12

The Hidden Zero Problem

Effective Altruism and Barriers to Marginal Impact

Mark Budolfson and Dean Spears

1. The hidden zero problem: An initial illustration

Suppose for the sake of argument that practitioners of effective altruism (EA) are completely correct about the amount of good done by their top charities. Is there is any further reason to worry that giving to these charities is not an effective way of doing good?

It turns out that there is, and a real-world illustration of the worry is based on the fact that several billionaires closely follow the recommendations of EA, and can credibly commit to "top up" the revenues of many of the charities that EA recommends, in order to ensure that those charities meet operating budget targets. These facts are readily knowable based on public information (see references later in this section). In some previous years, because the amount of plausible shortfall in all of the top-ranked EA charities *combined* was only several tens of millions of dollars per year, and because this group of billionaires had the capacity to top up such charities to erase *much more* than that level of shortfall and arguably seemed to commit to making sure that no real funding needs went unmet, during that time the expectation associated with donations to these leading EA charities of, say, a magnitude of $1,000 was arguably that no difference was made to the operations of the charity, and slightly less money was transferred from the billionaires to those charities.[1] If so, the expected effect of a donation to an EA recommended charity during this time period, even assuming the charities did just as much good as EA advisors claimed, would have been merely to transfer money to a billionaire in the United States, and accomplish nothing for the global poor.

In recent years, this situation has likely changed in connection with at least one and perhaps more leading EA charities, which may now have much more capacity to scale up operations quickly, as discussed near the end of this section. But an important philosophical point remains, namely that this example of how donations by normal people could have zero positive effect illustrates what we call the

[1] Here we bracket the possibility that the expectation could zero because it is knowable that e.g. donations less than or equal to $1,000 amount to insignificant digits in all of the relevant decision-making by charity organizations, billionaires, and others—compare Budolfson (2018).

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

hidden zero problem, which is that the marginal effect of an action often depends on a hidden parameter that is ignored in widespread EA analyses of efficacy, where that parameter might realistically have the value of zero in a way that ensures that individual actions are not efficacious. In the billionaires example, the hidden parameter is the marginal effect of a donation on the operating budget of a charity; the phenomena of billionaires topping up charities to predetermined targets illustrates how such a parameter could be zero even if all of the other parameters that EA evaluators track actually have the positive values that EA proponents claim—and if such a parameter is zero, then the marginal effect of a donation can be zero regardless of how positive the other parameters are the EA proponents track.

In the next section, we articulate this hidden zero problem more formally, and in later sections we provide a number of additional important examples of this problem for EA. We emphasize that there is no argument here that top-rated EA charities should not exist or should pursue other activities. We are instead focused narrowly on the question of what the expected effect is of an individual's donation, and what an ordinary individual donor has reason to do. In the rest of this section, we consider the dynamics of the particular billionaires example that we introduced above in more detail.

The possibility of billionaires standing ready to top up top-rated charities is occasionally acknowledged by EAs, but is then quickly dismissed as not relevant to reality.[2] The only commentator we know who has taken the issue seriously is Iason Gabriel. However, Gabriel believes that the billionaires example is not ultimately a big deal on the grounds that if individual donations do in fact reduce the amount that billionaires donate to top-rated EA charities, then that simply means that those billionaires will then donate the money saved to the next best charities instead—thereby ensuring that an individual's donation does have some significant positive marginal effect, albeit slightly less than the effect EA proponents claim.[3]

However, it is an empirical claim that billionaires have invested the same amount in other slightly less effective charities instead. Unfortunately, that claim appears false in light of publicly available information that shows that in the past, EA-directed billionaires have for principled reasons not been willing to redirect excess money to charities "further down the list". In what follows, we detail the publicly available evidence for this, which suggests that the billionaires problem may well have created a hidden zero in the recent past, even if the situation has now evolved.

The importance of the problem becomes clear from a careful study of what might be called the recent "pivot toward billionaires in EA", in which billionaires now dominate the funding for top EA charities, together with the fact that there

[2] Compare MacAskill (2015, p. 119).
[3] Gabriel (2016) introduces the billionaires problem to the literature.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

are only a handful of top ranked EA charities, many of which have a surprisingly low limit (by the organization's own account) to the resources it can absorb and genuinely turn into welfare gains. The leading example of the pivot toward billionaires is provided by Good Ventures, a foundation run by billionaires Cari Tuna and Facebook co-founder Dustin Moskovitz, which in the recent past, by its own claims, had so much money to invest that it could not find nearly enough opportunities to invest its vast resources consistent with the EA criteria it endorses as a constraint on making donations. In light of this, until recently it is possible that it reliably filled any *genuine* need for resources that could be converted into welfare gains by top EA charities.[4] (Throughout we understand "top-rated EA charities" to be the top charities recommended by the EA evaluator GiveWell.org.)

To understand in detail the way the pivot toward billionaires may have undermined the marginal effect of individual donations in previous years, it is important to see that Good Ventures *alone* represented over *two thirds* of all money moved to EA charities in 2015, as tracked by EA advisor givingwhatwecan.org.[5] Beyond this, the most important facts here (reported by Good Ventures, GiveWell, and others) are that:

(a) Good Ventures represents *only one* among a growing number of EA-focused "billionaires".[6]
(b) Good Ventures now funds and collaborates with the dominant EA advisor GiveWell in investment and strategic decision-making, and so Good Ventures makes its decisions in a way that perfectly tracks the dominant EA consensus.[7]
(c) Good Ventures *alone* has so much money that, by their own lights in several previous years they have been able *by themselves* to easily meet all of the funding needs of *all* of the charities that are deemed to be sufficiently effective to be worthy of investment on EA grounds, while still not being able to spend *nearly* as much money as they would like because they judge

[4] For more detail, a good place to start is two blog posts from GiveWell, Karnofsky (2015b), and Hassenfeld and Rosenberg (2015), and one blog post from Good Ventures, Karnofsky (2015a). A slightly older but important discussion superseded by the preceding is Karnofsky (2014).

[5] MacAskill (2016).

[6] For example, the Effective Altruism Global 2015 conference was advertised as "the largest ever convening of thought leaders, entrepreneurs, billionaires, CEOs, investors, and scientists, and more who are applying reason and data to tackle the world's biggest challenges", with a raffle competition to "win a ticket to EA Global (Effective Altruism Global) featuring Elon Musk". (Josh Jacobson, "Announcing the Doing Good Better Giveaway", Effective Altruism Forum, online at http://effective-altruism.com/ea/kn/announcing_the_doing_good_better_giveaway, accessed 8 April 2016 (same access date for other citations below unless context makes clear otherwise.)

[7] For example, a post on the Good Ventures website by Holden Karnofsky, at the time the director of both GiveWell and the Open Philanthropy Project, begins by stating that "Throughout the post, 'we' refers to GiveWell and Good Ventures, who work as partners on the Open Philanthropy Project", Karnofsky (2015a). As a result, we here sometimes use "GiveWell" to refer to what are, on paper, two organizations, GiveWell and the Open Philanthropy Project.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

that after their investments there are no more good EA opportunities for them to invest in.[8]

(d) In some previous years, arguably Good Ventures committed to meeting all the funding needs of the top-ranked EA charities that were essentially connected to those charities' actual activities of doing good.[9]

Given these publicly available facts, in some previous years it is arguable that one should have expected that among charities that are judged by EA to be top charities, any would-be shortfall in donations that would have any actual important impact on the operations of that charity would have been offset by funding from Good Ventures *alone*—which is, again, only one among a growing number of deep pockets that are closely following EA advice.

Further, contrary to the argument that the billionaires example is not a big deal, in the past Good Ventures has reported that it does not redirect excess money to projects "further down the list". On the contrary, Good Ventures—like many in the EA community—has explicitly endorsed the strategy of not redirecting money to charities further down the list, because it operates on the explicit principle that the next charities down the list are not worth giving to, and instead money is better saved or invested in other strategic initiatives—and those investments still leave Good Ventures in a position where it is unable to spend as much money as it would like.[10]

A further possibility is that billionaires might save the money that they do not donate today with the intention to donate to another high-quality charity later. However, even in a case where this is true, and even assuming inflation adjusted dollar-for-dollar substitution to later giving, this would not neutralize the billionaires example, because the effectiveness would still be substantially less than EA evaluations suggest for a number of reasons: the future investment is by hypothesis less effective (since otherwise the billionaire would donate now); wellbeing is improving quickly in poor countries, which may be expected to reduce the value of EA opportunities in the future; the marginal product of EA activities may be

[8] From the Good Ventures blog: "Good Ventures hopes to give away several billion dollars over the coming decades, which—when accounting for likely investment returns—would imply hundreds of millions of dollars per year in grants for an extended period of time at peak giving. In 2014, Good Ventures gave ~$15 million to GiveWell's top charities and an additional ~$8 million based on Open Philanthropy Project recommendations. In other words, their current level of giving is nowhere near where they hope it will eventually be" Karnofsky (2015a).

[9] For both of these aspects of their strategy, see Karnofsky (2015b). It is important to note that only a part of what GiveWell calls "room for funding" represents a need for funds that have an important impact on those charities actual activities of doing good—for discussion of this, see Hassenfeld and Rosenberg (2015).

[10] See Tuna (2015) announcing Good Ventures grants, which tracked the recommendations given to it by Hassenfeld and Rosenberg (2015) to focus on funding on only the top-rated EA charities, following the advice that it is better to save resources for future investments than invest in charities that are not top ranked.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

expected to decline as they become well-known and the world becomes richer with more altruism dollars to invest; the billionaire might not actually donate in the future for many reasons including death, decreased control over assets, new taxes or economic loss, or for any other reason.

In light of these considerations, together with other publicly available facts about the decision-making strategy of Good Ventures and other billionaires, we believe that in some previous years ordinary donations to top EA charities may not have done much good. If the donations of ordinary individuals accomplished anything, it may have been to reduce the amount that billionaires give to EA causes, increasing the bank account balances of these billionaires or their foundations. Thus, the billionaires example appears to be a significant problem: individual donations to top-rated EA charities may well have done no good for this reason at some times in the recent history of EA, and in fact may have done harm insofar as one agrees that a transfer from ordinary people to billionaires is harm—especially problematic if those ordinary people are misled about the nature of the transfer they are making.[11]

At the same time, these empirical dynamics are in flux, and the billionaires example could no longer be a problem if there is a change in capital allocation dispositions by billionaires, or if there is a large increase in the capacity of top charities to turn additional capital into wellbeing. For example, GiveWell indicates that starting in 2017 the charity GiveDirectly increased its capacity to turn capital into wellbeing, in a way that could arguably make the billionaires problem not as relevant to that specific charity (even if it remained relevant to other top EA charities).[12]

In what follows we set aside the specifics of the billionaires example, partly because the underlying empirical facts are unstable, for reasons just noted. Instead, we focus on explaining why there are likely to be many other hidden zero problems for EA elsewhere that arise from very different sources that we identify below, where those different sources are also more timeless and empirically stable than the billionaires problem. Thus, the billionaires problem provides a compelling

[11] For one way of developing a fairness-based objection to effective altruism on this sort of grounds, see Gabriel (under review).

[12] See the section on 'room for funding' in GiveWell's 2017 evaluation of GiveDirectly: https://www.givewell.org/charities/give-directly/january-2017-version#Roomformorefunding. Proponents of EA generally tend to put a more optimistic spin on room for funding and interaction with large donations; for recent discussion see: https://app.effectivealtruism.org/funds/why. A more pessimistic view is that room for funding estimates do not necessarily exclude amounts that EA evaluators know will be filled by Good Ventures or other billionaires, and beyond that, any gaps that remain by EAs' own lights also do not have nearly as high marginal product as the gaps they recommend the billionaires fill, partly because remaining gaps are based not on actual immediate need for funding for activities, but rather on increasingly speculative estimates of how strategic and capacity-building decisions in the further future might shake out differently if they have extra dollars now above and beyond what they actually have the capacity to use now—e.g. see http://blog.givewell.org/2015/11/18/our-updated-top-charities-for-giving-season-2015.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

and easy-to-understand initial illustration of a more fundamental and more timeless worry about the efficacy of EA donations, which is our focus in the remainder of the paper.

2. Analyzing the nature of the hidden zero problem, and the correct fundamental equation for EA vs. equations actually used in EA evaluation of charities

In this section, we articulate a fundamental analysis of the marginal effect of donations, which provides a more formal conceptualization of the hidden zero problem that was illustrated by the billionaires problem above. This analysis more clearly explains why donations that score very well on the existing metrics endorsed by EA might still have zero marginal effect (or net negative effects). By clearly distinguishing a number of distinct factors that are often ignored by EA, the equation also helps to clarify the logical space of factors relevant to the evaluation of charitable investments, as well as the logical space of objections to the effectiveness of specific charities.

Here is the equation we take to summarize the dynamics relevant to the marginal effect of a donation to a specific charity C to the lives saved by C:

$$\left(\frac{\Delta Lives\ Saved\ by\ C}{\Delta Donation\ to\ C}\right)=\left(\frac{\Delta Lives\ Saved\ by\ C}{\Delta Activity\ by\ C}\right)*\left(\frac{\Delta Activity\ by\ C}{\Delta Budget\ of\ C}\right)*\left(\frac{\Delta Budget\ of\ C}{\Delta Donation\ to\ C}\right)\cdots$$

(Correct EA)

The hidden zero problem arises from the possibility that one or more of the terms on the right-hand side could be zero, which would imply that the marginal effect of a donation (the left-hand side) to the lives saved by that charity is also zero regardless of how large the other terms are. More generally, the problem is one of "hidden elasticities": EA evaluations are generally blind to the fact that some terms in this equation are even relevant to a correct analysis of marginal impact—i.e. the right-most term. The ellipsis at the end indicates that in specific instances a complete equation will require a further multiplicative step each time the activity is passed along to another person or task along the chain from altruistic donor to final beneficiary. The billionaires problem illustrates how the expected change in budget per change in donation by ordinary non-billionaires could be zero, and how such a hidden zero could exist even if we assume that EA practitioners are entirely correct about the amount of good done by the charities they recommend. (Here and in what follows, for ease of exposition we use "lives saved" as intuitive shorthand for what ultimately makes for better or worse outcomes, so as to bracket the independently controversial issue of what should be valued and how.)

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

A further complication is that the equation above will not be fully correct insofar as there are spillovers from your donation to *C* onto the activities of other charities, and spillovers beyond *C* onto anything else that affects outcomes. To capture all those, one would have to calculate the change in good done due to everything other than *C* for a change in donation to *C*, and add those effects as in the right-most term here:

$$\left(\frac{\Delta Lives\ Saved}{\Delta Donation\ to\ C}\right) = \left(\frac{\Delta Lives\ Saved\ by\ C}{\Delta Donation\ to\ C}\right) + \left(\frac{\Delta Lives\ Saved\ other\ than\ by\ C}{\Delta Donations\ to\ C}\right)$$
$$\left(\text{Marginal Effect of a Donation}\right)$$

For example, Gabriel's reply to the billionaires problem can be understood as arguing that the right-most term added here is importantly positive because of the spillover of your donation to *C* onto the additional lives saved by the next best charities down the line. We've presented some reasons above for doubting that this specific spillover has the magnitude Gabriel assumes. More importantly, in the next section we'll cite arguments from Angus Deaton that the right-most term here is generally negative because of unintended side effects of charities beyond the lives they are directly focused on improving.[13]

In the rest of this section, we contrast the earlier equation Correct EA with a number of different equations that are often used in actual EA evaluations. This helps clarify why the dynamics behind the hidden zero problem matter, and why structuring analyses more intentionally on Correct EA can improve the accuracy of EA evaluations and EA thinking. In later sections, we provide more stable sources of hidden zero problems for EA beyond the billionaires problem, and we identify a number of different fundamental mechanisms that lead to these problems.

To begin, it is worth noting that there are bad methods of charity evaluation that should not be mistaken for EA evaluation. At the top of the list are evaluators such as Charity Navigator that base evaluations primarily on metrics such as percentage of budget spent on administrative expenses, which is inappropriate as any sort of measure of doing good. To see why this is inappropriate, consider a charity that does active harm with every dollar donated, but also spends a very low percentage of its budget on administrative expenses. This "charity" will be ranked very highly based on the percentage of its budget spend on administrative expenses. Now compare this to a second charity that must spend a higher percentage of its budget on administrative expenses, because this is necessary for it to operate in a domain where it then is able to do enormous net good per dollar with the rest of its budget. Obviously, the second charity would be engaged

[13] See factor (c) below.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

in more effective altruism than the first, even though the first would score better on the inappropriate metric of percentage of budget spend on administrative expenses.[14]

With this in mind, a first pass at a genuine metric for evaluating charities on effective altruist grounds, we might consider the following:

$$\left(\frac{\Delta Lives\ Saved}{\Delta Donation}\right)=\left(\frac{Total\ Lives\ Saved}{Total\ Budget}\right) \tag{EA1}$$

Equation EA1 could then be used to estimate the average cost per unit of good associated with different charities, which might then be used, in a particularly crude form of EA analysis.

A more detailed analysis might add an additional term that allows such an analysis to be more readily connected to empirical studies:

$$\left(\frac{\Delta Lives\ Saved}{\Delta Donation}\right)=\left(\frac{Total\ Activity}{Total\ Budget}\right)*\left(\frac{Total\ Lives\ Saved}{Total\ Activity}\right) \tag{EA2}$$

Using this equation EA2, the term Total Lives Saved/Total Activity might be investigated with RCTs and the like, and the term Total Activity/Total Budget can be estimated in a straightforward way.

To see the problem with equations EA1 and EA2, which might be called "average effect metrics", we need only note that marginal effect is not the same thing as average effect—where in connection with EA, we are interested in marginal effect, namely, the actual difference that would be made by additional investment in a charity.

At its current best, EA analyses sometimes rely on a more sophisticated equation than EA1 and EA2, where this more sophisticated equation does not simply equate the marginal effect of additional charity with the average effect. In particular, GiveWell and other leaders in current best practices for EA evaluation can be understood as aiming to use the following more sophisticated marginalist metric:

$$\left(\frac{\Delta Lives\ Saved}{\Delta Donation}\right)=\left(\frac{\Delta Lives\ Saved}{\Delta Activity}\right)*\left(\frac{\Delta Activity}{\Delta Budget}\right) \tag{EA3}$$

In this equation, the (marginal) effect of a donation is understood as the change in lives saved per change in activity (at the margin) (e.g. marginal lives saved per additional bed nets distributed) multiplied by the change in activity per change in

[14] Singer (2015); MacAskill (2015).

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

budget (at the margin). This equation is on the right track because it invokes actual elasticity terms (i.e. terms that quantify the percentage change in one variable that will result from a change in another) on the right-hand side of the sort relevant to marginal effects, which is an improvement over the explicitly averagist effect metrics of EA1 and EA2.[15]

However, even if one assumes for the sake of argument that EA is using EA3 and is entirely correct about the terms on its right-hand side, and is thus entirely correct about the good done by its top-rated charities, the hidden zero problem is that it could still be dubious that *donations* to those charities would do any good, because of the possibility that the term Δ *budget*/Δ *donation* could still be zero (i.e. that zero might be the correct value of that term in Correct EA above). Furthermore, EA evaluators' methods often invoke estimations and reasoning about the other elasticities in EA3 that make their actual method better represented by equation EA2 above. This is true, for example, as EA evaluators often rely on average effect metrics such as the total activity of an organization divided by its total budget as a proxy for the marginal effect of additional lives saved per additional budget. And note that despite frequent discussions of *crowdedness*, *tractability*, and *impact* by EA evaluators, those notions do not play much of a role in the actual spreadsheets where evaluations are performed—and even if they were incorporated into the spreadsheet fully, they would not remove the hidden zero worry that e.g. Δ *budget*/Δ *donation* could be zero. Finally, notions of *crowdedness*, *tractability*, and *impact* are in any event highly imperfect proxies for the marginalist notions they are intended to track, as one of us argues in another paper.[16] To verify that we are not being uncharitable or misunderstanding EA analyses, the reader can compare these claims to the actual spreadsheets used by GiveWell and other EA sources in charity evaluations.[17]

Having now analyzed the nature of the hidden zero problem and, more fundamentally, the marginal effect of donations and the problem of "hidden elasticities", in the remainder of this chapter we examine two of the elasticities in the right-hand side of the Correct EA equation in more detail. We highlight empirically stable mechanisms identified by economics and other disciplines that provide reason to worry that Δ Lives Saved/Δ Activity and Δ Budget/Δ Donation could be hidden zeros (or worse). We consider these in turn.

[15] For an introduction to the methods of leading EA evaluators, see: MacAskill (2015), https://www.givewell.org/how-we-work/criteria, https://www.givingwhatwecan.org/research/methodology, and http://www.openphilanthropy.org/research/our-process. Of particular interest are GiveWell's explicit cost-effectiveness calculations in spreadsheets available at: http://www.givewell.org/international/technical/criteria/cost-effectiveness/cost-effectiveness-models. The reader can judge the extent to which these EA evaluators are using methods more akin to EA1, EA2, EA3, or Correct EA—we submit that their methods are often closest to EA2. For more on the ethical dimension of the argument, see Singer 1972, Singer 2009, Lichtenberg 2013, Singer 2015, and Budolfson under review b.

[16] Budolfson (under review a).

[17] Givewell 2015 and Budolfson (under review a).

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

3. Arguments that Δ Lives Saved/Δ Activity could be a hidden zero or worse: Evidence that RCTs may not be representative of future results and other empirical considerations

Among the evidence that the EA community cites, randomized controlled trials (RCTs) are of central importance and are often cited by EA as the "gold standard" of evidence.[18] However, Nobel Laureate Angus Deaton, Nancy Cartwright, and others have offered a critique of conclusions about effectiveness that depend on the kind of quick reliance on RCTs that is common in the EA community.[19]

The core of the critique is that there is a large inferential gap between the RCTs that EA depends on, and the conclusions EA draws from them. The basic objection is that when EA concludes on the basis of an RCT that an intervention would be highly effective if scaled up and deployed widely, the following facts (*a*) and (*b*) generally prevent that conclusion from being supported by the evidence:

(a) We don't have reason to think the intervention is going to work even when scaled up within the location of the RCT, partly because the equilibrium that results from a very large number of such interventions might have very different properties from the one that emerges from a handful of such interventions in an RCT (this is one way RCTs, like other causally well-identified empirical studies, often lack external validity—in this case, by lacking generalizability to additional interventions in the same context).

(b) We don't have reason to think that such an intervention would have similar positive effects elsewhere (as opposed to negative effects) (this is another way RCTs often lack external validity—in this case, lack of generalizability to interventions in different contexts—i.e. it may not be generalizable to other populations/locations).

What works in one village might not work in a neighboring village, and it certainly might not work in another region where people have very different customs and societies, and where there are empirically different background facts. Instead, the intervention could do harm. For example, a program that is verified with an RCT to promote latrine use (rather than open defecation) in largely-Muslim Bangladesh could discourage latrine use in a Hindu part of neighboring India, just a few miles away.[20]

In this way, the truth in some cases could be worse than a hidden zero—instead, deploying the intervention could do net harm rather than merely no good, consistent with the internal validity of the RCT that is used by EA to conclude that it

[18] https://blog.givewell.org/2012/08/23/how-we-evaluate-a-study
[19] Cartwright and Hardie (2012); Deaton and Cartwright (2017).
[20] Coffey and Spears (2017).

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

would do good. In other words, (a) and (*b*) draw attention to ways in which Δ Lives Saved/Δ Activity could be a hidden zero or worse—or at least close to zero in a way that undermines EA evaluators' conclusions—consistent with RCT results such as those cited in connection with leading EA charities. For real-world examples, see debates about whether EA recommendations of deworming charities have been based on flawed inferences from RCTs,[21] whether EA recommendations on cash transfers have been based on flawed RCTs that ignored their longer-term negative side effects,[22] whether evidence-based policy recommendations on sanitation are based on flawed inferences from RCTs,[23] and others. To be sure, the problems of internal and external validity are well-understood (although not overcome) by the best econometric practitioners of the development economics literature. The current point is that limits to external validity and the barriers to generalization may not be explicit in any particular study, and that they are ordinarily overlooked in the actual practice of EA evaluations.[24]

Deaton also argues that an additional important factor operates through politics and institutional development:

> *(c)* we have reason to expect large-scale deployment of EA interventions to have negative side effects beyond (*a*) and (*b*) that cannot practically be measured by RCTs.

For example, Deaton believes that even public health interventions that genuinely save lives tend to have longer-term negative consequences by preventing the evolution of public health institutions and other stepping stones to good governance and self-sufficiency within the society that receives the EA treatment. Investments by charities also tend to unintentionally benefit powerful oppressors in society, who are often the main forces standing in the way of social progress. In this way, even the best large-scale interventions tend to retard an entire society's escape from deprivation, as these are the key factors for escape. If the cost of delaying an entire society's escape from deprivation in such a way were quantified, Deaton seems to believe that we should expect the harm done to outweigh the lives saved even by the most promising EA interventions.[25]

On the basis of all of these considerations, Deaton generally opposes the recommendations of EA evaluators, which are based on what he sees as overly quick inferences from RCTs—as Deaton puts it, "If it were so simple, the world would already be a much better place. Development is neither a financial nor a

[21] Humphreys (2015); Berger (2015).
[22] Haushofer and Shapiro (2018); Ozler (2018); compare the earlier short-run results in Haushofer and Shapiro (2016).
[23] Hammer and Spears (2016); Coffey and Spears (2018).
[24] Cartwright and Hardie (2012), Deaton and Cartwright (2017), Bates and Glennerster (2017).
[25] Deaton (2013).

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

technical problem but a political problem, and the aid industry often makes the politics worse."[26] Instead, he joins many other leading economists in arguing that the best bet to help the global poor is to try to change international policies that handicap their growth and equitable development, particularly agricultural and trade policies.[27]

A full empirical test of Deaton's conclusions is beyond the state of econometric science and the data available. So, we do not take a position here on Deaton's conclusions about what truly effective altruism would require. Here we merely note that Deaton, Cartwright, and others' objections to the use of RCTs identify timeless sources for potential hidden zeros or worse in the face of even well-conducted RCTs that EA evaluations take as the "gold standard" of evidence.[28]

4. Arguments that Δ Budget/Δ Donation could be a hidden zero: Principal-agent problems and other empirical considerations

Are there empirically stable reasons why Δ Budget/Δ Donation could be a hidden zero? In this section we draw on theoretical and empirical literature from economics to show that it is realistic that this could be a hidden zero even in a situation where an organization's budget is known not to be topped up to funding targets due to the incentives that fundraisers generally have.

Specifically, here we identify a novel mechanism for donation crowd-out: the principal-agent problem of an organization's fundraising. Principal-agent problems arise when principals (e.g. directors of an organization) can only imperfectly monitor the efforts of agents (e.g. employees, contractors)—which is almost always the case in an actual organization. Because agents often have different goals than principals, in these cases it is likely, other things being equal, that agents will be motivated to act in their own best interests, contrary to the goals of the organization that are defined by its principals.

In any sufficiently large development organization, to be a candidate for EA's attention, a managerial *principal* who is responsible for the overall direction of the organization is likely to cooperate with *agents* in the organization of multiple types: at least two types are program implementation agents and fundraising agents. It is a special property of international charities, unlike many businesses, that implementation and revenue-collecting agents can be different people, perhaps located on different continents, and never encountering one another in person.[29]

[26] Deaton (2015). [27] See Stiglitz (2003).
[28] For additional discussion, see Budolfson and Spears under review.
[29] Contrast this with the case of a retail business that is paid precisely when it provides a service to its customer, so fundraising and service provision are necessarily linked.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

In international development, the principal-agent challenges for implementation agents are well-known and well-studied.[30] Indeed, because implementation principal-agent relationships are often a part of the program design being evaluated as a part of a development project, the EA movement explicitly considers these relationships in selecting projects and they are at the heart of the public advocacy by proponents of evidence-based development policy.[31]

In contrast, the *fundraising* principal-agent problem receives little attention in the development economics literature, and almost no attention in the EA literature. However, agency problems may be at least as important in fundraising. In many charities, fundraising is done by dedicated staff who report to organization principals. Fundraisers are in some way incentivized to successfully raise funds. This incentive could take various forms:

- **Fixed target.** Fundraisers are paid a salary that is independent of the amount of money they raise, except that they are fired if they do not raise enough funds in a specific period.
- **Flexible target.** Fundraisers are paid a fixed salary, and the probability of being fired is decreasing in the amount of funds they raise.
- **Sharecropping.** Fundraisers "sharecrop" with the charity, keeping a fixed percentage of the funds they raise.
- **Billionaire's charade.** A billionaire has promised to ensure the fundraising operation meets the principal's target budget; the fundraising continues merely to save the billionaire some money and to preserve the appearance of a normal charity.

The consequences of the principal-agent arrangement for effective altruists depend on its details. For example, in the sharecropping case, the elasticity of the organization's budget with respect to a donation is less than one by the amount of the sharecropping. In the fixed target case the elasticity could approach zero: if effort is costly, then (abstracting away from risk aversion) fundraising agents would always collect precisely their target, and a surprise donation would be entirely captured by the fundraiser in the form of reduced effort, with no extra money passed on to the organization.[32] This would imply that in the fixed target case the marginal benefit of a donation in terms of lives saved is zero, no matter how effective the organization's program is at its development goals, just as in the billionaire's charade case.

[30] Chaudhury et al. (2006); see also World Bank (2004). [31] Banerjee and Duflo (2011).
[32] See the paper by James Snowden (Chapter 5 in this volume) for a perspective on risk aversion and effective altruism.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

An existing but young empirical literature has estimated the value of Δ Budget/Δ Donation for a number of different kinds of charities and other entities. Naturally, like any set of empirical studies, this literature contains research of varying persuasiveness and immediacy of application to the elasticities that EA evaluators need to know. The table below presents a set of estimates from the literature of the effect of revenue (of various sources available for empirical study) on organizations' budgets:

Source	Method	Elasticity
Andreoni et al (2014)	effect of UK government grants, matching on charity score	depends on size; >1 for smallest
Kingma (1989)	effect of government grants on donations to US public radio	0.865
Heutel (2014)	effect of private donations on US government grants	small, but evidence inconclusive
Andreoni and Payne (2011)	effect of government grants; panel data on US charities	0.25
Andreoni and Payne (2012)	effect of government grants; panel data on Canadian charities	0 (or negative)

This is not an exhaustive list, nor do we necessarily endorse the empirical methods of these papers. In particular, one inapplicability of many of the studies in the table is that they focus on government grants, rather than small private donations, because large grants are particularly amenable to the techniques of causal identification. These estimates may or may not generalize well to EA evaluation; assessing such generalizability would be an important goal of further investigation.

Despite those limitations, we believe three conclusions are clear from the table:

- Some estimated elasticities are much below 1 (where 1 would imply that an extra donation translates into an increase in the organization's budget exactly dollar-for-dollar); these studies therefore give evidence that the problem we highlight could be a large practical concern.
- The estimated elasticities vary radically across studies; these studies do not give us confidence that the elasticity is in fact any particular number.
- Some studies present evidence that the elasticity varies across organizations; this is theoretically expected, and suggests that EA evaluations need organization-specific estimates.

In particular, the empirical literature includes estimates of *zero*. In cases where this is true, the additional effect of a donation would be zero—no matter how effective an organization's programs are and no matter how rigorous and

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

generalizable the evidence of a program's effectiveness is—because the donation would have no effect on the budget or extent of the program implemented. This is not a mere theoretical possibility: it is quantitatively suggested by at least some of the empirical estimates in the literature. If these estimates should be considered wrong or inapplicable, it is important to understand why.[33]

5. Conclusion

EA evaluators point to many facts that seem to suggest opportunities for ordinary people to improve the lives of the world's poorest. But whether these are actual opportunities to improve lives depends on factors highlighted by the Correct EA equation above that have not previously been considered in EA analysis. If any one of the terms in that equation is a hidden zero, then the product is zero, and an altruistic gift is likely not effective. By examining two links in the chain of elasticities within the equation in detail (namely, the change in lives saved that results from change in activity, and the change in organizational budget that results from a change in donations) we have seen that theoretical and empirical literature from economics and other disciplines gives reason to be concerned that, in many cases of practical relevance, some of these terms are in fact hidden zeros, or worse. As a result, even if one agrees with the facts highlighted by existing EA evaluations, there is room to worry that donations to those charities might still do no good or even be harmful on balance.

In sum, the equations above describe the marginal effect of donations, and highlight neglected factors that are relevant to correct consequentialist analysis.[34]

[33] In addition to the references in the table, see also Andreoni and Payne 2003, Duncan 2004, Bernheim 1986, and Warr 1982. Since we first presented this paper, EA evaluators have introduced a crude estimate of the impact of EA funding on the revenues of EA charities and other charities. This is a positive step in the direction of capturing some of these dynamics; however, it is aimed at only one small class of potential hidden zeros, and does not attempt to quantify a large range of others, such as unintended negative side effects of the sort discussed by Deaton, or the interactions within EA funding discussed in the first section in connection with the pivot toward billionaires. Furthermore, even within the class at which it is aimed, it currently tends to be based on judgmental estimates of the relevant effects, rather than empirically quantified estimates. Nonetheless, it is a model of how EA estimates can be improved in practice. See https://blog.givewell.org/2018/02/13/revisiting-leverage/.

[34] Thanks to Elizabeth Ashford, Anne Barnhill, Alexander Berger, David Boonin, Luc Bovens, Emily Clough, Sarah Conly, Jonathan Courtney, Diane Coffey, David Faraci, Seb Farquhar, Iason Gabriel, Hilary Greaves, Michelle Hutchinson, Alison Jaggar, Peter Jaworski, Will MacAskill, Sarah McGrath, Theron Pummer, Rob Reich, Ben Sachs, James Snowden, Daniel Wodak, and audiences at Bowling Green, the London School of Economics, the University of St. Andrews conference on the philosophical foundations of effective altruism, the Georgetown Business school workshop on methodology in applied ethics, and students in Princeton's Introduction to Moral Philosophy course taught by Sarah McGrath.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

References

Andreoni, James, and Abigail Payne. 2003. "Do government grants to private charities crowd out giving or fund-raising?" *American Economic Review* 93 (30): 792–812.

Andreoni, James, and Abigail Payne. 2011. "Is crowding out due entirely to fundraising? Evidence from a panel of charities." *Journal of Public Economics* 95 (5): 334–43.

Andreoni, James, and Abigail Payne. 2012. "Crowding-out charitable contributions in Canada: New knowledge from the north." Working Paper No. w17635. *National Bureau of Economic Research*.

Andreoni, James, Abigail Payne, and Sarah Smith. 2014. "Do grants to charities crowd out other income? Evidence from the UK." *Journal of Public Economics* 114: 75–86.

Banerjee, Abhijit. and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. PublicAffairs.

Bates, Mary Anne, and Rachel Glennerster. 2017. "The Generalizability Puzzle". *Stanford Social Innovation Review*, Summer: 50–4.

Berger, Alexander. 2015. "New Deworming Reanalyses and the Cochrane Review." Givewell Blog. Available at http://blog.givewell.org/2015/07/24/new-deworming-reanalyses-and-cochrane-review.

Bernheim, Douglas. 1986. "On the Voluntary and Involuntary Provision of Public Goods." *American Economic Review* 76 (4): 789–93.

Budolfson, Mark. Under review a. "Utilitarian Virtues of Boring Low-Hanging Fruit, Even When Investing Many Millions."

Budolfson, Mark. Under review b. "Global Ethics and the Problem with Singer and Unger's Ethical Argument for an Extreme Duty to Provide Aid."

Budolfson, Mark. 2018. "The Inefficacy Objection to Consequentialism, and the Problem with the Expected Consequences Response." *Philosophical Studies* 176 (7): 1711–24.

Budolfson, Mark, and Dean Spears. Under review. "Mapping the Empirical Objections to Effective Altruism."

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. Oxford: Oxford University Press.

Chaudhury, Nazmul, et al. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *The Journal of Economic Perspectives* 20 (1): 91–116.

Coffey, Diane, and Dean Spears. 2017. *Where India Goes: Abandoned Toilets, Stunted Development and the Costs of Caste*. HarperCollins.

Coffey, Diane, and Dean Spears. 2018. "Implications of WASH Benefits Trials for Water and Sanitation." *Lancet Global Health* 6: 615.

Deaton, Angus. 2015. "Response to Effective Altruism." *Boston Review*. Available at http://bostonreview.net/forum/peter-singer-logic-effective-altruism.

Deaton, Angus. 2013. *The Great Escape*. Princeton University Press.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

Deaton, Angus, and Nancy Cartwright. 2017. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21.

Duncan, Brian. 2004. A Theory of Impact Philanthropy. *Journal of Public Economics* 88 (9–10): 2,159–80.

Gabriel, Iason. Under review. "Is Effective Altruism Fair to Small Donors?" 2016a typescript.

Gabriel, Iason. 2016. "Effective Altruism and its Critics." *Journal of Applied Philosophy* 34 (4). Page references are to the "online first" edition.

GiveWell.org. 2015. Spreadsheet Methodology. Available at http://www.givewell.org/files/DWDA%202009/Interventions/GiveWell_cost-effectiveness_analysis_2015.xlsx.

Hammer, Jeffrey, and Dean Spears. 2016. "Village Sanitation and Child Health: Effects and External Validity in a Randomized Field Experiment in Rural India." *Journal of Health Economics* 48: 135–48.

Hassenfeld, Elie, and Josh Rosenberg. 2015. "Our Updated Top Charities for Giving Season 2015." Givewell Blog. Available at http://blog.givewell.org/2015/11/18/our-updated-top-charities-for-giving-season-2015.

Haushofer, Johannes, and Jeremy Shapiro. 2016. "The Short-Term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya." *Quarterly Journal of Economics* 131 (4): 1,973–2,042.

Haushofer, Johannes, and Jeremy Shapiro. 2018. "The Long-Term Impact of Unconditional Cash Transfers: Experimental Evidence from Kenya." Working paper. Available at http://jeremypshapiro.com/papers/Haushofer_Shapiro_UCT2_2018-01-30_paper_only.pdf.

Heutel, Garth. 2014. "Crowding Out and Crowding in of Private Donations and Government Grants." *Public Finance Review* 42 (2): 143–75.

Karnofsky, Holden. 2015a. Should the Open Philanthropy Project be Recommending More/Larger Grants? Good Ventures. Available at http://www.goodventures.org/research-and-ideas/blog/should-the-open-philanthropy-project-be-recommending-more-larger-grants.

Karnofsky, Holden. 2015b. "Good Ventures and Giving Now vs. Later." GiveWell Blog. Available at http://blog.givewell.org/2015/11/25/good-ventures-and-giving-now-vs-later.

Karnofsky, Holden. 2014. "Donor Coordination and the 'Giver's Dilemma.'" GiveWell Blog. Available at http://blog.givewell.org/2014/12/02/donor-coordination-and-the-givers-dilemma.

Kingma, Bruce. 1989. "An Accurate Measurement of the Crowd-Out Effect, Income Effect, and Price Effect for Charitable Contributions." *Journal of Political Economy* 97 (5): 1,197–207.

Humphreys, Macartan. 2015. "What Has Been Learned from the Deworming Replications: A Nonpartisan View." Columbia. Available at http://www.columbia.edu/~mh2245/w/worms.html.

OUP CORRECTED PROOF – FINAL, 19/08/19, SPi

Lichtenberg, Judith. 2013. *Distant Strangers: Ethics, Psychology, and Global Poverty.* Cambridge University Press.

MacAskill, William. 2015. *Doing Good Better*. Guardian Faber.

MacAskill, William. 2016. Presentation at Yale University. 6 May 2016.

Ozler, Berk. 2018. "GiveDirectly: Three-Year Impacts, Explained." World Bank Blog. Available at http://blogs.worldbank.org/impactevaluations/givedirectly-three-year-impacts-explained.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1 (3): 229–43.

Singer, Peter. 2009. *The Life You Can Save*. Random House.

Singer, Peter. 2015. *The Most Good You Can Do*. Yale University Press.

Stiglitz, Joseph. 2003. *Globalization and its discontents*. Norton.

Tuna, Cari. 2015. "Our Grants to GiveWell's 2015 Recommended Charities." Available at http://www.goodventures.org/research-and-ideas/blog/our-grants-to-givewells-2015-recommended-charities.

Warr, Peter. 1982. "Pareto Optimal Redistribution and Private Charity." *Journal of Public Economics* 19 (1): 131–8.

World Bank. 2004. *World Development Report: Making Services Work for Poor People*. World Bank.