Why Not Effective Altruism?[*]

Richard Yetter Chappell

November 16, 2023

Abstract

Effective altruism sounds so innocuous—who could possibly be opposed to doing good, more effectively? Yet it has inspired significant backlash in recent years. This paper addresses some common misconceptions, and argues that the core "beneficentric" ideas of effective altruism are both excellent and widely neglected. Reasonable people may disagree on details of implementation, but all should share the basic goals or values underlying effective altruism.

Introduction

Effective altruism (EA) is defined as "the project of trying to find the best ways of helping others, and putting them into practice."[1] This involves trying to maximize the philanthropic return on our investments of moral effort, time, and resources. Those who self-identify

[*] Forthcoming in *Public Affairs Quarterly* (please cite final published version). Thanks to Susanne Burri, Oscar Delaney, Theron Pummer, Tina Rulli, Thomas Sinclair, Helen Yetter-Chappell, an anonymous referee, and participants on the EA Forum, for helpful discussion and comments. This work was supported by a University of Miami Fellowship in the Arts & Humanities.

[1] 'What is Effective Altruism', https://www.effectivealtruism.org/articles/introduction-to-effective-altruism, accessed 4/30/2023.

as pursuing this project—effective altruists, or "EAs"—have coalesced around four main cause areas as their current best guesses for doing good effectively,[2] but the details aren't essential, and could easily be revised in light of new evidence. What's essential to the project is instead a *cause agnostic* commitment to following the evidence where it leads, in service of most effectively helping others.[3]

Effective altruism is sometimes confused with utilitarianism. It shares with utilitarianism the innocuous claim that, all else equal, it's better to do more good than less. But EA does not entail utilitarianism's more controversial claims. It does not entail hegemonic impartial maximizing: the EA project may just be one among many in a well-rounded life. (Nothing in EA demands that you put *all* of your resources towards the EA project.) Nor does it entail the

[2] These are: (i) global health and development, especially via GiveWell-recommended organizations such as the Against Malaria Foundation; (ii) non-human animal welfare, especially factory farmed animals, but with increasing attention to wild animal suffering as a neglected area of concern; (iii) global catastrophic risks, especially pandemic and AI risk; and (iv) EA community building, to increase the resources being directed to the prior three cause areas. Compare the four "EA Funds" at https://www.givingwhatwecan.org/funds/effective-altruism-funds, accessed 4/30/2023.

[3] This broadly *welfarist* understanding of the goal to be pursued—"helping others", in the sense of *promoting their well-being*—may distinguish EA from other (as yet merely hypothetical) optimizing normative movements such as "Effective Justice" (Crisp and Pummer 2020), "Effective Aesthetics" (Chappell 2022b), and so on.

rejection of deontic constraints against instrumental harm. Deontologists can seek to help others effectively, within the constraints of permissibility. I've elsewhere described the underlying philosophy of effective altruism as "beneficentrism"—or "utilitarianism minus the controversial bits"—that is, "the view that promoting the general welfare is deeply important, and should be amongst one's central life projects."[4] Utilitarians take beneficentrism to be *exhaustive* of fundamental morality, but others may freely combine it with other virtues, prerogatives, or pro tanto duties.

EA, so understood, is theoretically rather tame. Some proponents claim that the EA project, or something like it, is morally obligatory,[5] but I make no such claim here. I just claim that it is *good,* and that participating in the project is morally *better* than failing to do so. This is enough to prompt disagreement from EA's critics, many of whom do not seem to regard it as morally good at all.[6]

More specifically, I claim that *some* form of the EA project has these moral qualities, and that pursuing the project *in the right*

[4] Chappell (2022a).

[5] E.g., Pummer (2023).

[6] See, e.g., Adams, Crary, and Gruen (2023), or the vituperative critics cited in McMahan (2016).

way is among the morally best things that one can do. Sometimes critics take themselves to be attacking a narrower target: "EA as it actually exists", or some such. But these critics tend to be ignorant of the actually-existing diversity of opinion and approaches within EA (it's a big tent!), and have a tendency to straw-man their targets. To address the narrow criticisms, one would need to get into the sociology of EA, and assess the extent to which critics' characterizations accurately describe their targets. Such sociological questions are of limited philosophical interest.[7] From the perspective of first-personal moral deliberation about how to live our lives, the important question is not whether we should blindly defer to actual EAs (of course we shouldn't), but whether there is

[7] If some EAs seem to be making a contingent mistake that's entirely incidental to their core principles, it strikes me as misleading to frame such incidental allegations as criticisms of *effective altruism* as such. Moreover, the attributions often don't even strike me as descriptively accurate, which makes them seem especially pointless to engage with. Two additional reasons for my philosophical focus: (1) The degree of hostility many critics express towards EA doesn't make sense if they agree with EA principles and simply disagree about how best to apply them. One doesn't see these critics say, "EA is a great idea, and here's how we could do it better." Their disagreement seems deeper than that, and what I'm interested in assessing here is whether such deeper disagreement is really reasonable. (2) Disagreements of principle strike me as more tractable than politically-tinged empirical disagreements. I know how to argue about ethical principles. I don't even know where to begin with someone who thinks that "abolishing capitalism" is our best bet for improving the world, or that answers to economic questions are best found by consulting "eco-feminists" rather than economists. So I think it best to bracket such matters (where we may simply "agree to disagree"), and instead seek common ground on a more theoretical level.

some version of the EA project that is worth pursuing. So that is the question that this paper will address.

This paper focuses on the four main ideas that have prompted backlash from EA's critics: moral prioritization, earning to give, billionaire philanthropy, and longtermism. In each case, I argue, there's room for reasonable disagreement about the details, but the core idea can't reasonably be rejected. Everyone should agree that the *core* moral claims of EA in these four areas are simply correct. I close by briefly considering a more political mode of critique.

1 Moral Prioritization

While many people are motivated to do (some) good, it's rare to optimize our moral efforts—even when some efforts could do orders of magnitude more good than others, at no greater cost.[8] The core idea of effective altruism is that we should generally try to achieve more good rather than less with our investments of moral effort, time, and money. This innocuous-sounding principle turns out to be surprisingly controversial. Here I'll quickly address five

[8] Ord (2013).

major reasons why some might balk at the idea of optimization.

First, it's *prescriptive*, and many may naturally prefer to retain personal discretion over how they direct their moral efforts and philanthropic resources. Moral prescriptions might be thought to entail social censure when violated, whereas you surely shouldn't be liable to social censure simply for doing good suboptimally.

But not all prescriptions entail social censure when violated. Consider the possibility of moral guidance within the supererogatory. Whatever you choose is wonderful—already above and beyond the call of duty—but some options may still be *better* than others. In such a case, a rational moral agent may be guided to select the best action; to pick a worse option would be, in a sense, a practical *mistake*. But it's not a "mistake" of the sort that others have standing to criticize the agent for. The agent may still be entirely praiseworthy; just not *as* praiseworthy as they could have been, given that some choices are undeniably better than others.

We may think of the minimal core of EA as being prescriptive in this weaker, censure-free sense.[9] After determining what's most

[9] It's an interesting open question, which I won't attempt to settle here, whether stronger claims of obligation may also be warranted. See Pummer (2023) for further discussion.

worth doing, it's a further question where to draw the line for moral criticism. Plausibly, *some* philanthropic choices are so egregiously wasteful, given the immense opportunity costs, that criticism *could* be warranted. But I agree that it generally isn't reasonable to criticize someone for supporting an excellent cause merely because it wasn't the *optimal* cause. This is structurally similar to how it may be reasonable to criticize a comparatively wealthy person who donates *nothing* to charity, but it doesn't seem reasonable to criticize everyone who fails to sacrifice the impartially optimal amount.[10] In both cases, we may nonetheless be drawn to the weak prescriptions of scalar consequentialism:[11] it is always better to do better, and we can recognize this without committing ourselves to any particular threshold for what we regard as an *unacceptably* poor showing.

Second, one might worry that attempts to optimize are self-defeating due to *measurability bias*: favoring easily verifiable and quantifiable small wins over the uncertain prospect of pursuing

[10] One important difference between the amount we donate and the effectiveness of our selections is that increasing the former is more costly to the giver. This creates a natural excuse for failing to give more. Failing to *direct* a donation optimally, by contrast, seems a more gratuitous form of moral suboptimality. For related discussion, see Pummer (2016). On the impermissibility of gratuitous suboptimality, see also Chappell (2019b).

[11] Norcross (2020).

larger-scale (e.g., political or institutional) reforms.[12] But this doesn't follow at all: if measurability bias is recognizably harmful, then optimizing requires us to counteract it. Indeed, EA funders aren't shy about supporting speculative efforts to reduce existential risk, when they judge this to offer a worthwhile bet *in expectation*. It's not easy to *quantify* the value of such efforts, but many EAs nonetheless support them because they believe that almost *any* reasonable assignment of values robustly yields the conclusion that such efforts can be extremely worthwhile.

This seems to be a common cause of confusion. Alice Crary, Lori Gruen, and Carol J. Adams, in an essay titled 'The predictably grievous harms of Effective Altruism',[13] claim that:

[12] For excellent discussion of this "institutional critique", see Berkey (2018). Broi (2019) offers a narrower interpretation of the critique, according to which Effective Altruists are contingently biased against the possibility that some causes may have *increasing* (rather than *diminishing*) marginal returns. I don't see any reason to think that this is a bias rather than simply a first-order empirical disagreement about particular cases. If large EA grantmakers were presented with compelling *evidence* that a large grant towards "systemic change" would have disproportionately large expected benefits, better than anything available to uncoordinated smaller donors, I have every expectation that they would fund it. Moreover, I fully expect that there would be significant efforts on the EA Forum to start co-ordinating smaller donors to get in on the opportunity. The notion that EA is "individualistic" in the sense of being *inherently opposed to coordination*, no matter the potential benefits, is just bizarre.

[13] https://blog.oup.com/2022/12/the-predictably-grievous-harms-of-effective-altruism/, accessed 5/4/2023.

See also Srinivasan (2015): "What's the expected marginal value of becoming an anti-capitalist revolutionary? To answer that you'd need to put a value and probability measure on achieving an unrecognizably different world—even, perhaps, on our becoming unrecognizably different sorts of people. It's hard

> To step inside the utilitarian frame is to accept that values that count as "good" can be abstractly quantified. Its methods leave it incapable of addressing historically sedimented structural injustices and intergenerational injuries, since these aren't the sorts of things that can be quantified by EA-style metrics.

Unfortunately, they never explain why they believe this. Any such estimate would of course be very rough and subject to dispute. But the same is true of estimating the value of the entire future of humanity, and hence the disvalue of premature extinction, and that hasn't stopped EAs from thinking about the latter. Some argument needs to be given to establish the strong claim that the harms of racism, for example, are strictly impervious to estimation when the harms of total human extinction are not. If the worry is

enough to quantify the value of a philanthropic intervention: how would we go about quantifying the consequences of radically reorganizing society?"

As explained below, at least a rough ballpark estimate in answer to these questions would seem necessary in order to have a justified belief that becoming an anti-capitalist revolutionary is *actually a good idea*. If you're truly *clueless* about the expected consequences of an action, it's hard to see much reason to do it. It would seem especially indefensible to pass up *saving someone's life* because you prefer to take a gamble that you don't even think is clearly positive in expectation. For more on cluelessness and expected value (including the important point that we may seek to maximize expected value without necessarily making explicit *calculations*), see 'The Cluelessness Objection' in Chappell, Meissner, and MacAskill (2023): https://www.utilitarianism.net/objections-to-utilitarianism/cluelessness/.

just that there is no credible estimate that would yield the desired result that fighting "historically sedimented structural injustices" should take priority over (say) saving lives from malaria, the critic's beef is not with measurability bias, but more fundamentally with the cause-neutral commitment to doing more good rather than less.

Some opposition may stem from assuming an unduly narrow view of what constitutes *evidence*. One could reasonably worry that strict constraints on allowable forms of evidence could easily prove counterproductive. If, e.g., we needed evidence from randomized controlled trials to justify our actions, there would be very little we could justifiably do. But of course allowable evidence should not be so constrained—any *epistemic reason* can count.[14]

Data from randomized controlled trials provides especially strong evidence, and is worth pursuing where possible, but justified credences may also be informed by weaker forms of evidence. For example, if one believes that civil rights activism has a track record of vastly improving society, one could appeal to that historical

[14] Many EAs made precisely this point during the pandemic, while public health authorities misleadingly claimed we had "no evidence" about whether infection granted immunity, or whether experimental vaccines were safe and effective; obviously base rates from past viruses and vaccines provides *some* evidence, however inconclusive and subject to revision.

evidence in making an inductive case for assigning comparably high expected value to comparable forms of activism today. Such estimates are likely to be highly uncertain and contestable; but that's just to say that it's highly uncertain and contestable how worthwhile it is to engage in activism today. And that is surely true. (Crucially, EA principles are open to uncertain and contestable means to doing the most good. It just depends on what your total evidence *truly* supports. And this can be hard to know!)

This suggests a simple dilemma for those who claim that EA is incapable of recognizing the need for "systemic change". Either their *total evidence* supports the idea that attempting to promote systemic change would be a better bet (in expectation) than safer alternatives, or it does not. If it does, then EA principles straightforwardly endorse attempting to promote systemic change. If it does not, then by their own lights they have *no basis* for thinking it a better option. In neither case does it constitute a coherent objection to EA principles.

Sometimes, the institutional critique is stated in ways that presuppose that "complicity" with suboptimal institutions *entails* net harm. For example, Adams, Crary, and Gruen (2023, xxv) write:

> EA's principles are actualized in ways that support some of the very social structures that cause suffering, *thereby* undermining its efforts to "do the most good." (emphasis added)

This is terrible reasoning. It's entirely possible—indeed, plausible—that you may do the most good by supporting some structures that cause suffering. For one thing, even the best possible structures—like democracy—will likely cause some suffering; it suffices that the alternatives are even *worse*. For another, even a suboptimal structure might be too costly, or too risky, to replace. So the fact that it "support[s] some of the very social structures that cause suffering" is *no reason at all* to think that EA fails to "do the most good."

But again, if there turns out to be good reason to believe that current EA priorities are actually doing *more* harm than good, then that's precisely the sort of thing that EA principles are concerned with (and that actual EAs are open to hearing, if presented with evidence—an epistemic task that the quoted authors never attempt).

Third, one might insist that remedying (local) social injustice should take priority over general beneficence. For example, some might think it more important for Americans to protect their compatriots from racism than to protect Nigerians from malaria, even if the latter efforts would do more to improve their beneficiaries' well-being. This would make most sense if one thought one had a *special obligation* to resist local injustice, much as parents plausibly have special obligations to care for their children.

But even if that's so, we could still think that EA principles provide the best account of our more general reasons of beneficence. So then we must ask how reasons of justice and beneficence relate. There's no essential conflict here: a conciliatory view might hold that one should satisfy one's special obligations (on the assumption that they are limited in scope, not excessively demanding), and that *after* that point it is especially morally excellent to pursue the project of effective altruism. I don't see any obvious reason for social justice advocates to oppose EA principles, so understood.

Of course, others of us may dispute the assumed priority of social justice, especially if it demands resources that could be used elsewhere to better effect.[15] Proponents of beneficence may argue that, whatever our local special obligations, it would surely be

[15] For further questioning of the common—but rarely supported—assumption that justice automatically takes priority over beneficence, see Barrett (2022).

indecent—unjust, even—to neglect the greater plight of the global poor. So there remain important disputes over moral priorities that I cannot settle here. While I personally see a lot of appeal to the utilitarian view that *it's always morally better to help people more than to do anything else that overall helps less*, it is at least worth noting the availability of more conciliatory positions for those who prioritize other values (within limits).

Fourth, one might think that optimizing approaches to beneficence unfairly "abandon" those who are less easily helped. A 2016 article in the magazine *Third Sector* vividly explained the concern as follows:[16]

> [Chief Executive of Oxfam GB] Goldring says it would be wrong to apply the EA philosophy to all of Oxfam's programmes because it could mean excluding people who most need the charity's help. For a certain cost, the charity might enable only a few children to go to school in a country such as South Sudan, where the barriers to school attendance are high, he says; but that does not mean it should work only in countries where the cost of

[16] http://www.thirdsector.co.uk/effective-altruism-will-donors-change-ways/fundraising/article/1384629, accessed 5/5/2023. See also Gabriel (2017).

> schooling is cheaper, such as Bangladesh, because that would abandon the South Sudanese children.

Of course, we can all agree that ideally we'd like to be able to help everyone. But the idea that optimized philanthropy entails greater "abandonment", when faced with resource constraints, is precisely backwards. When we do not have enough resources to help everyone, it is inevitable that *some* will be "abandoned", or left without the aid that they need. EA principles suggest that we should try to *minimize the burden* of that abandonment. For example, if Oxfam spends some of its budget educating 100 children in South Sudan, when they instead could have educated an additional 1000 children in Bangladesh, that choice means that a *greater* number of children (i.e., 900 more) will be abandoned. That seems objectionable.

The problem is that a focus on not abandoning any salient *groups* may entail abandoning a far greater number of *individuals*. Yet it is surely individuals, not groups, that *ultimately* matter. It would not be fair to the abandoned 1000 Bangladeshi children to prioritize a smaller number of South Sudanese children over them, merely because they belong to the same social group as some *other* children that you've already helped.[17]

As this example suggests, some resistance to moral prioritization may stem from a reluctance to face up to real trade-offs. We might feel better if we can delude ourselves into believing that, by helping every *group* (according to some salient partitioning of individuals into groups), we have thereby helped *everyone*.[18] But of course that isn't true—as we can emphasize by considering the *unhelped* Bangladeshi children as a separate (and larger) group that is "abandoned" by Oxfam's refusal to prioritize by cost-effectiveness—and the proper aim of moral action is not to delude ourselves into feeling better.

Fifth and finally, some might find prioritization politically, ideologically, or personally inconvenient. For example, Sanbonmatsu (2023, 211) laments "the over-valorization of billionaires and fi-

[17] Alternatively, if there is some reason why educating additional children in the same country has diminishing returns—e.g., perhaps a small core of educated citizens in a country greatly improves the prospects for future development—then that simply calls for revisiting which option is truly "optimal", all things considered.

[18] A related thought is that we have thereby given everyone more of a *chance* of aid, or perhaps a more *equal* chance of aid. But this also doesn't follow: it will depend on the precise details of the selection process. Worse, it's unclear why it would even *matter*: if we care about chances at all, why does "God's lottery" (Walden 2014), determining who is in the more cost-effective group to aid, not count as random enough? Now, I think these kinds of "fairness" intuitions could be an interesting place for critics of EA to focus, in developing a possible critique of welfarist prioritization. But it needs to be *developed.* Merely noting that helping some entails abandoning others is not yet an objection to helping more and abandoning fewer.

nanciers in EA discourse, and a corresponding undervalorization of grass-roots activists and radicals." It's surely an empirical possibility that generous rich people do *more good* than "grass-roots activists and radicals", but many seem uncomfortable acknowledging this as a moral possibility. Many authors in the same volume complain that they are no longer so competitive for funding when funders are guided by EA principles.[19] And wealthy academics in developed countries have obvious social and financial incentives to prefer moral ideologies that valorize *saying the right things* over *opening their wallets*.[20] So even if EA principles were entirely morally correct, we should still expect them to inspire backlash from those advantaged by more traditional conceptions of ethical life and decision-making.

[19] Sanders (2023, 7) even charges that failing to fund her work, as "a Black activist [working] in Black communities", is "upholding white supremacist ideas about which communities are worthy of support and which ones aren't."

[20] Cf. the well-established social phenomenon of "do-gooder derogation", as discussed in Minson and Monin (2012). We all know that vegans face a lot of unjustified hostility from omnivores suffering from cognitive dissonance. It would be extremely surprising if effective altruism failed to motivate similar unjustified hostility, since it is so plainly contrary to the material interests of the currently comfortable.

2 Earning to Give

One of the most distinctive and controversial recommendations to emerge from Effective Altruism is the idea that altruists should consider highly-paid jobs—such as in the financial sector, corporate law, or as cosmetic surgeons—over more conventionally "ethical" careers in the social sector. By earning extra money and then donating it, a career in cosmetic surgery could do more good—and thus, by EA lights, be more ethical—than working as a (less well-paid) family physician or social worker.[21]

This is obviously a very revisionary way of thinking about what constitutes an ethical career. But it's hard to deny the basic moral insight that helping people *indirectly* is still a morally good and worthwhile thing to do.[22] We're used to thinking of "altruistic careers" as ones that *directly* help people, but "earning to give" is, at least in principle, an equally legitimate way to do good via one's

[21] MacAskill (2015, 76–78).

[22] I'm inclined towards the stronger claim that it is *no less important*, in principle, than helping them directly. It would seem morally self-indulgent (rather than genuinely altruistic) to prefer to be personally closer to the moral action even at the cost of doing *less* good for the ultimate beneficiaries. But that isn't essential to my argument here. Even someone who gives *some* extra weight to more direct modes of helping should still recognize that indirect aid is *also* good, and could be morally better than a *sufficiently* less effective form of direct aid. I don't imagine anyone could seriously defend the view that direct assistance has *lexical* priority over more indirect ways of helping people, no matter how much more good we would achieve via the latter.

career (so long as one reliably follows through with the “giving” part).

Of course, there may be limits to this. A career as a hitman or drug dealer may violate deontic constraints, the wrongness of which cannot (by deontological lights) be “offset” by doing more good via one’s donations.[23] Moral theorists may argue about precisely which directly harmful careers could, or could not, be justified by indirectly saving more lives. But these edge cases are a distraction from the core idea, much as an excessive focus on the ethics of Robin Hoodery would be a distraction when evaluating the basic case for giving more to the poor. In both cases, we can simply limit our attention to increasing one’s donations *via permissible means*.

Rare exceptions aside, most careers are presumably permissible. The basic idea of earning to give is just that we have good moral reasons to prefer better-paying careers, *from among our permissible options*, if we would donate the excess earnings. There can thus be excellent altruistic reasons to pursue higher pay. This

[23] And even utilitarians can appeal to instrumental reasons to endorse commonsense constraints in practice. See ‘Respecting Commonsense Moral Norms’ in Chappell, Meissner, and MacAskill (2023): https://www.utilitarianism.net/utilitarianism-and-practical-ethics/#respecting-commonsense-moral-norms.

claim is both true and widely neglected. The same may be said of the comparative claim that one could easily have *more* moral reason to pursue "earning to give" than to pursue a conventionally "altruistic" career that more directly helps people. This comparative claim, too, is both true and widely neglected. Neither of these important truths is threatened by the claim that one should not pursue an *impermissible* career. The relevant moral claim is just that the *directness* of our moral aid is not intrinsically morally significant (or at least not of *overwhelming* moral significance), so a wider range of possible actions are worth considering, for altruistic reasons, than people commonly recognize.

3 Billionaire Philanthropy

Aside from prioritization and earning to give, another major source of backlash to effective altruism is the movement's courting of big donors. Of course, it makes sense that if billionaires exist, we should prefer that they spend their money in ways that effectively help others. And billionaires, notoriously, do exist. So we should prefer that they spend their money in ways that effectively help others.

Many regret that our political and economic system is set up in such a way as to allow such extreme inequality to arise in the first place. Rather than engaging in that first-order dispute, I want to suggest that it is strictly irrelevant to how we should assess EA principles.[24] There is nothing inconsistent about both (i) trying to change the system to make it more egalitarian, and (ii) until such a time as those efforts succeed, encouraging those with excessive wealth to dispose of it in better rather than worse ways.

EA explicitly acknowledges the fact that billionaire philanthropists are capable of doing immense good, not just immense harm. Some find this an inconvenient truth, and may dislike EA for highlighting it. But I do not think it is objectionable to acknowledge relevant facts, even when politically inconvenient.

Alternatively, if one believes that there are compelling arguments that billionaire philanthropy *necessarily* does more harm than good, then they might instead conclude that the best thing billionaires can do is voluntarily pay more taxes (i.e., donate to

[24] One might instead raise questions about how reliance on a small number of billionaire funders might distort EA *organizations*, culture, etc., but those sorts of concerns are beyond the scope of this paper.

the US Treasury). That would be a surprising result,[25] and I doubt that many actually believe it, but it is at least conceptually possible. But even that would be no *objection* to EA principles, but just a possible *implication* of them (when combined with unusual empirical assumptions). Unless critics seriously want billionaires to deliberately try to do *less* good rather than more, it's hard to make sense of their opposing EA principles on the basis of how they apply to billionaires.[26] Clear thinking requires us to acknowledge that political antipathy towards billionaires should not bleed into philosophical antipathy towards how EA principles apply to billionaires.

[25] The US budget makes no pretense of even *attempting* to impartially promote the good, and the politicians who haggle over its details are among the least-trusted members of society.

[26] Some argue that billionaires are not morally entitled to *discretion* over how to dispose of their fortunes, if illegitimate. See Cordelli (2017). But this point seems friendly to EA, insofar as it suggests that billionaires may be morally *required* to donate in the best way possible, rather than according to their personal whims or inclinations. Optimizing for justice rather than global welfare may lead to subtly different recommendations. But there seems likely to be significant overlap between the two goals. It would seem hard to deny that donating to classic EA recommendations in global health & development would do more to substantively promote justice than would donating to the US Treasury, for example.

4 Longtermism

One final source of backlash is *longtermism*: "the idea that positively influencing the longterm future is a key moral priority of our time." (MacAskill 2022, 4.) Some of the backlash is presumably due to the idea's endorsement by people, like Elon Musk, with whom the critics do not wish to associate. But I take it that "guilt by association" is not a philosophically reputable way to assess ideas. (Hitler's vegetarianism gives us no reason to torture animals.)

Many longtermists appeal to the potential "astronomical" value of the future (Bostrom 2003) to explain why reducing existential risks should be a top priority (insofar as we can identify feasible means to do so).[27] Many people find this intuitively weird: "How can you be thinking about future millennia when there are people suffering in the here and now?" But I think this reaction is not ultimately defensible.

Critics of effective global giving discourage giving equal consideration to the many distant needy when there are locals who would benefit from being prioritized instead. "Charity begins at home,"

[27] This may include via indirect means like basic research to improve our future epistemic position and capabilities for identifying and addressing possible existential risks.

these critics tell us.[28] But I trust that most readers of this paper are sufficiently cosmopolitan to agree that we should not ignore the greater plight of children dying of malaria overseas, merely because they are geographically distant from us. We can—and should—intellectually appreciate that "statistical lives" are every bit as real as the ones we see before our eyes. But distance in time seems no more intrinsically significant than distance in space. So we should not be moved by appeals to strictly prioritize the more easily identifiable individuals of the "here and now". We should want to help people, and bring about a better world, without (geographic or temporal) restriction.

A more serious distinction arises between *improving* lives vs *enabling them to exist* (for example, by averting a risk of human extinction prior to their conception). Some critics of longtermism draw on theses within population ethics to argue that human extinction would not be such a big deal.[29] If successful, these objections could provide principled grounds for giving less weight

[28] Or: "How can you be thinking about distant countries when there's a homeless person here right before your eyes!?" For further discussion of this critique, see Chappell (2019a).

[29] See, e.g., Setiya, 'The New Moral Mathematics', https://www.bostonreview.net/articles/the-new-moral-mathematics/, accessed 5/8/2023.

than longtermists do to the importance of averting extinction.[30]

While I think those objections fail, first note that even *they* do not rule out longtermism in its broadest sense. Longtermism is a big tent, which includes room for "asymmetric" views of population ethics on which additional miserable lives are bad but additional happy lives are merely neutral (rather than good). Such views still imply that we should be concerned about the risk of dystopian futures containing immense suffering (or "S-risks"). If there is a non-trivial chance of such S-risks eventuating, reducing these risks should plausibly be a key moral priority: astronomical suffering is not something to be viewed lightly, on any account.[31]

I'll now argue that specifically *life-affirming* longtermism (on which additional good lives are actually good, not merely neutral) should be widely accepted. This is because it offers significant theoretical benefits without corresponding costs.

To bring this out, it's important to distinguish the general life-

[30] Though Shulman and Thornley (forthcoming) argue that mitigating global catastrophic risks is still likely to be cost-effective even just considering the interests of people already alive today.

[31] A major cost of accepting the asymmetry is that it would seemingly imply—absurdly—that voluntary extinction would be the best possible outcome. Extinction would remove the risk of bad future lives, while the loss of good future lives doesn't count as a "loss" at all on this view. To avoid this implication, I think we should reject the asymmetry and instead acknowledge that (good) future lives have value, and the loss of that value would be a bad thing.

affirming view (that, all else equal, it is better for more happy lives to exist) from total utilitarianism. Philosophers often focus on the stark contrast between total utilitarianism (which treats creating happy lives as interchangeable with improving existing lives) and asymmetric person-affecting views (which deny that we have *any* non-instrumental reason to bring more lives, however happy, into existence). But neither of these extremes is commonsensical. A moderate alternative would combine the bare life-affirming view with extra weight for those who exist antecedently.[32] On such a view, we may have both (i) strong person-directed reasons to care especially about the well-being of antecedently existing individuals, and (ii) weaker impersonal reasons to promote value by bringing additional good lives into existence. When the amount of value at stake is sufficiently large, even reasons of the intrinsically weaker kind may add up to be very significant indeed. This can explain why avoiding human extinction should be a very high priority on a wide range of reasonable, life-affirming views, without depending on anything as extreme as total utilitarianism.

Some mistakenly fear that life-affirming longtermism entails

[32] McMahan (2013); Chappell (2017).

repugnant tradeoffs between quantity and quality of life, allowing any finite utopian population of flourishing lives to be outweighed by a sufficiently larger population of barely-positive lives.[33] But the answer to how to limit quantity-quality tradeoffs cannot be found by denying that additional lives are good at all. The same puzzle will simply re-emerge within a life, where it is undeniable that adding additional good moments is at least sometimes good (rather than strictly neutral).

If we find a principled way to balance quantity and quality within a life, the same principles could presumably be applied at the population level. If such principles can be found, then life-affirming longtermists can use them to avoid the repugnant conclusion. If no such principles are possible, *everyone* must make their peace with "repugnance". In neither case is anything gained by denying that new lives can have non-instrumental value.

Now consider the benefits of endorsing a life-affirming population ethics. It allows us to take at face value the commonsense claim that life can be *worth* celebrating. Unlike Benatar (2006), we need not regret each new life as a new site for potential suffering

[33] This is Parfit (1987)'s "repugnant conclusion".

with no moral upside. We get to recognize flourishing lives as genuinely (i.e., *absolutely*, not merely comparatively or conditionally) good.[34] This is far more morally respectable, I believe, than quasi-nihilistic alternatives that see all of existence, no matter how flourishing, as no better than an empty void.

Neutrality about future lives implies that utopia is (in prospect) no better than a barren, lifeless rock. It implies that the total extinction of all future value-bearers could be more than compensated for by throwing a good enough party for those who already exist. These strike me as *deeply* abhorrent claims, and I don't see any good reason to accept them.

We would all do better to embrace life-affirming longtermism. Only by appreciating that our children's lives are absolutely valuable, and capable of (constitutively, non-instrumentally) contributing to the world's being a better place, do we regard them with the full respect that they deserve. Of course we have additional—more partial—reasons to care for our children in particular. But it would be a mistake to entirely neglect their potential for partly constituting (or non-instrumentally contributing to) the value of the world;

[34] Cf. Frick (2020) for a population-ethical theory based on reasons that are conditional on an individual's existence.

a potential that is equally shared by all their possible descendants. Full appreciation and respect for their value as persons requires affirming this (commonly realized) potential for absolute value. It should not leave us coldly indifferent, and nor should the prospect of future life.

Of course, it remains an open question how to implement a concern for protecting future generations. You could accept life-affirming longtermism in principle while remaining highly uncertain about what should be done in practice. Longtermists can disagree about whether to prioritize (i) specific risk-mitigating interventions, or (ii) more general investigation into possible risks and responses, or (iii) more general societal (ethical, scientific, and economic) progress and capacity-building so that future generations can do a better job than we at tackling future problems. Maybe there are other options too. I leave open such questions of implementation. I'm merely arguing that we should all agree on the in-principle importance of the long-term future.

5 Political Critique

In her Foreword to Adams, Crary, and Gruen (2023, xi), Amia Srinivasan writes:

> Political critique does not, and should not, merely address what social and political movements say about themselves… [but also] what effects they systematically bring about in the world, which structures they tend to reinforce, and which people they empower and which they silence.

In this paper, I've argued that common *intellectual* critiques of effective altruist principles fail. I would consider the core EA principles to be truisms if they weren't so widely neglected, as they seem almost impossible to reasonably deny. But it's always possible that true claims might be used to ill effect in the world. Many objections to effective altruism, such as the charge that it provides "moral cover" to the wealthy, may best be understood in these *political* terms.

I don't think philosophers have any special expertise in adjudicating empirical disagreements over the expected consequences of accepting or rejecting different moral principles. So I'll merely note two general reasons for being wary of this charge.

First, I think we should have a strong default presumption in favour of truth and transparency. While it's always *conceivable* that "noble lies" could be justified, we should generally be very skeptical that lying about morality would *actually* be for the best.[35] In this particular case, it seems especially implausible that discouraging people from *trying to do good effectively* is a good idea. (To illustrate the risks, consider that if you convince just *one* person *not* to take a course of action—such as earning to give—that would have led to their donating an extra $50k per year to GiveWell's top charities, then you are causally responsible for approximately ten people's deaths per year. That's really bad!) I can't rule out the possibility that accepting EA principles would somehow cause more harm than good, but it sure would be surprising. So there's a high bar for allowing political judgments to override intellectual ones.

Second, political judgments seem especially prone to bias. It's striking that these "naïve utilitarian" calls for esotericism (effectively: lying about the truth of effective altruist principles) are

[35] Lazari-Radek and Singer (2010).

exclusively coming from non-utilitarians, i.e. those who aren't primarily concerned with promoting the impartial good to begin with. To bring out why I find this suspicious, consider how the political critique may be turned on its head. One obvious real-world effect of *denouncing* EA principles is that these denunciations provide "moral cover" for the morally complacent: those who do not wish to donate more of their money to effective charities, or rethink their choice of career, or entertain the possibility that wealthy philanthropists may warrant more esteem than they.

I worry that polemical public denunciations and stigmatizing dismissals of effective altruism will predictably have the result that fewer people perform acts of effective altruism, resulting in more unnecessary death and suffering. Yet the vituperative critics display no apparent concern about this grave moral risk. It's hardly wild speculation: the best available evidence suggests that every $5k donated to GiveWell's top charities saves a child's life on average.[36] And effective altruists give a *lot* to GiveWell's top charities.[37] Other EA cause areas are more speculative, but are

[36] https://www.givewell.org/impact-estimates#Impact_metrics_for_grants_to_GiveWells_top_charities, accessed 10/22/2023.

[37] Open Philanthropy alone regranted $350 million to GiveWell in 2022: https://www.openphilanthropy.org/research/update-on-our-planned-allocation-to-givewells-recommended-charities-in-2022/, accessed 10/22/2023. Small

widely judged to have even greater cost-effectiveness in expectation. Undermining all this without an extremely good basis seems incredibly irresponsible.

Of course, anyone can make mistakes, and it is important to be able to correct these when they occur. So it's a good thing that concrete criticism of current EA priorities tends to be warmly welcomed: EA organizations literally ran a "criticism contest" in which they gave out $120,000 in prizes to the best entries.[38] But that's a very different thing from polemical denunciations of effective altruism *as such.* EA's loudest critics don't come right out and *say*, "Stop saving lives!" But that's the obvious real-world effect that they systematically bring about.

As an academic, I think we should assess claims primarily on their epistemic merits, not their practical consequences. But insofar as the political critique disavows this academic norm, it must also expose *itself* to practical evaluation. And in this case, the harm it risks is clear and grave. Political opponents of effective altruism

donors collectively gave millions more.

[38] https://forum.effectivealtruism.org/posts/YgbpxJmEdFhFGpqci/winners-of-the-ea-criticism-and-red-teaming-contest, accessed 10/22/2023. Someone wrote a well-received criticism of the criticism contest, prompting Scott Alexander to write 'Criticism of criticism of criticism', https://www.astralcodexten.com/p/criticism-of-criticism-of-criticism, accessed 10/22/2023.

have very likely caused the deaths of a great many children.[39]

Conclusion

The answer to our title question, ‘Why not effective altruism?’, is that there’s no principled reason why not. We should all want to do more good rather than less, and use the best available evidence to guide our efforts. There’s plenty of room for reasonable disagreement about how best to pursue this humanitarian goal. But its in-principle desirability cannot reasonably be disputed.

I’ve argued that the core effective altruist principle of *moral prioritization* is especially indisputable. I’ve further argued that other ideas associated with EA, such as earning to give and life-affirming longtermism, have similarly compelling theoretical bases. One *could* reject those further ideas while still embracing the core of effective altruism. But I’ve suggested that there seems no good reason for such rejections. One may certainly reject the most radically utilitarian *interpretations* of those ideas in favour of more commonsensical variants. But one shouldn’t throw the baby out

[39] In the counterfactual sense that, had they not acted thus, those deaths would not have occurred. Which is not, of course, to claim that they are the *direct* cause of death.

with the bathwater. On the contrary, I've argued that *wholesale* rejection of effective altruist ideas and principles would itself be intellectually indefensible.

Some may nonetheless argue that we can have good *political* reasons to bury inconvenient (or "harmful") truths. I grant that this is possible, but I think we should have a high bar for endorsing such dishonesty. I also worry that it's far more likely that *denunciations* of effective altruism function to provide "moral cover" for the morally complacent. Doing more good may not be in our self-interest, after all. But it is worth doing, nonetheless.

References

Adams, Carol J., Alice Crary, and Lori Gruen, eds. 2023. *The Good It Promises, the Harm It Does: Critical Essays on Effective Altruism.* Oxford University Press.

Barrett, Jacob. 2022. "Social Beneficence." *GPI Working Paper No. 11-2022.*

Benatar, David. 2006. *Better Never to Have Been: The Harm of Coming into Existence.* Oxford: Oxford University Press.

Berkey, Brian. 2018. "The Institutional Critique of Effective Altruism." *Utilitas* 30 (2): 143–71. https://doi.org/10.1017/s0953820817000176.

Bostrom, Nick. 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308–14. https://doi.org/10.1017/s0953820800004076.

Broi, Antonin. 2019. "Effective Altruism and Systemic Change." *Utilitas* 31 (3): 262–76. https://doi.org/10.1017/s0953820819000153.

Chappell, Richard Yetter. 2017. "Rethinking the Asymmetry." *Canadian Journal of Philosophy* 47 (2): 167–77. https://doi.org/10.1080/00455091.2016.1250203.

———. 2019a. “Overriding Virtue.” In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer, 218–26. Oxford University Press.

———. 2019b. “Willpower Satisficing.” *Noûs* 53 (2): 251–65.

———. 2022a. “Beneficentrism.” https://rychappell.substack.com/p/beneficentrism.

———. 2022b. “The Nietzschean Challenge to Effective Altruism.” https://rychappell.substack.com/p/the-nietzschean-challenge-to-effective.

Chappell, Richard Yetter, Darius Meissner, and William MacAskill. 2023. *An Introduction to Utilitarianism*. www.utilitarianism.net.

Cordelli, Chiara. 2017. “Reparative Justice and the Moral Limits of Discretionary Philanthropy.” In *Philanthropy in Democratic Societies*, edited by Rob Reich, Chiara Cordelli, and Lucy Bernholz. UChicago Press.

Crisp, Roger, and Theron Pummer. 2020. “Effective Justice.” *Journal of Moral Philosophy* 17 (4): 398–415. https://doi.org/10.1163/17455243-20193133.

Frick, Johann. 2020. “Conditional Reasons and the Procreation

Asymmetry." *Philosophical Perspectives* 34 (1): 53–87. https://doi.org/10.1111/phpe.12139.

Gabriel, Iason. 2017. "Effective Altruism and Its Critics." *Journal of Applied Philosophy* 34 (4): 457–73. https://doi.org/10.1111/japp.12176.

Lazari-Radek, Katarzyna de, and Peter Singer. 2010. "Secrecy in Consequentialism: A Defence of Esoteric Morality." *Ratio* 23 (1): 34–58. https://doi.org/10.1111/j.1467-9329.2009.00449.x.

MacAskill, William. 2015. *Doing Good Better*. Penguin Random House.

———. 2022. *What We Owe the Future*. Basic books.

McMahan, Jeff. 2013. "Causing People to Exist and Saving People's Lives." *The Journal of Ethics* 17 (1-2): 5–35. https://doi.org/10.1007/s10892-012-9139-1.

———. 2016. "Philosophical Critiques of Effective Altruism." *The Philosophers' Magazine* 73: 92–99.

Minson, Julia A., and Benoît Monin. 2012. "Do-Gooder Derogation: Disparaging Morally Motivated Minorities to Defuse Anticipated Reproach." *Social Psychological and Personality Science* 3 (2): 200–207. https://doi.org/10.1177/19485506114156

95.

Norcross, Alastair. 2020. *Morality by Degrees: Reasons Without Demands*. Oxford University Press.

Ord, Toby. 2013. "The Moral Imperative Toward Cost-Effectiveness in Global Health." *Center for Global Development* 12. http://www.cgdev.org/content/publications/detail/1427016.

Parfit, Derek. 1987. *Reasons and Persons*. Oxford University Press.

Pummer, Theron. 2016. "Whether and Where to Give." *Philosophy and Public Affairs* 44 (1): 77–95. https://doi.org/10.1111/papa.12065.

———. 2023. *The Rules of Rescue: Cost, Distance, and Effective Altruism*. Oxford University Press.

Sanbonmatsu, John. 2023. "Effective Altruism and the Reified Mind." In *The Good It Promises, the Harm It Does: Critical Essays on Effective Altruism*, edited by Carol J. Adams, Alice Crary, and Lori Gruen. Oxford University Press.

Sanders, Brenda. 2023. "How Effective Altruism Fails Community-Based Activism." In *The Good It Promises, the Harm It Does: Critical Essays on Effective Altruism*, edited by Carol J. Adams, Alice Crary, and Lori Gruen. Oxford University Press.

Shulman, Carl, and Elliott Thornley. forthcoming. "How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role." In *Essays on Longtermism*, edited by Jacob Barrett, Hilary Greaves, and David Thorstad. Oxford University Press.

Srinivasan, Amia. 2015. "Stop the Robot Apocalypse." *London Review of Books* 37: 3–6. http://www.lrb.co.uk/v37/n18/amia-srinivasan/stop-the-robot-apocalypse.

Walden, Kenneth. 2014. "The Aid That Leaves Something to Chance." *Ethics* 124 (2): 231–41. https://doi.org/10.1086/673438.