Effective Altruism

International Encyclopedia of Ethics (forthcoming 2019)

Theron Pummer and William MacAskill

1. Introduction

Climate change is on course to cause millions of deaths and cost the world economy trillions of dollars. Nearly a billion people live in extreme poverty, millions of them dying each year of easily preventable diseases. Just a small fraction of the thousands of nuclear weapons on hair-trigger alert could easily bring about global catastrophe. New technologies like synthetic biology and artificial intelligence bring unprecedented risks. Meanwhile, year after year billions and billions of factory-farmed animals live and die in misery. Given the number of severe problems facing the world today, and resources required to solve them, we may feel at a loss as to where to even begin. The good news is that we can improve things with the right use of money, time, talent, and effort. These resources can bring about a great deal of improvement, or very little, depending on how they are allocated.

The effective altruism movement consists of a growing global community of people who use reason and evidence to assess how to do as much good as possible, and take action on this basis. Launched in 2011, the movement now has thousands of members, and influence over billions of dollars. The movement has substantially increased awareness of the fact that some altruistic activities are much more *cost-effective* than others, in the sense that they do much more good than others per unit of resource expended. According to the non-profit organization GiveWell, it costs around $3,500 to prevent someone dying of malaria by distributing bednets. By contrast, the U.S. government typically spends around $9.5 million to save a life (Merrill 2017).

Giving to charity is one sort of altruistic activity (*see* CHARITY; PHILANTHROPY). But there are plenty others. Many effective altruists aim to do as much good as possible in their career choices, and regard how much a given career would enable them to give to charity as only one of many relevant factors. In any case, assessing the cost-effectiveness of altruistic activities is no easy matter. Sometimes it requires expending time and effort in processing large and complex bodies of evidence. In other cases, the issue is instead determining how confident to be about the possible outcomes of activities in the absence of much if any evidence at all. Things are more complicated still insofar as what matters is the *difference* an activity makes, what it brings about that wouldn't have happened otherwise. Largely for these reasons, cost-effectiveness assessments are often centralized within the effective altruist community. For example, GiveWell is devoted to assessing charities that help those in extreme poverty in terms of the additional benefits delivered for each additional dollar donated. Animal Charity Evaluators has a similar mission with respect to charities focusing on animal welfare. Many effective altruists base their charitable donations very closely on the recommendations of such "meta-charities". In addition, the organization 80,000 Hours specializes in providing advice on choosing careers that enable one to do as much good as possible in one's working life.

In this entry, we discuss both the definition of effective altruism (§2) and objections to effective altruism, so defined (§3).

2. The Definition of Effective Altruism

Many take effective altruism to be synonymous with utilitarianism, the normative theory according to which an act is right if and only if it produces no less well-being than any available act (*see* UTILITARIANISM; WELL-BEING). This is a category mistake. Effective altruism is not utilitarianism, nor is it any other normative theory or claim. Instead, effective altruism is the *project* of using evidence and reason to try to find out how to do the most good, and on this basis trying to do the most good (MacAskill forthcoming). We here offer a series of considerations that support and clarify this definition of effective altruism. They are largely driven by how the term is used by leaders and members of the effective altruism community, as well as by the views and activities of those within the community. We are mindful of the possibility that effective altruism will evolve over time. The definition offered here should therefore be regarded as provisional.

The tentative aim of effective altruism is doing good in the sense of promoting well-being, with everyone's well-being counting equally. Effective altruism, then, is tentatively *welfarist* in that its tentative aim in doing good concerns promoting well-being only, and not, say, protecting biodiversity or conserving natural beauty for their own sakes (*see* WELFARISM). Since effective altruism is a project, rather than a normative claim, it is possible for one to both adopt this project as well as accept a non-welfarist conception of the good (or indeed to adopt multiple projects, some of which involve promoting welfarist good and some of which involve promoting non-welfarist good).

Effective altruism is also *impartial* in that its tentative aim is to promote well-being without regard to who in particular has a given amount of well-being, their race, gender, nationality, and so forth (*see* IMPARTIALITY). The outcome in which Anya has 1 unit of well-being, Basilio has 10 units, and Chloe has 100 units, is just as good, from an impartial point of view, as the outcome in which Chloe has 1 unit of well-being, Anya has 10 units, and Basilio has 100 units; moreover, we don't need to know anything further about Anya, Chloe, and Basilio to know that that's the case. (Precise well-being units are used here for illustrative purposes only; quantities of well-being might be imprecise.)

Impartiality about who in particular has a given amount of well-being is compatible with prioritizing benefits to the worse off (*see* PRIORITARIANISM). Suppose Anya has 1 unit of well-being and Basilio has 10 units. We can claim that getting an extra unit of well-being to Anya has greater priority, or does more to make the outcome better, than getting an extra unit to Basilio, because getting an extra unit to *anyone* who has 1 unit has greater priority than getting an extra unit to *anyone else* who has 10 units.

There are many different altruistic activities we can engage in, either directly or through charitable donations. They include, for example: feeding the hungry; protecting endangered species; mitigating climate change; reforming immigration policy; researching cures for illnesses; preventing sexual violence; aiding those in extreme poverty; eliminating factory farming; averting nuclear war. Effective altruism's impartiality drives what is sometimes referred to as its "cause neutrality", that is, choosing among the many various possible activities, causes, or problems on the basis of which enable us to do the most good with limited resources, rather than on the basis of (say) personal connections. Of course, because skills and experiences vary substantially across individuals, it is often the case that different people can do substantially more good by working on different problems. According an instrumental role for personal fit is consistent with acting impartially.

Effective altruism is also about prioritizing benefiting *more* people rather than fewer, at least when all else is equal. Suppose that, with a fixed amount of resources, we could either restore eyesight to two people or instead restore eyesight to twenty others. It is relatively uncontroversial that the latter would do more good, other things being equal. Next suppose we could either restore eyesight to a hundred people or instead save two others from dying of malaria. Even in the presence of full empirical information, it may not be obvious which would do more good. There is significant philosophical disagreement about whether and how benefits of various sizes for separate individuals evaluatively combine (*see* AGGREGATION).

On a simple additive view of aggregation, the good done by conferring a collection of benefits is given by the sum of the sizes of each of the benefits conferred. For example, if preventing a mild headache corresponds to a gain of 0.01 units of well-being, and preventing a death corresponds to a gain of 1,000 units of well-being, then on this simple additive view more good is done by saving 200,001 people from a mild headache than saving two from death. This view is highly controversial. Many doubt there is any number of people it would be better to save from a mild headache than save two from death. Still, most of us accept at least some more limited form of aggregation. Suppose we could either save two people from death or save two million others from losing their legs. Most accept that it would be better to save the many, if they were sufficiently numerous (see the survey data in Voorhoeve 2014: footnote 3).

The effective altruist project aligns with common sense in its incorporation of at least some limited form of aggregation (according to which it matters both how many individuals benefit and how much they benefit each). Which particular account of aggregation is correct, and more broadly which particular impartial welfarist conception of the good is correct, are notoriously difficult, yet extremely important, questions. Does impartiality extend, for example, to possible future persons (Parfit 1984: part four) or to sentient nonhuman animals (Singer 2009)? As these questions make clear, there are *philosophical* as well as empirical challenges to working out how to do the most good, impartially construed, per dollar (or other unit of resource) expended.

Given these various challenges, a strict view according to which only those who are *in fact* successful in their aims to promote the good qualify as engaging in effective altruism seems under-inclusive. Instead it seems we can engage in effective altruism even if we are not in fact doing as much good as possible (as specified by whichever impartial welfarist conception of the good is correct). If we are genuinely *trying* to do this, that is enough for us to be engaged in the project, though of course trying is not itself the project's aim. Its tentative aim is promoting as much well-being impartially construed as possible.

The effective altruist project has two aspects, one intellectual and the other practical. The intellectual aspect is trying to find out how to do the most good through the use of reason and evidence (involving empirical sciences as well as epistemology, ethics, and decision theory). The practical aspect is trying to do the most good, which consists in altruistic activity based on or informed by the intellectual aspect. Must one engage in both of these aspects of the project, at least to some extent, to engage in the project at all? One might engage in effective altruism by focusing almost exclusively on the intellectual aspect. Of course, some may find it odd that trying to find out how to do the most good without *any* intention of doing good (e.g., by sharing one's findings with those who will do good by using them) could count as engaging in effective altruism.

At the other extreme, suppose one bases one's altruistic activities on the advice of others completely blindly, and just happens to be promoting a lot of good by dint of following the

recommendations of effective altruists. Some may hold that, to be engaging in the project of effective altruism, one must use reason or evidence to some extent, for example in determining whom to ask for advice, so that one is justified to at least some extent in one's choice of altruistic activities. We need not here take a stand on such questions. But surely engaging in the intellectual aspect of effective altruism does not require one to conceive of one's altruistic activities as obligatory, or as doing good in any metaethically heavyweight sense of "good".

If *effective altruism* is the project of using evidence and reason to try to find out how to do the most good, and on this basis trying to do the most good, then what is an *effective altruist*? We do not believe that, to be an effective altruist, one must always be engaged in or motivated by effective altruism. Most if not all effective altruists have a variety of projects in their lives besides effective altruism, even if effective altruism is one of their most central projects. So it seems one need not be a perfect effective altruist to be an effective altruist. Similarly, occasionally running slower than you are able to does not disqualify you from being a fast runner. Being a fast runner is a matter of running fast enough, and being an effective altruist is a matter of engaging in effective altruism enough. Perhaps *significant* engagement is enough. If "enough" or "significant" are hopelessly arbitrary or vague, we can instead construe being an effective altruist as purely a matter of degree, corresponding to the degree to which one engages in effective altruism.

Finally, neither engaging in effective altruism nor being an effective altruist necessarily requires participating in the effective altruism community (where the latter characteristically involves collaborating with other effective altruists at meetings or events, publicly committing to doing the most good possible with a significant portion of the resources in one's possession, encouraging others to do likewise, etc.). There were many people around trying to find out how to do the most good, and trying to do it, millennia before the arrival of the effective altruism movement (*see* MOZI). And there are undoubtedly many people around today who qualify as effective altruists, yet are either unaware of the social movement or refrain from participating in it for various reasons. Indeed, most effective altruists would leave the movement if they came to believe that participating in it would stand in the way of their aim of doing the most good, and would proceed to pursue this aim in some perhaps less overt form.

3. Objections

While effective altruism is not itself a normative claim, it can be supported or opposed by various normative claims, no less than other projects and activities can be. We here discuss some possible objections to effective altruism. These are objections to effective altruism as defined above, rather than to the effective altruism movement in its present form.

First, it might be objected that it is morally permissible not to always try to promote as much good as possible (*see* AGENT-CENTERED OPTIONS). Whether or not this claim is true, it is no objection to effective altruism. That it is permissible not to always engage in some project does not itself count against doing so.

Second, one might claim that we are morally obligated not to always try to promote as much good as possible (*see* AGENT-CENTERED RESTRICTIONS). For example, we have obligations not to harm others, steal, lie, and so forth, and we have obligations to put our friends and family ahead of strangers. These obligations can often outweigh reasons or obligations to promote the good, and thus constitute *constraints* on promoting the good. Always trying to

promote the most good may involve harming others, stealing, lying, and neglecting friends and family. But an objection to always trying to promote the most good is no objection to trying to promote the most good in some but not all circumstances, or aspects of life. This is not a wholly satisfactory response, as it leaves it open that being a perfect effective altruist, or always engaging in effective altruism, would involve violating constraints. We therefore endorse a *qualified definition of effective altruism*, according to which effective altruism is the project of using evidence and reason to try to find out how to do the most good, and on this basis trying to do the most good, *without violating constraints*.

Third, one might object to effective altruism insofar as it is a wholly individualistic project, where the guiding question for each engaged in the project is "How can I do the most good, given what others are likely to do?" There seem to be many cases in which, even if we each do the most good we can individually, we do less good than we collectively could have done. Well-known coordination problems can give rise to such cases (Gibbard 1965). There may also be cases in which each individual member of some group cannot make any difference to the good, yet together the group can make a substantial difference (Nefsky 2011). Further, there are general and familiar Kantian worries of the form, "What if everyone did that?". For these various reasons, it may be undesirable to construe effective altruism as a solely individualistic project. We have deliberately left our definition of effective altruism underspecified with respect to individualism versus collectivism, leaving it open that the project be construed at least partly if not wholly collectively, so that its guiding question is "How can *we together* do the most good?". There is undoubtedly more work to be done in this area, both in theory and in practice (Dietz 2019; Collins forthcoming).

Fourth, one might object to effective altruism on epistemic grounds, claiming that we are utterly clueless about what activities will, in the indefinitely long run, do the most good (Lenman 2000). For example, we might be reasonably confident that deworming children will increase school attendance. How reasonably confident can we be about the further effects of increased school attendance (on who is born, or the state of the local economy, in the following year)? What about the effects of these further effects, and the effects of these further further effects, and so on? At some point effects become unforeseeable, in the sense that we have no reason whatsoever to regard them as more likely to result from some of our acts than from others. It is possible that, if you deworm a child, this will through some very long sequence of events result in the rise of a genocidal dictator in the next century. But we have no reason whatsoever to regard this remote effect as more likely to result from your deworming the child than from not deworming the child, and going to the movies instead. It may be true that, since the vast majority of the effects of our activities are unforeseeable, we are, overall, clueless about the effects of our activities. Even if this is true, and humbling, it is practically irrelevant. Our choice of activities can be guided by foreseeable effects, when we have some reason to regard them as more likely to result from some activities than others, but not by unforeseeable effects. The dark sea of unforeseeable effects is vast but also silent (Greaves 2016).

Fifth, one may object that, for any activity that has a very high probability of doing a lot of good, say, saving thousands of lives, there is some other activity with a very low probability of doing some sufficiently greater amount of good that effective altruism must rank higher. Consider *existential risks*, that is, risks of events that would either annihilate intelligent life on Earth or permanently and drastically curtail its potential (Bostrom 2003). Even setting aside the possibility that our distant descendants will settle the stars, Earth itself could host around 10^{16} (ten million billion) people over the next billion years. If their lives would go well were all these

people to exist, then ensuring their existence may do an astronomical amount of good. One might thus argue that engaging in an activity that even slightly reduces some particular existential risk does more good *in expectation* than engaging in an activity that will nearly certainly save a thousand lives today (*see* RISK). Our response to this objection is that questions of how to calculate expected goodness, aggregate well-being across separate individuals, and weight the well-being of possible future individuals, are all questions about which there is reasonable disagreement, both within philosophy and the effective altruist community. Effective altruism per se does not presuppose answers to these questions. Rather, working out the most plausible answers to these questions is itself part of the project of effective altruism (Beckstead 2013; Greaves and Ord 2017; MacAskill and Ord 2018).

Sixth, there are several axiological objections one might raise. It may be that there are no precisely measurable quantities of well-being. We here use a precise cardinal scale of well-being for the sake of illustration, but effective altruism does not essentially require this. If an ordinal ranking of activities in terms of well-being promoted is the most informative ranking obtainable, effective altruism would be about trying to find out and do those activities that are ranked highest. Next, consider activities that do quite different things: one prevents existing people dying of malaria, another prevents the existence of factory farmed animals, and a third ensures the existence of people in the far future. There may be many cases in which neither of two activities ranks higher than the other, in terms of well-being promoted, and yet they are not ranked equally either. We here somewhat hamfistedly refer to these as cases of "incommensurability" (*see* INCOMMENSURABILITY AND INCOMPARABILITY). Perhaps while there is commensurability between activities within categories like extreme poverty, animal suffering, and the far future, there is a degree of incommensurability across these categories, where the best in each category are incommensurable with each other even if the worst in each category is worse than the best in each of the others. Effective altruism would then be about trying to find out and do those activities that do no less good than those with which they are commensurable. The project would however have relatively little scope if there were rampant incommensurability, e.g., between any two activities that benefit separate individuals. But such rampant incommensurability seems independently implausible. Moreover, there may be ways of ranking incommensurable activities by appealing to tools for decision making under normative uncertainty (MacAskill 2016). A final axiological objection comes from those skeptical of the very notion of "the good, simpliciter" (Foot 1983; Thomson 2008). But effective altruism is (again, at least tentatively) about doing good in the sense of promoting well-being, or benefiting others. Any minimally plausible normative theory should be able to make sense of benefiting others, and have some nontrivial place for beneficence.

Continuing from this last point, we will close by very briefly noting a positive case for effective altruism. We find it hard to deny the *minimal view* that we have reason to benefit all others, regardless of their race, gender, nationality, and so forth, and more reason to benefit them more, and most reason to benefit them as much as possible, at least defeasibly and if other things are equal. This minimal view is "portable" in the sense that it can be found in a wide variety of normative theories (*see* CONSEQUENTIALISM; DEONTOLOGY; ROSS, W.D.; CONTRACTUALISM; VIRTUE ETHICS). While we believe the reasons of beneficence implied by the minimal view are of nontrivial weight, we need not here insist on how much weight they have relative to other reasons (for example, to benefit oneself or one's near and dear). There are further questions of whether and when these reasons of beneficence give rise to

obligations, and how such obligations would interact with moral options to sometimes do less than what is impartially best (Pummer 2016; Sinclair 2018). For now, we note only that the minimal view provides a relatively uncontroversial and ecumenical basis for effective altruism.

SEE ALSO: AGENT-CENTERED OPTIONS; AGENT-CENTERED RESTRICTIONS; AGGREGATION; CHARITY; CONSEQUENTIALISM; CONTRACTUALISM; DEONTOLOGY; IMPARTIALITY; INCOMMENSURABILITY AND INCOMPARABILITY; MOZI; PHILANTHROPY; PRIORITARIANISM; RISK; ROSS, W.D.; UTILITARIANISM; VIRTUE ETHICS; WELFARISM; WELL-BEING.

References:

80,000 Hours: https://80000hours.org/

Animal Charity Evaluators: https://animalcharityevaluators.org/

Beckstead, Nick 2013. *On the Overwhelming Importance of Shaping the Far Future*. PhD Thesis. Department of Philosophy, Rutgers University.

Bostrom, Nick 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas* vol. 15, pp. 308-14.

Collins, Stephanie forthcoming. "Beyond Individualism," in Hilary Greaves and Theron Pummer (eds.), *Effective Altruism: Philosophical Issues*. Oxford: Oxford University Press.

Dietz, Alexander 2019. "Effective Altruism and Collective Obligations," *Utilitas* vol. 31, pp. 106-15.

Foot, Philippa 1983. "Utilitarianism and the Virtues," *Proceedings and Addresses of the American Philosophical Association* vol. 57, pp. 273-83.

Gibbard, Allan 1965. "Rule-Utilitarianism: Merely an Illusory Alternative?" *Australasian Journal of Philosophy* vol. 43, pp. 211-20.

GiveWell: https://www.givewell.org/

Greaves, Hilary 2016. "Cluelessness," *Proceedings of the Aristotelian Society* vol. 116, pp. 311-39.

Greaves, Hilary and Ord, Toby 2017. "Moral Uncertainty about Population Axiology," *Journal of Ethics and Social Philosophy* vol. 12, pp. 135-67.

Lenman, James 2000. "Consequentialism and Cluelessness," *Philosophy and Public Affairs* vol. 29, pp. 342-70.

MacAskill, William 2016. "Normative Uncertainty as a Voting Problem," *Mind* vol. 125, pp. 967-1004.

MacAskill, William and Ord, Toby 2018. "Why Maximize Expected Choice-Worthiness?" *Noûs* doi: 10.1111/nous.12264, pp. 1-27.

MacAskill, William forthcoming. "The Definition of Effective Altruism," in Hilary Greaves and Theron Pummer (eds.), *Effective Altruism: Philosophical Issues*. Oxford: Oxford University Press.

Merrill, Dave 2017. "No One Values Your Life More Than the Federal Government," Bloomberg: https://www.bloomberg.com/graphics/2017-value-of-life/.

Nefsky, Julia 2011. "Consequentialism and the Problem of Collective Harm: A Reply to Kagan," *Philosophy and Public Affairs* vol. 39, pp. 364-95.

Parfit, Derek 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pummer, Theron 2016. "Whether and Where to Give," *Philosophy and Public Affairs* vol. 44, pp. 77-95.

Sinclair, Thomas 2018. "Are We Conditionally Obligated to Be Effective Altruists?" *Philosophy and Public Affairs* vol. 46, pp. 36-59.

Singer, Peter 2009. *Animal Liberation*, updated edition. New York: Harper.

Thomson, Judith 2008. *Normativity*. Peru, Illinois: Open Court.

Voorhoeve, Alex 2014. "How Should We Aggregate Competing Claims?" *Ethics* vol. 125, pp. 64-87.

Suggested Readings:

Arrhenius, Gustaf, Bykvist, Krister, Campbell, Tim, and Finneron-Burns, Elizabeth (eds.) forthcoming. *The Oxford Handbook of Population Ethics*. Oxford: Oxford University Press.

Barry, Christian and Øverland, Gerhard 2016. *Responding to Global Poverty: Harm, Responsibility and Agency*. Cambridge: Cambridge University Press.

Buchak, Lara 2013. *Risk and Rationality*. Oxford: Oxford University Press.

Greaves, Hilary and Pummer, Theron (eds.) forthcoming. *Effective Altruism: Philosophical Issues*. Oxford: Oxford University Press.

Hare, Caspar 2013. *The Limits of Kindness*. Oxford: Oxford University Press.

Illingworth, Patricia, Pogge, Thomas, and Wenar, Leif (eds.) 2011. *Giving Well: The Ethics of Philanthropy*. New York: Oxford University Press.

Kamm, Frances 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.

MacAskill, William 2015. *Doing Good Better: Effective Altruism and a Radical New Way to Make a Difference*. London: Faber & Faber.

MacAskill, William, Bykvist, Krister, and Ord, Toby forthcoming. *Moral Uncertainty*. Oxford: Oxford University Press.

Pummer, Theron forthcoming. *Effective Altruism for Everyone*. New York: Oxford University Press.

Singer, Peter 2015. *The Most Good You Can Do: How Effective Altruism Is Changing Ideas about Living Ethically*. London: Yale University Press.

Woodruff, Paul (ed.) 2018. *The Ethics of Giving: Philosophers' Perspectives on Philanthropy*. New York: Oxford University Press.