Effective altruism: an elucidation and a defence[1]

John Halstead,[2] Stefan Schubert,[3] Joseph Millum,[4] Mark Engelbert,[5] Hayden Wilkinson,[6] and James Snowden[7]

Abstract

In this paper, we discuss Iason Gabriel's recent piece on criticisms of effective altruism. Many of the criticisms rest on the notion that effective altruism can roughly be equated with utilitarianism applied to global poverty and health interventions which are supported by randomised control trials and disability-adjusted life year estimates. We reject this characterisation and argue that effective altruism is much broader from the point of view of ethics, cause areas, and methodology. We then enter into a detailed discussion of the specific criticisms Gabriel discusses. Our argumentation mirrors Gabriel's, dealing with the objections that the effective altruist community neglects considerations of justice, uses a flawed methodology, and is less effective than its proponents suggest. Several of the criticisms do not succeed, but we also concede that others involve issues which require significant further study. Our conclusion is thus twofold: the critique is weaker than suggested, but it is useful insofar as it initiates a philosophical discussion about effective altruism and highlights the importance of more research on how to do the most good.

[1] We are especially to grateful to Per-Erik Milam for his contribution to an earlier draft of this paper. For very helpful contributions and comments, we would like to thank Brian McElwee, Theron Pummer, Hauke Hillebrandt, Richard Yetter Chappell, William MacAskill, Pablo Stafforini, Owen Cotton-Barratt, Michael Page, and Nick Beckstead. We would also like to thank Rebecca Raible of GiveWell for her helpful responses to our queries. Finally, we are very grateful to Iason Gabriel for comments, criticisms and suggestions. The views expressed here are only those of the authors, and any mistakes are our own.

[2] Blavatnik School of Government, Oxford.
[3] Centre for Effective Altruism.
[4] National Institutes of Health.
[5] University of Maryland.
[6] Australian National University.
[7] GiveWell.

1. Introduction

In a recent paper in the *Journal of Applied Philosophy*, Iason Gabriel discusses a number of moral and methodological dimensions of effective altruism.[8] As philosophers who identify with effective altruism, we are pleased to see public discussion of the effective altruism community's views and actions, and grateful to Gabriel for prompting this conversation, as well as for his ultimate recommendation of support. However, we also believe that several of the objections that Gabriel discusses fail, and will attempt to show that in this paper.[9]

Our paper is structured as follows. In the next section, we argue against the view that effective altruism is predominantly utilitarianism applied to global poverty and health interventions supported by randomised controlled trials (RCTs) and disability-adjusted life year (DALY). Rather, effective altruism is much broader, from the point of view of ethics, cause areas, and methodology, respectively. We then enter into a more detailed discussion of Gabriel's arguments. Our argumentation mirrors Gabriel's structure, dealing with the objections that effective altruism neglects considerations of justice, uses a flawed methodology, and is less effective than its proponents suggest. We reject several of the objections, but also concede that many of the questions Gabriel raises are very difficult, and agree with him that considerable further research is required on them. Our conclusion is thus twofold: the critiques Gabriel raises are weaker than he suggests, but are useful insofar as they initiate a philosophical discussion about effective altruism and highlight the importance of more research on how to do the most good.

[8] Iason Gabriel, 'Effective Altruism and its Critics', *Journal of Applied Philosophy* Online First (2016): 1–17.
[9] It is not quite clear to what extent Gabriel himself endorses these objections. We discuss issues with interpreting Gabriel in the next section.

2. What is Effective altruism?

Gabriel distinguishes two versions of effective altruism - a 'thin' version and a 'thick' version. He defines the thin version as the view "...that 'we should do the most good we can' and that this involves using a substantial amount of our spare resources to make the world a better place...".[10] However, Gabriel thinks that the thin version does not quite capture what makes effective altruism "unique", and therefore opts to examine the thick version, according to which effective altruism:

> "...adopts a largely welfarist understanding of value. According to this view, good states of affairs are those in which suffering is reduced and premature loss of life averted. Second, it is broadly consequentialist, maintaining that we should do whatever maximizes the sum of individual welfare at all times. … Third, the movement takes 'a scientific approach to doing good', which means using tools such as cost-effectiveness analysis and randomization to help quantify and compare the impact of different interventions."[11]

Before we discuss this characterisation, it is useful to introduce a different distinction between two different ways of evaluating effective altruism. The first kind of evaluation focuses on the *definition of effective altruism*. There is now relatively broad consensus within the effective altruism community that effective altruism is using evidence and reason to figure out how to benefit others as much as possible, and taking action on that basis.[12] This

[10] Gabriel op. cit., p. 2.
[11] Gabriel op. cit., pp. 2-3.
[12] Will MacAskill, "Introducing CEA's guiding principles", 2017. <http://effective-altruism.com/ea/181/introducing_ceas_guiding_principles/>. Note that this post was written after Gabriel's article.

definition allows for a wide range of ethical aims, and methods to achieve those aims.

Another way of evaluating effective altruism focuses on the actions and recommendations of the effective altruist community.[13] It is possible that they in fact rest on more narrow ethical and methodological views than those expressed by the definition of effective altruism. Since Gabriel mainly seems to evaluate effective altruism using this second approach, we will focus on that, though we will also have some things to say about the first kind of evaluation. Note, though, that arguments against the actions and recommendations of the community are not necessarily valid against the idea of effective altruism as expressed by its definition.

With this clarified, we take issue with several parts of Gabriel's claims about the nature of effective altruism. Firstly, we should note that Gabriel does not always stick to his own definition of thick effective altruism. Most conspicuously, in his section "Is Effective Altruism Unjust?" he seems to assume that effective altruism is specifically utilitarian, rather than just "broadly consequentialist".[14] This is not correct if interpreted as a statement about the definition of effective altruism.

Moreover, while it is true that many leading figures within effective altruism are utilitarians, or are sympathetic to utilitarianism, there are also numerous people in the effective altruist community with other ethical outlooks. For example, the effective altruist group Giving What We Can has many high-profile nonconsequentialist, egalitarian, and/or prioritarian philosophers as members, such as Alex Voorhoeve, Adam Swift, Nir Eyal, and Ole Norheim. This ethical diversity is perhaps not surprising given that prominent normative

[13] Exactly how to define the effective altruist community is a tricky issue. It is generally agreed that to be a member of the effective altruism community, you have to live up to a certain set of criteria, including commitment to others and a scientific mindset (cf. MacAskill, 2017, op. cit). However, self-identification as a member of the effective altruism community also seems to be a necessary condition for membership. In practice, there is relative agreement within the effective altruist community on who counts as a member.
[14] Gabriel op. cit., p. 2.

arguments for many effective altruist actions are not dependent on *any* moral theory. As Jeff McMahan pointed out in a recent article, "Peter Singer's earliest arguments for a view even more radical than that of most effective altruists appealed in the first instance to a single widely held moral intuition and argued that consistency required those who accepted the intuition to give most of their wealth to the relief of extreme poverty".[15]

Second, Gabriel focuses nearly exclusively on charitable giving to alleviate global poverty and ill-health. Although Gabriel acknowledges that people in the effective altruist community do work on other cause areas, the reader is likely to come away with an inaccurate impression.[16] While it is true that these areas do receive a considerable amount of attention - particularly from organisations such as GiveWell and Charity Science - there are many effective altruist organisations working on other causes. For instance, the Future of Humanity Institute (University of Oxford) and the Machine Intelligence Research Institute focus on reducing the risks of human extinction, and Animal Charity Evaluators and Sentience Politics work on reducing animal suffering. Yet another effective altruist focus area concerns what in the community are sometimes called "meta" causes - ways to improve the world more indirectly, such as via increasing the effectiveness of direct efforts to improve the world. Organisations working within this area include the Center for Applied Rationality, which runs workshops in which participants learn to counter their cognitive biases, and 80,000 Hours, which gives evidence-based advice to young people who want to make a difference through their careers.

The variety of effective altruist organisations is mirrored by the variety of views among

[15] Jeff McMahan, 'Philosophical Critiques of Effective Altruism', 2016 <https://www.academia.edu/24333624/Philosophical_Critiques_of_Effective_Altruism>, p. 1.

[16] Gabriel op. cit., p. 13 and p. 15, fn 6. Analogously, if a paper purporting to criticize the environmentalist movement only focused on environmentalist work on climate change, and very briefly mentioned that environmentalists' work on other areas, then one would probably come away with an inaccurate view of environmentalism.

people who self-identify as members of the effective altruist community. In a 2015 survey of the priorities of members of the effective altruist community, no cause area was seen as the top priority by a majority of respondents. While it is true that global poverty was identified as their top priority by a plurality of respondents (37%), many other cause areas had considerable support, including animal welfare, the far future/existential risk, and improving rationality.[17]

Let us make one final clarification before moving on to the more specific criticisms of Gabriel's. In certain passages, it is not quite clear whether Gabriel finds the objections he discusses persuasive, or if he rather wants to highlight possible objections. In multiple places he contradicts earlier arguments of his against effective altruism, or seems to partly take them back. On the other hand, he also does write that "[t]his paper focuses on the thick version of effective altruism and demonstrates its weaknesses".[18] These mixed signals make the task of interpretation difficult. Our chief concern is, in any case, with the force of the arguments themselves, rather than with Gabriel's stance towards them.

3. Is Effective Altruism Unjust?

Gabriel begins by arguing that the thick version of effective altruism fails to pay sufficient heed to considerations of justice, including egalitarianism, prioritarianism, and rights.[19] As noted above, his argument proceeds on the assumption that thick effective altruism is a form of utilitarianism, which is inconsistent with the aforementioned principles. This is, as we have seen, not true of the definition of effective altruism, which is ecumenical between a range of

[17] Effective Altruism Hub, 'The 2015 Survey of Effective Altruists: Results and Analysis', <https://eahub.org/sites/eahub.org/files/SurveyReport2015.pdf>, p. 6.
[18] Gabriel op. cit., p. 3. Our emphasis.
[19] Gabriel op. cit., pp. 2-6.

moral theories.

However, in line with the discussion in the last section, it might be that his argument is directed against the actions and recommendations of the effective altruist community. It is a sociological fact that there are more members of the effective altruism community who believe in some of those moral theories - notably utilitarianism - than others.[20] Nonetheless, there is good reason to believe that effective altruist recommendations within global poverty and health are, in spite of that, not dependent on utilitarianism, but in fact would be endorsed by a range of moral theories. Gabriel's arguments appeal only to hypothetical (rather than actual) cases in which there is a supposed conflict between effective altruist recommendations and equality, priority, or rights and thus fail to show that effective altruist recommendations actually do rely on utilitarianism. Indeed, Gabriel notes that the recommendation of distributing malaria bed nets is robust across moral theories, including prioritarianism and rights-based theories.[21]

There are reasons to believe that Gabriel is right about this. The Against Malaria Foundation (AMF), which distributes malaria bed nets, is highly cost-effective[22] and benefits some of the worst-off people in the world: predominantly infants in Sub-Saharan Africa who would die because their families could not afford a protective bednet.[23] It is therefore reasonable to believe that AMF does very well on prioritarian, egalitarian, and sufficientarian criteria. Since the cost-effective global poverty charities that GiveWell has identified

[20] The 2015 Survey of Effective Altruists: Results and Analysis <https://eahub.org/sites/eahub.org/files/SurveyReport2015.pdf>.
[21] Gabriel op. cit., p. 15 fn. 25.
[22] GiveWell, 'Against Malaria Foundation (AMF)' <http://www.givewell.org/international/top-charities/amf>.
[23] Daniel Sharp and Joseph Millum, 'Prioritarianism for Global Health Investments: Identifying the Worst Off', *Journal of Applied Philosophy*, 2015; Lawrence M. Barat et al., 'Do Malaria Control Interventions Reach the Poor? A View Through the Equity Lens', *The American Journal of Tropical Medicine and Hygiene* 71, 2 suppl (2004): 174–8; Obinna Onwujekwe, Kara Hanson and Julia Fox-Rushby, 'Inequalities in purchase of mosquito nets and willingness to pay for insecticide-treated nets in Nigeria: challenges for malaria control interventions', *Malaria journal* 3, 1 (2004).

generally target the worst-off globally, this is likely to hold true for most of the rest of them.[24]

We turn now to rights. Gabriel asks us to imagine that people in the effective altruism community are faced with a choice over whether to fund a campaign for government regulation improving the working conditions of sweatshop workers. Assuming that the regulation would decrease workplace accidents but also decrease total welfare, Gabriel argues that most people in the effective altruism community would decline to fund the campaign.

As we understand Gabriel, his point is not that people in the effective altruism community would violate rights by declining to fund the campaign. Rather, his point is that effective altruism does not sufficiently take rights violations into account in cause and intervention selection.

Gabriel bases his assessment of what people in the effective altruism community would do on William MacAskill's argument that consumers should not boycott sweatshop goods. But this is a quite different issue from whether campaigns to improve factory safety should be supported.[25] Furthermore, while MacAskill does argue that we ought not to boycott sweatshops, his argument is not that doing so maximises total utility even at the expense of reduced workplace safety. Instead his argument is that sweatshop jobs are typically better for the sweatshop workers themselves than the jobs they would have if the sweatshop did not exist, which include "backbreaking, low-paid farm labour, scavenging, or unemployment".[26] Data on workplace safety is limited, but a 2013 survey in Guatemala found that 43% of surveyed sweatshop workers thought their current job was safer than the previous one, with only 10% saying it was less safe.[27] If sweatshops are indeed safer than the alternatives, then

[24] For discussion of this question, see James Snowden, "Is Charity About Helping the Poorest?" <https://www.givingwhatwecan.org/post/2016/06/is-charity-about-equality/> .
[25] William MacAskill, *Doing Good Better : Effective Altruism and a Radical New Way to Make a Difference* (London: Guardian Books, 2015), chap. 8.
[26] MacAskill, 2015, op. cit., p. 160.
[27] J. R. Clark and Benjamin Powell, 'Sweatshop Working Conditions and Employee Welfare: Say It Ain't Sew', *Comparative Economic Studies* 55, 2 (2013): 343–57.

the force of rights-based critiques of the purchase of sweatshop goods would seem to be lessened: it could even be argued that it is commendable from some rights-based views.[28] This arguably renders irrelevant the claim, made by Gabriel in a footnote,[29] that the right to workplace safety cannot be forfeited: this is not a case in which workers can choose between forfeiting and not forfeiting their right to safety.

However, it is true that MacAskill has argued that some people should pursue morally controversial careers, such as working for a petrochemical company, and donate their earnings to effective charities.[30] He defends this position on the basis that because one's labour is replaceable, the harm would have occurred anyway, and that the harm is foreseen rather than intended. Thus, he contends that nonconsequentialists should endorse earning to give in such morally controversial careers, even though they may violate rights. There may be reasonable disagreement on this issue. This said, it is worth noting that we are not aware of any effective altruist organisations who recommended rights violations.

We have shown that when it comes to global poverty interventions, effective altruist recommendations are more robust across moral theories than Gabriel's arguments imply (though we should emphasize that more research is needed). When we move beyond global poverty, however, moral disagreement makes more of a difference. There is substantial disagreement among people in the effective altruism community regarding *cause selection* – whether one ought to work on, e.g. global poverty, animal suffering, or existential risk. It is plausible that these cause selection disagreements are to a large extent due to moral disagreements (e.g. regarding how to value animal suffering relative to human suffering, or

[28] See the rights-based defence of sweatshops in Benjamin Powell and Matt Zwolinski, 'The Ethical and Economic Case Against Sweatshop Labor: A Critical Assessment', *Journal of Business Ethics* 107, 4 (2011): 449–72.
[29] Gabriel, n21.
[30] William MacAskill, 'Replaceability, Career Choice, and Making a Difference', *Ethical Theory and Moral Practice* 17, 2 (2014): 269–83.

how to value possible future persons), but more research is needed to give a precise account of the extent to which specific effective altruist recommendations are dependent on moral theory choice.

Gabriel argues that the effective altruism community should heed the issue of moral disagreement; that it should "address the problem posed by competing values by providing advice that is sensitive to different value systems, enabling advisees' values to bear on personal cause selection".[31] We can only agree. In fact, the Centre for Effective Altruism has provided a cause prioritization tool which does precisely that.[32]

4. Is Effective Altruism Blind?

In the second section of his article, Gabriel discusses several objections to the effective altruism community's methods: the claims that they suffer from observation bias, quantification bias, and instrumental bias, respectively.

Observation bias

Gabriel contends that the effective altruism community focuses overly narrowly on RCTs and that this "has led the movement to ignore other less tangible opportunities to do good".[33] He points out that RCTs have a number of familiar limitations. They are costly to run and can only be carried out on a relatively small-scale. This scope-restriction means they are not suited to evaluating country-scale initiatives, nationwide advocacy programs, or projects that function over a longer time period.

There are two problem with Gabriel's argument. The first is that the effective altruism

[31] Gabriel op. cit., p. 6.
[32] <https://www.effectivealtruism.org/cause-prioritization-tool/>.
[33] Gabriel op. cit., p. 7.

community does not, in fact, narrowly rely on RCTs, and the second is that Gabriel does not provide sufficient evidence that insofar as the effective altruism community does rely on RCTs, this reliance is unwarranted.

With respect to the first point, consider the Open Philanthropy Project, which GiveWell has set up in conjunction with the philanthropic foundation Good Ventures. The Open Philanthropy Project explicitly declares that it is "open to supporting high-risk, high-reward work, as well as work that could take a long time to pay off" and that it is willing to take "action in the face of uncertainty"; i.e. even if it lacks anything remotely resembling RCT-style evidence.[34] The Open Philanthropy Project offers grants and gives recommendations in a variety of cause areas, including US criminal justice reform, immigration policy, and global catastrophic risk reduction - none of which are supported by evidence from RCTs.[35]

On the other hand, it is true that GiveWell does rely to a quite large extent on RCTs (though it does not do so exclusively).[36] Overall, there is some amount of disagreement within the effective altruist community regarding how to compare interventions backed by rigorous evidence ("proven interventions") with speculative interventions, but there seems to be a consensus that it would not be rational to rule out all interventions whose effectiveness have not been demonstrated in RCTs as a matter of principle.

In fact, towards the end of his methodological section, Gabriel himself acknowledges that "[e]ffective altruists have ... tried to incorporate new forms of evidence and analysis into

[34] Open Philanthropy Project, 'Vision & Values', 2016 <http://www.openphilanthropy.org/about/vision-and-values>. See also Holden Karnofsky, 'Hits-based Giving' *Open Philanthropy Project* <http://www.openphilanthropy.org/blog/hits-based-giving>.
[35] http://www.openphilanthropy.org
[36] For example, some of the evidence for deworming comes from a natural experiment. See Hoyt Bleakley, 'Disease and Development: Evidence from Hookworm Eradication in the American South', *The Quarterly Journal of Economics* 122, 1 (2007): 73–117.

their thinking"[37] and goes on to briefly discuss how the effective altruist community in fact does use forms of evidence other than RCTs.[38]

Turning to our second point, Gabriel merely conjectures that insofar as the effective altruist community does rely on RCTs, this reliance causes it to unduly overlook effective giving opportunities. Much more needs to be done to successfully defend this conclusion than giving general arguments regarding the costs and limitations of RCTs. We would require a rigorous argument indicating that there are charities which produce greater expected value than currently recommended charities, in spite of the fact that the evidence in support of them is weaker.

Quantification bias

Next, Gabriel alleges that the effective altruist community suffers from 'quantification bias'. His focus here is the effective altruist community's significant reliance on the DALY metric, which he thinks excludes important considerations, such as the value of hope.[39] He also mentions that cost-effectiveness analysis ignores the significance of 'iteration effects'. We are unsure what this means, but he may be referring to indirect effects, which we discuss in the next subsection.

The first thing to say is that this argument rests on the premise that the effective altruist community does in fact rely to a significant extent on the DALY metric. But as Gabriel later notes, even the effective altruist organization which is perhaps most strongly associated with the use of quantitative metrics, GiveWell, has somewhat distanced itself from the DALY metric: "generally [we] make at least some attempt to convert impact into units

[37] Gabriel op. cit., p. 10.
[38] See our comments about how best to interpret Gabriel in section 2.
[39] Hence his criticisms do not concern quantification in general, but rather the DALY metric, and the effective altruism community's use of it, in particular.

of... DALYs, a common metric in public health, though we do not always find these units helpful or make them a key input into our recommendations."[40] Elsewhere in the community, there is still less reliance on the DALY metric: the Open Philanthropy Project makes relatively little use of it, and it is almost never used by those working on existential risk and animal welfare.

With this clarified we can return to Gabriel's contention that the DALY metric excludes other important contributors to subjective wellbeing, such as hope.[41] It is widely acknowledged within the effective altruist community that the DALY metric is at best an approximate measure of subjective wellbeing as it only counts health effects, which are but one determinant of wellbeing. This is why many people in the effective altruist community are supportive of calls within moral philosophy and welfare economics for the use of a Wellbeing Adjusted Life Year (WALY) metric, which would include health and non-health components of well-being.[42] However, because the WALY metric has yet to be fully worked out and, unlike the DALYs metric, is not widely used, it is arguably not yet workable. We may therefore have to wait until this is rectified before using WALYs.[43]

This said, it is possible that the discrepancy between DALYs and WALYs sometimes gets us to the wrong answers about how to do the most good. Gabriel explores this issue by discussing how we ought to prioritise HIV/AIDS funding between antiretrovirals and

[40] GiveWell 'Cost-effectiveness', <http://www.givewell.org/international/technical/criteria/cost-effectiveness>.
[41] Gabriel op. cit., p. 8. At the start of the section on quantification bias, Gabriel argues that the effective altruism community's focus on the DALY measure has led it to exclude values other than subjective well-being. He then seems to go on to argue that the effective altruism community's disregard of the value of hope is evidence for that thesis. However, when explaining the value of hope, he says that it is valuable precisely because it improves subjective well-being: "Hope is valuable in this context either because it leads people to feel better about their lives or because it simply is feeling better about one's life."
[42] MacAskill, 2015, op. cit., chap. 2. For discussions of WALYs, see, for example, John Broome, *Weighing Lives* (Oxford: Oxford University Press, 2004); Paul Dolan and Daniel Kahneman, 'Interpretations Of Utility And Their Implications For The Valuation Of Health', *The Economic Journal* 118, 525 (2008): 215–34.
[43] However, for a thorough attempt to operationalise the WALY metric see John Bronsteen, Christopher Buccafusco and Jonathan S. Masur, 'Well-being analysis vs. cost-benefit analysis', *Duke Law Journal* 62, 1603 (2013).

condoms, assuming that the latter would avert more DALYs per dollar. He argues that funding antiretrovirals may nevertheless be justified because "it may be better to live in a society where one can hope to receive medical treatment if one is sick than to live in one where the largest numbers of people get treated overall".[44] Thus, the health-based gain provided by condoms (which may be captured by both the DALY and the WALY metric) may be outweighed by the hope-based gain of antiretrovirals (which may only be captured by the WALY metric). While this could in principle be true, Gabriel fails to give us any evidence, but merely conjectures that this is so. Thus while we agree that the DALY metric is a less than perfect guide to well-being, we also want to emphasise that only careful research on the DALY and WALY burden of different diseases can show whether the use of one rather than the other makes a difference to how we ought to allocate resources in some particular case.

Instrumental bias

Gabriel concludes this section by arguing that the effective altruist community has an 'instrumental bias'.[45] This objection to effective altruism, which solely concerns the community's work on global poverty and health, comprises several distinct claims:[46]

> (1) The overly narrow cost-effectiveness analysis used by the effective altruist community ignores values associated with community participation, such as autonomy and self-respect.
>
> (2) Overly narrow cost-effectiveness analysis ignores important instrumental side-effects of community involvement, such as community buy-in and improvements

[44] Gabriel op. cit., p. 8
[45] Gabriel op. cit., p. 8.
[46] Gabriel op. cit., pp. 8-10.

in human-capital.

(3) There is a strong *pro tanto* reason to favour democratic government leadership over private philanthropy when it comes to addressing social needs because:

a. Politicians are accountable to the electorate for the decisions they make, unlike private foundations.

b. Unlike democratically produced policy, private philanthropy is not under pressure to meet demands of public justification.

c. Under certain circumstances service provision by non-state actors can diminish state capacity, impacting disproportionately on the worst-off.

d. Many of the worst development failures have occurred at the hands of experts freed from democratic oversight.

Before we go into the substantive arguments, it should be noted that as we understand Gabriel's argument, only claim (1) could be thought a criticism of the effective altruist community on the grounds that it neglects the purported intrinsic value of democracy.[47] The others are criticisms on instrumentalist grounds - grounds that concern the quality of outcomes.

Reasonable people will differ on claim (1) and obviously we cannot settle that issue here. It is, however, worth making one general point with regard to the effective altruist community's relationship with values such as community participation and democracy. Since the recipients of effective aid are disproportionately those who currently lack political power, it is reasonable to think that providing highly-effective aid to improve health or alleviate poverty is likely to improve political equality, autonomy, and self-respect, rather than the

[47] It is also debatable whether self-respect is really related to the intrinsic value of democracy. Rather, this seems to be a contingent side-effect of democracy.

converse.

Turning to claim (2), it should be noted that there is disagreement within the effective altruism community about how many indirect effects we should consider when evaluating an intervention. It is true that GiveWell only focuses on the direct impact of interventions and that they therefore ignore the myriad indirect effects of interventions.[48] One reason is that indirect effects are difficult and expensive to estimate and that making a robustly positive difference has, in their view, a better track record than trying to make a potentially larger but highly uncertain difference. Another reason is their belief that human empowerment is likely to have large positive indirect ("flow-through") effects, which means that interventions with large direct effects also are likely to have large indirect effects.[49] This means, in turn, that we can focus on direct effects, as they are a sufficiently good proxy for the total effects of interventions. However, others in the effective altruism community disagree with this line of argument, arguing that indirect effects are likely to be too strong relative to direct effects for this to be a sound strategy.[50] Since this issue involves huge questions in social science and moral philosophy,[51] again, we cannot hope to settle it here, but instead leave it to future discussion.

Claim (3a) contends that democratic government is accountable to the electorate, unlike effective altruist foundations, and that this is a reason to favour service provision by the former rather than the latter. However, as we discuss below, the charities favoured by the effective altruist community largely operate in weak state settings, which in any case often

[48] GiveWell, 'Cost-effectiveness' op. cit.
[49] GiveWell confirmed this in personal email correspondence, 29th September 2016. For discussion see Holden Karnofsky, 'The moral value of the far future' *The GiveWell Blog*, 2014 <http://blog.givewell.org/2014/07/03/the-moral-value-of-the-far-future/>; Holden Karnofsky, 'Flow-through effects' *The GiveWell Blog*, 2013 <http://blog.givewell.org/2013/05/15/flow-through-effects/>
[50] See Paul Christiano, 'On Progress and Prosperity' *Effective Altruism Forum*, 2014 <http://effective-altruism.com/ea/9f/on_progress_and_prosperity/>.
[51] For discussion see for example Hilary Greaves, 'The Social Disvalue of Premature Deaths' in I. Hirose and A. Reisner (eds.) Weighing and Reasoning (Oxford University Press, 2015): 72–86.

have very weak democratic accountability. Moreover, there may be inefficiencies associated with state provision in weak state settings, such as corruption, which could outweigh the potential accountability gains.

Regarding claim (3b), it should first be said that it is highly controversial whether public or private action must meet demands of public justification.[52] Second, it is difficult to see how most of the activities of the effective altruist community that relate to global poverty and health, such as preventing malaria and giving out unconditional cash transfers, could fail to meet standards of public justification. It is plausible that all reasonable theories of justice agree that all people should have a minimally decent standard of living.[53] If so, then no reasonable theory of justice could disagree with most of the actions of the effective altruist community that aim to alleviate global poverty and ill-health.

In support of claim (3c), Gabriel cites a piece by Emily Clough in the Boston Review arguing that the effective altruist community neglects the unintended negative effects of aid on political institutions.[54] This is a very complex issue which we cannot hope to settle conclusively here, but it is worth noting that that there are some problems with Clough's argument. As Hauke Hillebrandt pointed out in his response to Clough in the same publication, even if claim (3c) is true in some cases, it is not true of the interventions supported by the effective altruist community.[55] As Clough herself concedes, "In the short term... in weak-state settings [where the charities recommended by the effective altruist

[52] For a version of public justification, see for example Charles Larmore, 'The Moral Basis of Political Liberalism', *The Journal of Philosophy* 96, 12 (1999): 599–625. For criticism of public justification see for example David Enoch, 'Against public reason' in D. Sobel, P. Vallentyne and S. Wall (eds.) *Oxford Studies in Political Philosophy*. Volume 1 (Oxford: Oxford University Press, 2015).
[53] For a defence of this claim see John Rawls, *The Law of Peoples : With, The Idea of Public Reason Revisited* (Cambridge, Mass; London: Harvard University Press, 1999).
[54] Emily Clough, 'Effective Altruism's Political Blind Spot', *Boston Review*, 2015 <https://bostonreview.net/world/emily-clough-effective-altruism-ngos>.
[55] Hauke Hillebrandt, 'Effective Altruism, Continued: On Measuring Impact', *Boston Review*, 2015 <http://bostonreview.net/blog/hauke-hillebrandt-giving-what-we-can-effective-altruism-impact>.

community operate] the marginal likelihood of harm to state institutions from NGO service is clearly swamped by welfare gains".[56] Indeed, Hillebrandt presents evidence that development assistance for health by nonprofits has actually increased domestic government health spending.[57]

Finally, while (3d) may be true, it does not follow that the interventions the effective altruist community favours run the risk of causing harm in this way and Gabriel makes no attempt to show that that is the case.

5. Is Effective Altruism Effective?

Gabriel concludes his paper by discussing two objections to the effective altruist community's claims to effectiveness: that they do not properly take counterfactual effects of donations into account, and that they do not to a sufficient degree take action to initiate systemic change.

Counterfactuals

Gabriel's argument concerning the counterfactual impact of individual donors raises some intriguing issues for the effective altruist community.[58] The effective altruist community is well-known for arguing that for a person to have impact, it must be the case that her contribution would not have been made by someone else if she had chosen to act differently.[59] However, Gabriel argues that counterfactual impact arguments could also undermine the

[56] Emily Clough, 'Response to Hauke Hillebrandt', *Boston Review*, 2015 <http://bostonreview.net/blog/emily-clough-response-hauke-hillebrandt>.
[57] We do not find Clough's rebuttal in her reply persuasive for the reasons laid out by Hillebrandt in the comments to her response.
[58] Gabriel op. cit., pp. 11–12.
[59] MacAskill, 'Replaceability, Career Choice, and Making a Difference' op. cit.

claim that individual donors to effective charities have impact in the way one might naively assume.

The effective altruist community is now made up of a wide range of actors – from college students to multibillion dollar philanthropic organisations like Good Ventures. Gabriel argues that if a small donor stopped donating to an effective charity, these organisations would take up the slack. Therefore, a current small donation does not increase the charity's budget at all: the elasticity of the charity's budget to a current small donation is zero. Therefore, if these individual donors have any impact, it is not through preventing malaria or parasitic disease, but through some other mechanism. More formally, the argument is:

> (1) For any current small donor, D, if D stopped donating $X to the currently recommended charities, then a large donor would increase funding by $X.
>
> (2) If it is true that if D stopped donating $X to the currently recommended charities, then a large donor would increase funding by $X, then D does not have impact in the way that the effective altruist community suggests.
>
> (3) Therefore, current small donors do not have impact in the way that the effective altruist community suggests.

According to premise 1, due to the behaviour of large donors, the elasticity of a charity's budget to a current small donor's donation is zero. However, there is reason to believe that in fact large donors would not behave in the way suggested and that this mechanism therefore would not make the elasticity zero (though there may be other mechanisms which do; see below). In defending premise 1, Gabriel only specifically mentions Good Ventures. Good

Ventures is currently the most important effective altruist donor, as of 2015, accounting for around two thirds of the money moved by GiveWell, which itself moves the most money of the effective altruist organisations.[60] However, GiveWell explicitly instructs Good Ventures not to take up the slack left by small donors, on the basis that doing so would reduce overall donor interest in effective giving.[61]

Instead, GiveWell advises Good Ventures to pledge to commit a 'fair share' of the funding gap, which is determined independently of other funders' behaviour. Furthermore, GiveWell have told us that they have not yet worked with any other donors comparable in size or mission to Good Ventures and similarly able to fill most existing gaps, and that they would give such donors the same advice.[62] This suggests that the mechanism posited by Gabriel would not occur and that we should at least reserve judgement on premise 1 until further evidence becomes available.

Moreover, as Gabriel appears to accept, if the argument were sound it would not establish that donating to currently recommended charities is not the most effective way to do good. It would establish that a small donation of $X would not have counterfactual impact *via the currently most effective charities*. However, the donation would increase the funds available to large foundations and, provided the large foundations were committed to effective altruist principles, $X would be shifted to the most effective *counterfactually unfunded* charity, whether available now or in the future. If so, the small donation would counterfactually cause the maximal possible increase in marginal net benefits, which is what

[60] Tyler Heishman, 'GiveWell's money moved and web traffic in 2015' *The GiveWell Blog*, 2016 <http://blog.givewell.org/2016/05/13/givewells-money-moved-web-traffic-2015/>.
[61] Holden Karnofsky, 'Good Ventures and Giving Now vs. Later' *The GiveWell Blog*, 2015 <http://blog.givewell.org/2015/11/25/good-ventures-and-giving-now-vs-later/>. See also Holden Karnofsky, 'Donor coordination and the "giver's dilemma"' *The GiveWell Blog*, 2014 <http://blog.givewell.org/2014/12/02/donor-coordination-and-the-givers-dilemma/>. See also the main article and Holden Karnofsky's comments in response to Ben Hoffman, 'GiveWell and the problem of partial funding' *Effective Altruism Forum* <http://effective-altruism.com/ea/17e/givewell_and_partial_funding/>.
[62] *GiveWell*, personal email correspondence, 17th June 2016.

effective altruist donors should ultimately aim to do. For the reasons laid out in the preceding counterfactual argument, it need not always be true that the only way to meet this ultimate aim is to counterfactually cause the maximal possible marginal increase in the spending of the currently most effective charities.

Another potential worry mentioned by Gabriel is that since there are few effective counterfactually unfunded opportunities in the short to medium-term future, individuals will not in fact have counterfactual impact even by the more indirect mechanism set out in the previous paragraph. It is obviously very difficult to say whether this is true, but there are some reasons to doubt it.

First consider the funding needs of currently recommended charities. Gabriel states that currently recommended charities have limited room for more funding, roughly in the tens of millions.[63] The funding gap situation has changed since Gabriel wrote his piece. For the coming year, the combined funding gap of all of GiveWell's currently recommended charities probably exceeds $150m.[64] Nonetheless, it is true that a billionaire donor could, if they were willing to do so, fill the current funding gap of currently recommended charities.

Looking to the medium-term, on some estimates, the funding gap over the next fifteen years for very effective possible interventions within global health alone (that is, not including cash transfers and other non-health interventions) is in the hundreds of billions, which far exceeds the funding capacity of all existing effective altruist donors.[65] However, at present there does appear to be a shortage of highly effective charities to implement some of these possible interventions, so the scale of the funding gap for possible effective

[63] Gabriel op. cit., p. 16 n45.
[64] Figures are the aggregate funding gap of recommended charities, available on the GiveWell website. See <http://www.givewell.org/charities/top-charities>.
[65] Dean T. Jamison et al., 'Global health 2035: a world converging within a generation', *The Lancet* 382, 9908 (2013): 1898–1955.

interventions does not guarantee that donors will be able to find effective charities in the medium term future. There is therefore work to be done in setting up new highly effective charities. Effective altruist groups appear to be making progress in this area. For example, after an examination of the public health cost-effectiveness literature, the effective altruist organisation Charity Science recently set up Charity Science: Health, which uses mobile phones to remind people of their vaccination deadlines.[66] The evidence suggests that this charity could have comparable cost-effectiveness to GiveWell-recommended charities. Moreover, the pressure to create such charities is partially driven by the availability of funds, which may be an indirect benefit of donating, even if doing so fills all current funding gaps. This is one way in which the counterfactual impact of donations might be *greater* than it at first appears.

In general, more research is needed on several issues pertaining to the counterfactual impact of donations. There are a number of mechanisms bearing on the counterfactual impact of donations and further research needs to examine these carefully.[67]

Systemic change

Gabriel's final argument contends that the effective altruist community ignores or even obstructs systemic institutional change that would, according to the objection, actually do the most good. Proponents of this objection appear to use the term 'systemic change' in at least two different ways. Some, such as Amia Srinivasan, Matthew Snow, and Brian Leiter, use it

[66] See http://www.charitysciencehealth.com/

[67] On this see for example Mark Budolfson and Dean Spears, 'Effective Altruism, Marginal Impact, and Fundraising: Weak Links in Effective Altruism's Chain'; Carl Shulman, 'Annual "splitting" of funding gaps can be partial funging when gaps carry over across years' Reflective Disequilibrium, 2016 <http://reflectivedisequilibrium.blogspot.co.uk/2016/08/annual-splitting-of-funding-gaps-can-be.html>.

to refer specifically to overthrowing global capitalism,[68] whereas others use it to refer more generally to any kind of political change. Gabriel appears to have the second sense in mind, and so our argument will focus on the broader kind of systemic change.

Gabriel gives the following argument for the view that the effective altruist community ignores, and stand in the way of, systemic change:

> (1) Due to the psychological framing effects of their approach to advocacy, the effective altruist community is likely to believe in a moral hierarchy between given and receiver, unlikely to develop an accurate understanding of the systemic causes of poverty, and unlikely to work for systemic change.[69]
>
> (2) Due to their commitment to cause neutrality, the effective altruist community is unlikely to demand systemic change and is poorly equipped to bring it about.[70]
>
> (C) The effective altruist community ignores, and stand in the way of, systemic change.

However, after presenting this argument, he also says that "in fact, effective altruists are increasingly interested in systemic change...",[71] which, at the very least, appears to be tension with (1), (2) and (C). Thus, Gabriel's stance towards this argument is somewhat unclear (cf

[68] Amia Srinivasan, 'Stop the Robot Apocalypse' *London Review of Books*, 24 September 2015, 37, 18: 3–6; Matthew Snow, 'Against Charity', *Jacobin*, 2015 <https://www.jacobinmag.com/2015/08/peter-singer-charity-effective-altruism/>; Brian Leiter, 'McMahan's defense of "effective altruism" (EA) against the philosophical critics' *Leiter Reports: A Philosophy Blog*, 2016 <http://leiterreports.typepad.com/blog/2016/04/mcmahans-defense-of-effective-altruism-against-the-philosophical-critics.html>.

[69] Gabriel op. cit., p. 12.

[70] Gabriel op. cit., pp. 12-13. The cause neutrality notion has been used in several different senses within the effective altruism community. However, it is clear that Gabriel means what Stefan Schubert calls "cause-impartiality": selecting causes without prejudice, based on impartial estimates of impact. That is also how we use the term below. Cf. Stefan Schubert, "Understanding cause-neutrality", 2017. <https://www.centreforeffectivealtruism.org/blog/understanding-cause-neutrality/>.

[71] Gabriel op. cit., p. 13.

our general comments on how to interpret Gabriel in section 2). In any case, we will rebut the argument.

In our view, the papers cited in support of (1) - by Hattori, Darnton and Kirk, and Vohs, Meade and Goode, respectively - do not provide adequate evidence for that claim.[72] Hattori conjectures that aid legitimises a moral hierarchy between giver and receiver, lessens pressure on the systemic causes of poverty, and undermines the autonomy of recipients, but only provides general theoretical arguments to that effect. Little empirical evidence showing that aid really has these detrimental effects is presented. The report by Darnton and Kirk performs a discourse analysis of a staged conversation between at most 21 people (none of whom identify as being part of the effective altruist community).[73] The discourse analysis notes that many participants used terms such as 'aid' and 'charity', and conjectures that usage of these terms in part causes the participants to be more likely to hold what the authors believe to be false beliefs about poverty.[74] However, little evidence is given for these conjectures; for example, no attempt is made to consider the possibility that the causation runs in the other direction (from allegedly false beliefs to usage of particular terms), or that there is no causal connection between the two.

Finally, Gabriel misinterprets the conclusions of the Vohs et al paper. The study purports to show that being primed with money in various ways – including reading sentences with monetary connotations, having money in one's field of vision, and being given money – dampens altruistic inclinations. It does *not*, *pace* Gabriel, purport to show that

[72] Tomohisa Hattori, 'The Moral Politics of Foreign Aid', *Review of International Studies* 29, 2 (2003): 229–47; Andrew Darnton and Martin Kirk, *Finding Frames: New Ways to Engage the UK Public in Global Poverty* (London, 2011); Kathleen D. Vohs, Nicole L. Mead and Miranda R. Goode, 'The Psychological Consequences of Money', *Science* 314, 5802 (2006): 1154–6. Gabriel also cites Martin Kirk, 'Beyond Charity: Helping NGOs Lead a Transformative New Public Discourse on Global Poverty and Social Justice', *Ethics & International Affairs* 26, Special Issue 02 (2012): 245–263, but the only relevant evidence appealed to here is the Darnton and Kirk report.

[73] Darnton and Kirk op. cit., pp. 2–3.

[74] Darnton and Kirk op. cit., chap. 4.

focusing on donating money dampens altruistic inclinations, and it therefore does not show that focusing on donations in some way makes systemic change harder to achieve. Moreover, numerous high-power replications of currency priming studies, many of which involve priming manipulations that were identical to or similar to the manipulations that were used by Vohs et al, have failed to replicate the original results.[75] In addition, Vohs et al observed multiple null effects but did not report any in the published paper. According to Rohrer et al, "these numerous null effects increase the chance that the money priming phenomena reported by [Vohs et al and others] are not real".[76]

These issues aside, the effective altruist community is, as Gabriel later says, doing significant work on systemic change. Due to its commitment to cause neutrality, effective altruism has always been open to systemic change, in principle. Cause neutrality requires that support of a cause is based solely on the basis of the amount of good it does, rather than on the basis of some other factor, such as personal connection to the cause. People in the effective altruist community only support malaria charities insofar as they produce the most good impartially conceived, and do not support them for agent-relative reasons, such as that they had a sibling who died from malaria. If it turned out that, for instance, supporting a charity working to change international trade rules maximised expected value, then cause neutrality would dictate that we ought to give to that charity rather than one working in global health.

Claim (2) denies the foregoing line of argument. Gabriel argues that the effective altruist community's commitment to cause neutrality may make it too eager to change

[75] See Richard A. Klein et al., 'Investigating Variation in Replicability: A "Many Labs" Replication Project', *Social Psychology* 45, 3 (2014): 142–52; Doug Rohrer, Harold Pashler and Christine R. Harris, 'Do subtle reminders of money change people's political views?', *Journal of Experimental Psychology: General* 144, 4 (2015): e73–85.

[76] Rohrer, Pashler and Harris op. cit., p. e75. Vohs responds in Kathleen D. Vohs, 'Money priming can change people's thoughts, feelings, motivations, and behaviors: An update on 10 years of experiments.', *Journal of Experimental Psychology: General* 144, 4 (2015): e86–93.

priorities so that it gives up too soon on organisations pushing for systemic change or misses new opportunities that may arise. Gabriel believes that this may "detract from [effective altruism]'s effectiveness".[77] Note, however, that cause neutrality alone is not sufficient to establish this conclusion. If systemic change is indeed the most effective kind of intervention, then cause neutrality simply requires that we support systemic change organisations. Thus one would need to add some other premise to establish that the effective altruist community is inclined to give up too soon on organisations pushing for systemic change. Gabriel does not establish that the effective altruist community is short-sighted in this way, and the fact that it includes organisations working on difficult long-term causes such as human rationality and existential risk would seem to be some indication to the contrary.

Indeed, there is clear and mounting evidence that the effective altruist community is concerned with systemic change *in practice*. Effective altruist organisations such as the Future of Humanity Institute and Sentience Politics pursue effective altruist causes through policy work. Arguably, the most important effective altruist group in this area is Open Philanthropy Project, which since 2012 has given out $87m in grants to groups working for political change in areas such as immigration policy, farm animal welfare, global catastrophic risks, and criminal justice reform.[78] In light of this, the claim that the effective altruism community ignores systemic change seems mistaken.

Proponents of the systemic change objection, whether understood as an objection regarding political change broadly construed or as one regarding anti-capitalism specifically, thus far have approached the issue in the wrong way. Figuring out which organisations are the most effective is extremely difficult: there are no obvious or easy answers. Therefore, to

[77] Gabriel op. cit., p. 13.

[78] See <http://www.openphilanthropy.org/giving/grants>. Gabriel acknowledges the work of the Open Philanthropy Project in a footnote (Gabriel op. cit., p. 17 fn. 53).

show that the effective altruist community is unduly neglecting organisations that work for systemic change, one would need to show, or at least provide some indicative considerations for, the following: for some *specific* systemic change-focused organisation, according to the best available evidence and arguments on the likelihood of the organisation bringing about systemic change and the benefits of such change, the expected value of supporting the organisation exceeds that of supporting organisations currently recommended by the effective altruist community which do not work for systemic change, such as the Against Malaria Foundation. It is clearly inadequate merely to conjecture that the effective altruist community neglects systemic change, just as it would be inadequate for the effective altruist community merely to conjecture that the Against Malaria Foundation is more effective than charities working in other areas, such as microfinance.

6. Conclusions

We hope that this paper has helped to clarify the nature of effective altruism. On the ethical side, effective altruism is not a purely utilitarian affair. Regarding evidence, they use other sources than RCTs. And when it comes to cause areas, global poverty and health is but one of several cause areas. The debate on effective altruism's pros and cons would gain from an appreciation of this breadth.

In addition to this, we have rebutted several of the specific criticisms raised by Gabriel - though we do not claim to have provided a comprehensive treatment of all of them. More research is needed on a number of issues, such as the extent to which specific effective altruist recommendations are dependent on moral theory choice; how to choose between proven and speculative interventions; alternatives to the DALY metric; to what extent the

effective altruism community should take indirect effects into account when estimating the impact of interventions; the myriad mechanisms affecting the counterfactual impact of donations; and on whether the effective altruism community should support charities pushing for systemic change.

Figuring out how to do the most good is an ongoing project. In tackling it, we are inevitably confronted with a host of difficult moral and empirical questions. One of our aims here has been to point out some of the questions that need to be answered, and to make clear how much research will be needed to answer them.