philosophy & Public Affairs* 1 (1972): 229–43. Singer describes himself as an effective altruist and is viewed as an inspiration by many of its founding members. The argument proposed by Singer and EAs is largely assistance-based rather than contribution-based. We do not enter into the debate about the relative merits of the two arguments. For the interested reader, we recommend Christian Barry and Gerhard Øverland, *Responding to Global Poverty: Harm, Responsibility, and Agency* (Oxford: Oxford University Press, 2016).
13. Singer, "Famine, Affluence, and Morality," p. 231. Singer distinguishes between a number of different versions of the principle.
14. Many have criticized Singer for overstating the extent of his conclusions from the Shallow Pond, noting that the example instead justifies a version of the following principle: "If we can prevent something (very) bad from happening at a minimal cost to ourselves, and others, then we ought to do it."
15. Ibidem.
16. For a similar discussion, see Theron Pummer, "Risky Giving," *The Philosophers' Magazine* 73.2 (2016): 62-70.
17. There are some who would argue that the magnitude of the harm should be irrelevant. The most famous example of such an argument in the case of certain rather than probabilistic harm would be John M. Taurek, "Should the Numbers Count?" *Philosophy and Public Affairs* 6.4 (1977): 293-316. We side with the critique of this argument presented by Derek Parfit, "Innumerate Ethics," *Philosophy and Public Affairs* 7.4 (1978): 285-301. However, we do not pursue the argument any further here.
18. Tom Dougherty, "Rational Numbers: A non-consequentialist explanation of why you should save the many and not the few," *The Philosophical Quarterly* 63.252 (2013): 413-427. For a compelling case that most deontologists should have no problem using expected utility models when dealing with risky choices, see Seth

Lazar and Chad Lee-Stronach, "Axiological Absolutism and Risk," *Noûs* 53.1 (2019): 97-113.

19. The process generating these outcomes is usually described as random and the entire set of possible outcomes is described as a "random variable." In the case of height, this may represent the genetic and epigenetic factors that generate the combinations of possible human heights. In the case of populations of cities, this may be the combination of economic, social, and political variables that leads to a concentration of individuals in certain urban areas and not others. And in the case of opportunities for doing good, this may simply reflect the underlying distribution of income, economic opportunity, stochastic shocks (e.g. earthquakes, war, etc.). For our purposes, we consider these processes to be a black box that can be usefully described as random and approximated through the tools of statistical inference.
20. The example of the dice roll is highly artificial because the number of possible outcomes is fixed and known in advance. Many of the examples considered include either unbounded distributions or distributions with unknown bounds (e.g. populations, income, wealth, etc.). Although the paper does not discuss the specific problems of statistical inference that emerge when dealing with unbounded distributions, these problems exist and become particularly severe when discussing probability distributions with heavy tails and non-existent higher moments. For the interested reader, we recommend Sergey Foss, Dmitry Korshunov, and Stan Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions* (NY: Springer, 2013).
21. Given that the distribution of height differs by gender, we can focus on the height of men.
22. https://ourworldindata.org/human-height
23. The standard deviation measures the amount of dispersion in the data. A small standard deviation indicates a compact dataset, while a large standard deviation indicates high amounts of dispersion. The standard deviation is calculated as the square root of the variance. The variance is the average of the squared differences from the mean.
24. Tables 1 and 2 draw inspiration from the discussion in Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable* (NY: Random House, 2007): 231-233.
25. A more technical implication of the same two facts is that normal distributions are well-characterized by only two variables: their mean and standard deviation.
26. The tail index is often referred to as a survival function. In the engineering cases from which the terminology emerged, the survival function measures the probability that a unit survives beyond time t. In the cases of interest to us, the survival function represents the probability that a value larger than a given x (or smaller than a given $-$x) can still be found in the given population. A rough way of putting it is that the survival function measures the probability that extremely large (or extremely small) values "survive" or remain available.
27. In the case of the Pareto or power law probability distribution described in Figure 2 above, the right tail index is: $\Pr(X > x) = \left(\frac{x}{x_{min}}\right)^{-(\alpha-1)}$ for $x > x_{min}$, where $\alpha = 1.1$ and $x_{min} = 1$

28. There is a lively debate about this in the economics literature. Scholars differ in terms of what counts as a city, what is the lowest bound of the power law (i.e. x_{min}), and which statistical techniques should be used for the estimation. Note that we are here simply assuming that a power law constitutes a good fit (though a number of other heavy tailed distributions could provide a better fit for the data). In Table 2 above, we assume the probability distribution of cities above a certain population threshold follows the power law $p(X = x) = Cx^{-\propto}$, where $C = (\propto -1){X_{min}}^{(\propto -1)}$, $\alpha = 2.3$ and $x_{min} = 40{,}000$. These estimates come from Mark E.J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics* 46.5 (2005): 323-351. This means that the probability that a city is larger than a given population threshold is equal to $p(X > x) = (\frac{x}{x_{min}})^{-(\propto -1)}$.
29. New York City is the only 99.9-percentile city in our dataset, with 8,398,748 inhabitants as of July 1, 2018. All of the 89 cities in the 80-to-90-percentile range combined have a total population of 15,640,446, slightly less than twice as much. So New York has 46.5 times the population of the average city in the 80-to-90-percentile range. So, supposing charities are like cities, given a choice between a charity which you think is somewhere in the 80-to-90-percentile range and a charity which you think is in the 99.9th percentile, the latter choice will do 4,550% more good per dollar.
30. For details about their approach, see the website <https://80000hours.org/>, as well as Benjamin Todd, *80,000 Hours: Find a Fulfilling Career that Does Good* (CreateSpace Independent Publishing Platform, 2016).
31. Meta-charities such as GiveWell and Giving What We Can provide rankings of charities based on impact assessments rather than the more typically used financial metrics of Charity Navigator and others. For more details, see <https://www.givewell.org/about/>
32. https://concepts.effectivealtruism.org/concepts/prioritization-research/
33. According to the Open Philanthropy Project, it adopts a strategy of hits-based-giving "where a small number of enormous successes account for a large share of the total impact — and compensate for a large number of failed projects" <https://www.openphilanthropy.org/blog/hits-based-giving> Holden Karnofsky (the author of the blog post) argues that there is some degree of overlap with the principles of for-profit investing and provides a link to an article called "Black Swan Farming" that provides the following two key principles of start-ups investing: "(1) that effectively all the returns are concentrated in a few big winners, and (2) that the best ideas look initially like bad ideas." <http://paulgraham.com/swan.html>
34. In some cases, the sample average either takes a very large sample or never converges in value to the true average. In some extreme cases, the mean does not even exist. Even if it does, the central limit theorem does not predict rapid convergence for heavy-tailed variables in the same way that it does for thinner tailed ones.
35. MacAskill, *Doing Good Better*, p. 50.
36. The probability distribution does not have to be symmetrical.
37. Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (New York: Hachette Books, 2020).

38. Holden Karnofsky, "A Conflict of Bayesian Priors", *The GiveWell Blog*, <https://blog.givewell.org/2009/12/05/a-conflict-of-bayesian-priors/> We believe that the underlying assumptions about the shape of the distribution of interventions by effectiveness is one potential source of the conflict of Bayesian priors referenced in the article.
39. MacAskill, *Doing Good Better*, p. 93.
40. Rob Reich, *Just Giving: Why Philanthropy Is Failing Democracy and How It Can Do Better*, Princeton University Press, 2018; Emma Saunders-Hastings, "Plutocratic Philanthropy," *The Journal of Politics* 80.1 (2018), pp. 149-161; Ryan Pevnick, "Democratizing the Nonprofit Sector", *Journal of Political Philosophy* 21.3 (2013), pp. 260–82. For a dissent, see *Against Against Billionaire Philanthropy*. Scott Alexander, "Against Against Billionaire Philanthropy." Slate Star Codex, September 28, 2019. https://slatestarcodex.com/2019/07/29/against-against-billionaire-philanthropy/.
41. Sarah Reckhow and Jeffrey W. Snyder, "The expanding role of philanthropy in education politics," *Educational Researcher 43* (2014), pp. 186–195, at p. 191.
42. We use institutional critique and structural critique as synonyms. We see no principled distinction between the two terms as deployed in the literature.
43. Angus Deaton, "The Logic of Effective Altruism: Response", *Boston Review* <http://bostonreview.net/forum/logic-effective-altruism/angus-deaton-response-effective-altruism>
44. Technological research might be an exception to this rule, though typically "if you didn't invent it, someone else would have" applies.
45. Owen Cotton-Barratt, "Prospecting for Gold", Effective Altruism Handbook, 2nd edition (Oxford, 2016), pp. 16-44, especially pp.19-22; accessed October 14th, 2019 at <https://effectivealtruism.org/handbook>. The HTH is also mentioned as one of three key features of the EA movement in other places on the EA blog. See, for example: <https://app.effectivealtruism.org/groups/resources/effective-altruism-community-building>
46. "Introduction to Effective Altruism." Effective Altruism. Centre for Effective Altruism, June 22, 2016. https://www.effectivealtruism.org/articles/introduction-to-effective-altruism/ and "The Effective Altruism Handbook." Effective Altruism. Accessed November 1, 2019. https://www.effectivealtruism.org/handbook/.
47. Holden Karnofsky, "Hits-Based Giving."
48. Daniele Lantagne, G. Balakrish Nair, Claudio F. Lanata and Alejandro Cravioto, "The Cholera Outbreak in Haiti: Where and How Did It Begin?" *Current Topics in Microbiology and Immunology* 379 (2014), pp. 145-164.
49. UN Report of the Special Rapporteur on extreme poverty and human rights, 2016.https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=20794&LangID=E
50. Joseph A. Lewnard, Marina Antillón, Gregg Gonsalves, Alice M. Miller, Albert I. Ko, Virginia E. Pitzer, "Strategies to Prevent Cholera Introduction during International Personnel Deployments: A Computational Modeling Analysis Based on the 2010 Haiti Outbreak", *PLoS Med* 13.1 (2016), e1001947.
51. TED Radio Hour, "How Can Deserts Turn into Grasslands?" *15 Nov 2013*.

52. Jennifer Rubenstein, "The Misuse of Power, Not Bad Representation: Why It Is Beside the Point that No One Elected Oxfam", *The Journal of Political Philosophy*, 22.2 (2014), pp. 204–230.
53. Nathan Nunn and Nancy Qian, "US Food Aid and Civil Conflict", *American Economic Review* 104.6 (2014), pp. 1630-66.
54. If the true underlying distribution of effectiveness did not have a heavy right tail, it would be extremely unlikely for there to be any examples of this magnitude in the entire world. By analogy, suppose we established radio contact with an alien civilization, and asked them how tall their tallest individual was. If their tallest individual was thousands of meters tall, then either the distribution of alien heights is heavy-tailed, or the aliens are a race of giants. If we independently knew that most of these aliens were only a few meters high, that would almost conclusively rule out the second hypothesis, leaving only the first. Thus, precisely because these cherry-picked examples are orders of magnitude more effective than most interventions, their existence is evidence for the HTH.
55. Idem, p. 50.
56. Toby Ord, "The Moral Imperative toward Cost-Effectiveness in Global Health", *Center for Global Development* 2013. <http://www.tobyord.com/research> A revised version of the essay is also available in *Effective Altruism: Philosophical Issues*.
57. Idem, p. 2.
58. Idem, p. 3.
59. Cotton-Barratt, "Prospecting for Gold", pp. 39-40.
60. Mitchell, GE and Calabrese, TD, "Proverbs of nonprofit financial management", *The American Review of Public Administration* 49.6 (2018), pp. 649-661; Chikoto, GL and Neely, DG (2014). "Building nonprofit financial capacity: The impact of revenue concentration and overhead costs", *Nonprofit and Voluntary Sector Quarterly*, 43 (2014), pp. 570-588; Greenlee, JS and Trussel, JM, "Predicting the financial vulnerability of charitable organizations", *Nonprofit Management and Leadership* 11 (2000), pp. 199-210.
61. Lecy, JD and Searing, AM, "Anatomy of the nonprofit starvation cycle: An analysis offalling overhead ratios in the nonprofit sector", *Nonprofit and Voluntary Sector Quarterly* 44 (2015), pp. 539-563; Gregory, AG and Howard, D, "The nonprofit starvation cycle", Stanford Social Innovation Review (2009), pp. 49-53.
62. The intuition relies on the Central Limit Theorem (CLT) According to the CLT, the probability distribution of the sum of any n independent, identically-distributed random variables with mean μ and variance σ^2 approaches a normal probability distribution as $n \to \infty$. For proof and discussion, see Larry A. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (New York: Springer, 2004), 77-82. For a discussion of the difficulties with applying the CLT to empirical phenomena, see Aidan Lyon, "Why are normal distributions normal?" *British Journal of Philosophy of Science* 65 (2014): 621-649.
63. The tail is particularly heavy when the variance is larger than the mean.
64. Lyon, "Why are normal distributions normal?", 630-633. For further discussion of the process generating a log-normal distribution and illustrations from multiple domains, see Eckhart Limpert, Werner A. Stahel, and Markus Abbt,

"Log-normal distributions across the sciences: keys and clues," *BioScience* 51.5 (2001): 341-352.

65. For a summary of the primary mechanisms generating power laws, see Newman, "Power Laws, Pareto distributions, and Zipf's law."
66. Technically log-normal will only occur if the variables are all positive. If some of the variables might be negative, then the distribution will be a combination of a log-normal distribution and another log-normal distribution multiplied by (-1). The resulting distribution will have two heavy tails instead of one.
67. Effectiveness is measured in lives saved per $10,000 here. We have intentionally chosen a simplistic example to illustrate the point.
68. It is instructive to consider what happens if the underlying distribution is normal, but with very high variance. For example, suppose that the standard deviation is 1000 lives. We would expect most of our red dots to be between -1000 and 1000 where 68.27% of our values are located (or at least between -2000 and 2000 where 95.45% of the data is located). What this means is that if instead our random sample is scattered across many orders of magnitude (as in Figure 4a), that would constitute evidence against a normal probability distribution.