A research agenda for the

GLOBAL PRIORITIES INSTITUTE

Hilary Greaves and Will MacAskill

December 2017

GLOBAL PRIORITIES INSTITUTE

Abstract

This document lays out a research agenda for the Global Priorities Institute. The primary intended audience is academics who wish to work at, or in collaboration with, the Institute. In general, we have tried to list topics that we believe are particularly important, tractable and/or neglected, rather than all topics of interest to the Institute.

The central focus of GPI is on theoretical issues that arise for actors who wish to use some of their scarce resources to do as much good as possible. Within that, we distinguish between three broad research areas. The first is *cause prioritisation*. Cause prioritisation is the research field that addresses the question of what problem or problems altruistic agents ought to focus their scarce resources on. The most common answers to this question within the effective altruism community are global health and development, farm animal welfare, and reduction of existential risks (with a particular focus on AI). The second is *cross-cutting considerations*. These include theoretical issues that arise for altruistic agents engaged in what we might call 'means prioritisation' (i.e. choosing among available means for tackling a specified cause), and also theoretical issues that arise whether prioritisation is between rival causes or between rival means (or both). The third is the scope of *effective altruism*. This is the research field that asks to what extent effective altruism should be part of an individual's life, or be part of the decisions of large organisations such as corporations and governments.

Table of Contents

Executive Summary **4**

Introduction **8**

GPI's mission 8

GPI's research strategy 8

Cause Prioritisation **10**

The long-termism paradigm 10

The value of the future 12

Indirect effects 14

Animal welfare 15

Cross-cause comparisons 17

Cross-cutting considerations **20**

Altruistic economics 20

Altruistic coordination theory 21

Altruistic decision theory 22

Doing good now vs doing good later 23

Mission hedging 24

Epistemological issues 25

Diversification 28

Distribution of cost-effectiveness across interventions 29

Moral Uncertainty 30

Indirect justifications of decision norms 31

The Scope of Effective Altruism 33

The obligation to engage in effective altruism 33

Cluelessness 34

Effective Altruism as a political philosophy 34

Existing academic research that has broad relevance for GPI **36**

Further EA Reading **39**

Executive Summary

There are many problems in the world. Because resources are scarce, we cannot solve them all. We therefore need to prioritise among those problems if we are to have the largest possible impact.

The effective altruism community currently prioritises global health and development, farm animal welfare, and reduction of existential risks (with a particular focus on AI). Are these really the correct problems to focus on? What might we be missing? And what framework ought we to use to determine which problems are highest-priority?

We think that the following areas are of particular interest:

The *long-termism paradigm*. Let us define *long-termism* as the view that the primary determinant of the value of the actions we take today is the effect of those actions on the very long-term future. This is a very popular view within the EA community, and has enormous importance if correct. This warrants much more work to make this argument rigorous, explore to what extent this view is justified on moral views other than total utilitarianism, and work out what exactly its implications are.

The *value of the future*. As above, long-termism is often thought to lead to the conclusion that we ought to prioritise extinction risk reduction. This presupposes that the future will be good. But one can at least imagine some scenarios in which we should expect the future to contain more bad than good. How should we assess that possibility? What is the likelihood of good outcomes versus bad outcomes? And how should we weight very bad futures compared to very good futures - should we treat the best possible future as equally as good as the worst possible future, or should we give more weight to bad scenarios?

Indirect effects. Effective altruists generally assume that, in evaluating interventions, we should in principle take into account *all* welfare-relevant effects of those interventions. That is, we should include not just direct effects, like the impact on school attendance from deworming school children, but also indirect effects, like the impacts on population size, economic growth, and government activity. This seems highly plausible, although it also seems somewhat in tension with popular claims in health ethics/economics. If we should take indirect effects into account, then, ultimately, we need to assess the impact of our actions from now until the end of time. How should we do this?

Animal welfare. One distinctive aspect of the EA community is its focus on improving the welfare of non-human animals. This focus raises a number of interesting and unresolved theoretical questions, including about the ways in which we can improve the lives of non-human animals and how we ought to prioritise between interventions that improve human lives and interventions that improve non-human animal lives.

Cross-cause comparisons. Comparative cost-effectiveness analysis is relatively straightforward when the interventions being compared are sufficiently similar, for example two different ways of relieving blindness, or two different ways of increasing the number of years children spend in school. But this is only a very narrow class of prioritisation decisions. We also need to be able to compare very different interventions in terms of 'amount of good done per dollar spent', for example programs focused on preventing blindness versus programs increasing child test scores. This requires constructing an adequate common measure of 'good done', to create a common numerator, in a principled way.

Even when a group of altruistic individuals have chosen a problem, or number of problems, to focus on, there are still many open theoretical questions that they face if they wish to do the most good. Some of these concern what individuals ought to do — under what conditions they should try to do good right away, and under what conditions should they invest their resources in order to do more good later. Some of these questions concern how groups on altruistic individuals can ensure that collectively they have the largest possible impact.

Within this, we think that the following areas are particularly interesting and important:

Altruistic economics. Economic theory normally proceeds either (a) assuming nothing substantive about individuals' preferences (assuming only structural conditions, e.g. that preferences are complete and transitive), or (b) assuming that preferences are in some sense 'self-interested' (e.g. that an individual's utility depends only on her own consumption/leisure/etc.). It may be that interesting new phenomena arise, and interesting new results can be established, when we assume instead that individuals are perfectly altruistic (for example, that each individual's utility function is simply the utilitarian one).

Altruistic coordination theory. Given multiple actors deciding how to distribute resources (for example money, but also perhaps labour) for altruistic purposes, how will they, or should they, act? The puzzle is cleanest in the case where they have slightly different values leading them to value different opportunities differently – for example if two donors agree on the first-best use of money but disagree on the second-best, they each prefer that the other fully funds the first-best use. Variations of it deal with cases with multiple donors, or where there are also empirical disagreements, or private information, or comparative advantage of different actors contributing to different projects.

Altruistic decision theory. Various apparently altruistic and apparently very reasonable behaviours seem puzzling on closer inspection assuming that the agent is attempting to maximise expected good. These puzzling behaviours include many that card-carrying EAs find themselves drawn to, such as donating to more than one charity, and avoiding supporting work on x-risk mitigation on grounds of 'risk aversion'. The same behaviours might make a lot more sense assuming a less pure form of altruism (the most obvious alternative being: a 'warm glow' theory of motivation), or assuming deviations from expected utility theory that are arguably irrational (such as ambiguity aversion).

Doing good now vs doing good later. If an altruist wants to do good, she faces the question of when to do good. With her money, she could donate right away, or she could invest the money at a later date, or she could take out a loan in order to give more now. With her time, she could try to get a high-impact job right away, or she could spend time getting further education or job training, in order to have a larger impact later on. Under what conditions should direct intervention be attempted earlier vs. later?

Mission hedging. For some strategies, there is covariance between the amount of resources you control and the cost-effectiveness of the opportunities you have available to you. Examples include earning to give by by founding an AI company if you are aiming to donate to AI safety and investing in oil companies if you are aiming to donate the returns of your investment to climate change mitigation efforts.

Epistemological issues. Figuring out how to do the most good is very difficult, and often it seems that subtle differences in epistemology would lead one to quite different conclusions. These include differences in responses to paucity of hard evidence, in level of trust in abstract arguments leading to counterintuitive conclusions, and in the relative weight placed on different types of evidence.

Diversification. What rationales are there, either for the individual or for the EA community/world as a whole, to diversify across causes/interventions, rather than simply identifying the intervention with the highest expected cost-effectiveness and supporting exclusively that intervention?

Distribution of cost-effectiveness across interventions.
It is a platitude within the EA community that the cost-effectiveness of interventions within a single cause area typically varies by many orders of magnitude. How strong is the evidence for this claim, and what can we establish about the shape of distributions of cost-effectiveness more generally, both within and between causes?

Moral uncertainty. When effective altruists attempt to compare the importance of different problems, or the effectiveness of different interventions, they typically default to using a utilitarian axiology. But, even if you are sympathetic to a utilitarian axiology, it would clearly be overconfident to be certain in that axiology. So, plausibly, we should try to incorporate moral uncertainty into our reasoning when we prioritise among problems. It remains underexplored, however, what implications the fact of moral uncertainty has. How do practical conclusions change when we incorporate reasonable moral uncertainty into our analysis?

Indirect justifications of decision norms. Fundamentally, we assume, prioritisation of all forms (whether among problems or among interventions) should be via expected value theory: we simply seek those interventions that have highest expected value. But this injunction is extremely abstract and general. For practical purposes, it is useful also to have some more easily applicable principles for identifying interventions that are likely to have high expected value. The typical way of prioritising among problems in the effective altruism community is to assess them in terms of their importance (how many individuals does this problem affect, and by how much), their tractability (how

much progress can we make on this problem with a given unit of resources), and their neglectedness (how many resources are put towards this addressing this problem already). What is the status of this framework? Is it the right one to use?

The primary focus of our research agenda is on the question of how one can use a given unit of resources to do as much good as possible. However, there are also important questions about the nature and strength of the motivation for and/or the moral imperative towards caring about that question. Is everyone required to dedicate their lives to effective altruism? If not, to what extent do we have obligations to engage in effective altruism? And to what extent ought considerations of what will do the most good influence the decisions of governments, as opposed to private individuals?

The obligation to engage in effective altruism. This topic concerns whether effective altruism is simply a beneficial project that one might or might not choose to engage in, or whether stronger things can be said in its favour from the point of view of moral philosophy. Questions of this type form the main focus of most of the existing commentary on effective altruism among moral philosophers.

Cluelessness. Many people who would otherwise be inclined towards EA-like behaviour refrain from such behaviours because of epistemic concerns: they feel that they are simply too clueless about which well-meaning actions would in the end do net good versus harm, and how much, for it to 'make sense' to expend significant resources on altruistically-meant interventions. (The source of this worry is closely related to issues of 'indirect' and/or long-run effects: perhaps a randomised controlled trial can give us a pretty good idea of a particular immediate consequence of a given intervention, but what about e.g. the knock-on implications for political structures, individuals not treated, long-run trajectories of economic development, population size and environmental degradation, and any effects on x-risk?)

Effective altruism as a political philosophy. Effective altruism typically concerns itself with the decisions of individuals. But there's no principled reason why this should be so. We could ask how corporations could use their resources to do the most good. And, more interestingly, we could ask how governments could use their resources to do the most good.

Introduction

GPI's mission

Despite the intellectual roots of the effective altruism (EA) movement in philosophy and in economics and the massive growth of the movement outside academia, effective altruism is currently recognised in academia only to a very limited extent. With the partial exception of moral philosophy, most academics have neither heard of EA nor come across many of its core ideas, and of those that have, a significant proportion are hostile. Just about all students who come across EA do so via the work of EA non-profit organisations, rather than through the academic literature. Of the little academic literature that does exist focusing specifically on EA, a significant proportion is of rather poor quality, and most of it consists of 'external critique' (that is, articles of the form "what is wrong with effective altruism"), rather than 'internal' engagement with and development of the issues that become urgent once one takes the EA project seriously. All of this contrasts radically with, for example, climate change, where there is a large community of experts and a vast body of extremely high-quality in-depth work, and anyone unfamiliar with at least the basics of the climate change debate is regarded as something of an ignoramus.

At the same time, we don't think that this situation is inevitable. The core ideas of EA are powerful and defensible, academics are fundamentally responsive to arguments and, within philosophy at least, we have already seen significant interest in EA ideas from the very best academics. We therefore think that it is feasible to significantly alter the current state of affairs for the better.

The core idea behind GPI is that, given that effective altruism aims to achieve the status of the most influential new philosophy of the 21st Century, doing for the pursuit of good what the Enlightenment did for the pursuit of truth, it is important for these ideas to get serious academic attention. This motivates the following

> **Mission**: To develop the intellectual roots of the effective altruism movement with the level of rigour and detail that is characteristic of academia, to gain widespread acceptance of the core tenets of EA throughout academia, and to harness the brainpower of academia to tackle research questions that are important by EA lights.

GPI's research strategy

The first step towards this mission is setting out a research agenda, which is what this document provides. This agenda is not a complete list of possible topics, but should give the reader a sense of our focus areas. In general we have tried to list topics that we believe are particularly important, tractable and/or neglected. However, we acknowledge that research output and quality is highly dependent on the expertise and level of personal interest of the researcher. If someone could do interesting research on topics not listed here that are of clear relevance to GPI's mission, we would warmly welcome that.

The central focus of GPI is what we call 'global priorities research': theoretical issues that arise in response to the question, "With a given unit of resources, what should we do with those resources if our aim is to do the most good?"

Within that overarching research question, we distinguish between three broad research areas. The first is *cause prioritisation*. Cause prioritisation is the research field that addresses the question of what problem or problems altruistic agents ought to focus their scarce resources on. The most common answers to this question within the effective altruism community are global health and development, farm animal welfare, and reduction of existential risks (with a particular focus on AI). The second is *cross-cutting considerations*. These include theoretical issues that arise for altruistic agents engaged in what we might call 'means prioritisation' (i.e. choosing among available means for tackling a specified cause), and also theoretical issues that arise whether prioritisation is between rival causes or between rival means (or both). The third is the *scope of effective altruism*. This is the research field that asks to what extent effective altruism should be part of an individual's life, or be part of the decisions of large organisations such as corporations and governments.

In general, the questions on this research agenda are ones that are currently relatively neglected within academia. However, there are also some areas that are highly pertinent to our concerns and for which there is already a well-developed academic literature. Because of considerations of diminishing marginal value, together with the fact that a key part of our strategy is to *redirect* the attention of other academics where necessary, we don't plan to directly contribute to these existing literatures as part of core GPI research. However, we will need to ensure that we are familiar with them, so that we can (i) benefit from their insights, (ii) avoid reinventing the wheel, and (iii) engage appropriately with other academics who might have shared interests and/or be in a position to contribute to our research and strategy. We list topics that have this status under 'existing research to engage with.'

A particularly promising vein of research topics are those that engage with beliefs that are (i) widely held within the effective altruism community but (ii) have never been defended at length or with high levels of rigour. One could engage with these beliefs either by making the belief precise and the defence of them rigorous; or by criticising them. Either type of project would be appealing to us.

Cause Prioritisation

There are many problems in the world. Because resources are scarce, we cannot solve them all. We therefore need to prioritise among those problems if we are to have the largest possible impact.

The effective altruism community currently prioritises global health and development, farm animal welfare, and reduction of existential risks (with a particular focus on AI). Are these really the correct problems to focus on? What might we be missing? And what framework ought we to use to determine which problems are highest-priority?

We think that the following areas are of particular interest.

The long-termism paradigm

Let us define *long-termism* as the view that the primary determinant of the value of the actions we take today is the effect of those actions on the very long-term future. This is a very popular view within the EA community, and has enormous importance if correct. This warrants much more work to make this argument rigorous, explore to what extent this view is justified on moral views other than total utilitarianism, and work out what exactly its implications are.

Those in the effective altruism community typically believe that long-termism entails that the most cost-effective activities are those that mitigate existential risks, where an existential risk is a risk that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development. Within that category, the primary focuses are on reducing the risk of extinction from novel pathogens and from artificial intelligence.

Potential research projects in this area:

- Rigorously state and assess the argument in favour of long-termism. What about the idea that, because of ongoing risk of extinction from 'unknown unknowns', we should discount value in the future by an annual rate? What about the 'doomsday' argument that we should believe it's astronomically unlikely that we would find ourselves right at the beginning of the distribution of a vast number of people across time?
- Explore the extent to which long-termism is justified under a wide variety of different moral views, including different population ethics. To what extent do non-totalist axiologies, and in particular person-affecting approaches to population ethics, bring us to similar conclusions about long-termism or the importance of existential risk reduction?

- Assess the claim that long-termism leads to the conclusion that existential risk as the most pressing problem to work on. Does it lead to the conclusion that reducing extinction risk is the most pressing problem to work on? Should we focus on 'trajectory changes' (that is, smaller but very long-lasting improvements to total value achieved at every time) instead? Might risk aversion or ambiguity aversion be grounds for prioritising causes that don't try to shape the long-run future?
- Assess whether we ought to maximise expected value for very small probabilities of very large amounts of value. If not, what ought we to do instead?
- Assess whether it's a defensible position that we ought to try to bring about an astronomically large finite amount of value in the future, but not an infinitely large amount of value. If this is not a defensible position, is this a reductio of the idea that we ought to try to bring about an astronomically large finite amount of value, or an argument that we really should be pursuing infinite amounts of value? If the latter, what are the best activities to pursue?
- Assess the case in favour of speeding up technological progress, rather than of (directly) trying to avoid extinction.
- It's extremely difficult to know the causal effects of our actions or to predict technological and political changes over a large timescale. What activities (other than reducing extinction risk) look very good across a wide variety of future scenarios, or in a wide variety of circumstances in which we turned out to be wrong? Might these 'broad' approaches to existential risk reduction be more effective than 'narrow' approaches (such as working on technical AI safety)?
- Many people in the effective altruism community believe that climate change poses a very low probability of existential risk. Assess whether this is correct.
- If 99% of the human race were killed, how likely would it be that something similar to modern civilisation would be rebuilt? If the human race were made extinct but other life continued on earth, how likely is it that other intelligent life would develop and create a technologically advanced civilisation? How likely is it that other intelligent life with spring up elsewhere in the accessible universe and spread across the stars?
- Are there existential risks that we're currently missing or neglected?

Existing academic literature:

- Martin Weitzman, On modeling and interpreting the economics of catastrophic climate change, *Review of economics and statistics* 91 (2009): 1-19.
- Charles Jones, Life and growth, Journal of political economy 124 (2016): 539-578.
- Yew-Kwang Ng, The importance of global extinction in climate change policy, *Global policy* 7 (2016): 315-322.
- John Broome, The most important thing about climate change, in Jonathan Boston, Andrew Bradstock and David Eng (eds.), *Public policy: why ethics matters* (Canberra, A.C.T.: ANU E Press, 2010), pp. 101-116.

- Nick Bostrom, Astronomical waste: The opportunity cost of delayed technological development, *Utilitas* 15 (2003): 308-314.
- Jason Matheny, Reducing the risk of human extinction, *Risk analysis* 27 (2007): 1335-1344.
- Nick Beckstead, On the overwhelming importance of shaping the far future (PhD diss., Rutgers University, 2013).

Existing EA discussion:

- Nick Beckstead, A proposed adjustment to the astronomical waste argument.
- Nick Beckstead, How to compare broad and targeted attempts to shape the far future.
- Nick Beckstead, The long-term significance of reducing global catastrophic risks.
- Ben Todd, If you want to do good, here's why future generations should be your focus.
- Amanda Askell, Common objections to Pascal's Wager.

The value of the future

As above, long-termism is often thought to lead to the conclusion that we ought to prioritise extinction risk reduction. This presupposes that the future will be good. But one can at least imagine some scenarios in which we should expect the future to contain more bad than good. How should we assess that possibility? What is the likelihood of good outcomes versus bad outcomes? And how should we weight very bad futures compared to very good futures - should we treat the best possible future as equally as good as the worst possible future, or should we give more weight to bad scenarios?

Even if we believe that the future is good, we should bear in mind that not all good outcomes are the same. For example, from a total hedonistic utilitarian perspective, the vast majority of possible 'good' future for the human race are, comparatively speaking, of almost no value compared to the very best possible future. Might it therefore be more important to ensure that the future is as good as possible in the worlds in which the human race does not go extinct, than to reduce the probability of extinction?

What's more, different axiologies differ radically in terms of what the best possible future looks like: is it flourishing lives among flesh-and-blood agents, or blissful hedonic states run by computer programs, or something else again? This raises both philosophical and practical worries. Philosophical, because, if we think the long-run future is of overwhelming importance, then it's important to work out what ultimate state we should be aiming for. Practical because, even though there is a current convergence between the goals of individuals with different sets of values in EA, this convergence might dissipate at a crucial juncture, such as at the point of development of superintelligence, when the designers are choosing what values to code in to their AI. Are there convergent instrumental goals that many different moral views would agree on? Given axiological uncertainty, can we make any claims about what sort of future we should try to aim for?

Potential research projects in this area:

- Assess whether it might be more important to ensure that future civilisation is good, assuming we don't go extinct, than to ensure that future civilisation happens at all.
- Assess the expected value of the continued existence of the human race. Might this expected value be negative, or just unclear? How do our answers to these questions vary if we (i) assume classical utilitarianism; (ii) assume non-utilitarian or non-consequentialist moral theories; (iii) fully take moral uncertainty into account?
- Should we be more concerned about avoiding the worst possible outcomes for the future than we are for ensuring the very best outcomes occur (whether because the worst outcomes are worse than the best outcomes are good, or because they're more neglected)? If so, what activities would be best?
- To what extent does the idea of option value give us strong reason to prevent human extinction even if we're unsure about the sign of the value of the future? What's the chance that the people making the decision in the future about how to use our cosmic endowment are such that we would be happy, now, to defer to them?
- A number of people in EA have suggested the idea of the Long Reflection: a period of tens of thousands of years, where humans have much greater cognitive abilities than today, which we can dedicate to working out what is ultimately of value, before we embark on spreading to the stars. Does this idea make sense? If so, what are the conditions that we should bake into the long reflection?
- Some have suggested a 'Grand Bargain' among different axiologies, where the universe is divided up among different plausible moral views, and those different parts are optimised in the way that's best from each of those different moral views. Might this be a plausible goal to aim for?

Existing academic literature:

- Thomas Hurka, Asymmetries in value, *Noûs* 44 (2002): 199-223.
- David Benatar, *Better never to have been: the harm of coming into existence*. Oxford: Oxford University Press, 2008.
- Arthur Schopenhauer, On the sufferings of the world, in *Essays and aphorisms*. Translated by R. J. Hollingdale. London: Penguin, 1970.
- Steven Pinker, The better angels of our nature: why violence has declined. New York: Viking Books, 2011.

Existing EA discussion:

- Paul Christiano, Why might the future be good?
- Ben West, An argument for why the future may be good

- Carl Shulman, Spreading happiness to the stars seems little harder than just spreading
- Carl Shulman, Are pain and pleasure equally energy-efficient?
- Brian Tomasik, Risks of astronomical future suffering
- David Althaus and Lukas Gloor, Reducing risks of astronomical suffering: a neglected priority
- Michael Dickens, Is preventing human extinction good?
- Nick Beckstead and Carl Shulman, Will the future be good? [unpublished]
- William MacAskill, Human extinction, asymmetry, and option value
- Toby Ord, Existential risk and existential hope

Indirect effects

Effective altruists generally assume that, in evaluating interventions, we should in principle take into account *all* welfare-relevant effects of those interventions. That is, we should include not just direct effects, like the impact on school attendance from deworming schoolchildren, but also indirect effects, like the impacts on population size, economic growth, and government activity. This seems highly plausible, although it also seems somewhat in tension with popular claims in health ethics/economics. If we should take indirect effects into account, then, ultimately, we need to assess the impact of our actions from now until the end of time. How should we do this?

Potential research projects in this area:

- Assess the case for taking indirect effects into account in cost-effectiveness and cost-benefit analysis. In particular, how does this case relate to the popular idea that it would be morally inappropriate for healthcare prioritisation to take into account anything other than the patient's direct 'medical need' for the intervention being evaluated?
- It's commonly believed within the EA community that the long-run effects of our actions are typically of much greater expected impact than the short-run effects. Assess whether this is true.
- What are the long-term effects of small benefits now, such as saving a life or improving the conditions of caged hens? How great in value are these changes? Might they be great enough to rival extinction risk mitigation efforts?
- To what extent should indirect effects decrease the estimated variance in cost-effectiveness of different programs?
- A reasonably common view within the EA community is that taking long-run effects into account strengthens the case for focusing on improving human lives rather than improving non-human animal lives. Assess whether this is true.
- In comparing cause areas, when are the comparisons driven primarily by short-run effects, and when by differences in long-run effects? Given the vastness of the future, it is clear that in an *objective* sense the answer should be that almost always the long-run effects are vastly more important. On the other hand, short-run effects are often better predictable, and for the purposes of expected value

comparisons, this consideration pushes in the other direction. Which consideration dominates in which situations?

Existing academic literature:

- Dan W. Brock, Separate spheres and indirect benefits, *Cost effectiveness and resource allocation* 1 (2003): 1-12.
- Kasper Lippert-Rasmussen and Sigurd Lauridsen, Justice and the allocation of healthcare resources: should indirect, non-health effects count?, *Medicine, healthcare and philosophy* 13 (2010): 237-246.
- Jessica Du Toit and Joseph Millum, Are indirect benefits relevant to health care allocation decisions?, *The journal of medicine and philosophy* 41 (2016): 540-557.

Existing EA discussion:

- Hilary Greaves, Indirect effects
- Toby Ord, The value of very long-reaching effects
- Owen Cotton-Barratt, Human and animal interventions: the long-term view
- Holden Karnofsky, Flow-through effects
- Robert Wiblin, Making sense of long-term indirect effects
- Paul Christiano, My outlook
- Peter Hurford, Five ways to handle flow-through effects
- Tobias Baumann, Uncertainty smooths out differences in impact
- Brian Tomasik, Charity cost-effectiveness in an uncertain world
- Jonah Sinick, Robustness of cost-effectiveness estimates and philanthropy
- Brian Tomasik, Charity cost-effectiveness in an uncertain world
- Holden Karnofsky, Carl Shulman, Robert Wiblin, Paul Christiano, and Nick Beckstead, Flow through effects conversation
- Eliezer Yudkowsky, Flow-through is not a good defense of any EA project...

Animal welfare

One distinctive aspect of the EA community is its focus on improving the welfare of non-human animals. This focus raises a number of interesting and unresolved theoretical questions, including about the ways in which we can improve the lives of non-human animals and how we ought to prioritise between interventions that improve human lives and interventions that improve non-human animal lives.

Some of the issues that people in the EA community think about, such as wild animal suffering, have had very

little academic discussion indeed.

Potential research projects in this area:

- What sorts of entities have moral value? Humans, presumably. But what about non-human animals? Insects? The natural environment? Artificial intelligence? And how can we make well-being comparisons across them? (Obviously, this goes beyond specifically animal welfare issues, but the case of animals is a natural place to start this investigation.) How should we make inter-species comparisons of wellbeing? Is brain size a reasonable proxy? If not, how can we do better?
- Where is the 'zero level' for wellbeing? Which farm animals have lives that are net positive vs net negative? Do wild animals have lives that are net positive or net negative? What are the implications of different population axiologies for this question?
- Economic models typically represent animal welfare, if at all, only to the extent that it is represented in human preferences. Develop a rigorous economic model that embraces anti-speciesism, and work through how much difference this makes to the important conclusions such models are used to support, for example within agricultural economics.
- To what extent would changes to the farm production of one animal affect the numbers of other (farmed and wild) animals born? What are the differences between the welfare consequences of eating farmed meat vs. hunted meat?
- A number people in the EA community worry about the 'rich meat eater' problem: that saving lives, or boosting economic growth, leads to greater meat production. This could significantly decrease the net positive impact of saving lives or improving economic growth, or could even means that saving lives has net negative value. A counterargument is that saving lives has many positive indirect effects, too, such as contribution to innovation and the building of infrastructure. Which of these considerations is larger? How, in general, should we think about the impact on animals of improving human lives?
- Some people in the effective altruism community think that improving the living conditions of non-human animals in the wild is among the most important causes. Assess the case for this. What tractable activities are there in this area?

Existing academic literature:

- Gaverick Matheny and Kai M. A. Chan, Human diets and animal welfare: the illogic of the larder, *Journal of agricultural and environmental ethics* 18 (2005): 579-594.
- Yew-Kwang Ng, Towards welfare biology: evolutionary economics of animal consciousness and suffering, *Biology and philosophy* 10 (1995): 255-285.
- Tyler Cowen, Policing nature, *Environmental ethics* 25 (2003): 169-182.
- Jeff McMahan, The moral problem of predation, in Andrew Chignell, Terence Cuneo and Matt

Halteman (eds.), *Philosophy comes to dinner: arguments about the ethics of eating* (London: Routledge, 2015).

- Nicolas Delon and Duncan Purves, Wild animal suffering is intractable (draft).
- Tatjana Višak and Robert Garner (eds.), The ethics of killing animals (Oxford: Oxford University Press, 2015).
- Gaverick Matheny, Least harm: a defense of vegetarianism from Steven Davis's omnivorous proposal, *Journal of agricultural and environmental ethics* 16 (2003): 505–511.

Existing EA discussion:

- Rossa O'Keeffe-O'Donovan and Eva Vivalt, Agricultural economics and animals
- Carl Shulman, How are brain mass (and neurons) distributed among humans and the major farmed land animals?
- Carl Shulman, Trends in farmed animal life-years per kg and per human in the United States
- Carl Shulman, Various functional forms for brain-weighting wild insects and farmed land animals favor the former
- Carl Shulman, Vegan advocacy and pessimism about wild animal welfare
- Carl Shulman, Some considerations for prioritization within animal agriculture
- Brian Tomasik, Is brain size morally relevant?
- Luke Muehlhauser, Report on consciousness and moral patienthood
- Brian Tomasik, The importance of wild animal suffering
- Michael Dickens, Why the Open Philanthropy Project should prioritize wild animal suffering
- Toby Ord, Crucial considerations for animal welfare
- Robin Hanson, Why meat is moral, and veggies are immoral

Research to engage with:

- F. Bailey Norwood and Jayson Lusk, Compassion, by the pound: the economics of farm animal welfare. Oxford: Oxford University Press, 2011.
- An extensive literature in agricultural economics

Cross-cause comparisons

Comparative cost-effectiveness analysis is relatively straightforward when the interventions being compared are sufficiently similar, for example two different ways of relieving blindness, or two different ways of increasing the number of years children spend in school. But this is only a very narrow class of prioritisation decisions. We also need to be able to compare very different interventions in terms of 'amount of good done per dollar spent', for example programs focused on preventing blindness versus programs increasing child test scores. This requires constructing an adequate common measure of 'good done', to create a common numerator, in a principled way.

A preliminary step in this direction is to construct a practically usable index of well-being. Traditional economic models assume that well-being depends only on 'consumption', but this is either obviously false or so highly aggregated as to be of little direct use in practical applications: what we need is a tool that helps to integrate considerations of money, political conditions, health, education level, climate, and so forth. Tools doing a restricted version of this are already well-developed in the health economics literature ('quality-adjusted life years', or 'QALYs'). But to compare non-health interventions, and to include non-health effects even of health interventions, we need to generalise these tools. At present, there are only a very few preliminary attempts in this direction.

Even if we manage to make comparisons across very different types of well-being impacts, we still face problems. Crucially, we need to make comparisons when the *beneficiaries* are very different: we could benefit present-day humans or non-human animals or future people; or we could do good by bringing about good things that (arguably) don't involve benefiting individuals, such as by adding people with positive well-being to the world.

Potential research questions:

- What are the ways in which existing measures of 'good' (e.g. wealth created, QALYs) are limited? What would a metric that did not suffer from these problems look like?
- What common yardsticks can we use as proxies to compare between very different types of intervention, in terms of their long-run impact?
- When engaging in cost-benefit analysis, economists use increase in wealth as the metric by which to compare the impact of very different sorts of intervention. To what extent is this a good proxy for long-run impact?
- If we fully take moral uncertainty into account, at what point should we be indifferent between benefitting someone by 1 unit of wellbeing and bringing someone into existence with n units of wellbeing?
- How should we weigh benefits to humans against benefits to non-human animals?

Existing academic literature:

- Richard Cookson *et al.*, Years of good life based on income and health: Re-engineering cost-benefit analysis to examine policy impacts on well-being and distributive justice, *CHE research paper* 132 (2016).
- Marc Fleurbaey, Equivalent income, in Matthew D. Adler and Marc Fleurbaey (eds.), *The Oxford handbook of well-being and public policy* (Oxford: Oxford University Press, 2016), pp. 453-475.
- Richard Layard et al., What predicts a successful life? A life-course model of well-being, *The economic journal* 124 (2014): F720-F738.

Existing EA discussion:

- GiveWell, How GiveWell and mainstream policymakers compare the "good" achieved by different program
- Nick Bostrom, Crucial considerations and wise philanthropy
- Carl Shulman, What proxies to use for flow-through effects?
- Carl Shulman, Turning log-consumption into a [crude] measure of short-run human welfare
- Owen Cotton-Barratt and Eva Vivalt, Multidimensional measures
- Owen Cotton-Barratt and Eva Vivalt, WALYs

Existing research to engage with:

- Multidimensional outcome analysis. There is a substantial literature in health economics on quality-adjusted life years (QALYs). A smaller literature in development economics focuses on constructing 'multidimensional poverty indices'. Here the driving idea is that poverty is traditionally measured only in terms of low income, but that other factors (e.g. longevity, literacy) should be taken into account too. (This idea is sometimes associated with Sen's "capabilities approach".) There are obvious analogies between the task of constructing a QALY measure and the task of constructing a MPI, but at least in practice, there are also significant differences between the way the two constructions are carried out.

Cross-cutting considerations

Even when a group of altruistic individuals have chosen a problem, or number of problems, to focus on, there are still many open theoretical questions that they face if they wish to do the most good. Some of these concern what individuals ought to do — under what conditions they should try to do good right away, and under what conditions should they invest their resources in order to do more good later. Some of these questions concern how groups on altruistic individuals can ensure that collectively they have the largest possible impact.

Within this, we think that the following areas are particularly interesting and important.

Altruistic economics

Economic theory normally proceeds either (a) assuming *nothing* substantive about individuals' preferences (assuming only structural conditions, e.g. that preferences are complete and transitive), or (b) assuming that preferences are in some sense 'self-interested' (e.g. that an individual's utility depends only on her own consumption/leisure/etc.). It may be that interesting new phenomena arise, and interesting new results can be established, when we assume instead that individuals are perfectly altruistic (for example, that each individual's utility function is simply the utilitarian one).

One might well, for example, encounter epistemic problems: prices will no longer reliably signal the information they are meant to, if the preferences that people act on (the altruistic ones) diverge radically from their *self-interested* preferences ('insofar as I only care about myself, I prefer…').

If so, this area is likely to contain low-hanging fruit, since economics is not normally interested in questions of completely pure altruism (implicitly viewing this as psychologically so unrealistic as not to be worth thinking about).

Potential research topics in this area:

- What is the right thing to say about the Fundamental Theorems of Welfare Economics, when we start talking about a community of pure altruists instead of a community of self-interest maximisers?
- How can a community of altruists with different and empirical views gain from trade? What are the challenges for moral trade that go beyond the challenges for ordinary trade, and can they be overcome?
- Is it ethical to 'offset' the harm that you cause?

Existing academic literature:

- Toby Ord, Moral trade, *Ethics* 126 (2015): 118-138.

Existing EA discussion:

- Ben Garfinkel, What is the relationship between effective altruism and economics?
- Paul Christiano, Certificates of impact
- Michael Page, Certificates of impact
- Paul Christiano, Repledge++
- Carl Shulman, Donor lotteries: demonstration and FAQ

Altruistic coordination theory

Given multiple actors deciding how to distribute resources (for example money, but also perhaps labour) for altruistic purposes, how will they, or should they, act? The puzzle is cleanest in the case where they have slightly different values leading them to value different opportunities differently – for example if two donors agree on the first-best use of money but disagree on the second-best, they each prefer that the other fully funds the first-best use. Variations of it deal with cases with multiple donors, or where there are also empirical disagreements, or private information, or comparative advantage of different actors contributing to different projects.

Tools from game theory, bargaining theory and mechanism design should be applicable to analyse at least some versions of the questions.

Potential research topics in this area:

- Are there institutions or mechanisms we can design which help improve allocative efficiency? Might 'certificates of impact' serve this role?
- Does the idea of comparative advantage make sense for a community of altruists?
- What does game theory look like among a community of altruists? For example, assuming identical moral and empirical beliefs, such a community cannot face Prisoner's Dilemmas, but they can face Stag Hunt.

Existing academic literature:

- Dominik Peters, Economic design for effective altruism, in Jean-François Laslier et al. (eds.), *Future of economic design*. Forthcoming.

Existing EA discussion:

- Rossa O'Keeffe-O'Donovan, Economics of career choice
- Luke Muehlhauser, Technical and philosophical questions that might affect our grantmaking, section 4
- Holden Karnofsky, Donor coordination and the "giver's dilemma"
- Holden Karnofsky, Good Ventures and giving now vs later (section Coordination issues)
- The Open Philanthropy Project, A conversation with Professor S. Nageeb Ali
- Owen Cotton-Barratt and Zachary Leather, Donor coordination under simplifying assumptions
- Benjamin Todd, The value of coordination
- Benjamin Todd, How can we best work together as a community?
- Jess Whittlestone, Building an effective altruism community

Altruistic decision theory

Various apparently altruistic and apparently very reasonable behaviours seem puzzling on closer inspection *assuming* that the agent is attempting to maximise expected good. These puzzling behaviours include many that card-carrying EAs find themselves drawn to, such as donating to more than one charity, and avoiding supporting work on x-risk mitigation on grounds of 'risk aversion'. The same behaviours might make a lot more sense assuming a less pure form of altruism (the most obvious alternative being: a 'warm glow' theory of motivation), or assuming deviations from expected utility theory that are arguably irrational (such as ambiguity aversion).

Possible research projects:

- Map out the various forms of altruism, such as 'pure' altruism, and 'warm glow' altruism.
- Which decision-theoretic models rationalise which behaviours? Which are the genuine psychological mechanisms at work?
- To what extent to the behaviours in question manifest deviations from ideal rationality and/or deviations from pure altruism?

Existing academic literature:

- Daniel Batson, *Altruism in humans*. New York: Oxford University Press, 2011.
- Daniel Batson et al., Empathy and altruism, in C. R. Snyder and Shane J. Lopez (eds.). *Oxford handbook of positive psychology*, 2nd ed. (Oxford: Oxford University Press, 2009), pp. 417-427.
- Jon Elster, The Valmont effect: the warm-glow theory of philanthropy, in Patricia M. L. Illingworth, Thomas Pogge and Leif Wenar (eds.), *Giving well: the ethics of philanthropy* (New York: Oxford University Press, 2011), pp. 67-83.

- James Andreoni, Privately provided public goods in a large economy: the limits of altruism, *Journal of public economics* 35 (1988): 57-73.
- James Andreoni, Giving with impure altruism: applications to charity and Ricardian equivalence, *Journal of political economy* 97 (1989): 1447-1458.
- James Andreoni, Impure altruism and donations to public goods: a theory of warm-glow giving, *The economic journal* 100 (1990): 464-477.

Doing good now vs doing good later

If an altruist wants to do good, she faces the question of *when* to do good. With her money, she could donate right away, or she could invest the money at a later date, or she could take out a loan in order to give more now. With her time, she could try to get a high-impact job right away, or she could spend time getting further education or job training, in order to have a larger impact later on. Under what conditions should direct intervention be attempted earlier vs. later?

When we look at how philanthropic actors implicitly answer this question, the results are mixed. Individual donors typically give a certain amount each year. Foundations and universities are typically set up in perpetuity. Individuals who are trying to do good with their career typically get at least an undergraduate education, if they can. Governments typically borrow money in order to spend more now (though often they can recoup some of that borrowing through higher taxation).

Within the EA community, there is far from consensus on the issue. There are many different considerations in this area that point in different directions, and it's not at all clear how to weigh those considerations against one another.

Potential research topics:

- What are the considerations that are relevant to the giving now versus later question? Can we build a model to quantitatively represent these considerations?
- Should we use an exponential discount rate, or some other function? Is there reason to think that we're in an unusual time with respect to how quickly we ought to discount future donations?
- To what extent does a continuous-time version of the giving now or later question look similar to the "job search" literature in economics, where individuals have to decide whether to take a job or keep looking?
- How does our answer to the now vs later question change depending on whether we're talking about money or (various form of) time?
- How does our answer vary depending on the problem we're trying to address?
- Is there any justification for universities or foundations existing in perpetuity?

- Within the EA community, typically small donors give as they earn, whereas large donors currently save much of their planned donations. Does this division make sense?
- In addition to building a model by which to answer the question, can we use existing data to quantitatively address the giving now or later question for specifical cause areas?
- If discount rates differ significantly across cause areas, might this provide an opportunity for people across different cause area to engage in moral trade across time?

Existing academic literature:

- William MacAskill, When should an effective altruist donate? (draft; see also his Princeton presentation)
- Jed Emerson, Jay Wachowicz and Suzi Chun, Social return on investment: exploring aspects of value creation in the nonprofit sector, in Teresa Moore (ed.), *Social purpose enterprises and venture philanthropy in the new millennium* (San Francisco: The Roberts Foundation, 2000), vol. 2, pp. 132-173.
- Peter Frumkin, Strategic giving: the art and science of philanthropy. Chicago: Chicago University Press, 2006.
- Renée A. Irvin, Endowments: stable largesse or distortion of the polity?, *Public administration review* 67 (2007): 445-457.
- Paul Jansen & David Katz, For nonprofits, time is money, *The McKinsey quarterly* 1 (2002): 124-133.
- Michael Klausner, When time isn't money: foundation payouts and the time value of money, *Stanford social innovation review* 1 (2003): 51-59.
- Cliff Landesman, When to terminate a charitable trust?, *Analysis* 55 (1995): 12-13.
- Dan Moller, Should we let people starve—for now?, *Analysis* 66 (2006): 240-247.

Existing EA discussion:

- Owen Cotton-Barratt and Benjamin Todd, Give now or later? What to do when the order of your actions matters
- Paul Christiano, Giving now vs. later
- Paul Christiano, The best reason to give later
- Robin Hanson, Parable of the multiplier hole
- Carl Shulman, Social rate of return
- Carl Shulman, High social returns are rarely sustainable
- Julia Wise, Giving now vs. later: a summary
- Draft structure for a paper on giving now vs later

Mission hedging

For some strategies, there is covariance between the amount of resources you control and the cost-effectiveness

of the opportunities you have available to you. Examples include earning to give by by founding an AI company if you are aiming to donate to AI safety and investing in oil companies if you are aiming to donate the returns of your investment to climate change mitigation efforts.

Potential research topics in this area:

- Precisely define the concept of mission hedging, outline the areas where it might apply, and assess its importance.
- What are the concrete implications of mission hedging? What activities would one undertake that would be different if one didn't mission hedge?

Existing academic literature:

- Brigitte Roth Tran, Divest, disregard, or double down?, *Finance and economics discussion series* 42 (2017).

Existing EA discussion:

- Hauke Hillebrandt, Socially responsible investing (draft)
- Kyle Boghosian, Selecting investments based on covariance with the value of charities

Epistemological issues

Figuring out how to do the most good is very difficult, and often it seems that subtle differences in epistemology would lead one to quite different conclusions. These include differences in responses to paucity of hard evidence, in level of trust in abstract arguments leading to counterintuitive conclusions, and in the relative weight placed on different types of evidence.

One common view in the EA community is that we should favour interventions that have more evidential support, all else being equal. On the face of it, this conflicts with expected value theory (if "all else being equal" means: "expected value being equal"). On the other hand, it also seems reasonable. What is the correct response to this tension?

Another related strand of disagreement within the EA community is to what extent one should place weight on one's idiosyncratic 'inside view' judgments, rather than deferring to the views of the majority of experts on the issue. All other things being equal, the latter idea seems to push against 'weirder' beliefs, such as that reducing AI risk might be astronomically important, or that reducing wild animal suffering might be among the most important causes. But is this just timidity?

Finally, there is an open question on how much weight to put on different *types* of evidence: evidence from randomised controlled trials, versus on theoretical models, versus philosophical argument.

Potential research topics in this area:

- Is it rationally permissible to be ambiguity averse? If so, does this give a good argument for preferring activities like global health improvements rather than working on AI safety?
- Is it correct that most interventions are fairly ineffective? If so, then is it the case that interventions that are supported only by speculative evidence will generally have lower *expected* value than that of interventions supported by more solid evidence?
- What's the base rate for positive impact activities compared to neutral and negative impact activities? How common are situations in which most ways of acting do harm, and under which conditions is this the case? What implications do these facts have for what problems we ought to focus on?
- Should we have a 'prior' over impact, such that it's astronomically unlikely that we could have the sort of positive impact that it seems we can have by reducing existential risk if total utilitarianism is correct? What bearing does this have on the value of long-run future focused activities?
- Those in the effective altruism community often have unusual views. To what extent should we be exceptionally epistemically modest? Should disagreement among peers lead us to decrease our credence in such views? Should we have the same levels of epistemic modesty about unusual moral views as we do about unusual empirical views?
- How important is the distinction between 'sequence' thinking and 'cluster' thinking? What's
- How much weight should we place on philosophical arguments? Can we perform a 'pessimistic meta-induction,' arguing that, because most philosophical arguments in the past have been mistaken, we should place very little weight on them?

Existing EA discussion:

- Amanda Askell, Seminar presentation on speculative vs robust evidence
- Holden Karnofsky, Maximising cost-effectiveness via critical enquiry
- Holden Karnofsky, Sequence thinking vs. cluster thinking
- Holden Karnofsky, Modelling extreme model uncertainty
- Nick Beckstead, Common sense as a prior
- Robert Wiblin, Is it fair to say that most social programmes don't work?
- Greg Lewis, In defence of epistemic modesty
- Tobias Baumann, Uncertainty smooths out differences in impact
- Jonah Sinick, Many weak arguments vs one relatively strong argument

Research to engage with:

The epistemology literature on peer disagreement

- Adam Elga, Reflection and disagreement, *Noûs* 41 (2007): 478-502.
- David Christensen, Epistemology of disagreement: the good news, *Philosophical review* 116 (2007): 187-217.
- Richard Feldman and Ted A. Warfield (eds.), *Disagreement* (Oxford: Oxford University Press, 2010).
- David Christensen, Disagreement as evidence: the epistemology of controversy, *Philosophy compass* 4 (2009): 1-12.
- Alastair Wilson, Disagreement, equal weight, and commutativity, *Philosophical studies* 149 (2010): 321-326.
- David Christensen and Jennifer Lackey (eds.), *The epistemology of disagreement*: new essays (Oxford: Oxford University Press, 2013).

The literature in development economics on randomised controlled trials

The problem of external validity

- Tessa Bold et al. Scaling up what works: experimental evidence on external validity in Kenyan education, *Center for Global Development working paper* 321 (2013): 1-48.
- Eva Vivalt, How much can we generalize from impact evaluations?, forthcoming.
- Rajeev Dehejia, Cristian Pop-Eleches and Cyrus Samii, From local to global: external validity in a fertility natural experiment, *IZA Discussion Papers* 9300 (2015).

New approaches to drawing inferences out of sample

- Michael Gechter, Generalizing the results from social experiments: theory and evidence from Mexico and India, forthcoming.
- Todd J. Kowalski et al., Impact of hair removal on surgical site infection rates: a prospective randomized noninferiority trial a prospective randomized noninferiority trial, *Journal of the American College of Surgeons* 223 (2016): 704-711.
- Sylvain Chassang, Gerard Padró I Miquel and Erik Snowberg, Selective trials: a principal-agent approach to randomized controlled experiments, *American economic review* 102 (2012): 1279-1309.

Critical literature challenging the hegemony of RCTs

- Angus Deaton and Nancy Cartwright, Understanding and misunderstanding randomized controlled trials (draft).

Literature in social science more generally on qualitative evidence

- Gary Goertz and James Mahoney, *A tale of two cultures: qualitative and quantitative research in the social sciences.* Princeton: Princeton University Press, 2012.
- David Freedman, On types of scientific enquiry: the role of qualitative reasoning, in Janet M. Box-Steffensmeier, Henry E. Brady and David Collier (eds.), *The Oxford handbook of political methodology*

(Oxford: Oxford University Press, 2008), pp. 300-318.

- Jason Seawright, Multi-method social science: combining qualitative and quantitative tools. Cambridge: Cambridge University Press, 2016.
- Andrew Bennett and Jeffrey T. Checkel, Process tracing: from philosophical roots to best practices, in Andrew Bennett and Jeffrey T. Checkel (eds.), *Process tracing: from metaphor to analytic tool* (Cambridge: Cambridge University Press, 2014), pp. 3-37.

Diversification

What rationales are there, either for the individual or for the EA community/world as a whole, to diversify across causes/interventions, rather than simply identifying the intervention with the highest expected cost-effectiveness and supporting exclusively that intervention?

Possibilities here include: diminishing marginal returns of resources (directed towards a single cause area or intervention) to impartial value; diminishing marginal returns of progress in a single cause area to decision-maker's utility even in the absence of any such diminishing returns to impartial value; information value of investing in interventions; moral uncertainty.

Research topics in this area:

- What are the potential reasons for diversifying? Which, if any, are successful for individuals? For a large foundation? For the effective altruism community as a whole?
- How do the considerations in favour of or against diversifying apply when we consider how to allocate human resources rather than financial resources?
- To what extent should a large foundation diversify across different 'worldviews'? To what extent does normative uncertainty provide support for such diversification?
- How great is the difference in effectiveness between the best charities and typical charities? How fast do returns diminish within a cause area? Is the effectiveness among charities fat-tailed? (These questions also have implications for the extent to which we should prioritise further research over 'direct intervention'; map those implications.)
- How do the following compare? The variation of effectiveness of different organisations implementing (roughly) the same intervention; the variation of average effectiveness for different interventions in the same cause/domain; the variation of average effectiveness between different causes/domains?
- To what extent should those in the EA community try to diversify across time?

Existing academic literature:

- James Snowden, Does risk aversion give an agent with purely altruistic preferences a good reason to donate to multiple charities?

Existing EA discussion:

- James Snowden, Should we give to more than one charity?
- Holden Karnofsky, Worldview diversification
- Ben Kuhn, How many causes should you give to?
- Brian Tomasik, When should altruists be financially risk-averse?
- Carl Shulman, Salary or startup? How do-gooders can gain more from risky careers
- Owen Cotton-Barratt, What does (and doesn't) AI mean for effective altruism?

Distribution of cost-effectiveness across interventions

It is a platitude within the EA community that the cost-effectiveness of interventions within a single cause area typically varies by many orders of magnitude. How strong is the evidence for this claim, and what can we establish about the shape of distributions of cost-effectiveness more generally, both within and between causes?

Possible research projects in this area:

- Establish more rigorously what can be said about typical distributions of cost-effectiveness, both within and between causes.
- Precisely what is the relationship between distributions of cost-effectiveness on the one hand, and the relative expected value of exploration vs. exploitation on the other?
- How does variation of cost-effectiveness within a cause compare to variation of cost-effectiveness between causes? This relates to the issue of diversification, above: If it's the case that the variance of the distribution of interventions within a cause is comparable with the variance of the distribution of causes, then we will get the conclusion that we should diversify across causes.

Existing academic literature:

- Toby Ord, The moral imperative towards cost-effectiveness in global health, *Center for Global Development*, 2013.

Existing EA discussion:

- Owen Cotton-Barratt, Distributions of effectiveness
- Jeff Kaufman, The unintuitive power laws of giving

Moral Uncertainty

When effective altruists attempt to compare the importance of different problems, or the effectiveness of different interventions, they typically default to using a utilitarian axiology. But, even if you are sympathetic to a utilitarian axiology, it would clearly be overconfident to be *certain* in that axiology. So, plausibly, we should try to incorporate moral uncertainty into our reasoning when we prioritise among problems. It remains underexplored, however, what implications the fact of moral uncertainty has. How do practical conclusions change when we incorporate reasonable moral uncertainty into our analysis?

Potential research projects in this area:

- What are the implications of an appropriate treatment of moral uncertainty for the question of what problem we should be focusing on? How robust are the existing choices of most important problems (global poverty, farm animal welfare, and existential risk reduction) under different moral views?
- To what extent, in general, is the sum total of human welfare a good proxy for what's of value? If we moved to an egalitarian or prioritarian axiology, or if we assigned intrinsic value to some non-welfarist goods like natural beauty and art, would that change our conclusions much? What about if we incorporated ideas of justice?
- How does moral uncertainty change our assessment of the value of the long-run future?
- Under moral uncertainty, do some moral views with very high stakes swamp the expected value calculation? If so, which views are they?
- How should we think about the risks of doing harm in the course of doing good? What sorts of harms are morally permissible (from a non-consequentialist perspective), and which aren't?
- How likely is it that we've missed out on the most important cause? What are some contenders for Cause X?
- It's likely that we're making deep conceptual or moral mistakes (as almost all our forebears had done). Can we do value-free cause-prioritisation, where we decide between causes while making almost no commitments about what the correct moral view is?
- To what extent does moral uncertainty give us reason to pursue 'meta' activities, such as further research into moral philosophy?
- Does the idea of 'moral option value' give us a reason to prevent human extinction, even if we currently think that continued human existence is bad in expectation? What other implications does

this idea have?

Existing academic literature:

- William MacAskill, Krister Bykvist and Toby Ord, *Moral uncertainty*. Oxford: Oxford University Press, 2018 (see esp. chs. 8-9 and the conclusion).
- Hilary Greaves and Toby Ord, Moral uncertainty about population axiology (draft)
- Evan Williams, Promoting value as such, *Philosophy and phenomenological research* 87 (2013): 392-416.

Existing EA discussion:

- Toby Ord, The moral imperative towards cost-effectiveness
- Amanda Askell, The moral value of information
- William MacAskill and Benjamin Todd, Is it ever okay to take a harmful job in order to do more good? An in-depth analysis

Indirect justifications of decision norms

Fundamentally, we assume, prioritisation of all forms (whether among problems or among interventions) should be via expected value theory: we simply seek those interventions that have highest expected value. But this injunction is extremely abstract and general. For practical purposes, it is useful also to have some more easily applicable principles for identifying interventions that are likely to have high expected value.

The typical way of prioritising among problems in the effective altruism community is to assess them in terms of their importance (how many individuals does this problem affect, and by how much), their tractability (how much progress can we make on this problem with a given unit of resources), and their neglectedness (how many resources are put towards this addressing this problem already). What is the status of this framework? Is it the right one to use?

The prevalence of the 'ITN' this framework might seem surprising, and raises some questions. Precisely why should we use these criteria, rather than simply trying to assess the expected value directly? cost-effectiveness of different interventions? Except in those cases where one can have a proportional impact on a problem, why should one care about importance of the problem itself? If one can have an influence over how resources addressing a problem are used, might neglectedness be a bad thing? Why is this framework not used in fields outside of effective altruism?

Potential research projects in this area:

- Formally define the importance, tractability and neglectedness criteria, and make the case why these criteria map on to expected value maximisation.cost-effectiveness.
- Question the importance of these criteria, or explore the conditions under which these are or are not a useful set of criteria for estimating cost-effectiveness.
- What other 'rules of thumb', if any, might have a similar status vis-a-vis expected value theory in the EA context? Is 'cluster thinking' an example?

Existing EA discussion:

- William MacAskill, Doing good better: effective altruism and how you can make a difference. New York: Penguin, 2015, ch. 10
- EA concepts, Importance, tractability, neglectedness framework
- Open Philanthropy Project, Cause selection
- 80,000 Hours, How to compare different global problems in terms of impact
- Owen Cotton-Barratt, Factoring cost-effectiveness
- Paul Christiano, A conversation on cause prioritization research
- Owen Cotton-Barratt and Daniel Kokotajlo, How can we help the world? A flowchart
- Gregory Lewis, Beware surprising and suspicious convergence

Research to engage with:

The methodology of cost-effectiveness analysis: Cost-effectiveness and cost-benefit analysis are the economists' standard tools for project evaluation. Several aspects of the methodology of CBA and CEA, however, are contested, often for reasons that tap into fundamental normative controversies. Examples include the choice of a discount rate (to trade of costs/benefits incurred earlier against those incurred later), and the use or not of "distributional weights" (to account for the fact that a marginal dollar is worth more to a poor person than to a rich person).

- J-PAL, Conducting cost-effectiveness analysis.
- HM Treasury, The green book: appraisal and evaluation in central government. London: TSO, 2013.
- Amartya Sen, The discipline of cost-benefit analysis, *Journal of legal studies* 29 (2000): 931-952.

The Scope of Effective Altruism

The primary focus of our research agenda is on the question of how one can use a given unit of resources to do as much good as possible. However, there are also important questions about the nature and strength of the motivation for and/or the moral imperative towards caring about that question. Is everyone required to dedicate their lives to effective altruism? If not, to what extent do we have obligations to engage in effective altruism? And to what extent ought considerations of what will do the most good influence the decisions of governments, as opposed to private individuals?

The obligation to engage in effective altruism

This topic concerns whether effective altruism is simply a beneficial project that one might or might not choose to engage in, or whether stronger things can be said in its favour from the point of view of moral philosophy. Questions of this type form the main focus of most of the existing commentary on effective altruism among moral philosophers.

Potential research topics in this area:

- If it's the case that the long-run effects of one's actions are much larger in impact than the short-run effects, this seems to strengthen the case for there being strong duties of beneficence, simply because altruistic actions do so much more good than we might have thought. Assess whether this implication really holds.
- Non-consequentialist views often make 'emergency situation' provisos, where they tend to make recommendations in a more consequentialist manner (such as permitting rights violations or making acts of altruism obligatory). To what extent is it justified to think that we are living in an 'emergency situation'?
- If there is an obligation to engage in effective altruism, what is the nature of that obligation? Should all our resources be spent in whichever way would do the most good? Is the idea of conditional obligation compelling?
- Do obligations of beneficence require cause-impartiality?
- Even if beneficence is only one of many competing obligations in our lives, is it still the case that with respect to the reasons of beneficence that we have, we ought to try to do the most good?
- What are the best arguments for the claim that it's impermissible to use a significant part of one's resources in whatever way does the most good?

Existing academic literature:

- Peter Singer, *The life you can save: acting now to end world poverty*. New York: Random House, 2009.
- Peter Singer, *The most good you can do: how effective altruism is changing ideas about living ethically*. New

Haven: Yale University Press, 2015.

- Theron Pummer, Whether and where to give, *Philosophy & public affairs* 44 (2017): 77-95
- Joe Horton, The all or nothing problem, *The Journal of Philosophy* 114 (2016): 94-104.
- Theron Pummer, People and charitable causes are importantly different things.
- Andreas Mogensen, Should we prevent optimific wrongs, *Utilitas* 28 (2015): 215-226.

Cluelessness

Many people who would otherwise be inclined towards EA-like behaviour refrain from such behaviours because of epistemic concerns: they feel that they are simply too clueless about which well-meaning actions would in the end do net good versus harm, and how much, for it to 'make sense' to expend significant resources on altruistically-meant interventions. (The source of this worry is closely related to issues of 'indirect' and/or long-run effects: perhaps a randomised controlled trial can give us a pretty good idea of a particular immediate consequence of a given intervention, but what about e.g. the knock-on implications for political structures, individuals not treated, long-run trajectories of economic development, population size and environmental degradation, and any effects on x-risk?)

Potential research topics in this area:

- Is cluelessness-induced inaction rational?
- If it is rational, what it the theory of rationality that describes it? If it is not rational, why does that phenomenon occur?

Existing academic literature:

- Hilary Greaves, Cluelessness
- James Lenman, Consequentialism and Cluelessness

Existing EA discussion:

- Amanda Askell, The moral value of information

Effective Altruism as a political philosophy

Effective altruism typically concerns itself with the decisions of individuals. But there's no principled reason why this should be so. We could ask how corporations could use their resources to do the most good. And, more interestingly, we could ask how governments could use their resources to do the most good.
The first set of questions we could address in this area, then, is how the EA framework changes when we consider

the resources that governments have available: government budgets, legislative power, and ability to influence the actions of other states.

The second set of questions we could address is to what extent such considerations ought to play a role in government decision-making.

Potential research projects in this area:

- What are the most pieces of legislation that the (US or UK) government could enact that would have the greatest positive impact?
- To what extent ought the government to take actions that are better for the world even if they conflict with the 'will of the people'?
- Most of the individuals who are impacted by government decisions are people in the future or non-human animals. They do not get a vote, nor do they participate in markets. To what extent does this provide an argument against both the free market and statist political philosophies? Is there a better alternative?
- What is the best system of government from the perspective of impartial welfarism?
- How promising is futarchy - Robin Hanson's view that we ought to 'vote on values but bet on beliefs' - as a way of structuring society?

Existing academic literature:

- John Rawls, *A theory of justice*. Cambridge, Massachusetts: Belknap Press, 1971, esp. sect. 44.
- Brian Barry, Sustainability and intergenerational justice, Theoria 65 (1997): 43-64.
- Sue Donaldson and Will Kymlicka, *Zoopolis: a political theory of animal rights*. Oxford: Oxford University Press, 2013.
- Robert Goodin, *Utilitarianism as a public philosophy*. Cambridge: Cambridge University Press, 1995.

Existing EA discussion:

- The Open Philanthropy Project, U.S. Policy
- Holden Karnofsky, The Role of Philanthropic Funding in Politics
- Eliezer Yudkowsky, Politics is the Mindkiller
- Scott Alexander, Beware Systemic Change
- Robin Hanson, Futarchy
- Rob Reich, The logic of effective altruism

Existing academic research that has broad relevance for GPI

In addition to the topic-specific relevant areas we mention above, we here highlight some more general questions that are of relevance to many different topics on GPI's research agenda. (Again, these are areas of research that we aim to be familiar with, but not directly to contribute to as a central part of our work.)

Population ethics

Relevant to:

- Existential risk reduction
- Farm animal welfare
- Saving lives vs improving lives

Examples of relevant publications

- Derek Parfit, *Reasons and persons*. Oxford: Oxford University Press, 1984, part 4.
- Gustaf Arrhenius, *Population ethics: the challenge of future generations* (draft)
- Hilary Greaves, Population axiology, Philosophy compass 12 (2017): 1-15.

Risk aversion & ambiguity aversion

Relevant to:

- Preventing existential risks vs preventing near-term suffering
- Low risk high reward activities vs safe bets

Examples of relevant publications

On risk aversion

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

On ambiguity aversion

- Daniel Ellsberg, Risk, ambiguity, and the Savage axioms, *Quarterly journal of Economics* 75 (1961): 643–69
- Normative discussion of ambiguity aversion
- Nabil I. Al-Najjar and Jonathan Weinstein, The ambiguity aversion literature: a critical assessment, *Economics and philosophy* 25 (2009): 249-284, and references therein
- Other papers in the same special issue of *Economics and Philosophy* as Al-Najjar and Weinstein 2009
- (There is also an extensive experimental literature, focussing on the descriptive adequacy of models of ambiguity aversion, as opposed to considerations of whether or not ambiguity aversion is rational.)

Moral uncertainty

Relevant to:

- What cause is most important
- Whether and in what way it's permissible to cause harm in the course of doing good
- The extent of our obligations to engage in effective altruism

Examples of relevant publications

- Ted Lockhart, *Moral uncertainty and its consequences*. Oxford: Oxford University Press, 2000.
- Jacob Ross, Rejecting ethical deflationism, *Ethics* 116 (2006): 742-768.
- Andrew Sepielli, Along an imperfectly-lighted path (PhD diss., Rutgers University, 2010).
- Elizabeth Harman, Does moral ignorance exculpate?, *Ratio* 24 (2011): 443-468.
- William MacAskill, The infectiousness of nihilism, *Ethics* 123 (2013): 508-520.
- Andrew Sepielli, Moral uncertainty and the principle of equal weight among moral theories, *Philosophy and phenomenological research* 86 (2013): 580-589.
- Johan E. Gustafsson and Olle Torpman, In defence of My Favourite Theory, *Pacific philosophical quarterly* 95 (2014): 159-174.
- Brian Weatherson, Running risks morally, *Philosophical studies* 167 (2014): 141-163.
- Elinor Mason, Moral ignorance and blameworthiness, *Philosophical studies* 172 (2015): 3037-3057.
- Owen Cotton-Barratt, William MacAskill and Toby Ord (MS) Normative uncertainty, intertheoretic comparisons, and variance normalisation

Value of information

Relevant to:

- Giving now vs later
- The 'option value' argument for preventing human extinction

Examples of relevant publications:

- I. J. Good, On the principle of total evidence, *The British journal for the philosophy of science* 17 (1967): 319-321.
- Louis Eeckhoudt and Philippe Godfroid, Risk aversion and the value of information, *The Journal of Economic Education* 31 (2000): 382-388.
- Kenneth J. Arrow and Anthony C. Fisher, Environmental preservation, uncertainty, and irreversibility, in *Classic papers in natural resource economics* (London: Palgrave Macmillan, 1974), pp. 76-84.

Discounting

Relevant to:

- The case for and implications of long-termism

- Giving now vs later

Examples of relevant publications:

- Christian Gollier. *Pricing the planet's future: the economics of discounting in an uncertain world.* Princeton: Princeton University Press, 2012.
- Hilary Greaves, Discounting for public policy: a survey, *Economics and philosophy* 33 (2017): 391-439.

Critiques and analyses of altruistic motivation from economics and from psychology

Relevant to:

- Topics on 'The scope of effective altruism'

Examples of relevant publications

- Richard Povey, The limits to altruism (draft)
- Richard Povey, The socially optimal level of altruism (draft)
- Economics literature on models of charitable giving: See e.g. the work of James Andreoni (http://econweb.ucsd.edu/~jandreon/)
- Barbara A. Oakley, *Pathological altruism*. Oxford: Oxford University Press, 2011.
- Paul Bloom, *Against empathy: the case for rational compassion*. New York: Harper Collins, 2016.
- Naza Ashraf and Oriana Bandiera, Altruistic capital, *American economic review* 107 (2017): 70-75

Further EA Reading

A good introductory overview of the theoretical side of global priorities research is Prospecting for Gold by Owen Cotton-Barratt.

The most important websites to get up to speed on current thought and debates in the EA community are as follows:

- Givewell.org, and their blog
- Openphilanthropy.org, and their blog
- 80000hours.org, and their blog
- http://globalprioritiesproject.org/
- concepts.effectivealtruism.org/
- https://www.effectivealtruism.org/articles/
- Effective-altruism.com, though this also contains discussion of effective altruism community issues that aren't as relevant to the effective altruism research agenda
- reducing-suffering.org/
- https://foundational-research.org
- https://rationalaltruist.com/
- http://reflectivedisequilibrium.blogspot.co.uk/
- www.lesserwrong.com/, though this also contains discussion of issues concerning rationality that aren't as relevant to the effective altruism research agenda

Finally, here is an incomplete list of some of the most important articles and blog posts from the EA community that are relevant to GPI's research agenda (many of which are also mentioned above):

- 80,000 Hours How to compare different global problems in terms of impact
- 80,000 Hours List of the most urgent global issues
- Scott Alexander Ethics offsets
- Scott Alexander Nobody is perfect, everything is commensurable
- David Althaus and Lukas Gloor Reducing risks of astronomical suffering
- Nick Beckstead On the overwhelming importance of shaping the far future
- Nick Beckstead A proposed adjustment to the astronomical waste argument
- Nick Bostrom 3 ways to advance science
- Nick Bostrom Crucial considerations and wise philanthropy
- Paul Christiano Astronomical waste
- Paul Christiano Influencing the far future
- Paul Christiano Neglectedness and impact

- Paul Christiano Pressing ethical questions
- Paul Christiano Replaceability
- Paul Christiano The best reason to give later
- Paul Christiano The efficiency of modern philanthropy
- Owen Cotton-Barratt How valuable is movement growth?
- Owen Cotton-Barratt and Ben Todd Give now or later?
- Katja Grace Cause Prioritization Research
- Katja Grace Estimation Is the Best We Have
- Robin Hanson Marginal charity
- Robin Hanson Parable of the multiplier hole
- Holden Karnofsky Flow-through effects
- Holden Karnofsky Hits-Based Giving
- Holden Karnofsky Passive vs. rational vs. quantified
- Holden Karnofsky Sequence thinkings vs. cluster thinking
- Holden Karnofsky Your Dollar Goes Further Overseas
- Holden Karnofsky Why we can't take expected value estimates literally even when they're unbiased
- Holden Karnofsky Worldview diversification
- Jeff Kaufman Altruism isn't about sacrifice
- Jeff Kaufman The Unintuitive Power Laws of Giving
- Ben Kuhn A critique of effective altruism
- Greg Lewis Beware Surprising and Suspicious Convergence
- Toby Ord The Moral Imperative Towards Cost-Effectiveness
- Carl Shulman Are pain and pleasure equally energy efficient?
- Carl Shulman & Nick Beckstead A Long-run Perspective on Strategic Cause Selection and Philanthropy
- Jonah Sinick Many Weak Arguments vs. One Relatively Strong Argument
- Scott Siskind Dead children currency
- Scott Siskind Efficient charity
- Ben Todd The value of coordination
- Brian Tomasik Charity Cost Effectiveness in an Uncertain World
- Brian Tomasik The Haste Consideration Revisited
- Brian Tomasik Two-envelopes problem for brain size and moral uncertainty
- Brian Tomasik Why charities don't differ astronomically in expected cost-effectiveness
- Brian Tomasik How the simulation argument dampens future fanaticism
- Ben West Another critique of effective altruism
- Robert Wiblin How to create the world's most effective charity

globalprioritiesinstitute.org