CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Better vaguely right than precisely wrong in effective altruism: the problem of marginalism

Nicolas Côté[1],* and Bastian Steuwer[2]

[1]University of Toronto, Toronto, Canada and [2]Rutgers University, NJ, USA
*Corresponding author. Email: nico.cote@utoronto.ca

Abstract
Effective altruism (EA) requires that when we donate to charity, we maximize the beneficial impact of our donations. While we are in broad sympathy with EA, we raise a practical problem for EA, which is that there is a crucial empirical presupposition implicit in its charity assessment methods which is false in many contexts. This is the presupposition that the magnitude of the benefits (or harms) generated by some charity vary continuously in the scale of the intervention performed. We characterize a wide class of cases where this assumption fails, and then draw out the normative implications of this fact.

1. Introduction

The effective altruism (EA) movement has found great success in recent decades challenging common-sense intuitions about charity. Common sense would have it that you can do what you want with your own money, including giving it away to whatever charity you prefer to donate it to, provided you don't harm anyone in doing so. But for EA, doing no harm isn't enough. We must aim to maximize the beneficial impact of our donation, however much or little we donate (see Singer 2009, 2015; MacAskill 2015; Pummer 2016; Horton 2017. Precursors of EA can be found in Singer 1972 and Unger 1996).

This is the cardinal claim of EA, and in practice it has had incredibly restrictive applications: the effective altruist charity evaluator Give Well (2020a) rates a mere nine charities as donation-worthy, while The Life You Can Save (2021) lists a barely more generous 25. Many critics of EA object that this is too restrictive, that it disallows too much ordinary and (apparently) blameless charitable conduct.

While we think that these critics capture something morally important, we also believe that effective altruists are right that cost-effectiveness must guide our donation choices. If our choices are to be guided at all by considerations of what is morally desirable and by what best helps those we want to help, we must show due

respect for the fact that some charities *do* do much more for those in need than others.[1] Besides, all the bellyaching over the restrictive implications of EA overlooks the fact that these restrictions are simply the consequence of an incontestably attractive feature of EA, namely, its great action-guiding power. EA provides a very simple, evidence-based, and at first blush easy to apply method for assessing the donation-worthiness of any charity. It is a feature rather than a bug of this method that it generates a restricted menu of permissible options for individuals to choose from as they please. With thousands of charities in the world one could support it is impossible to make well-informed and rational philanthropic decisions unless this gigantic option space is restricted to a manageable menu of options.

Nevertheless, in this paper we do wish to raise a problem for effective altruism, which is that there is a crucial empirical presupposition implicit in the charity assessment methods of EA – at least, as EA has been applied so far to the field of aid and development – which is false in many contexts. Specifically, we contend that charity evaluators and philosophical defenders of EA have always implicitly assumed that the magnitude of the benefits (or harms) generated by some charity vary continuously in the scale of the intervention performed: i.e. tiny, incremental differences in levels of charity funding or in the scale of the aid intervention perform cannot generate huge differences in the size of the benefits (or harms) generated. This is true in the case of some aid interventions, such as interventions aimed at distributing insecticide-sprayed bednets, but is demonstrably false in the case of many others, notably of interventions that are aimed at bringing about what can be labelled as systemic change. This includes, but is not limited to, overcoming oppressive cultural norms and changing political or economic institutions. Consequently, the effective altruist method of charity assessment must be revised, on pain of providing poor advice where the latter sort of interventions is concerned. Unfortunately, since there exists no cheap and easy way of measuring the cost-effectiveness of charities which provide discontinuous (or 'lumpy') benefits, our critique may greatly undermine the action-guiding power of EA.

Our critique differs thereby from (although is related to) the familiar complaint that EA is blind to systemic change or political action (Herzog 2015; Lichtenberg 2015; Srinivasan 2015; Syme 2019). Unlike these critics, we believe that EA is not fundamentally incompatible with caring about such interventions. However, it is one thing to acknowledge that systemic change interventions are in principle open to be embraced by EA and another thing to take seriously the evaluation of such interventions. Unlike proponents of EA such as Singer (2009: 115–117, 2015: 157–164), Wiblin (2015), Rubinstein (2016: 517–518) and Berkey (2018), we believe that EA needs to be reformed in order not to be biased against organizations that provide lumpy benefits on the path to fundamental change for the better. However, this reform will come at the cost of moving EA closer to the common-sense position it wanted to overcome.

---

[1]This need not be based in a consequentialist injunction to do the most good. Few non-consequentialists would deny that we ought to help more rather than fewer people when we can. The classic is Taurek (1977). For some non-consequentialist arguments on saving the greater number see Otsuka (2004), Kamm (2007: Chs 1–2) and Liao (2008).

Our argument is structured as follows. In section 2 we will lay out in greater detail the target of our argument including the practical procedure which charity evaluators have followed to operationalize EA in the context of global poverty. It is not hard to show that this procedure presupposes a continuous relationship between the impact size and the scale of aid and development interventions. In section 3 we explain how this assumption fails in a broad range of cases, and how this breaks the charity assessment method of EA. We close this section by noting that there is a solution concept that does in principle permit accurate estimates of the cost-effectiveness of aid interventions which provide discontinuous benefits: the Shapley value. However, applying this solution requires far more information than we ever have access to in aid and development contexts. In section 4 we respond to the argument that, despite the observations made in section 3, effective altruists should double-down on their method of impact assessment. Section 5 concludes that we ought simply to acknowledge that our best methods of assessing 'the best charities' are much less discriminating than we would like.

2. Effective altruism in theory and in practice

EA is a broad tent, and different authors have very different views on what commitments it entails. To some extent this is due to the fact that EA is (at least) as much a social movement as a philosophical movement. This diversity of views and practices can make EA an elusive target to criticize. In this section we therefore proceed by defining our target more precisely. Our definition captures an important current in EA's approach to global poverty – the initial impetus for EA.[2] In particular, our target definition captures a revisionary approach to charitable giving which departs from prevailing common sense. Therefore, for simplicity, we will refer to our target simply as EA. In the conclusion we return to the question what our arguments entail for the broader EA movement.

A good starting point for defining our target is the definition used by the Centre for Effective Altruism (2021): 'effective altruism is about using evidence and reason to figure out how to benefit others as much as possible, and taking action on that basis'. MacAskill unpacks this definition which is aimed for a general audience and highlights three features. The definition is non-normative, it is science-aligned, and it is maximizing. MacAskill (2019: 12–14) adds the requirements that EA is impartial – which we accept – and welfarist – which we do not include in our definition.

Let us take each of the components in turn. By non-normative MacAskill (2019: 15–17) means that EA asserts neither a moral obligation to give anything to charity nor a moral obligation to give effectively. While this makes EA a broad tent, we fear that it compromises too much of EA's critical character. EA started out as a challenge to common-sense approaches and common sense plainly accepts that

---

[2]This is still the case today. As a quick indicator 13 out of the 15 chapters in the recent edited volume *Effective Altruism: The Philosophical Issues* (Greaves and Pummer 2019) deal with such questions. The remaining two chapters include one on the definition of EA (MacAskill 2019) and one on long-term interventions (Beckstead 2019).

one can set oneself the personal project to give effectively. We propose instead a weak characterization of EA which retains some normative core. Following work by Pummer (2016) and Horton (2017), we remain agnostic on the question of *how much* EA requires us to donate – *if at all!* All we require of EA is that it implies that we ought to donate effectively *if* we donate. This minimal requirement does not rule out stronger versions of EA that impose demanding unconditional obligations to give.[3] Our concern is not with how much people should donate, but with whether the assessment methods employed by effective altruists provide good guidance as to where they should donate if they choose to donate.

Likewise, to avoid ruffling any feathers, we wish to remain agnostic about what is to count as good. The utilitarian roots of EA have led many effective altruists to embrace both a welfarist understanding of goodness and to embrace a view that ignores distributive questions. We do not take welfarism to be central to EA, indeed we believe that welfarism would be harmful in building the broad tent EA wants to build. So justice, freedom, virtue, biotic health, etc. are all on the table as possible final ends to be promoted. Our criticism in this paper is therefore largely independent from the question of what counts as 'good'.

This leaves us with the two components of the definition which MacAskill (2019: 14–15) describes as uncontroversial: science-alignment and maximization. EA is science aligned in the sense that it (proudly) locates itself within the 'evidence-based' movement in social science, and insists (plausibly, given the high stakes of aid and development) that it is morally irresponsible to donate to some charity if you lack solid evidence that it is effective. 'Solid evidence' here indicates a high evidentiary threshold. In particular, claims about impact size not to be robustly established by studies capable of controlling for selection bias. In practice this usually means EA won't recommend a charity unless randomized controlled trials (RCTs) have demonstrated its effectiveness (see GiveWell's 2016a stated procedure on evidence assessment). There are important exceptions – GiveWell (2013) briefly listed VillageReach among its top charities on the basis of convincing but non-randomized studies – but not many: all of GiveWell's top nine charity recommendations are supported by multiple RCTs. This is very much in keeping with a strain of thinking within the evidence-based movement which views RCTs as the gold standard for establishing the causal efficacy of some intervention and estimating its average impact size (Banerjee and Duflo 2011; Karlan and Appel 2011; Leigh 2018),[4] on the grounds that RCTs control for all confounding factors and therefore register only the causal influence of the intervention (see, however, Deaton and Cartwright 2018 for criticism of this assumption). On top of high-quality micro studies, charity evaluators also insist on high-quality regression analyses measuring the macro-level effect of programmes on variables of interest (e.g. poverty levels, health outcomes) as interventions found to be promising by controlled studies are scaled up and

---

[3]For example, Berkey (2020) argues that EA cannot avoid appealing to demanding unconditional obligations.

[4]See also the overview of the debate between 'thinking big' and 'thinking small' in Cohen and Easterly (2009).

performed in different locations. This is needed to establish that the observed causal relation generalizes beyond the initial context, as sometimes a locally effective intervention becomes ineffective when scaled up (e.g. providing better education to individual students is good for them, but expanding education for everyone will have lesser effects due to the fact that education is partly a positional good; if education mainly helps in securing a government job, increased education won't increase the amount of government jobs). These are heavy evidentiary burdens to discharge, and unsurprisingly they rule out most charities.

The maximization criterion is straightforward: in deciding where to donate, it is not enough to do some good. Rather, 'we should do the most good we can' (Singer 2015: vii). More specifically, we should donate only to the most cost-effective charities, i.e. 'the one[s] that help[] the most people the greatest amount per dollar' (Alexander 2013). Once we've eliminated from consideration all charities for which we lack solid evidence of effectiveness, we rank those that remain by how much value (as measured by improvements on health, poverty, etc.) we can expect them to produce from additional funding.

Strictly speaking, we're looking to compare the *expected marginal rates of return* on additional donations of each charity, and keep only those charities that have the highest expected return. And in practice, effective altruists have followed a very simple heuristic for measuring expected rates of return, which we will refer to as 'myopic marginalism'. Here is one way. Start by calculating the past rate of return on donations. This is easy enough: simply divide the total size of the benefit generated by some intervention by the total cost of the programme. This first measure is a bit crude, since it only tells you about average return on donation, not the return on the last dollar, but it is used by EAs and it does tell you something (see MacAskill 2015; Open Philanthropy Project 2017; GiveWell 2020*b*; Giving What We Can 2021; and especially GiveWell's 2021 explicit cost-effectiveness calculations in spreadsheets). A more sophisticated measure becomes possible if you have a time-series plotting the evolution of the programme's costs and benefits: instead of looking at total costs and benefits, look at the ratio of the most recent change in size of the benefits to the change in the costs of the programme. This measure does give you the marginal return on the last dollar spent (see Budolfson and Spears 2019 for further discussion). It is then predicted that the rate of return on the next dollar you donate to some organization will be very similar to the rate of return on the last dollar, up to however much more room the charity has for additional funding. With this information in hand, it is child's play to identify the elite group of charities that will maximize the impact of the next dollar you donate (up to however much additional room for funding each charity has). And so, in just two easy steps, we've winnowed the space of charities worth considering donating to to just a handful, greatly simplifying the decision problems of donors.

There are many steps in the decision procedure we've described that one might take issue with. Critics of EA have, for example, criticized the optimizing logic of EA, its over-reliance and over-insistence on RCTs, and its use of cost-effectiveness analysis, which makes no room for permissible partiality and is claimed to overweigh the value of a statistical life. We take issue with none of this in the present paper, and will focus only on the inadequacy of myopic marginalism. As

we will now see, myopic marginalism only yields accurate estimates of cost-effectiveness when the benefits of an intervention are continuous in its scale, because only then can we use past returns as a reliable guide to future returns. This would not be a problem if most or all cases were of this kind, but as we'll show there are many cases where this does not hold.

3. Marginalism and the problem of lumpiness

To get an intuitive handle on the problem, consider the following example. Suppose a farmer, Mnemosyne, needs to move a lump of hardened magma, so that she can till her land. The lump is too heavy for one person to lift on their own, but two people are just strong enough to move it the necessary distance. So Mnemosyne calls her wife, Scathach, and together they move the lump. In this example, what is the marginal productivity of each farmer? As the first worker on the scene, Mnemosyne's marginal product is zero, and as the second worker on the scene, Scathach's marginal product is one lifted lump. But it is not as though Scathach is a better worker than her wife: both were necessary to the lifting of the lump. Now suppose Mnemosyne has a lot of lumps lying around, and wants to hire workers to lift them. If Mnemosyne were a myopic marginalist, she would erroneously conclude from the fact that the first worker has a null marginal productivity that the next worker will also have a null marginal productivity. She would then never hire any workers at all, since she expects no output from any number of workers, and leave the lumps to litter her field. The myopia in myopic marginalism is apparent: when worker productivity is 'lumpy' and instead of continuous, it is not the marginal productivity of each worker that you need to look at; instead, you need to reason in terms of how productive variously sized worker coalitions would be, then hire the coalition which gives you best value for money. This will be the coalition whose 'Shapley value' (see Hart 1989) maximizes Mnemosyne's payoff. Roughly speaking, for any given coalition of individuals engaged in a cooperative enterprise, the Shapley value assigns to each individual the share of the gains of cooperation that corresponds to their expected contribution in a random ordering of all the individuals involved. So in the case of Scathach and Mnemosyne, each receives an equal Shapley value, because if you average their marginal contributions to lump-lifting over all the number of ways in which the coalition could form (namely, Mnemosyne arrives first, or Scathach arrives first) you would arrive at the same number.

In the context of charitable giving, we may put this point sharply in perspective by considering the incentive structures that support the reproduction of oppressive cultural norms. Numerous oppressive social practices – e.g. FGM, child labour and (formerly) footbinding – persist due to coordination problems: roughly, so long as other members of society participate in the practice, it is in your interest to do so as well, but everyone would be better off if everyone stopped participating in the practice (see Mackie (1996) on the issue of footbinding and FGM, and see Todaro and Smith (2015: Ch. 8) and Basu (1999) on the issue of child labour). This is what is sometimes called the social conformity trap (Andreoni *et al.* 2017), and it is a characteristic feature of a certain class of games where there

are multiple Nash equilibria in pure strategy, such as the so-called stag hunt game (Cooper 1999: Ch. 1). Note, however, that the more people cease conforming, the smaller the benefits of conformity, and the smaller the costs of non-conformity. Eventually, there comes a point where defection becomes sufficiently widespread that the expected benefits of conformity over non-conformity taper down to zero, and as soon as that point is reached, everyone who has not yet defected will instantly defect, pushing us towards the optimum equilibrium. This is exactly what happened in the case of footbinding: a practice that had lasted for a thousand years ended within a generation as soon as a large enough minority stopped doing it (Mackie 1996: 1006). Crucially for us, the existence of multiple equilibria entails a discontinuous relation between the scale of interventions aimed at breaking social conformity traps and the benefits they provide.

To see how this all works, it may be helpful to examine one example in detail. Take the practice of female genital mutilation (FGM). The worst kind of FGM is infibulation, in which the clitoris, labia minora, and outer walls of the labia majora are removed (Mackie 1996: 1002). It is usually performed by adult women on girls around 8 years old, and it causes serious lifelong health problems. Urination and menstruation become painful and difficult, infections are common, as are stones in the urethra and bladder (Nussbaum 1999: 120); subsequent to the operation sexual intercourse becomes traumatically painful, and childbirth requires introcision and subsequent resewing of the genital area (Mackie 1996: 1003). And yet, FGM persists due to cultural norms: men will refuse to marry non-infibulated women, and women have no prospects outside of marriage (Mackie 1996: 1004). Moreover it is believed that being infibulated is healthier, cleaner and more beautiful than not being infibulated, and infibulation is thought medically necessary.

Incredibly, these norms make it *worse* to be non-infibulated so long as everyone else is infibulated, because in these communities marriage is so crucial to women's prospects that it is worse to be healthy but celibate than to be married and in pain. And this is the key point: the only reason it is better to be infibulated in these communities is that being the only non-mutilated woman is competitively disadvantageous: if mothers stopped infibulating their daughters, the latter would suffer no competitive disadvantage with regards to their peers when looking for a husband, and then all could marry without suffering excruciating pain. (The same holds, of course, too if prospective husbands stopped preferring infibulated partners over non-infibulated partners.) And presumably no woman would then want to be the only one to undergo FGM, because that would be to choose marriage and pain over painless marriage (notice that this implies that it is better for those who do conform that sufficiently many others conform too, because only then is undergoing FGM advantageous in any way).

We can model formally the payoffs associated with FGM. In effect, women in communities which practice FGM have exactly two choices; call these 'Don't Conform' and 'Conform'. If all women choose Don't Conform then they obtain the best possible payoff. In contrast, if all women choose Conform then they get a worse payoff, but it is still the second-best they can get away with. However, when only a few women choose 'Don't Conform', the women who choose 'Conform' get comparatively better payoffs, whereas the women who choose 'Don't Conform' are punished for departing from the cultural norm, and obtain

Table 1. The Social Conformity Trap

	Don't Conform	Conform
Don't Conform	4, 4	−3, −4
Conform	−4, −3	−1, −1

a payoff that is significantly worse than what they would have obtained had they chosen 'Conform'. This is the characteristic payoff structure of a close variant of what is known as the stag hunt game. Formally, in a simple setting where there are only two players in the game, the payoffs associated with each player's choice in each possible setting can be represented by Table 1, where the numbers represent the utilities associated with each payoff.

What's important to realize here is that if each player believes the other will conform to the social practice, the best strategy is to conform as well. And if conformity is a stringently enforced social expectation, if everyone in your community has adhered to these standards in the past, then the reasonable belief to have is that everyone will conform. This is how people can get stuck in inferior equilibria. And indeed, Table 1 characterizes the payoff structure not just of FGM, but also of domestic servitude of women under patriarchy, beauty norms, religious conformity, and many other oppressive cultural practices. Note that if we generalize the above matrix to a situation with N players, then the benefits of either strategy will be increasing in the number of players who adopt that strategy: the more people conform, the costlier it is to not conform, but the more people defect, the less costly it is. And for any plausible model of an N-player coordination game of this sort, there will be some number $n < N$ such that not conforming is the optimal strategy whenever you expect at least n other players not to conform.

How does the existence of multiple equilibria 'lumpify' the benefits of breaking social conformity traps? Consider that the first few defectors will not, on balance, benefit from it; in fact, short of potentially expensive compensating offsets, they will be strictly worse off for taking their moral stand. But the social benefit of convincing the nth woman to boycott the practice, whose defection pushes other people's expectations to the tipping point, are immense, because this will rapidly cascade into the defeat of the social practice, hugely improving everyone's welfare, and greatly improving the state of justice, freedom and equality. In other words, the marginal benefit of convincing the first $n−1$ players to defect is in every case negative (at least in terms of welfare), and the marginal benefit of convincing the nth player to defect is *the sum total of all the benefits yielded by the ending of the practice*. This problem generalizes. Tipping points aren't limited to oppressive social practices, but are a characteristic feature of *all* games with multiple equilibria.

Any charity whose work involves pushing people from a lower to a higher equilibrium will generate lumpy benefits. And many charities' work does involve this, as many are engaged in the project of challenging oppressive cultural norms. This is at the heart of Girls not Brides' and Amnesty International's

work, for example. This is a feature that those charities share with many activist movements: progressive activism does tremendous good when it effects substantive political change (e.g. the civil rights movement), but very little if it fails (e.g. the Occupy movement), and meanwhile, whether or not it succeeds, mass activism consumes considerable time and resources.

For an effective altruist charity evaluator, this presents a conundrum: what is the marginal value of donating an extra \$10 to an organization such as Girls not Brides or Amnesty International who provide 'lumpy' benefits? Or the marginal value of the donation implicit in taking part in activism rather than working? As explained earlier, the effective altruistic method for computing the marginal value of a donation is to look at past returns on donations to various organizations; it is then predicted that the marginal rate of return on the next dollar you donate to some organization will be very similar to the rate of return on the last dollar. This decision procedure is fairly easy to apply, since knowledge of past returns is cheap to procure (assuming that you've run your studies and are keeping books, all the data should be in), and NGOs are usually happy to tell charity evaluators how much they can scale up their operations with additional funding. Indeed this is what makes marginalism attractive: it is informationally undemanding yet inferentially powerful. But past returns on investment is only a guide to returns on future investment if the relation between returns and investment is continuous. And in the case of games with multiple equilibria, this crucial presupposition fails: the return on the next dollar you spend may be *hugely* greater than the return on the last dollar you spent.

And unlike in the case of Mnemosyne and Scathach, here the Shapley value cannot come to our rescue, because we face problems of incomplete information that preclude its computation. Computing Shapley values requires much more knowledge than mere knowledge of past returns on donation: it requires knowledge of the total output of various hypothetical coalitions. Mnemonsyne was only able to identify the worker coalition that gave her best value for money because she knew from the get-go that it takes exactly two workers to lift one lump. But in games with multiple equilibria, we don't know ahead of time what the magical value of n is which pushes expectations to the tipping point. All we know is that it lies between one and the total number of players, which is too bare a guide on which to base Shapley value estimates. Heck, in some cases we won't know ahead of time what game form we're in, and thus how many possible higher equilibria there are. Simply shifting one's attention from one individual's marginal impact to the impact of various donor coalitions is not going to be useful here.[5]

One can try to guesstimate how many equilibria there are and what n's value is by looking at past cases of successful collective action to defeat social norms that are similar to the kinds one is looking to overcome, but such guesstimates are unlikely to be very reliable or cheap to procure. For one thing, the historical record on successful attempts at overthrowing oppressive cultural norms may be sparse or

---

[5]This explains why an appeal to collective obligations (as in Dietz (2019: 109–112) is not going to help with this problem. Relatedly, the idea of pooling the resources of effective altruists to support the solutions that are marginally best at this higher level of resources is explicitly embraced by Berkey (2019: 7–8).

non-existent (China is the only country where footbinding was ever practiced, so there would have been no inductive base from which to infer n's value). Further, the value of n may be quite sensitive to local attitudes and conditions (e.g. payoffs may differ according to religiosity), and likewise the factors that drive expectations of conformity may vary from place to place, which makes it difficult to draw reliable inferences from other cases about 'what works' even in cases where something did work. And anyway, estimates are likely to suffer from reference class problems, as judgements about which cases are similar enough to new cases of interest to permit inferences are highly contestable (is footbinding structurally similar enough to FGM as a practice to permit inferences about how large a minority needs to defect for the social practice to fall? Is the practice of FGM in Afghanistan similar enough to the practice of FGM in Eritrea?). In sum, when lumpy benefits are at issue, cost-effectiveness analysis becomes too informationally greedy to yield reliable assessments. This is not to suggest that there are not some rare examples in which we have such information. In some infrastructure projects we have fairly good information about the necessary fixed costs to complete the project. Such projects are lumpy. Building a water treatment plant provides slightly lumpy benefits, because you get no benefits from it at all unless you invest enough to cover its fixed costs. But beyond that point, the relation between variable costs and total benefits is smooth: for every extra day you pay your employees and keep your plant running, you get an equal amount of sanitation. If you know the fixed costs, then it is trivial to extend myopic marginalism to this case. It is notable that the example of the water plant is another case of direct aid. The important goods we described, like breaking social conformity traps, are lumpy all the way down, and they pose more serious problems: there are only discrete tipping points at which large benefits accrue, and no 'investment floors' past which further investment or cooperation begins to yield smooth returns. In these cases, you do need to compute a Shapley value to figure out the value of an intervention, and the problems of incomplete information must be squarely faced. Unlike in the case of direct aid investments it is much more complicated to find out where the tipping points are.

4. Doubling down on myopic marginalism

What is the appropriate way to respond to all these observations? One possible response is that if the cost-effectiveness of lumpy-benefit providers such as Girls Not Brides cannot reliably be measured, however cost-effective they might in fact be at any level of funding, then tough luck for child brides and victims of FGM. Being well-intentioned in trying to overcome FGM and child marriage is not enough, we need to show that Girls Not Brides and similar organizations stand a good chance of actually changing things. There are reasons to be sceptical. Many well-intentioned charities do no or very little good in the world. For example, research by the US-based Coalition for Evidence-Based Policy indicates that up to 75% of all social programmes that were tested for effectiveness had no measurable positive effect – though some had measurably harmful ones (GiveWell 2019). All the

while the stakes in development and aid are high. Money sent to ineffective charities cannot be used elsewhere. And EA is reminding us that there are charities that we *know* to be highly effective in saving many, possibly thousands of infants from easily preventable causes of death.

We aren't terribly impressed by the flat response that 'we've shown effectiveness, while you haven't'. This response suggests that the failure of some aid programme to demonstrate cost-effectiveness according to EA's method should make us highly confident that it is in fact less cost-effective than some programme which has demonstrated its effectiveness by those standards. But such high confidence is only warranted if the test employed permits accurate measurements of both programmes' effects. In general, assume we have a measurement tool that is highly accurate in measuring A-type interventions but poor in measuring B-type interventions, i.e. with a probability close to 1 it would fail to register the effectiveness of B-type interventions even when these are effective. This measurement tool cannot give us a good guidance for comparing A-type interventions with B-type interventions. The judgement that A-type interventions fare better than B-type interventions is not due to anything about the interventions, but an artefact of our assessment methods. Myopic marginalism falls into this trap. Its impact assessment method is highly inaccurate when benefits are lumpy: if some programme is not at tipping point scale, then myopic marginalism will inevitably evaluate it as not cost-effective, regardless of how much more cost-effective it would be at the tipping point scale than any other programme.

Still, you might think that what counts in favour of EA's recommendations is not that they are demonstrably more cost-effective but rather that we have better information about their cost-effectiveness. Our inability to accurately estimate the cost-effectiveness of lumpy-benefit-providing would then be in itself a good reason to err on the side of caution and only donate to the EA-approved programmes, since we at least know that those ones do a lot of good. This would be an example of what in decision theory is called 'ambiguity aversion' or 'aversion to severe uncertainty': you have no particular reason for thinking that *f*-ing is better than *y*-ing, but your greater uncertainty about the chances of *y*'s possible outcomes leads you to prefer *f*-ing. Aversion to severe uncertainty of this sort is explicitly endorsed by GiveWell (2016*b*), who say 'we generally prefer to give where we have strong evidence that donations can do a lot of good rather than where we have weak evidence that donations can do far more good'.

In response, we accept that this is one reason to prefer the EA approved programmes. An agent who is averse to severe uncertainty may, for this reason, systematically select the recommendations of EA rather than interventions that provide lumpy benefits. This suggestion is noteworthy since it goes against the grain of the utilitarian origins of EA, for utilitarianism is traditionally not sympathetic to ambiguity aversion. This is for two simple reasons: first, one of the strongest and most widely cited arguments for utilitarianism are the so-called 'impartial observer' theorems by John Harsanyi and later economists (Harsanyi 1955; Hammond 1981; Fishburn 1984; Broome 1991), which purport to show that any person who is impartial, sympathetic and rational must rank alternative courses of action by utilitarian standards. But crucially the notion of rationality required to obtain these results rule out ambiguity-aversion on the

part of the social evaluator. Second, we can show that if a utilitarian is ambiguity-averse with respect to individual prospects (i.e. they prefer to give individuals unambiguous benefits to ambiguous ones), then utilitarians will sometimes have to choose between one course of action which is guaranteed to yield greater total well-being, and another which is guaranteed to give each individual a better (because less ambiguous) prospect – so either utilitarians must make everyone worse off in expectation, or they must fail to maximize total well-being (Rowe and Voorhoeve 2018: 262–265). Ambiguity-neutral utilitarianism, by contrast, will never face this choice: maximizing total well-being is always consistent with choosing what is expected to be best for everyone.

In any case, the appeal to aversion to severe uncertainty does not help the case of EA. EA wants to say that only a small number of charities are donation-worthy. Its power of discriminating among the vast number of charities in the world is its greatest asset. Aversion to severe uncertainty is, however, at best rationally permissible. To yield the prescriptions EA makes for the charitable giving of others, the proponent of EA would need to argue the substantially stronger claim that one is rationally required to be averse to severe uncertainty. Otherwise, EA cannot tell an agent who is neutral to severe uncertainty that they are making a mistake in donating to lumpy-benefiting-providing organizations. Nor is aversion to severe uncertainty somehow constitutive of caring about giving effectively: provided your attitude to severe uncertainty is rationally permissible, what counts as 'the' effective intervention sensitively depends on just what your attitude is. If you're permissibly averse, the effective intervention is intervention that strikes the best balance between maximizing expected benefits and reducing severe uncertainty, while if you're permissibly neutral the effective intervention is the intervention that simply maximizes expected benefits. We take it that both aversion and neutrality with respect to severe uncertainty are rationally permissible attitudes, so neither can be uniquely constitutive of caring about effectiveness. The bottom line is that effectiveness is not linked to any particular attitude towards severe uncertainty.

Second, the argument from severe uncertainty would have more force if EA approved aid programmes and programmes which provide lumpy benefits were all in the business of providing the same sorts of benefits. Ambiguity aversion is most sensible in cases such as the Ellsberg paradox (Ellsberg 1961), where you have a choice between an ambiguous lottery (unknown probabilities) and an unambiguous lottery (known probabilities), but both lotteries return cash prizes. For one thing, if the kinds of goods you have a choice between are so different as to be difficult to compare in value precisely, then even if you are ambiguity averse, your ambiguity aversion may fail to support any determinate ranking among your options.[6] This contrasts with Ellsberg paradox cases, where any

---

[6]To take an extreme example, suppose there are four goods: A, B, C, D, you prefer A to B, and C to D, but you cannot rank A and C, or A and D, or B and C, or B and D. Now suppose you have a choice between a lottery that gives you a 0.5 chance of A and a 0.5 chance of C and an ambiguous lottery that gives you an unknown chance p of C and a chance $1-p$ of D. In this case, no matter how ambiguity averse you are, it is impossible to rank these two lotteries against one another, because for any level of ambiguity aversion, there will always be an infinity of utility functions consistent with your preferences that will rank one lottery over the other.

positive level of ambiguity aversion will require you to choose the less ambiguous lottery. But let's set this point aside.

The appeal of ambiguity aversion is the appeal of 'playing it safe'. But while we may want to play it safe if we have to select based on poor information when the outcomes are always money, things can look very different if we are talking about taking a chance on poor information to promote goods of justice and liberty versus taking a bet on surer information to promote the rather different goods of health and income. Charitable giving is not just a one-off decision. It is also not just a decision done by a single agent. If followers of EA are convinced that they should be averse to severe uncertainty, then they would all systematically choose direct aid programmes designed to improve health and income. Other values like justice and liberty would, routinely, get the short end of the stick. This makes the case of severe uncertainty aversion in EA different from that of a simple Ellsberg paradox. In a simple Ellsberg paradox, there is just a single choice in which we might want to play it safe. But playing it safe over and over again, all the while there is a bias in what the safe option is, is less compelling. Here the appeal of playing it safe competes with the appeal of promoting under-served values.

It is relevant in this context that EA-approved interventions are (almost) all providers of direct aid, i.e. private goods such as income, medications and bednets, whose consumption is linked to improvements in consumption levels or health.[7] Direct aid is favoured by EA in large part because direct act typically has continuous marginal benefits. Providing one bednet is not too different from providing two bednets, and so on. The improvements brought about by direct aid obviously matter, and are always a major aim of interventions aimed at ending oppressive cultural practices – the best reason for ending infibulation is that it is traumatically painful and cruelly impairs women's lifetime health. But there are special reasons to care about ending oppression that go beyond their immediate benefits in terms of health and income: oppressive social practices such as FGM and child marriage are high crimes against their victims, not mere misfortunes, like malaria; they crush individual liberty, and create classes of dominators and dominated. We have compelling interests in halting high crimes, protecting liberty, and undermining structures of domination that are not reducible to our other compelling interests in relieving the burdens of poverty and ill health. (Note that we are not begging the question against welfarist construals of EA. It might be, for example, that liberty and non-domination are both components of or contributors to well-being and so their value is reducible to that of well-being; we just need to deny that it is reducible to the value of health and income.) If these ends can *only* be advanced by donating to EA non-approved causes, that is a reason to take risks on such causes by donating to them. And not to put too fine a point on it, but the end of infibulation would be a very good thing, just as the end of footbinding was a good thing, and as the end of just any oppressive social practice would be a good thing. So if things go well in your support of some NGO devoted to challenging these practices, things

---

[7] All of GiveWell's charities are of this kind. The Life You Can Save recommends Development Media International, an organization for improving health literacy and Oxfam much of whose work is on advocacy and not direct aid. More recently it added two organizations working on climate advocacy.

will go *very* well indeed, and this again gives you reason to take a chance on such causes.

This means that we can have reason to choose interventions with lumpy, discontinuous benefits, precisely because they serve different ends from the direct aid interventions favoured by EA. There is another reason why we may prefer to donate to organizations carrying out such interventions. We earlier mentioned that it is well-known that many charities fail to provide any substantial benefits at all. It is another well-known fact that aid interventions that do a great deal of good in the short term can do immense harm over the long term. Against such a background, one might reasonably be concerned with giving greater weight to ruling out interventions that have a risk of making things worse than making sure to identify interventions that are effective in making things better. Consider an analogy. Refusing to convict criminal defendants unless their guilt is established beyond a reasonable doubt implies that many criminals will walk, but that is appropriate given the stakes in criminal defence trials. Likewise, refusing to recommend donating to some organization on the ground that there is a risk of perpetuating harm implies refusing to recommend many charities that are in fact donation-worthy, but that is appropriate given the stakes in aid and development.

It is worth pointing out that because programmes aimed at undermining oppressive cultural practices are, unlike the GiveWell approved programmes, *not* in the business of providing direct aid, they don't raise many of the concerns about harmful effects of direct aid. As de Waal (1997) has documented, disaster relief aid to war zones, or even countries suffering peacetime famines (e.g. North Korea in the 90s) can only reach their intended recipients with the consent of local warlords and tyrants. Such aid saves many lives, de Waal shows, but Angus Deaton and others have argued that by providing support to oppressive regimes, such aid prolongs political violence and repression, undermining long-term development and costing more lives down the line.

Concerns about the risks of aid are not limited to war zones or authoritarian regimes. Aid can have negative consequences for state capacity. Effective state institutions are extremely important for development, and weak institutions can lead to poverty traps (Sen 1999: 111–203; Acemoglu and Robinson 2012; Deaton 2013: 267–324). In countries with weak institutions (which are often major recipients of aid), the provision of aid can lead to a substitution effect: when NGOs provide goods and services that the state could itself deliver through taxation, the state can scale back its own efforts to provide these goods (or doesn't build up the capacity to deliver them), so the short-term effect of the intervention is a wash, and its long-term effect is seriously negative because it undermines state institutions (Wenar 2011: 115, 125–126; Deaton 2013: 292–294; Acemoglu 2015; Clough 2015). Substitution effects are a real concern for direct aid programmes such as the ones recommended by GiveWell (acknowledged today by most NGOs and charity evaluators), since they do provide goods which the government also provides. But they're not a concern for charities that challenge social conformity traps, because reforming oppressive norms and practices is different. Either governments of developing countries have no interest in providing them, or if they are interested, they often need the

support of civil society organizations. Building an NGO-run school can crowd out public investment in education, but advocating alongside the state against child marriage is more likely to be mutually reinforcing. It is somewhat ironic, therefore, that the appeal of EA's focus on direct aid lay in a form of playing it safe (aversion to severe uncertainty) given that a different way of caring about playing it safe (aversion to causing harm) counsels in the opposite direction.

5. Conclusion

So where does this leave us? The impact assessment methods of EA, at least as those have been developed so far, provide us with no reliable means to compare the cost-effectiveness of programmes whose benefits are continuous in the scale of the programme with the cost-effectiveness of programmes that provide lumpier benefits. Aversion to severe uncertainty may give one reason to prefer supporting only the EA-approved charities, but other reasons counsel in favour of organizations that provide lumpier benefits. These organizations provide benefits (when these accrue) that no direct aid programmes can and which are tremendously valuable. In addition, such efforts don't carry the usual risk factors that have aid sceptics so worried. The conclusion to draw from this, we submit, is that our comparative assessments of donation worthiness are bound to be incomplete. While there may be EA reasons to favour donating to GiveWell-approved programmes (e.g. aversion to severe uncertainty), there are also EA reasons to favour groups such as Girls Not Brides that aim to provide lumpy (but important) benefits, and by EA's own lights none of these reasons are decisive. Therefore it must be permissible to donate to either kind of charity.

The result is that the principles of EA are less action-guiding than its proponents have let on. What does this mean for EA? Have we provided a refutation or reform of EA? This is largely a terminological dispute. It is important to recognize that our critique pushes EA closer to the highly permissive common-sense position. EA, insofar as it is still distinctive, is less of a revolution in thinking about how to give than a minor adjustment. This conclusion may sound a bit defeatist. But this is not so devastating. For one thing, our critique does not strip EA of all its bite: the Make a Wish foundation, local churches in rich communities, and wealthy universities provide none of the goods and raise none of the measurement challenges we've been concerned to highlight in this paper, so the effective altruist critique of such organizations applies with full force. And besides, we still have some ways of distinguishing donation-worthy-lumpy-benefit-providers from donation-unworthy ones. We can look at whether some organization is well-organized, transparent, sets goals for itself which it is good at meeting, and goals whose attainment seems like a plausible means for reaching desired ends. These are the sorts of features that generally characterize successful interventions, so we may be more confident in the effectiveness of organizations which have those features than in that of organizations which lack them. Likewise, in the case of interventions aimed at breaking social conformity traps specifically, lab work by Andreoni *et al.* (2017) suggests that opinion polls, which publicize people's preferences for change, technologies which expedite the

spread of information, and the offsetting of the non-conformity costs of first-movers through rewards for boycotting behaviour can help to accelerate social change. This suggests that interventions which provide such services are likelier to succeed than those which do not. Further, NGOs which aim to provide more important lumpy benefits will generally be worthier of support. And it is worth remembering that even if our best attempts at computing Shapley values are going to be unreliable, even unreliable estimates may reliably favour some programmes over others.

These considerations will help us winnow the space of donation-worthy-lumpy-benefit-providers. They won't give us a cookbook recipe for choosing between those that are left and those approved of by myopic marginalism. Rather than a single instruction they provide us with a way of thought about charitable giving. If you are concerned about the relative paucity of information that comes with difficult to assess interventions, then you may favour direct aid. If you are concerned that particular causes and ends are underrepresented, for example that aid is overly focused on health and wealth, then you may favour organizations such as Girls Not Brides. If you are concerned about the adverse long-term effects of aid, then such civil society organizations will be a good choice, too. EA can help guide our behaviour by eliminating some options and telling us which considerations are good reasons to support one among the many left. That's the best we can do.

How you react to this finding depends on your temperament. If you were initially attracted to EA because it seemed highly objective and didn't let people's values (and potential biases) guide donation behaviour, then our results should dampen your enthusiasm for EA. EA won't provide you with a solid and purely objective formula for how to give. But you might instead see this result as a blessing in disguise: if you worried that EA generated excessively restrictive recommendations, then this newfound permissiveness will be welcome news to you.

References

Acemoglu D. 2015. Response to the logic of effective altruism. Boston Review, 1 July. <https://bostonreview.net/forum/logic-effective-altruism/daron-acemoglu-response-effective-altruism>.

Acemoglu D. and J.A. Robinson 2012. *Why Nations Fail. The Origins of Power, Prosperity, and Poverty.* New York, NY: Crown Publishing.

Alexander S. 2013. *Efficient Charity – Do Unto Others.* The Centre for Effective Altruism. <https://www.effectivealtruism.org/articles/efficient-charity-do-unto-others/>.

Andreoni J., N. Nikiforakis and S. Siegenthaler 2017. Social change and the conformity trap. Unpublished manuscript. <https://pdfs.semanticscholar.org/7f3d/21b0de0353e50df0ae344b274b594228f839.pdf>.

Banerjee A.V. and E. Duflo 2011. *Poor Economics: Rethinking Poverty and the Ways to End it.* Gurgaon: Random House India.

Basu K. 1999. Child labor: cause, consequence, and cure, with remarks on international labor standards. *Journal of Economic Literature* 37, 1083–1119.

Beckstead N. 2019. A brief argument for the overwhelming importance of shaping the far future. In *Effective Altruism: Philosophical Issues,* ed. H. Greaves and T. Pummer, 80–98. Oxford: Oxford University Press.

Berkey B. 2018. The institutional critique of effective altruism. *Utilitas* 30, 143–171.

Berkey B. 2019. Collective obligations and the institutional critique of effective altruism: a reply to Alexander Dietz. *Utilitas* **31**, 326–333.

Berkey B. 2020. Effectiveness and demandingness. *Utilitas* **32**, 368–381.

Broome J. 1991. *Weighing Goods. Equality, Uncertainty, and Time*. Oxford: Oxford University Press.

Budolfson M. and D. Spears 2019. The hidden zero problem: effective altruism and barriers to marginal impact. In *Effective Altruism: Philosophical Issues*, ed. H. Greaves and T. Pummer, 184–201. Oxford: Oxford University Press.

Centre for Effective Altruism. 2021. CEA's Guiding Principles. <https://www.centreforeffectivealtruism.org/ceas-guiding-principles/>.

Clough E. 2015. Effective altruism's political blind spot. Boston Review, 14 July. <http://bostonreview.net/world/emily-clough-effective-altruism-ngos>.

Cohen J. and W. Easterly, eds. 2009. *What Works in Development? Thinking Big and Thinking Small*. Washington, DC: Brookings Institution Press.

Cooper R. 1999. *Coordination Games*. Cambridge: Cambridge University Press.

Deaton A. 2013. *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, NJ: Princeton University Press.

Deaton A. and N. Cartwright 2018. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* **210**, 2–21.

Dietz A. 2019. Effective altruism and collective obligations. *Utilitas* **31**, 106–115.

Ellsberg D. 1961. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* **75**, 643–669.

Fishburn P.C. 1984. On Harsanyi's utilitarian cardinal welfare theorem. *Theory and Decision* **17**, 21–28.

Giving What We Can. 2021. Charity Evaluation: How to Find an Effective Charity. <https://www.givingwhatwecan.org/research/methodology>.

GiveWell. 2013. VillageReach. <https://www.givewell.org/international/charities/villagereach#Doesitwork>.

GiveWell. 2016a. Our Principle for Assessing Evidence. <https://blog.givewell.org/2012/08/17/our-principles-for-assessing-evidence/#Properties>.

GiveWell. 2016b. Why We Can't Take Expected Value Estimates Literally (Even When They're Unbiased). <https://blog.givewell.org/2011/08/18/why-we-cant-take-expected-value-estimates-literally-even-when-theyre-unbiased/>.

GiveWell. 2019. Guest Post: Proven Programs are the Exception, not the Rule. <https://blog.givewell.org/2008/12/18/guest-post-proven-programs-are-the-exception-not-the-rule/>.

GiveWell. 2020a. Our Top Charities. <https://www.GiveWell.org/charities/top-charities>.

GiveWell. 2020b. Our Criteria for Top Charities. <https://www.givewell.org/how-we-work/criteria>.

GiveWell. 2021. GiveWell's Cost-Effectiveness Analyses. <http://www.givewell.org/international/technical/criteria/cost-effectiveness/cost-effectiveness-models>.

Greaves H. and T. Pummer, eds. 2019. *Effective Altruism: The Philosophical Issues*. Oxford: Oxford University Press.

Hammond P.J. 1981. Ex-ante and ex-post welfare optimality under uncertainty. *Economica* **48**, 235–250.

Harsanyi J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* **65**, 309–321.

Hart S. 1989. Shapley value. In *The New Palgrave: Game Theory*, ed. J. Eatwell, M. Milgate and P. Newman, 210–216. London: Palgrave Macmillan.

Herzog L. 2015. (One of) effective altruism's blind spot(s), or: why moral theory needs institutional theory. Justice Everywhere. <http://justice-everywhere.org/international/one-of-effective-altruisms-blind-spots-or-why-moral-theory-needs-institutional-theory/>.

Horton J. 2017. The all or nothing problem. *Journal of Philosophy* **114**, 94–104.

Kamm F.M. 2007. *Intricate Ethics*. Oxford: Oxford University Press.

Karlan D. and J. Appel 2011. *More Than Good Intentions. How a New Economics Is Helping to Solve Global Poverty*. London: Penguin Books.

Leigh A. 2018. *Randomistas. How Radical Researchers Are Changing Our World*. New Haven, CT: Yale University Press.

Liao S.M. 2008. Who is afraid of numbers? *Utilitas* **20**, 447–461.

Lichtenberg J. 2015. Peter Singer's extremely altruistic heirs. The New Republic, 30 November. <https://newrepublic.com/article/124690/peter-singers-extremely-altruistic-heirs>.

MacAskill W. 2015. *Doing Good Better: Effective Altruism and a Radical New Way to Make a Difference.* London: Guardian Books.

MacAskill W. 2019. The definition of effective altruism. In *Effective Altruism: Philosophical Issues*, ed. H. Greaves and T. Pummer, 10–28. Oxford: Oxford University Press.

Mackie G. 1996. Ending footbinding and infibulation: a convention account. *American Sociological Review* **61**, 999–1017.

Nussbaum M. 1999. *Sex and Social Justice.* New York, NY: Oxford University Press.

Open Philanthropy Project 2017. Cause Selection. <http://www.openphilanthropy.org/research/cause-selection>.

Otsuka M. 2004. Skepticism about saving the greater number. *Philosophy & Public Affairs* **32**, 413–426.

Pummer T. 2016. Whether and where to give. *Philosophy & Public Affairs* **44**, 77–95.

Rubinstein J.C. 2016. The lessons of effective altruism. *Ethics & International Affairs* **30**, 511–526.

Rowe T. and A. Voorhoeve 2018. Egalitarianism under severe uncertainty. *Philosophy & Public Affairs* **46**, 239–268.

Sen A. 1999. *Development as Freedom.* Oxford: Oxford University Press.

Singer P. 1972. Famine, affluence, and morality. *Philosophy & Public Affairs* **1**, 229–243.

Singer P. 2009. *The Life You Can Save: How to Play Your Part in Ending World Poverty.* London: Picador.

Singer P. 2015. *The Most Good You Can Do: How Effective Altruism is Changing Ideas About Living Ethically.* New Haven, CT: Yale University Press.

Srinivasan A. 2015. Stop the robot apocalypse. *London Review of Books* **37**, 3–6.

Syme T. 2019. Charity vs. revolution: effective altruism and the systemic change objection. *Ethical Theory and Moral Practice* **22**, 93–120.

Taurek J.M. 1977. Should the numbers count? *Philosophy & Public Affairs* **6**, 293–316.

The Life You Can Save 2021. Best Charities. <https://www.thelifeyoucansave.org/best-charities>.

Todaro M.P. and S.C. Smith 2015. *Economic Development* (12th edn). New York, NY: Pearson.

Unger P. 1996. *Living High and Letting Die: Our Illusion of Innocence.* New York, NY: Oxford University Press.

de Waal A. 1997. *Famine Crimes: Politics and the Disaster Relief Industry in Africa.* Bloomington, IN: Indiana University Press.

Wenar L. 2011. Poverty is no pond. In *Giving Well: The Ethics of Philanthropy*, ed. P. Illingworth, T. Pogge and L. Wenar, 104–132. Oxford: Oxford University Press.

Wiblin R. 2015. Effective altruists love systemic change. 80,000 Hours Blog. <https://80000hours.org/2015/07/effective-altruists-love-systemic-change/>.

Nicolas Côté is an SSHRC post-doctoral research fellow at the University of Toronto. His doctoral research focused on axiological questions concerning the measurement of freedom and its value, as well as on the formal characterization of deontological moral theories and their behaviour under uncertainty. His present research focuses on the ethics of decision-making under severe uncertainty.

Bastian Steuwer is a post-doctoral associate at the Center for Population-Level Bioethics at Rutgers University. His research focuses on distributive justice, health priority-setting, the morality of saving people from harm, the ethics of risk, and discrimination. Email: steuwer@cplb.rutgers.edu