A RESEARCH AGENDA FOR THE

GLOBAL PRIORITIES INSTITUTE

Hilary Greaves, William MacAskill,
Rossa O'Keeffe-O'Donovan and Philip Trammell

February 2019

We acknowledge Pablo Stafforini, Aron Vallinder, James Aung, the Global Priorities Institute Advisory Board, and numerous colleagues at the Future of Humanity Institute, the Centre for Effective Altruism, and elsewhere for their invaluable assistance in composing this agenda.

Table of Contents

Introduction

GPI's vision and mission

There are many problems in the world. Because resources are scarce, it is impossible to solve them all. An actor seeking to improve the world as much as possible therefore needs to prioritise, both among the problems themselves and, relatedly, among means for tackling them.

This task of prioritisation requires careful analysis. Some opportunities to do good are vastly more cost-effective than others. But identifying which are the better opportunities requires grappling with a host of complex questions - questions about how to evaluate different outcomes, how to predict the effects of actions, how to act in the face of uncertainty, how to identify more practically usable proxies for the criteria we ultimately care about, and many other topics.

In practice, at present, only a relative minority of actors (whether individual or institutional) make their decisions explicitly and significantly based on consideration of the question of 'which option would do the most impartial good?', even when the actions in question are nominally altruistically motivated. There are many reasons for this. Some, of course, concern constraints imposed by politics, or other limits of motivation. But in significant part the stumbling block is that *we simply do not have enough information or understanding about what it would look like* to determine priorities and actions on the basis of a scientific assessment of the amount of good (all things considered, in the long run, and in impartial terms) that the candidate options can reasonably be expected to do. In this situation, it is natural for decision-makers to use other, sometimes quite unrelated criteria for the purpose of practical decision-making.

A significant exception to this general tendency is found in the effective altruism movement. Over the past ten years or so, this growing movement has devoted a rapidly increasing flow of resources, both intellectual and financial, to the enterprise of doing good as effectively as possible. For example, the Open Philanthropy Project has made more than 500 philanthropic grants with a total worth of more than $500m since 2012, and 80,000 Hours has tracked thousands of people who have made significant changes to their career plans based on its research and recommendations. The movement has developed numerous novel and exciting ideas, and has been audacious in pushing forward the implementation of those ideas. However, due to a lack of suitably rigourous foundational research, many of the ideas in question are not yet mainstream in academic circles.

The Global Priorities Institute exists to develop and promote rigourous, scientific approaches to the question of how appropriately motivated actors can do good more effectively. Our core belief is that the existence of a wide base of high-quality research on these questions, and (relatedly) an increased focus on those questions within academia, is a prerequisite for the widespread adoption of an effectiveness-based approach to global prioritisation.

This line of thought motivates the following vision and mission:

GPI's Vision

A world in which global priorities are set by using evidence and reason to determine what will do the most good.

GPI's Mission

To conduct and promote world-class, foundational academic research on how most effectively to do good.

GPI's research agenda

The central focus of GPI is what we call 'global priorities research': research into issues that arise in response to the question, 'What should we do with a given amount of limited resources if our aim is to do the most good?' This question naturally draws upon central themes in the fields of economics and philosophy.

Thus defined, global priorities research is in principle a broad umbrella. Within that umbrella, this research agenda sets out the more specific research themes that GPI is particularly interested in at the present time.

The document is structured as follows.

Section 1 outlines what we call the longtermism paradigm. This paradigm centres around the idea that because of the potential vastness of the future portion of the history of sentient life, it may well be that the primary determinant of which actions are best is the effects of those actions on the very long-run future, rather than on more immediate considerations. Because these ideas seem plausible, seem likely to have fairly radically revisionary implications if correct, and are currently quite neglected, **this is the main focus of GPI's own research** (at the time of writing and, we predict, for at least the next two years). **We are particularly keen to hear from other researchers who share this interest.**

Section 2 concerns general issues in cause prioritisation. This covers issues that are not specific to a longtermist point of view, but that arise for agents engaged in an exercise of global prioritisation.

Appendix A indicates additional areas of possible research that would further GPI's mission, but that GPI itself is not working on now or for the immediately foreseeable future, for reasons of capacity and focus. Appendix B indicates areas of existing academic literature that serve as particularly relevant background for the topics on this research agenda. Appendix C contains links to additional informal discussion of the themes discussed in this research agenda.

The intended audience for this document is academics (especially, but not only, in economics and philosophy) who are potentially interested in working with GPI, whether as GPI researchers or as external collaborators, or who are otherwise interested in the same mission.

1. The longtermism paradigm

As noted above, an actor seeking to improve the world as much as possible with limited resources needs to prioritise: which problems should she focus on and which steps should she take to address those problems, to the exclusion of others?

Key to GPI's approach to this question is what we call *the longtermism paradigm*. This paradigm has two key components. First, insofar as consequences matter to the value of actions, *all* the consequences of one's actions matter, and not only those that are in any specified sense 'direct'. Second, all consequences (of a given type) matter *equally*: a given harm or benefit, say, matters to the same extent regardless of where or when in space and time it occurs.

This paradigm has potentially radical implications. Given how long sentient life could potentially survive for, it suggests that the primary determinant of the value-differences among the best actions we could take today could well be the effects of those actions on the very long-term future, rather than on any effects within (say) our own lifetimes. In stark contrast, mainstream economics and policy research typically takes the perspective that improving the course of the far future is not tractable enough to tackle directly. Instead, it is generally believed that the best way to impact the future is to promote some programme of economic development or growth.

This contrast warrants much more research to work out the articulation, evaluation, implications and implementation of longtermist ideas in global prioritisation.

1.1 Articulation and evaluation of longtermism

Let us define *longtermism* as the view that the primary determinant of the differences in value of the actions we take today is the effect of those actions on the very long-term future. This view is supported by plausible arguments, and has widespread significance if correct. This warrants much more work to articulate, evaluate and work out the implications of a longtermist view.

Potential research projects:

- It is natural to think that in evaluating interventions, we should in principle take into account *all* welfare-relevant effects of those interventions, not only those that are in some specified sense 'intended' or 'direct'. For example, in the evaluation of a school-based deworming programme, we should not only count the direct effects of the treatment on the health or schooling of treated children, but also indirect effects, including side-effects of the intervention (for example, the effects of the distribution of medicine on local politics) and knock-on effects that are causally downstream of the immediately intended effect (such as later-life outcomes for the treated children, spillover effects on non-treated children, and impacts on population size, economic

growth, and government activity). The argument that we should value these effects, however, seems somewhat in tension with the common view in medical ethics 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 (Kamm 1993; Brock 2003; Lippert-Rasmussen and Lauridsen 2010; Du Toit and Millum 2016). How is this tension best resolved? (Mogensen MS) (INFORMAL: Greaves 2015)

PHIL - MEDICAL ETHICS

- There is already a substantial literature (on both sides) evaluating the claim that one should adopt a zero rate of pure time preference in public policy evaluation (Greaves 2017). However, given the importance of this claim to the longtermism paradigm, research that changes the balance of arguments on this question could still be high value. What more, if anything, can be said on the matter?

  PHIL - ETHICS OF DISCOUNTING **ECON** - DISCOUNTING

- Assuming both that indirect effects should be counted and that future welfare should not be discounted, provide a rigourous articulation of the case for thinking that the primary determinant of value-differences between the best actions available to us today is the expected effects of those actions on the very far future (Bostrom 2003) (INFORMAL: Karnofsky 2014; Todd 2017). How sensitive is this argument to variations in other evaluative assumptions over which there is reasonable disagreement? (Beckstead 2013; Beckstead forthcoming) (INFORMAL: Ord 2017; Sittler 2018)

  PHIL - ETHICS OF DISCOUNTING **ECON** - DISCOUNTING

- To what extent do considerations of saturation (for example, the possibility that utility as a function of consumption is bounded) constrain the possibilities for leveraging the vastness of the future to identify actions with extremely high value?

  PHIL - ETHICS OF DISCOUNTING **ECON** - DISCOUNTING, WELFARE ECONOMICS

- Should altruists in general be moved primarily by explicit considerations of long-run impact, or are such efforts intractable? (INFORMAL: Tomasik 2013; Bostrom 2014)

  PHIL - POPULATION ETHICS **ECON** - DISCOUNTING, TIME-SERIES ECONOMETRICS, MACROECONOMIC THEORY

Existing academic literature:

- Beckstead, Nicholas. 'A Brief Argument for the Overwhelming Importance of Shaping the Far Future'. In *Effective Altruism: Philosophical Issues*, edited by Theron Pummer and Hilary Greaves. Oxford: Oxford University Press, forthcoming.
- ———. 'On the Overwhelming Importance of Shaping the Far Future'. PhD dissertation. New Brunswick: Rutgers University, 2013.
- Bostrom, Nick. 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development'. *Utilitas* 15, no. 3 (2003): 308–14.
- Brock, Dan W. 'Separate Spheres and Indirect Benefits'. *Cost Effectiveness and Resource Allocation* 1, no. 4 (2003).

- Greaves, Hilary. 'Discounting for Public Policy: A Survey'. *Economics & Philosophy* 33, no. 03 (2017): 391–439.
- ———. 'Discounting Future Health'. In *Global Health Priority-Setting: Cost-Effectiveness and Beyond*, edited by Ruger and Verguet Otterson Millum Johansson Jamison Emanuel Norheim. Oxford: Oxford University Press, forthcoming.
- Kamm, Frances M. *Morality, Mortality; Volume I: Death and Whom to Save From It*. Oxford: Oxford University Press, 1998.
- Lippert-Rasmussen, Kasper, and Sigurd Lauridsen. 'Justice and the Allocation of Healthcare Resources: Should Indirect, Non-Health Effects Count?' *Medicine, Health Care and Philosophy* 13, no. 3 (2010): 237–46.
- Mogensen, Andreas L. 'Meaning, medicine and merit'. Manuscript in preparation.
- Toit, Jessica du, and Franklin Miller. 'The Ethics of Continued Life-Sustaining Treatment for Those Diagnosed as Brain-Dead'. *Bioethics* 30, no. 3 (2016): 151–58.

Existing informal discussion:

- Nick Bostrom, Crucial considerations and wise philanthropy, 9 July 2014
- Hilary Greaves, Repugnant interventions, 15 August 2015
- Holden Karnofsky, The Moral Value of the Far Future, 3 July 2014
- Toby Ord, Why the long-term future of humanity matters more than anything else, and what we should do about it, 6 September 2017
- Thomas Sittler, The expected value of the long-term future, 2018
- Benjamin Todd, Presenting the long-term value thesis, 24 October 2017 - a
- Benjamin Todd, Why despite global progress, humanity is probably facing its most dangerous time ever, October 2017 - b
- Brian Tomasik, Charity cost-effectiveness in an uncertain world, 28 October 2013

1.2 Sign of the value of the continued existence of humanity

Longtermism is often thought to lead to the conclusion that we ought to prioritise extinction risk reduction. This presupposes that the expected value of continued human existence is positive. But one can at least imagine some scenarios, and at least some value systems, in which we should expect humanity's future to contain more bad than good. Before engaging in more fine-grained cause prioritisation across efforts to improve or extend the expected course of human civilisation, therefore, it is important to consider the sign of its expected value.

Potential research projects:

- Assess the expected value of the continued existence of the human race. Might this expected value be negative, or just unclear (INFORMAL: Christiano 2013; West 2017)? How do our answers to these questions vary if we (i) assume utilitarianism (INFORMAL: Shulman 2012; Dickens 2015); (ii) assume a non-utilitarian axiology (INFORMAL: Greaves, 2016; Brauner and Grosse-Holz 2018); (iii) fully take axiological uncertainty into account (Greaves and Ord 2017; MacAskill MS-b)?

  PHIL - MORAL UNCERTAINTY, POPULATION ETHICS **ECON** - MEASUREMENT, MODEL UNCERTAINTY

- Assuming that there is a single, context-independent welfare level corresponding to a life's having zero contributive value to social welfare (Broome 2004: ch. 10), what kinds of lives have zero welfare in this contributive sense?

  PHIL - POPULATION ETHICS **ECON** - WELFARE ECONOMICS

- 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 (MacAskill MS-a)? 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?

  PHIL - MORAL UNCERTAINTY **ECON** - VALUE OF INFORMATION, INTERGENERATIONAL GOVERNANCE

- 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, because avoidance of the bad outcomes is more neglected, or because bad outcomes should be weighted more than good outcomes when other relevant things are equal) (Hurka 2010) (INFORMAL: Althaus and Gloor 2018; Tomasik 2018)? If so, what activities would be best? (MacAskill MS-a) (INFORMAL: Gloor 2018)

  PHIL - MORAL UNCERTAINTY, DECISION THEORY **ECON** - CATASTROPHIC RISK

Existing academic literature:

- Broome, John. *Weighing Lives*. Oxford; New York: Oxford University Press, 2004.
- Greaves, Hilary, and Toby Ord. 'Moral Uncertainty About Population Axiology'. *Journal of Ethics & Social Philosophy* 12, no. 2 (2017): 135–67.
- Hurka, Thomas. 'Asymmetries In Value'. *Noûs* 44, no. 2 (2010): 199–223.
- MacAskill, William. 'Human Extinction, Asymmetry, and Option Value'. Manuscript in preparation - a.
- ———. 'Practical Ethics Given Moral Uncertainty'. Manuscript in preparation - b.
- MacAskill, William, Krister Bykvist and Toby Ord. *Moral Uncertainty*. Oxford: Oxford University Press, forthcoming.
- Greaves, Hilary. 'Optimum Population Size'. In *Oxford Handbook of Population Ethics*, edited by Gustaf Arrhenius, Krister Bykvist and Tim Campbell. Oxford: Oxford University Press, forthcoming.

Existing informal discussion:

- David Althaus and Lukas Gloor, Reducing Risks of Astronomical Suffering: a Neglected Priority, 2018
- Jan Brauner and Friederike Grosse-Holz, The expected value of extinction risk reduction is positive, 2018
- Paul Christiano, Why might the future be good?, 2013
- Michael Dickens, Is Preventing Human Extinction Good?, 2015
- Lukas Gloor, Cause prioritization for downside-focused value systems, 2018
- Hilary Greaves, Extinction risk and population ethics, 2016
- Carl Shulman, Spreading happiness to the stars seems little harder than just spreading, 2012
- Carl Shulman, Are pain and pleasure equally energy-efficient?, 2012
- Brian Tomasik, Risks of Astronomical Future Suffering, 2018
- Ben West, An Argument for Why the Future May Be Good, 2017

1.3 Mitigating catastrophic risk

It is often assumed that human civilisation is on a likely course to produce vast amounts of value over the course of the future. If this is correct, then it may be uniquely important from a longtermist perspective to minimise the risk of catastrophes, such as near-term human extinction, that could derail this course. The precise implications of this argument, however, warrant further scrutiny.

Potential research projects:

- Is there a fruitful notion of 'existential' risk (Bostrom 2002, Ord forthcoming) that is broader than the notion of extinction risk? What is the most fruitful such generalisation (Cotton-Barratt and Ord 2015)?

  ECON - CATASTROPHIC RISK

- Does longtermism lead to the conclusion that reducing existential risk should be the highest priority (Bostrom 2013)? Does it further lead to the stronger conclusion that reducing *extinction* risk should be the highest priority (Matheny 2007) (INFORMAL: Todd 2017)? Alternatively, should we focus on macroeconomic 'trajectory changes' (that is, smaller but very persistent/long-lasting improvements to total value achieved at every time), or other ways of increasing the expected value of the far future conditional on the survival of humanity, instead of on reducing particular large risks? (Beckstead 2013; Ng 2016; Méjean et al. 2017)

  ECON - GROWTH, CATASTROPHIC RISK, MACROECONOMIC THEORY

- What do the most plausible person-affecting views in population ethics say about the value of reducing extinction risk? (INFORMAL: Greaves 2016)

  PHIL - POPULATION ETHICS

- Mitigation of catastrophic risk is sometimes a matter of an extraordinarily small chance of generating extraordinarily high value. Is expected utility theory the correct approach for dealing with decisions of this character (Bostrom 2009; Tarsney 2018) (INFORMAL: Yudkowsky 2013)? Does any plausible alternative lead away from the idea that the opportunities in question are among the best from an ex ante evaluative standpoint (INFORMAL: Karnofsky 2011)?

  PHIL - DECISION THEORY **ECON** - DECISION THEORY

- A catastrophic risk can be called 'existential' to the extent that it threatens a large, permanent negative shock to the subsequent growth path. An even more precise characterisation of this property may be valuable. How can we best model the magnitude of the permanent costs associated with a given risk? (Ord forthcoming)

  ECON - CATASTROPHIC RISK, TIME-SERIES ECONOMETRICS, MACROECONOMIC THEORY

- How, concretely, should we adapt (endogenous) growth models to weigh the benefits that growth may pose for the long term against the catastrophic risks that may come with technological development? (Jones 2016)

  ECON - GROWTH, CATASTROPHIC RISK

- To date, most of the work in economics concerning long-term catastrophic risk mitigation has focused on climate change. To what extent does climate change pose a genuinely existential threat (Méjean et al. 2017, Ord forthcoming) (INFORMAL: Halstead, 2018)? How do the risks of climate change and the benefits from mitigating them compare with more neglected risks (Martin and Pindyck 2015, 2017; Ord forthcoming) (INFORMAL: Duda 2016)?

  ECON - CATASTROPHIC RISK, ENVIRONMENTAL ECONOMICS

Existing academic literature:

- Beckstead, Nicholas. 'On the Overwhelming Importance of Shaping the Far Future'. PhD dissertation. New Brunswick: Rutgers University, 2013.
- Bostrom, Nick. 'Existential Risk Prevention as Global Priority'. *Global Policy* 4, no. 1 (2013): 15–31.
- ———. 'Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards'. *Journal of Evolution and Technology* 9 (2002).
- ———. 'Pascal's Mugging'. *Analysis* 69, no. 3 (2009): 443–45.
- Cotton-Barratt, Owen, and Toby Ord. 'Existential Risk and Existential Hope: Definitions'. *Future of Humanity Institute*, Technical Report #2015-1.
- Méjean, Aurélie, Antonin Pottier, Stéphane Zuber and Marc Fleurbaey. 'Intergenerational Equity under Catastrophic Climate Change'. Working paper, 2017.
- Jones, Charles I. 'Life and Growth'. *Journal of Political Economy* 124, no. 2 (2016): 539–78.

- Martin, Ian W. R., and Robert S. Pindyck. 'Averting Catastrophes: The Strange Economics of Scylla and Charybdis'. *American Economic Review* 105, no. 10 (2015): 2947–85.
- ———. 'Averting Catastrophes that Kill'. Working paper, 2017.
- Matheny, Jason G. 'Reducing the Risk of Human Extinction'. *Risk Analysis* 27, no. 5 (2007): 1335–44.
- Millner, Antony. 'On Welfare Frameworks and Catastrophic Climate Risks'. *Journal of Environmental Economics and Management* 65, no. 2 (2013): 310–25.
- Ng, Yew-Kwang. 'The Importance of Global Extinction in Climate Change Policy'. *Global Policy* 7, no. 3 (2016): 315–22.
- Ord, Toby. *Existential Risk*. London: Bloomsbury, forthcoming.
- Weitzman, Martin L. 'On Modeling and Interpreting the Economics of Catastrophic Climate Change'. *Review of Economics and Statistics* 91, no. 1 (2009): 1–19.

Existing informal discussion:

- Roman Duda, Climate change (extreme risks), 2016
- Hilary Greaves, Extinction risk and population ethics, 2016
- John Halstead, Is climate change an existential risk?, 2018
- Holden Karnofsky, Why We Can't Take Expected Value Estimates Literally (Even When They're Unbiased), 2011
- Toby Ord, The timing of labour aimed at reducing existential risk, 2014
- Benjamin Todd, Why despite global progress, humanity is probably facing its most dangerous time ever, October 2017
- Eliezer Yudkowsky, Pascal's Muggle: Infinitesimal Priors and Strong Evidence, 2013

1.4 Other ways of leveraging the size of the future

The 'size' of the future may present us with other ways, beyond mitigating catastrophic risks, of producing vast amounts of value. In particular, we may be able to produce lasting technological or civilisational 'trajectory changes' whose expected long-term value exceeds that of existential risk mitigation. This warrants putting thought into identifying promising trajectory-change opportunities, and developing a framework for prioritising among them.

Potential research projects:

- Besides mitigation of catastrophic risk, what other kinds of 'trajectory change' or other interventions might offer opportunities with very high expected value, as a result of the potential vastness of the future? Can we construct a useful taxonomy for

thinking about these? (Bostrom 2005) (INFORMAL: Duda 2017; Duda 2018; Beckstead 2014; Whittlestone 2017; Baum et al. 2019)

- Technological developments in the recent past, such as stem cell research, have opened possibilities about whose moral value there is wide disagreement. It seems plausible that technological developments over the coming century, such as machine intelligence (Bostrom 2016), brain emulation (Sandberg and Bostrom 2008; Hanson 2016) or atomically precise manufacturing (Drexler 1987) will create many more such morally contentious opportunities, at much higher stakes—our responses to them even contributing substantially, perhaps, to the moral value of the future. What high-stakes moral conflicts are most likely to arise with emerging technologies, and how should the global community resolve them? (Bostrom 2005)

  ECON - TECHNOLOGICAL DEVELOPMENT, POLITICAL ECONOMY, BARGAINING THEORY, MECHANISM DESIGN

- For what kinds of philanthropic interventions do we expect effects to 'wash out' over very long timescales rather than to persist? Are their long-run effects typically of much greater expected value (whether positive or negative) than their short-run effects, taking into account both the vastness of the future and the generally greater uncertainty of effects that are more causally remote? (Beckstead 2013) (INFORMAL: Beckstead 2013)

  PHIL - ETHICS OF CHAOS **ECON** - TIME SERIES ECONOMETRICS, MACROECONOMIC THEORY

- Let *finitism* be the claim that, even if we ought perhaps to aim to bring about an astronomically large finite amount of value in the future, we ought not to aim explicitly to bring about an infinitely large amount of value. Is finitism defensible? If it is not defensible, 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, how do we compare outcomes involving possibilities of infinite quantities of value, in order to decide which such outcomes to pursue? (Vallentyne and Kagan 1997; Basu and Mitra 2003; Vallentyne and Lauwers 2004; Zame 2007; Asheim 2010; Bostrom 2011; Arntzenius 2014) (INFORMAL: West 2015)

  PHIL - DECISION THEORY, INFINITE ETHICS **ECON** - INTERGENERATIONAL EQUITY

Existing academic literature:

- Asheim, Geir B. 'Intergenerational Equity'. *Annual Review of Economics* 2, no. 1 (2010): 197–222.
- Arntzenius, Frank. 'Utilitarianism, Decision Theory, and Eternity'. *Philosophical Perspectives* 28, no. 1 (2014): 31–58.
- Basu, Kaushik, and Tapan Mitra. 'Aggregating Infinite Utility Streams with Intergenerational Equity: The Impossibility of Being Paretian'. *Econometrica* 71, no. 5 (2003): 1557–63.
- Baum, S. D. et al. 'Long-Term Trajectories of Human Civilization'. *Foresight*, forthcoming.

- Beckstead, Nicholas. 'On the Overwhelming Importance of Shaping the Far Future'. PhD dissertation. New Brunswick: Rutgers University, 2013.
- Bostrom, Nick. 'Infinite Ethics'. *Analysis and Metaphysics*, no. 10 (2011): 9–59.
- ———. *Superintelligence: Paths, Dangers, Strategies*. First edition. Oxford: Oxford University Press, 2014.
- ———. 'Technological Revolutions: Ethics and Policy in the Dark'. In *Nanoscale*, edited by Nigel M. de S. Cameron and M. Ellen Mitchell. Hoboken, NJ: John Wiley & Sons, 2007. 129–52
- ———. 'The Future of Human Evolution'. In *Death and Anti-Death*, edited by Charles Tandy. Ann Arbor: Ria University Press, 2005. 339–71.
- Drexler, K. Eric. *Engines of Creation: The Coming Era of Nanotechnology*. New York: Random House, 1987.
- Hanson, Robin. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. First Edition. Oxford: Oxford University Press, 2016.
- Lauwers, Luc, and Peter Vallentyne. 'Infinite Utilitarianism: More Is Always Better'. *Economics & Philosophy* 20, no. 2 (2004): 307–30.
- Sandberg, Anders, and Nick Bostrom. 'Whole Brain Emulation: A Roadmap'. *Future of Humanity Institute*, Technical Report #2008-3.
- Vallentyne, Peter, and Shelly Kagan. 'Infinite Value and Finitely Additive Value Theory'. *The Journal of Philosophy* 94, no. 1 (1997): 5–26.
- Zame, William R. 'Can intergenerational equity be operationalized?' *Theoretical Economics* 2, no. 2 (2007): 187–202.

Existing informal discussion:

- Nick Beckstead, A Proposed Adjustment to the Astronomical Waste Argument, 2013
- Nick Beckstead, A relatively atheoretical perspective on astronomical waste, 2014
- Paul Christiano, Against moral advocacy, 2013
- Roman Duda, Building effective altruism, 2017
- Roman Duda, Global priorities research, 2018
- Carl Shulman, Spreading happiness to the stars seems little harder than just spreading, 2012
- Ben West, Problems and Solutions in Infinite Ethics, 2015
- Jess Whittlestone, Improving institutional decision-making, 2017

1.5 Intergenerational governance

Many of the long-term plans made by present philanthropists and policymakers are vulnerable to being altered or undone by future generations. On the other hand, the value of future human civilisation will likely be determined primarily by future policy decisions. In evaluating the long-term consequences of our actions, therefore, we must reckon carefully with questions of how to influence the behaviour of future policymakers, and how to 'coordinate' optimally in the face of constraints on that influence.

Potential research projects:

- Economic research into the role of institutions is one field that directly attempts to influence the long term. Certain institutions, such as 'inclusive' governments, appear to be associated both with substantial increases in economic growth, across many generations, and with decreases in the probability of events (such as wars) that may be associated with catastrophic risk (Acemoglu et al. 2005). How can we estimate the effectiveness of various institution-building efforts on the long term?

  ECON - GROWTH, CATASTROPHIC RISK, INSTITUTIONAL ECONOMICS, ECONOMIC HISTORY

- When faced with an important, irreversible decision with respect to which one will soon learn relevant information, it is rational to preserve 'option value'; to delay the decision until after the information has been acquired (Bishop 1982; Dixit and Pindyck 1994). In delaying an important social decision intergenerationally, however, we may worry that future generations' values and preferences will differ from our own (INFORMAL: Hanson 2018). Facing this tradeoff, under what circumstances should we 'principals' defer irreversible decisions to better-informed future 'agents'? (MacAskill, MS) (INFORMAL: Brauner and Grosse-Holz 2018)

  ECON - INTERGENERATIONAL GOVERNANCE, MECHANISM DESIGN, VALUE OF INFORMATION

- If we could ensure that our descendants would carry out our plans from their improved informational position, deferring irreversible decisions to them would offer us the best of both worlds. Can long-term inter-generational mechanisms be designed so as to enable this possibility? What might they look like? (Bostrom 2006) (INFORMAL: Tomasik 2018)

  ECON - INTERGENERATIONAL GOVERNANCE, MECHANISM DESIGN, VALUE OF INFORMATION

- The idea of the *long reflection* is that of a long period—perhaps tens of thousands of years—during which human civilisation, perhaps with the aid of improved cognitive ability, dedicates itself to working out what is ultimately of value (INFORMAL: MacAskill 2018; Lewis 2018). It may be argued that such a period would be warranted before deciding whether to undertake an irreversible decision of immense importance, such as whether to attempt spreading to the stars. Do we find ourselves, or are we likely to find ourselves, in a situation where a 'long reflection' would in fact be warranted? If so, how should it be implemented?

  PHIL - MORAL UNCERTAINTY **ECON** - INTERGENERATIONAL GOVERNANCE, MECHANISM DESIGN, VALUE OF INFORMATION

- Do 'broad' approaches to improving the far future (such as promoting good institutions or global peace) tend to be more or less effective, in expectation, than 'narrow' approaches (such as working on reducing the risk of bioengineered pandemics)? (INFORMAL: Beckstead 2013)

Existing academic literature:

- Acemoglu, Daron, Simon Johnson and James A. Robinson. 'Institutions as a Fundamental Cause of Long-Run Growth'. *Handbook of Economic Growth*, 1A (2005): 385–472.
- Bishop, Richard C. 'Option Value: An Exposition and Extension'. *Land Economics* 58, no. 1 (1982): 1–15.
- Bostrom, Nick. 'What is a Singleton?' *Linguistic and Philosophical Investigations* 5, no. 2 (2006): 48–54.
- Cotton-Barratt, Owen. 'Allocating Risk Mitigation across Time'. *Future of Humanity Institute*, Technical Report #2015-2.
- Dixit, Avinash and Robert S. Pindyck. *Investment Under Uncertainty*. Princeton: Princeton University Press, 1994.
- Kimball, Miles S. 'Making Sense of Two-Sided Altruism'. *Journal of Monetary Economics* 20, no. 2 (1987): 301–26.
- MacAskill, William. 'Human Extinction, Asymmetry, and Option Value'. Manuscript in preparation.

Existing informal discussion:

- Nick Beckstead, How to compare broad and targeted attempts to shape the far future, 2013.
- Jan Brauner and Friederike Grosse-Holz, The expected value of extinction risk reduction is positive, 2018
- Robin Hanson, On value drift, 2018
- Greg Lewis, The not-so-Long Reflection?, 2018
- William MacAskill, Our descendants will probably see us as moral monsters. What should we do about that?, 19 January 2018
- Toby Ord, The timing of labour aimed at reducing existential risk, 2014
- Brian Tomasik, Will Future Civilization Eventually Achieve Goal Preservation?, 2018

1.6 Economic indices for longtermists

It is standard practice in economics to evaluate policies, explicitly or implicitly, on the basis of their expected short-term impact on total economic output. It is also standard, though less

common, to evaluate policies on the basis of their expected short-term impact on economic indices designed to correspond more closely with human welfare, such as the human development index (HDI). From a longtermist perspective, however, the true measure of a policy's success is its impact on the long-term prospects of human civilisation. We must therefore ask how well the former indices track the latter objective, and, perhaps, how to construct and implement economic indices that track the latter objective more closely.

Potential research projects:

- Much government policy, economic research, and philanthropic activity is intended ultimately to increase the general rate of economic growth. Economic growth could be extremely beneficial, from a long-term perspective, as it promises to improve the entire course of the future. However technology-driven growth may raise existential risks, due for example to nuclear accidents, engineered pandemics or artificial superintelligence (INFORMAL: Yudkowsky 2013), and growth in general may have other negative effects (for instance, risks to human life (Jones 2016), climate change (IPCC 2014), or meat consumption (INFORMAL: Bogosian 2015)). How radically do these drawbacks render growth an imperfect proxy for expected long-term wellbeing? Is the correlation between consumption growth and long-term wellbeing even positive, given the current drivers of growth, from a geographical, sectoral and technological perspective? (Friedman 2006; Cowen 2007; Tomasik 2013; Cowen 2018) (INFORMAL: Beckstead 2014)

  ECON - GROWTH, MACROECONOMIC MEASUREMENT

- Of the comprehensive macroeconomic indices already available to us, which serve best as proxies for long-term expected global welfare (including but not limited to considerations of existential risks)? What would be the broad policy implications of targeting such indices instead of GDP per capita?

  ECON - GROWTH, MACROECONOMIC MEASUREMENT

- Are there any promising proxies for long-term wellbeing *not* already tracked as macroeconomic indices (INFORMAL: Shulman 2013; Bostrom 2014)? If so, how could these proxies be formalised and measured, and what would be the broad policy implications of targeting them instead of GDP per capita?

  ECON - GROWTH, MACROECONOMIC MEASUREMENT

Existing academic literature:

- Cowen, Tyler. 'Caring About the Distant Future: Why It Matters and What It Means'. *The University Of Chicago Law Review* 74, no. 5 (2007): 5–40.
- Cowen, Tyler. *Stubborn Attachments*. San Francisco: Stripe Press, 2018.
- Friedman, Benjamin M. *The Moral Consequences of Economic Growth*. New York: Vintage Books, 2006.
- IPCC. *Climate change 2014: Impacts, adaptation, and vulnerability. Part A: Global and sectoral aspects. Contribution of Working Group II to the Fifth Assessment Report of the*

Intergovernmental Panel on Climate Change. Cambridge, UK and New York, NY, USA: Cambridge University Press, 2014.

- Jones, Charles I. 'Life and Growth'. *Journal of Political Economy* 124, no. 2 (2016): 539–78.

Existing informal discussion:

- Nick Beckstead, How much can a long-run perspective help with strategic cause selection?, 9 July 2014
- Nick Bostrom, Crucial considerations and wise philanthropy, 9 July 2014
- Kyle Bogosian, Quantifying the Impact of Economic Growth on Meat Consumption, 2015
- Carl Shulman, What proxies to use for flow-through effects?, 2013
- Brian Tomasik, Differential Intellectual Progress as a Positive-Sum Project ('Economic growth' subsection), 2013
- Eliezer Yudkowsky, Do Earths with slower economic growth have a better chance at FAI?, 2013

1.7 Moral uncertainty for longtermists

Estimates of the value of an intervention are sensitive not only to uncertainty about the intervention's empirical consequences, but also to uncertainty about the normative criteria by which to evaluate consequences in general. This 'moral uncertainty' may prove particularly important from a longtermist perspective, as we may often have to choose between interventions whose short-term consequences are of similar value across all plausible moral theories, but whose long-term consequences differ substantially across plausible moral theories.

Potential research projects:

- Are there convergent instrumental goals that many different axiologies would agree on? Given axiological uncertainty, can we make any claims about what sort of future we should try to aim for? (Greaves and Ord 2017; MacAskill MS-b) (INFORMAL: Pummer 2015; Leech 2018)

  PHIL - MORAL UNCERTAINTY

- Under moral uncertainty, do some axiological views with very high stakes swamp the expected value calculation? If so, which views are they? What is the best way to deal with this 'fanaticism' issue? (Ross 2006; MacAskill and Ord 2018; Cotton-Barratt and Greaves MS) (INFORMAL: MacAskill 2018-a)

  PHIL - MORAL UNCERTAINTY, DECISION THEORY **ECON** - MODEL UNCERTAINTY, DECISION THEORY

- How likely is it that civilisation will converge on the correct moral theory given enough time? What implications does this have for cause prioritisation in the nearer term? (INFORMAL: MacAskill 2018-b)

  PHIL - MORAL UNCERTAINTY

- How likely is it that the correct moral theory is a 'Theory X', a theory radically different from any yet proposed? If likely, how likely is it that civilisation will discover it, and converge on it, given enough time? While it remains unknown, how can we properly hedge against the associated moral risk? (INFORMAL: MacAskill 2018-b)

  PHIL - MORAL UNCERTAINTY ECON - MODEL UNCERTAINTY, HEDGING

Existing academic literature:

- Greaves, Hilary, and Toby Ord. 'Moral Uncertainty About Population Axiology'. *Journal of Ethics & Social Philosophy* 12, no. 2 (2017): 135–67.
- Greaves, Hilary, and Owen Cotton-Barratt. 'A Bargaining-Theoretic Approach to Moral Uncertainty'. Manuscript in preparation.
- MacAskill, William. 'Practical Ethics Given Moral Uncertainty'. Manuscript in preparation.
- MacAskill, William, and Toby Ord. 'Why Maximize Expected Choice-Worthiness?' *Noûs*, 14 July 2018.
- Ross, Jacob. 'Rejecting Ethical Deflationism'. *Ethics* 116, no. 4 (2006): 742–68.

Existing informal discussion:

- John Halstead, Moral uncertainty and climate change, May 2017
- Holden Karnofsky, Update on Cause Prioritization at Open Philanthropy, 2018
- Gavin Leech, Existential risk as common cause, 2018
- William MacAskill, Our descendants will probably see us as moral monsters. What should we do about that?, 19 January 2018 - a
- William MacAskill, Should we expect moral convergence?, 2018 - b
- Theron Pummer, Moral Agreement on Saving the World, 2015

1.8 Longtermist status of interventions that score highly on short-term metrics

It is sometimes argued that interventions designed to score highly on short-term metrics—such as cost-effective poverty alleviation programmes—are also typically the actions with the best expected long-term consequences. If that is correct, then longtermism (even if true) has little practical significance. It is therefore important to evaluate this argument.

Potential research projects:

- Is there any motivation for prioritising interventions that score highly on short-term metrics that is respectable from a longtermist perspective? (INFORMAL: Karnofsky 2014; Tomasik 2015)
- To what extent should a worry of 'suspicious convergence' (INFORMAL: Lewis 2016) incline us against the hypothesis that the interventions that have the best short-termist motivation also fare well by longtermist lights?
- What are the long-term effects of interventions that seem particularly high-priority from a short-term perspective, such as saving human lives (INFORMAL: Karnofsky 2013) or improving the conditions of caged hens (Matheny and Chan 2005) (INFORMAL: Shulman 2013)? What is the sign of these effects, and how substantial are they? Under what conditions, if any, might they exceed the expected long-term impacts of (other) efforts aimed explicitly at improving the long term?

ECON - TIME SERIES ECONOMETRICS, STRUCTURAL MODELLING, FORECASTING

Existing academic literature:

- Matheny, Gaverick and Kai Chan. 'Human Diets and Animal Welfare: The Illogic of the Larder'. *Journal of Agricultural and Environmental Ethics* 18, no. 6 (2005): 579–94.

Existing informal discussion:

- Holden Karnofsky, Flow-through effects, 2013
- Holden Karnofsky, The Moral Value of the Far Future, 2014
- Greg Lewis, Beware suspicious and surprising convergence, 2016
- Carl Shulman, Vegan advocacy and pessimism about wild animal welfare, 2013
- Brian Tomasik, Should Altruists Focus on Reducing Short-Term or Far-Future Suffering?, 2015

2. General issues in global prioritisation

Even after developing a rigourous framework for prioritising causes from a longtermist perspective, there are still many open theoretical questions that one faces if one wishes to do the most good. Some of these questions, for example, concern how best to aggregate weak or heterogeneous evidence. Others concern timing: under what conditions we should try to do good right away, and under what conditions we should invest in order to do more good later. Still others concern how groups of altruistic individuals can coordinate to maximise their collective impact. The following areas of research strike us as particularly interesting and important.

2.1 Decision-theoretic issues

The framework of expected utility theory sometimes produces deeply counterintuitive conclusions, especially when we are faced with the prospect of extremely low-probability, high-magnitude payoffs. When faced with the possibility of infinite payoffs, the expected utility framework breaks down altogether. These and other decision-theoretic problems are of particular interest to individuals or organisations trying to do good, whose concerns may extend beyond the relatively local scope for which standard decision theory has been developed, and warrant the development of nonstandard decision-theoretic solutions.

Possible research projects:

- Faced with the task of comparing actions in terms of expected value, it often seems that the agent is 'clueless': that is, that the available empirical and theoretical evidence simply supplies too thin a basis for guiding decisions in any principled way (Lenman 2000; Greaves 2016) (INFORMAL: Tomasik 2013; Askell 2018). How is this situation best modelled, and what is the rational way of making decisions when in this predicament? Does cluelessness systematically favour some types of action over others?

  PHIL - EPISTEMOLOGY, DECISION THEORY **ECON** - DECISION THEORY, BEHAVIOURAL ECONOMICS

- One common view is that we should favour interventions that have more evidential support, all else being equal. On the face of it, this conflicts with the maximisation of expected value if one would prefer an intervention with much stronger evidence but a (possibly infinitesimally) small reduction in expected value (if 'all else being equal' means: 'expected value being equal'). On the other hand, it also seems reasonable to place some value on the uncertainty of an intervention. What is the correct response to this mean-variance tradeoff? (Askell, forthcoming) (INFORMAL: Hurford 2013)

  PHIL - EPISTEMOLOGY, DECISION THEORY **ECON** - VALUE OF INFORMATION

- If most interventions are indeed fairly ineffective, 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?

  PHIL - DECISION THEORY, EPISTEMOLOGY **ECON** - VALUE OF INFORMATION, BAYESIAN UPDATING

- Should an actor have a prior belief over the distribution of his possible impact (INFORMAL: Karnofsky 2011) such that it's astronomically unlikely that he could have the sort of positive impact that it seems one can have by reducing existential risk if total utilitarianism is correct? What bearing does this have on the expected value of activities aiming to improve the long-run future?

  PHIL - DECISION THEORY, EPISTEMOLOGY **ECON** - BAYESIAN UPDATING

- To what extent should we be 'risk averse' in our approach to doing good, and what are the implications of reasonable risk aversion? (Quiggin 1982; Buchak 2013; Greaves et al. MS)

  PHIL - DECISION THEORY **ECON** - DECISION THEORY

- What are the implications of ambiguity aversion (whether rational or not) for the project of doing good? (Rowe and Voorhoeve 2018)

  PHIL - DECISION THEORY **ECON** - DECISION THEORY

- Often it seems that subtle differences in epistemology would lead one to quite different conclusions concerning which interventions have the highest expected impartial value. These include differences in responses to paucity of hard evidence, in level of trust in abstract arguments leading to counterintuitive conclusions, in responses to interpersonal disagreement, and in the relative weight placed on different types of evidence. To what extent should this lack of robustness move us away from simply maximising expected value with respect to whatever credences we happen (now) to have? Is there a plausible alternative approach? (INFORMAL: Karnofsky 2016)

  PHIL - DECISION THEORY, EPISTEMOLOGY

Existing academic literature:

- Askell, Amanda. 'Evidence Neutrality and the Moral Value of Information'. In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer. Oxford: Oxford University Press, forthcoming.
- Bostrom, Nick. 'Pascal's Mugging'. *Analysis* 69, no. 3 (2009): 443–45.
- Buchak, Lara. *Risk and Rationality*. Oxford: Oxford University Press, 2013.
- Greaves, Hilary, Andreas L. Mogensen and William MacAskill. 'Longtermism for Risk Averse Altruists'. Manuscript in preparation.
- Greaves, Hilary. 'Cluelessness'. *Proceedings of the Aristotelian Society* 116 (2016): 311–39.
- Jansen, C., G. Schollmeyer, and T. Augustin. 'Concepts for Decision Making under Severe Uncertainty with Partial Ordinal and Partial Cardinal Preferences'. *International Journal of Approximate Reasoning* 98 (2018): 112–31.
- Lenman, James. 'Consequentialism and Cluelessness'. *Philosophy & Public Affairs* 29, no. 4 (2000): 342–70.

- Quiggin, John. 'A Theory of Anticipated Utility'. *Journal of Economic Behavior & Organization* 3, no. 4 (1982): 323–43.
- Tarsney, Christian. 'Exceeding Expectations: Stochastic Dominance as a General Decision Theory'. Working paper, 2018.

Existing informal discussion:

- Amanda Askell, Tackling the ethics of infinity, being clueless about the effects of our actions, and having moral empathy for intellectual adversaries, 2018
- Amanda Askell, Seminar presentation on speculative vs robust evidence
- Tobias Baumann, Uncertainty smooths out differences in impact, 2017
- Peter Hurford, Why I'm Skeptical About Unproven Causes (And You Should Be Too), 2013
- Holden Karnofsky, Why we can't take expected value estimates literally (even when they're unbiased), 18 August 2011
- Holden Karnofsky, Worldview diversification, 2016
- Brian Tomasik, Charity cost-effectiveness in an uncertain world, 28 October 2013

2.2 Epistemological issues

Thinking about global prioritisation, particularly (although not only) within the longtermist paradigm, tends to rely on heavily philosophical considerations and to reach some surprising and counterintuitive conclusions. We must therefore assess the extent to which this unusual circumstance should undermine our confidence in the conclusions in question.

Potential research projects:

- To what extent should an actor should place weight on her own idiosyncratic 'inside view' judgments, rather than deferring to the views of the majority of peers/experts on the issue? (Elga 2007; Christensen 2007; Christensen 2009; Feldman and Warfield 2010; Wilson 2010; Christensen and Lackey 2013) (INFORMAL: Beckstead 2013; Lewis 2017)

  PHIL - MORAL UNCERTAINTY **ECON** - VALUE OF INFORMATION, BAYESIAN UPDATING
- How much weight should we place on philosophical arguments? Is there a sound 'pessimistic induction' against placing much weight on them, assuming that most philosophical arguments in the past have been mistaken?

  PHIL - EPISTEMOLOGY, METAPHILOSOPHY
- What mechanisms can induce individuals to report their moral views honestly to each other?

  PHIL - EPISTEMOLOGY, MORAL UNCERTAINTY **ECON** - GAME THEORY, MECHANISM DESIGN

- Should one have the same levels of epistemic modesty about unusual moral views as one should about unusual empirical views?

PHIL - EPISTEMOLOGY, MORAL UNCERTAINTY

Existing academic literature:

- Christensen, D. 'Epistemology of Disagreement: The Good News'. *Philosophical Review* 116, no. 2 (2007): 187–217.
- Christensen, David. 'Disagreement as Evidence: The Epistemology of Controversy'. *Philosophy Compass* 4, no. 5 (2009): 756–67.
- Christensen, David, and Jennifer Lackey, eds. *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press, 2013.
- Elga, Adam. 'Reflection and Disagreement'. *Noûs* 41, no. 3 (2007): 478–502.
- Feldman, Richard, and Ted A. Warfield, eds. *Disagreement*. Oxford: Oxford University Press, 2010.
- Wilson, Alastair. 'Disagreement, Equal Weight, and Commutativity'. *Philosophical Studies* 149 (2010): 321–6.

Existing informal discussion:

- Amanda Askell, The Moral Value of Information, 2017
- Nick Beckstead, Common sense as a prior, 2013
- Holden Karnofsky, Maximising cost-effectiveness via critical enquiry, 2011
- Holden Karnofsky, Sequence thinking vs. cluster thinking, 2014 - a
- Holden Karnofsky, Modelling Extreme Model Uncertainty, 2014 - b
- Greg Lewis, In defence of epistemic modesty, 2017
- Jonah Sinick, Many Weak Arguments vs. One Relatively Strong Argument, 2013
- Benjamin Todd, Is it fair to say that most social programmes don't work?, 2017

2.3 Discounting

An actor aiming to do good faces two central timing questions.

First: When should she put her resources to philanthropic use? With her money, she could donate right away, or she could invest the money in order to donate 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. Even if her goal is to maximise total, non-discounted welfare across time, therefore, she must determine whether the returns to financial investments are high enough to warrant the associated delays.

Note that, when we look at how philanthropic actors in fact choose to discount, the results seem to depend largely on their institutional setting. Small individual donors typically give a certain amount of their income each year. Foundations and universities are typically set up in perpetuity. Governments typically borrow money in order to spend more now. These differences may reflect important differences in the constraints these actors face. Governments, for instance, can largely repay the costs of their 'social investments' through higher tax revenues in the future. This raises the question of how attitudes to discounting should differ for altruistic actors across these institutional contexts.

Second: If philanthropic interventions promise different payoff schedules, how should an agent compare payoffs which will accrue at different time periods? When payoffs consist of increases to human consumption, it makes sense to discount them to the extent that beneficiaries in the future will be wealthier (or poorer). Other kinds of payoffs, however—such as decreases in existential risk—presumably can be most naturally discounted on the basis of other heuristics.

Potential research projects:

- Positive returns on investment, and increasing information about where to give, constitute important reasons to consider waiting before committing resources to philanthropic use; rising global output, and accordingly declining opportunities for cost-effective philanthropy, constitute important reasons to consider committing resources earlier. More thoroughly, what are the considerations that are relevant to the question of when to donate? Can we build a quantitative economic model to represent and weigh these considerations? (Weitzman 1998; Gollier 2004; Fleurbaey and Zuber 2015; MacAskill MS) (INFORMAL: Christiano 2013a; Christiano 2013b; Wise 2013; Cotton-Barratt and Todd 2015; Todd 2017)

  ECON - DISCOUNTING, VALUE OF INFORMATION, FORECASTING

- Some (e.g. Parfit 2011) have argued that the present is an unusual time with respect to how quickly we ought to discount future donations. Is this correct?

  PHIL - ETHICS OF DISCOUNTING **ECON** - DISCOUNTING

- How might 'search theory', in which individuals have to decide whether to commit to taking some opportunity or hold out for a better opportunity (Mortensen 1986), shed light on the question of philanthropic discounting and when to do good?

  ECON - DISCOUNTING, VALUE OF INFORMATION, SEARCH THEORY

- How does the proper approach to philanthropic discounting depend on whether we are considering monetary investments or investments in human capital? What relevant restrictions apply in one case but not the other? For example, it is much more difficult to 'borrow' human capital than it is to borrow for a monetary investment.

  ECON - DISCOUNTING, HUMAN CAPITAL

- How do discount rates, and discount risks, currently differ across high-priority cause areas? To what extent are these differences and risks great enough to warrant placing high value on the 'liquidity' of capital to be put to philanthropic use? For instance, should altruistic agents earn to give, or learn broadly useful skills, instead of

specialising in a field that will likely soon be sub-optimal? (Cotton-Barratt 2015) (INFORMAL: Ord 2014)

ECON - DISCOUNTING, HUMAN CAPITAL

- Is there any justification for the observed tendency of smaller donors to give as they earn, while larger donors save and give later? Is there any justification for universities or foundations existing in perpetuity?

  ECON - DISCOUNTING, INSTITUTIONAL ECONOMICS

- When making their investment decisions, private investors typically discount monetary returns not only for temporally neutral reasons, such as the prospect of higher personal consumption in the future, but also for reasons of pure time preference. As a result, market interest rates should be expected to exceed the rate at which the marginal utility of consumption is declining. Does this imply that, under ordinary circumstances, temporally neutral altruists should save rather than give?

  ECON - DISCOUNTING, INTERGENERATIONAL EQUITY, FINANCIAL ECONOMICS

- Policymakers typically discount dollar-valued social costs and benefits not only for temporally neutral reasons, such as the prospect of higher average consumption in the future, but also to incorporate citizens' pure time preferences and as a reflection of short-term political incentives. How would policy recommendations change on evaluating social costs and benefits from a temporally neutral perspective (as explored in the optimal taxation context, for example, by Barrage (2018))? How might a patient agent provide incentives for an impatient government to implement policy consistent with placing a higher valuation on the future? How might an unusually patient government provide incentives for future (possibly impatient) governments to continue to make future-oriented investments?

  ECON - DISCOUNTING, SOCIAL CHOICE THEORY, POLITICAL ECONOMY, INTERGENERATIONAL EQUITY, INTERGENERATIONAL GOVERNANCE, MECHANISM DESIGN

- How should we construct a long-term discount schedule, from a temporally neutral perspective, in light of the profile of current and emerging existential risks (and our schedule of opportunities to reduce them)?

  ECON - DISCOUNTING, CATASTROPHIC RISK, INTERGENERATIONAL EQUITY

- What is the 'exchange rate', in a given economy, between consumption and moral value? Note that some consumption today consists of activity that is likely of negative value, such as inhumane animal agriculture. Other consumption, such as pain relief, may have much more positive value than is typically appreciated. How severely do such considerations render Ramsey-discounted consumption an imperfect proxy for moral value? How should we expect the weight of such considerations to change in the future?

  ECON - DISCOUNTING, CATASTROPHIC RISK, INTERGENERATIONAL EQUITY

Existing academic literature:

- Barrage, Lint. 'Be Careful What You Calibrate for: Social Discounting in General Equilibrium'. *Journal of Public Economics* 160 (2018): 33–49.

- Cotton-Barratt, Owen. 'Allocating Risk Mitigation across Time'. *Future of Humanity Institute*, Technical Report #2015-2.
- Emerson, Jed, Jay Wachowicz and Suzi Chun. 'Social Return on Investment: Exploring Aspects of Value Creation in the Nonprofit Sector'. *Social Purpose Enterprises and Venture Philanthropy in the New Millennium* 2 (2000): 132–73.
- Fleurbaey, Marc, and Stéphane Zuber. 'Discounting, Risk and Inequality: A General Approach'. *Journal of Public Economics* 128 (2015): 34–49.
- Frumkin, Peter. *Strategic Giving: The Art and Science of Philanthropy*. Chicago: University of Chicago Press, 2006.
- Gollier, Christian. 'Maximizing the Expected Net Future Value as an Alternative Strategy to Gamma Discounting'. *Finance Research Letters* 1, no. 2 (2004): 85–9.
- Greaves, Hilary. 'Discounting for Public Policy: A Survey'. *Economics & Philosophy* 33, no. 03 (2017): 391–439.
- Irvin, Renée A. 'Endowments: Stable Largesse or Distortion of the Polity?' *Public Administration Review* 67, no. 3 (2007): 445–57.
- Jansen, Paul, and David Katz. 'For Nonprofits, Time Is Money'. *The McKinsey Quarterly* 1 (2002): 124–33.
- Klausner, Michael D. 'When Time Isn't Money: Foundation Payouts and the Time Value of Money'. *SSRN Electronic Journal*, 2003.
- Landesman, Cliff. 'When to Terminate a Charitable Trust?' *Analysis* 55, no. 1 (1995): 12–13.
- MacAskill, William. 'When Should an Effective Altruist Donate?' Manuscript in preparation.
- Méjean, Aurélie, Antonin Pottier, Stéphane Zuber and Marc Fleurbaey. 'Intergenerational Equity under Catastrophic Climate Change'. Working paper, 2017.
- Moller, D. 'Should We Let People Starve - for Now?' *Analysis* 66, no. 3 (2006): 240–7.
- Mortensen, Dale. 'Job Search and Labor Market Analysis'. *Handbook of Labor Economics* 2 (1986): 849–919.
- Nordhaus, William D. 'A Review of the *Stern Review on the Economics of Climate Change*', *Journal of Economic Literature* 45, no. 3 (2007): 686–702.
- Parfit, Derek. *On What Matters*. Oxford: Oxford University Press, 2011.
- Stern, Nicholas. *The Economics of Climate Change*. Cambridge, UK: Cambridge University Press, 2006.
- Tarsney, Christian. 'Does a Discount Rate Measure the Costs of Climate Change?' *Economics & Philosophy* 33, no. 03 (2017): 337–65.

- Weitzman, Martin L. 'Why the Far Distant Future Should Be Discounted at Its Lowest Possible Rate'. *Journal of Environmental Economics and Management* 36, no. 3 (1998): 201–8.

Existing informal discussion:

- Paul Christiano, Giving now vs. later, 2013a
- Paul Christiano, The best reason to give later, 2013b
- Owen Cotton-Barratt and Benjamin Todd, Give now or later? What to do when the order of your actions matters, 2015
- Owen Cotton-Barratt, What does AI mean for EA movement building?, 2017
- Robin Hanson, Parable of the Multiplier Hole, 2010
- Toby Ord, The timing of labour aimed at reducing existential risk, 2014
- Benjamin Todd, Should you wait to make a difference?, 2017
- Julia Wise, Giving now vs. later: a summary, 2013

2.4 Diversification and hedging

What reasons are there, either for an individual philanthropist or for the global community of philanthropic actors, to diversify across causes/interventions, rather than simply identifying the intervention with the highest expected cost-effectiveness and supporting exclusively that intervention? Likewise, what reasons are there for philanthropic investors to diversify or hedge, instead of simply choosing the investments with highest expected return?

Possible justifications for diversification across causes and interventions include diminishing marginal returns of resources to impartial value within a given cause area or intervention; the information value of executing interventions; and moral uncertainty.

Relatedly, while investing for future giving, philanthropists may be able to maximise their impact by hedging their investments appropriately. For example, if an organisation wants to invest in renewable energy to reduce greenhouse gas emissions, they might hedge by also investing in oil companies: in the case that fossil fuels become unexpectedly profitable (for example because of discoveries of large new oil reserves), the organisation will then have more resources available to invest in renewable energy. More generally, investors should pick assets in part on the basis of their 'philanthropic beta': the association between the asset's value and the ease with which resources can be put to doing good.

Potential research projects:

- What are the potential reasons for diversifying investments across philanthropic causes? Which, if any, validly apply to individuals? To a large foundation? To the worldwide community of altruistic actors as a whole? (Snowden forthcoming)

ECON - DIVERSIFICATION

- How, if at all, do the considerations for or against diversifying across philanthropic causes differ when we consider how to allocate human capital resources rather than financial resources?

  ECON - DIVERSIFICATION, HUMAN CAPITAL

- To what extent should a large foundation diversify across different 'worldviews' (INFORMAL: Karnofsky 2016)? To what extent does moral uncertainty provide support for such diversification?

  PHIL - MORAL UNCERTAINTY **ECON** - DIVERSIFICATION

- Within the cause areas judged to be of exceptionally high priority, how quickly do we expect returns to diminish? (INFORMAL: Shulman 2014)

  ECON - APPLIED MICROECONOMICS

- Philanthropists face uncertainty about the rate at which doing good will grow more costly (INFORMAL: Christiano 2013). This rate is likely not perfectly correlated with market interest rates (or with the discount rates facing other philanthropists, given cause area disagreements). Should an investing philanthropist therefore sign 'charitable discount rate swaps', paying a sum if his discount rate is higher than expected (e.g. if some vaccine is developed more quickly than expected), in exchange for payment if it is lower? What other financial instruments might be used to hedge philanthropic risks? How might such arrangements best be implemented, given the cause areas that seem to be of highest priority?

  ECON - FINANCIAL ECONOMICS, HEDGING, DISCOUNTING

- Some investments' returns covary with the cost-effectiveness of high-priority philanthropic opportunities. 'Mission hedging' (Tran 2017) is the practice of exploiting this covariance. How important is mission hedging? How might it best be implemented, given the cause areas that seem to be of highest priority?

  ECON - FINANCIAL ECONOMICS, HEDGING

- As outlined above, when one has a well-defined mission (say, environmentalism), one can mission hedge (say, by investing in oil companies). But when one's mission is more open-ended, hedging may still be possible. For example, one might think that, under most choices of cause area, philanthropic resources will go further when the market is doing poorly. In that case, market beta—an asset's excess return per unit of market excess return, as given by the Capital Asset Pricing Model (Sharpe 1964; Lintner 1965)—is serving as a proxy for 'philanthropic beta'. How well does market beta serve as a good proxy for philanthropic beta in the face of cause uncertainty? Might other market indices serve as better proxies?

  ECON - FINANCIAL ECONOMICS, MODERN PORTFOLIO THEORY, HEDGING

- Individual philanthropic investors, too small to affect the marginal return within a given cause area, may have reason to invest risk-neutrally. Even so, it may be important for the cause area's funders to ensure that they collectively diversify. Is this a practical issue in any high-priority cause areas today? If so, how might it be resolved? (INFORMAL: Shulman 2012; Tomasik 2013)

  ECON - FINANCIAL ECONOMICS, DIVERSIFICATION

Existing academic literature:

- Elton, Edwin, Martin Gruber, Stephen Brown, and William Goetmann. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, 2009.
- Lintner, John. 'The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets'. *Review of Economics and Statistics* 47, no. 1 (1965): 13–39.
- Sharpe, William F. 'Capital asset prices: A theory of market equilibrium under conditions of risk'. *Journal of Finance* 19, no. 3 (1964): 425–42.
- Snowden, James. 'Should we give to more than one charity?' In *Effective Altruism: Philosophical Issues*, edited by Theron Pummer and Hilary Greaves. Oxford: Oxford University Press, forthcoming.
- Tran, Brigitte Roth. 'Divest, Disregard, or Double Down?' *Finance and Economics Discussion Series* 2017, no. 042 (2017).

Existing informal discussion:

- Kyle Bogosian, Selecting investments based on covariance with the value of charities, 2017
- Paul Christiano, Doing now vs. later, 2013
- John Halstead and Hauke Hillebrandt, Impact investing is only a good idea in specific circumstances, 2018
- Holden Karnofsky, Worldview Diversification, 2016
- Ben Kuhn, How many causes should you give to?, 2014
- Carl Shulman, Salary or startup? How do-gooders can gain more from risky careers, 2012
- Carl Shulman, It's harder to favor a specific cause in more efficient charitable markets, 2014
- Brian Tomasik, When Should Altruists Be Financially Risk-Averse?, 2013

2.5 Distributions of cost-effectiveness

Estimates of the effects of different interventions in different settings indicate that cost effectiveness can vary significantly, sometimes by multiple orders of magnitude, even within a given cause area. If so, this is important, because it pushes towards optimising for effectiveness over increasing the amount of resources going toward a cause. However, there is currently rather little rigourous investigation of the properties of the relevant cost-effectiveness distributions.

Potential research projects:

- Establish more rigourously and more generally what can be said about typical distributions of cost-effectiveness, both within and between causes, and (within a single cause) both between interventions and between different organisations implementing 'the same' intervention in different settings. (Ord 2013) (INFORMAL: Kaufman 2013; Kaufman 2015; Cotton-Barratt 2017)

  ECON - PROGRAMME EVALUATION, EXTERNAL VALIDITY

- How much of the variation of estimated cost effectiveness within a cause area is driven by differences in empirical settings or implementation between different evaluations (Meager 2018; Vivalt 2015)? How does variation of cost-effectiveness within a cause compare to variation of cost-effectiveness between causes (Vivalt 2019)? What are the implications for the case for diversification of cause areas and interventions?

  ECON - PROGRAMME EVALUATION, EXTERNAL VALIDITY, BAYESIAN UPDATING

- How does the estimated distribution of cost effectiveness affect the trade-off between the informational value of evaluating slightly different interventions in different settings versus the value created by implementing effective interventions given the existing state of knowledge? (INFORMAL: Askell 2017)

  ECON - VALUE OF INFORMATION, BAYESIAN UPDATING

- What's the base rate probability that an intervention with given features has positive, neutral or negative impact? How common are situations in which most ways of acting do harm, and which factors make this case more likely? What implications do these facts have for which problems we ought to focus on? (INFORMAL: Todd 2017)

  ECON - VALUE OF INFORMATION, BAYESIAN UPDATING

- How does the inclusion of indirect effects affect the estimated variance in cost-effectiveness across interventions?

  ECON - PROGRAMME EVALUATION, ECONOMETRIC THEORY, STRUCTURAL MODELLING

Existing academic literature:

- Meager, Rachael. 'Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature'. Working paper, 2018.
- Ord, Toby. 'The Moral Imperative Towards Cost-Effectiveness in Global Health'. *Center for Global Development*, 2013.
- Tengs et al, Tammy O. 'Five-Hundred Life-Saving Interventions and Their Cost-Effectiveness'. *Risk Analysis* 15, no. 3 (1994): 369–90.
- Vivalt, Eva. 'Heterogeneous Treatment Effects in Impact Evaluation'. *American Economic Review, Papers and Proceedings* 105, no. 5 (2015): 467–70.
- ———. 'How Much Can We Generalize from Impact Evaluations?' Working paper, 2019.

Existing informal discussion:

- Amanda Askell, The Moral Value of Information, 2017

- Benjamin Todd, Is it fair to say that most social programmes don't work?, 2017
- Owen Cotton-Barratt, Distributions of cost-effectiveness, 2017
- Holden Karnofsky, A conflict of Bayesian priors?, 2009
- Jeff Kaufman, The Unintuitive Power Laws of Giving, 2013
- Jeff Kaufman, Effectiveness: Gaussian?, 2015
- Toby Ord, Taking charity seriously, 2013
- Brian Tomasik, Why Charities Usually Don't Differ Astronomically in Expected Cost-Effectiveness, 2013

2.6 Modelling altruism

Economic theory typically proceeds either (a) making minimally substantive assumptions 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 his own consumption and leisure). Existing research shows that interesting new results can be established when we expand the domain of preferences to include the utility of others. However, this literature considers a relatively narrow domain of problems, and there is scope to further explore the implications of modelling agents as at least partially altruistic.

Potential research projects:

- Are there settings in which agents have other-regarding preferences, and are either short-lived or have a non-zero discount rate, and are therefore unable to achieve a socially optimal outcome, for example because they are unable to commit to 'punish' defectors to sustain an equilibrium (Rabin 1993; Povey 2014; Povey 2015)? What are the characteristics of these settings (Bolton and Ockenfels 2000)? Can we design mechanisms to overcome these challenges? Do these results have practical implications for decision-makers?

  ECON - GAME THEORY, MECHANISM DESIGN

- How should we adapt key economic models to account for altruistic individuals with other-regarding preferences (Bergstrom 2002, Sobel 2005)? Under what assumptions do key results, such as the Fundamental Theorems of Welfare Economics, still hold (Schall 1972; Pollack 1976; Rotemberg 2003)? In cases that they do not, can analogous results be derived?

  ECON - MICROECONOMIC THEORY

- Is there a theoretical 'optimal' level of altruism in relevant settings (Povey 2015)? Do these results provide practical insights or implications for agents attempting to do good?

  ECON - MICROECONOMIC THEORY, GAME THEORY

- Improve our understanding of the various motivations for apparently altruistic acts, for example 'pure' altruism or 'warm glow' altruism (Andreoni 1990; Ashraf and Bandiera 2017). Which characteristics of individuals or the choices that they face are associated with different types of apparently altruistic acts?

ECON - BEHAVIOURAL ECONOMICS

Existing academic literature:

- Andreoni, James. 'Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving'. *The Economic Journal* 100, no. 401 (1990): 464–77.
- Ashraf, Nava, and Oriana Bandiera. 'Altruistic Capital'. *American Economic Review* 107, no. 5 (2017): 70–5.
- Bergstrom, Theodore. 'Systems of Benevolent Utility Functions'. *Journal of Public Economic Theory* 1, no. 2 (1999): 71–100.
- Bolton, Gary E., and Axel Ockenfels. 'ERC: A Theory of Equity, Reciprocity, and Competition'. *American Economic Review* 90, no. 1 (2000): 166–93.
- Fehr, E., and K. M. Schmidt. 'A Theory of Fairness, Competition, and Cooperation'. *Quarterly Journal of Economics* 114, no. 3 (1999): 817–68.
- Pollack, Robert. 'Interdependent Preferences'. *American Economic Review* 66, no. 3 (1976): 309–20.
- Povey, Richard. 'The Limits to Altruism - A Survey'. Working paper, 2014.
- ———. 'The Socially Optimal Level of Altruism'. Working paper, 2015.
- Rabin, Matthew. 'Incorporating Fairness into Game Theory'. *American Economic Review* 83, no. 5 (1993): 1281–302.
- Rotemberg, Julio J. 'The Benevolence of the Baker: Fair Pricing under Threat of Customer Anger'. Working paper, 2003.
- Schall, Lawrence D. 'Interdependent Utilities and Pareto Optimality'. *Quarterly Journal of Economics* 86, no. 1 (1972): 19–24.
- Sobel, Joel. 'Interdependent Preferences and Reciprocity'. *Journal of Economic Literature* 43, no. 2 (2005): 392–436.

2.7 Altruistic coordination

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, cases where there are also empirical

disagreements, 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 these questions.

Potential research projects:

- What are the implications of comparative advantage for a community of altruists, who may be heterogeneous in terms of resources, skills, information and values (INFORMAL: Todd 2018a)?

  ECON - MICROECONOMIC THEORY

- How should game theoretic models be applied to analyse decisions faced by a community of altruists? For example, altruists with similar moral and empirical beliefs may face coordination problems similar to the 'stag hunt' game, whereby they can achieve a larger 'prize' if they coordinate, relative to working individually. (INFORMAL: Karnofsky 2014; Karnofsky 2015; Kuhn 2015; Ali 2016; Cotton-Barratt and Leather 2016; Todd 2018b)

  ECON - GAME THEORY

- How can results from the mechanism design literature help altruistic individuals and organisations to coordinate in a more effective manner (Andreoni 1998; Bracha et al. 2011; Conitzer and Sandholm 2011; Peters MS)? For example, among people with similar altruistic goals, each charitable act resembles the provision of a public good. However, in cases where individuals have (heterogeneous) private beliefs and/or information, they may have an incentive to mis-report these, in order to achieve an outcome closer to the one they prefer (Gibbard 1973; Satterthwaite 1977; Myerson 1979). Which mechanisms can induce individuals to report their beliefs about charitable interventions (approximately) truthfully (Myerson and Satterthwaite 1983; Li 2017)?

  ECON - MECHANISM DESIGN, SOCIAL CHOICE THEORY

- How can a community of altruists with different moral and empirical views gain from trade? Do traditional challenges in trade extend to the case of moral trade (Ord 2015) (for example, the Myerson-Satterthwaite theorem, according to which efficient trade cannot take place if two parties have private, stochastic valuations over the traded good)? What are the challenges for moral trade that go beyond the challenges for ordinary trade, and can they be overcome? (INFORMAL: Tomasik 2013; Oesterheld 2018)

  PHIL - MORAL UNCERTAINTY **ECON** - GAME THEORY, MECHANISM DESIGN

- Are there institutions or mechanisms we can design to help improve the allocative efficiency of altruists' resources among altruists?

  ECON - MECHANISM DESIGN

Existing academic literature:

- Andreoni, James. 'Toward a Theory of Charitable Fund-Raising'. *Journal of Political Economy* 106, no. 6 (1998): 1186–213.

- Bracha, Anat, Michael Menietti, and Lise Vesterlund. 'Seeds to Succeed?' *Journal of Public Economics* 95, no. 5–6 (2011): 416–27.
- Conitzer, Vincent, and Tuomas Sandholm. 'Expressive Markets for Donating to Charities'. *Artificial Intelligence* 175, no. 7–8 (2011): 1251–71.
- Gibbard, Allan. 'Manipulation of Voting Schemes: A General Result'. *Econometrica* 41, no. 4 (1973): 587–601.
- Li, Shengwu. 'Obviously Strategy-Proof Mechanisms'. *American Economic Review* 107, no. 11 (2017): 3257–87.
- Myerson, Roger B. 'Incentive Compatibility and the Bargaining Problem'. *Econometrica* 47, no. 1 (1979): 61–73.
- Myerson, Roger B, and Mark A. Satterthwaite. 'Efficient Mechanisms for Bilateral Trading'. *Journal of Economic Theory* 29, no. 2 (1983): 265–81.
- Ord, Toby. 'Moral Trade'. *Ethics* 126, no. 1 (2015): 118–38.
- Peters, Dominik. 'Economic Design for Effective Altruism'. Working paper, 2017.
- Satterthwaite, Mark Allen. 'Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions'. *Journal of Economic Theory* 10, no. 2 (1975): 187–217.

Existing informal discussion:

- S. Nageeb Ali, A conversation with Professor S. Nageeb Ali, 2016
- Paul Christiano, Certificates of Impact, 2014
- Paul Christiano, Repledge++, 2016
- Owen Cotton-Barratt and Zachary Leather, Donor coordination under simplifying assumptions
- Max Dalton, Mechanism design for altruistic cooperation, 2018
- Ben Garfinkel, What is the relationship between effective altruism and economics?
- Holden Karnofsky, Donor coordination and the "giver's dilemma", 2014
- Holden Karnofsky, Good Ventures and giving now vs. later (Section Coordination issues), 2015
- Ben Kuhn, Solving donation coordination problems, 2015
- Rossa O'Keeffe-O'Donovan, Economics of career choice
- Caspar Oesterheld, Multiverse-Wide Cooperation via Correlated Decision Making, 2018
- Michael Page, Certificates of Impact
- Carl Shulman, Donor lotteries: demonstration and FAQ, 2016
- Brian Tomasik, Gains from Trade through Compromise, 2013

- Benjamin Todd, Should you play to your comparative advantage when choosing your career?, 2018 - a
- Benjamin Todd, Doing good together — how to coordinate effectively, and avoid single-player thinking, 2018 - b
- Jess Whittlestone, Building an Effective Altruism Community

2.8 Individual vs institutional actors

In addition to asking how individuals can do good effectively, and to what extent they ought to, we can ask the analogous questions about larger entities, such as governments, philanthropic foundations, corporations and international institutions. This might in principle lead to different answers, since these larger entities have resources that are generally inaccessible to private individuals. These resources may allow them to make large lump-sum investments or to influence or create markets, and may imply a different approach to risk and diversification in investments. These resources include vastly greater budgets. Governments may also leverage legislative power and attempt to relatively direct opportunities to influence the actions of other states, either directly or through international organisations, treaties and agreements. Corporations and governments also play different roles in society to those played by private individuals (for instance, they bear special relationships to (respectively) their shareholders and citizens).

Potential research projects:

- Should organisations with access to a large amount of resources seek to do good in a way that is fundamentally different to individuals? Should these organisations assess expected value, risk and/or diversification in a different way to individuals when evaluating opportunities to do good? (Kagan 2011; McMahan 2016; Kissel 2017; Collins forthcoming) (INFORMAL: Karnofsky 2013; Reich 2015; Greaves 2017)

  PHIL - DECISION THEORY **ECON** - DECISION THEORY, MODERN PORTFOLIO THEORY

- What is the optimal design of international institutions that are formed to increase global public goods or decrease global public bads? Can institutions be designed to overcome the participation constraints and incentive compatibility constraints of potentially self-interested nation states, while achieving globally socially optimal investments? Can such mechanisms be enforced?

  ECON - GAME THEORY, MECHANISM DESIGN, OPTIMAL TAXATION

- To what extent ought a government to take actions that are better for the world even if they conflict with the preferences of, and/or are worse for, their own citizens (Goodin 1995)? What about the relationship between corporate philanthropy and shareholder preference/interest?

  PHIL - POLITICAL PHILOSOPHY, DUTIES OF BENEFICENCE, BUSINESS ETHICS

- 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 statist political

philosophies, perhaps analogous to the ways in which market failures justify deviations from a free market? Are there better alternatives? (Rawls 1971; Barry 1997; Donaldson and Kymlicka 2013)

PHIL - POLITICAL PHILOSOPHY ECON - POLITICAL ECONOMY, SOCIAL CHOICE THEORY

- What is the best feasible voting system from the perspective of impartial welfarism? For example, what impact should we expect quadratic voting (Lalley and Weyl 2018), futarchy (Hanson 2013) or approval voting (Brams and Fishburn 2007) to have on social welfare?

ECON - POLITICAL ECONOMY, SOCIAL CHOICE THEORY

Existing academic literature:

- Barry, Brian. 'Sustainability and Intergenerational Justice'. *Theoria* 44, no. 89 (1997): 43–64.
- Brams, Steven, and Peter C. Fishburn. *Approval Voting*. Springer Science & Business Media, 2007.
- Collins, Stephanie. 'Beyond Individualism'. In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer. Oxford: Oxford University Press, forthcoming.
- Donaldson, Sue, and Will Kymlicka. *Zoopolis: A Political Theory of Animal Rights*. Oxford; New York: Oxford University Press, 2013.
- Goodin, Robert. *Utilitarianism as a Public Philosophy*. Cambridge, UK: Cambridge University Press, 1995.
- Hanson, Robin. 'Shall We Vote on Values, But Bet on Beliefs?' *Journal of Political Philosophy* 21, no. 2 (2013): 151–78.
- Kagan, Shelly. 'Do I Make a Difference?' *Philosophy & Public Affairs* 39, no. 2 (2011): 105–41.
- Kissel, Joshua. 'Effective Altruism and Anti-Capitalism: An Attempt at Reconciliation'. *Essays in Philosophy* 18, no. 1 (2017): 1–23.
- Lalley, Steven P., and E. Glen Weyl. 'Quadratic Voting: How Mechanism Design Can Radicalize Democracy'. *AEA Papers and Proceedings* 108 (2018): 33–7.
- Maskin, Eric S. 'Mechanism Design: How to Implement Social Goals'. *American Economic Review* 98, no. 3 (2008): 567–76.
- McMahan, Jeff. 'Philosophical Critiques of Effective Altruism'. *The Philosophers' Magazine*, no. 73 (2016): 92–9.
- Rawls, John. *A Theory of Justice*. Cambridge, MA: Belknap Press, 1971.

Existing informal discussion:

- Scott Alexander, Beware Systemic Change, 2015

- Hilary Greaves, The collectivist critique of the effective altruist movement, 23 May 2017
- The Open Philanthropy Project, U.S. Policy, 2018
- Holden Karnofsky, The Role of Philanthropic Funding in Politics, 2013
- Rob Reich, The Logic of Effective Altruism, 2015

Bibliography

- Acemoglu, Daron, Simon Johnson and James A. Robinson. 'Institutions as a Fundamental Cause of Long-Run Growth'. *Handbook of Economic Growth*, 1A (2005): 385–472.
- Andreoni, James. 'Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving'. *The Economic Journal* 100, no. 401 (1990): 464–77.
- ———. 'Toward a Theory of Charitable Fund-Raising'. *Journal of Political Economy* 106, no. 6 (1998): 1186–213.
- Arntzenius, Frank. 'Utilitarianism, Decision Theory, and Eternity'. *Philosophical Perspectives* 28, no. 1 (2014): 31–58.
- Asheim, Geir B. 'Intergenerational Equity'. *Annual Review of Economics* 2, no. 1 (2010): 197–222.
- Ashraf, Nava, and Oriana Bandiera. 'Altruistic Capital'. *American Economic Review* 107, no. 5 (2017): 70–5.
- Askell, Amanda. 'Evidence Neutrality and the Moral Value of Information'. In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer. Oxford: Oxford University Press, forthcoming.
- Barrage, Lint. 'Be Careful What You Calibrate for: Social Discounting in General Equilibrium'. *Journal of Public Economics* 160 (2018): 33–49.
- Barry, Brian. 'Sustainability and Intergenerational Justice'. *Theoria* 44, no. 89 (1997): 43–64.
- Basu, Kaushik, and Tapan Mitra. 'Aggregating Infinite Utility Streams with Intergenerational Equity: The Impossibility of Being Paretian'. *Econometrica* 71, no. 5 (2003): 1557–63.
- Baum, S. D., et al. 'Long-Term Trajectories of Human Civilization'. *Foresight*, forthcoming.
- Beckstead, Nicholas. 'A Brief Argument for the Overwhelming Importance of Shaping the Far Future'. In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer. Oxford: Oxford University Press, forthcoming.
- ———. 'On the Overwhelming Importance of Shaping the Far Future'. PhD dissertation. New Brunswick: Rutgers University, 2013.
- Bergstrom, Theodore. 'Systems of Benevolent Utility Functions'. *Journal of Public Economic Theory* 1, no. 2 (1999): 71–100.
- Bishop, Richard C. 'Option Value: An Exposition and Extension'. *Land Economics* 58, no. 1 (1982): 1–15.
- Bolton, Gary E, and Axel Ockenfels. 'ERC: A Theory of Equity, Reciprocity, and Competition'. *American Economic Review* 90, no. 1 (2000): 166–93.

- Bostrom, Nick. ‘Astronomical Waste: The Opportunity Cost of Delayed Technological Development’. *Utilitas* 15, no. 3 (2003): 308–14.
- ———. ‘Existential Risk Prevention as Global Priority’. *Global Policy* 4, no. 1 (2013): 15–31.
- ———. ‘Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards’. *Journal of Evolution and Technology* 9 (2002).
- ———. ‘Infinite Ethics’. *Analysis and Metaphysics*, no. 10 (2011): 9–59.
- ———. ‘Pascal’s Mugging’. *Analysis* 69, no. 3 (2009): 443–5.
- ———. *Superintelligence: Paths, Dangers, Strategies*. First edition. Oxford: Oxford University Press, 2014.
- ———. ‘Technological Revolutions: Ethics and Policy in the Dark’. In *Nanoscale*, edited by Nigel M. de S. Cameron and M. Ellen Mitchell. Hoboken, NJ: John Wiley & Sons, 2007. 129–52.
- ———. ‘The Future of Human Evolution’. In *Death and Anti-Death*, edited by Charles Tandy, 339–371. Ann Arbor: Ria University Press, 2005. 339–71.
- ———. ‘What is a Singleton?’ *Linguistic and Philosophical Investigations* 5, no. 2 (2006): 48–54.
- Bracha, Anat, Michael Menietti, and Lise Vesterlund. ‘Seeds to Succeed?’ *Journal of Public Economics* 95, no. 5–6 (2011): 416–27.
- Brams, Steven, and Peter C. Fishburn. *Approval Voting*. Springer Science & Business Media, 2007.
- Brock, Dan W. ‘Separate Spheres and Indirect Benefits’. *Cost Effectiveness and Resource Allocation* 1, no. 4 (2003).
- Broome, John. *Weighing Lives*. Oxford; New York: Oxford University Press, 2004.
- Buchak, Lara. *Risk and Rationality*. Oxford: Oxford University Press, 2013.
- Christensen, D. ‘Epistemology of Disagreement: The Good News’. *Philosophical Review* 116, no. 2 (2007): 187–217.
- Christensen, David. ‘Disagreement as Evidence: The Epistemology of Controversy’. *Philosophy Compass* 4, no. 5 (2009): 756–67.
- Christensen, David, and Jennifer Lackey, eds. *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press, 2013.
- Collins, Stephanie. ‘Beyond Individualism’. In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer. Oxford: Oxford University Press, forthcoming.
- Conitzer, Vincent, and Tuomas Sandholm. ‘Expressive Markets for Donating to Charities’. *Artificial Intelligence* 175, no. 7–8 (2011): 1251–71.

- Cotton-Barratt, Owen. 'Allocating Risk Mitigation across Time'. *Future of Humanity Institute*, Technical Report #2015-2.
- Cotton-Barratt, Owen, and Toby Ord. 'Existential Risk and Existential Hope: Definitions'. *Future of Humanity Institute*, Technical Report #2015-1.
- Cowen, Tyler. 'Caring About the Distant Future: Why It Matters and What It Means'. *The University Of Chicago Law Review* 74, no. 5 (2007): 5–40.
- Cowen, Tyler. *Stubborn Attachments*. San Francisco: Stripe Press, 2018.
- Dixit, Avinash and Robert S. Pindyck. *Investment Under Uncertainty*. Princeton: Princeton University Press, 1994.
- Donaldson, Sue, and Will Kymlicka. *Zoopolis: A Political Theory of Animal Rights*. Oxford; New York: Oxford University Press, 2013.
- Drexler, K. Eric. *Engines of Creation: The Coming Era of Nanotechnology*. New York: Random House, 1987.
- Elga, Adam. 'Reflection and Disagreement'. *Noûs* 41, no. 3 (2007): 478–502.
- Elton, Edwin, Martin Gruber, Stephen Brown, and William Goetmann. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, 2009.
- Emerson, Jed, Jay Wachowicz and Suzi Chun. 'Social Return on Investment: Exploring Aspects of Value Creation in the Nonprofit Sector'. *Social Purpose Enterprises and Venture Philanthropy in the New Millennium* 2 (2000): 132–73.
- Fehr, E., and K. M. Schmidt. 'A Theory of Fairness, Competition, and Cooperation'. *Quarterly Journal of Economics* 114, no. 3 (1999): 817–68.
- Feldman, Richard, and Ted A. Warfield, eds. *Disagreement*. Oxford: Oxford University Press, 2010.
- Fleurbaey, Marc, and Stéphane Zuber. 'Discounting, Risk and Inequality: A General Approach'. *Journal of Public Economics* 128 (2015): 34–49.
- Frumkin, Peter. *Strategic Giving: The Art and Science of Philanthropy*. Chicago: University of Chicago Press, 2006.
- Gibbard, Allan. 'Manipulation of Voting Schemes: A General Result'. *Econometrica* 41, no. 4 (1973): 587–601.
- Gollier, Christian. 'Maximizing the Expected Net Future Value as an Alternative Strategy to Gamma Discounting'. *Finance Research Letters* 1, no. 2 (2004): 85–9.
- Goodin, Robert. *Utilitarianism as a Public Philosophy*. Cambridge, UK: Cambridge University Press, 1995.
- Greaves, Hilary. 'Climate Change and Optimum Population'. *The Monist* 102, no. 1 (2019): 42–65.
- ———. 'Cluelessness'. *Proceedings of the Aristotelian Society* 116 (2016): 311–39.

- ———. ‘Discounting for Public Policy: A Survey’. *Economics & Philosophy* 33, no. 03 (2017): 391–439.
- ———. ‘Discounting Future Health’. In *Global Health Priority-Setting: Cost-Effectiveness and Beyond*, edited by Ruger and Verguet Otterson Millum Johansson Jamison Emanuel Norheim. Oxford: Oxford University Press, forthcoming.
- ———. ‘Optimum Population Size’. In *Oxford Handbook of Population Ethics*, edited by Gustaf Arrhenius, Krister Bykvist and Tim Campbell. Oxford: Oxford University Press, forthcoming.
- Greaves, Hilary, and Toby Ord. ‘Moral Uncertainty About Population Axiology’. *Journal of Ethics & Social Philosophy* 12, no. 2 (2017): 135–67.
- Greaves, Hilary, Andreas L. Mogensen and William MacAskill. ‘Longtermism for Risk Averse Altruists’. Manuscript in preparation.
- Hanson, Robin. ‘Shall We Vote on Values, But Bet on Beliefs?’ *Journal of Political Philosophy* 21, no. 2 (2013): 151–78.
- ———. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. First Edition. Oxford: Oxford University Press, 2016.
- Hurka, Thomas. ‘Asymmetries In Value’. *Noûs* 44, no. 2 (2010): 199–223.
- IPCC. *Climate change 2014: Impacts, adaptation, and vulnerability. Part A: Global and sectoral aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge, UK and New York, NY, USA: Cambridge University Press, 2014.
- Irvin, Renée A. ‘Endowments: Stable Largesse or Distortion of the Polity?’ *Public Administration Review* 67, no. 3 (2007): 445–57.
- Jansen, C., G. Schollmeyer and T. Augustin. ‘Concepts for Decision Making under Severe Uncertainty with Partial Ordinal and Partial Cardinal Preferences’. *International Journal of Approximate Reasoning* 98 (2018): 112–31.
- Jansen, Paul, and David Katz. ‘For Nonprofits, Time Is Money’. *The McKinsey Quarterly* 1 (2002): 124–33.
- Jones, Charles I. ‘Life and Growth’. *Journal of Political Economy* 124, no. 2 (2016): 539–78.
- Kagan, Shelly. ‘Do I Make a Difference?’ *Philosophy & Public Affairs* 39, no. 2 (2011): 105–41.
- Kamm, Frances M. *Morality, Mortality; Volume I: Death and Whom to Save From It*. Oxford: Oxford University Press, 1998.
- Kimball, Miles S. ‘Making Sense of Two-Sided Altruism’. *Journal of Monetary Economics* 20, no. 2 (1987): 301–26.
- Kissel, Joshua. ‘Effective Altruism and Anti-Capitalism: An Attempt at Reconciliation’. *Essays in Philosophy* 18, no. 1 (2017): 1–23.

- Klausner, Michael D. 'When Time Isn't Money: Foundation Payouts and the Time Value of Money'. *SSRN Electronic Journal*, 2003.
- Lalley, Steven P., and E. Glen Weyl. 'Quadratic Voting: How Mechanism Design Can Radicalize Democracy'. *AEA Papers and Proceedings* 108 (2018): 33–7.
- Landesman, Cliff. 'When to Terminate a Charitable Trust?' *Analysis* 55, no. 1 (1995): 12–13.
- Lauwers, Luc, and Peter Vallentyne. 'Infinite Utilitarianism: More Is Always Better'. *Economics & Philosophy* 20, no. 2 (2004): 307–30.
- Lenman, James. 'Consequentialism and Cluelessness'. *Philosophy and Public Affairs* 29, no. 4 (2000): 342–70.
- Li, Shengwu. 'Obviously Strategy-Proof Mechanisms'. *American Economic Review* 107, no. 11 (2017): 3257–87.
- Lintner, John. 'The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets'. *Review of Economics and Statistics* 47, no. 1 (1965): 13–39.
- Lippert-Rasmussen, Kasper, and Sigurd Lauridsen. 'Justice and the Allocation of Healthcare Resources: Should Indirect, Non-Health Effects Count?' *Medicine, Health Care and Philosophy* 13, no. 3 (2010): 237–46.
- MacAskill, William. 'Human Extinction, Asymmetry, and Option Value'. Manuscript in preparation.
- ———. 'Practical Ethics Given Moral Uncertainty'. Manuscript in preparation.
- ———. 'When Should an Effective Altruist Donate?' Manuscript in preparation.
- MacAskill, William, Krister Bykvist and Toby Ord. *Moral Uncertainty*. Oxford: Oxford University Press, forthcoming.
- MacAskill, William, and Toby Ord. 'Why Maximize Expected Choice-Worthiness?' *Noûs*, 14 July 2018.
- Martin, Ian W. R., and Robert S. Pindyck. 'Averting Catastrophes: The Strange Economics of Scylla and Charybdis'. *American Economic Review* 105, no. 10 (2015): 2947–85.
- ———. 'Averting Catastrophes that Kill'. Working paper, 2017.
- Maskin, Eric S. 'Mechanism Design: How to Implement Social Goals'. *American Economic Review* 98, no. 3 (2008): 567–76.
- Matheny, Gaverick, and Kai Chan. 'Human Diets and Animal Welfare: The Illogic of the Larder'. *Journal of Agricultural and Environmental Ethics* 18, no. 6 (2005): 57–94.
- Matheny, Jason G. 'Reducing the Risk of Human Extinction'. *Risk Analysis* 27, no. 5 (2007): 1335–44.

- McMahan, Jeff. 'Philosophical Critiques of Effective Altruism'. *The Philosophers' Magazine*, no. 73 (2016): 92–9.
- Meager, Rachael. 'Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature'. Working paper, 2018.
- Méjean, Aurélie, Antonin Pottier, Stéphane Zuber, and Marc Fleurbaey. 'Intergenerational Equity under Catastrophic Climate Change'. Working paper, 2017.
- Millner, Antony. 'On Welfare Frameworks and Catastrophic Climate Risks'. *Journal of Environmental Economics and Management* 65, no. 2 (2013): 310–25.
- Moller, D. 'Should We Let People Starve - for Now?' *Analysis* 66, no. 3 (2006): 240–7.
- Mortensen, Dale. 'Job Search and Labor Market Analysis'. *Handbook of Labor Economics* 2 (1986): 849–919.
- Myerson, Roger B. 'Incentive Compatibility and the Bargaining Problem'. *Econometrica* 47, no. 1 (1979): 61–73.
- Myerson, Roger B, and Mark A. Satterthwaite. 'Efficient Mechanisms for Bilateral Trading'. *Journal of Economic Theory* 29, no. 2 (1983): 265–81.
- Ng, Yew-Kwang. 'The Importance of Global Extinction in Climate Change Policy'. *Global Policy* 7, no. 3 (2016): 315–22.
- Nordhaus, William D. 'A Review of the Stern Review on the Economics of Climate Change'. *Journal of Economic Literature* 45, no. 3 (2007): 686–702.
- Ord, Toby. 'Moral Trade'. *Ethics* 126, no. 1 (2015): 118–38.
- ———. 'The Moral Imperative Towards Cost-Effectiveness in Global Health'. *Center for Global Development*, 2013.
- ———. *Existential Risk*. London: Bloomsbury, forthcoming.
- Parfit, Derek. *On What Matters*. Oxford University Press, 2011.
- Peters, Dominik. 'Economic Design for Effective Altruism'. Working paper, 2017.
- Pollack, Robert. 'Interdependent Preferences'. *American Economic Review* 66, no. 3 (1976): 309–20.
- Povey, Richard. 'The Limits to Altruism - A Survey'. Working paper, 2014.
- ———. 'The Socially Optimal Level of Altruism'. Working paper, 2015.
- Quiggin, John. 'A Theory of Anticipated Utility'. *Journal of Economic Behavior & Organization* 3, no. 4 (1982): 323–43.
- Rabin, Matthew. 'Incorporating Fairness into Game Theory'. *American Economic Review* 83, no. 5 (1993): 1281–302.
- Rawls, John. *A Theory of Justice*. Cambridge, MA: Belknap Press, 1971.
- Ross, Jacob. 'Rejecting Ethical Deflationism'. *Ethics* 116, no. 4 (2006): 742–68.

- Rotemberg, Julio J. 'The Benevolence of the Baker: Fair Pricing under Threat of Customer Anger'. Working paper, 2003.
- Sandberg, Anders, and Nick Bostrom. 'Whole Brain Emulation: A Roadmap'. *Future of Humanity Institute*, Technical Report #2008-3.
- Satterthwaite, Mark Allen. 'Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions'. *Journal of Economic Theory* 10, no. 2 (1975): 187–217.
- Schall, Lawrence D. 'Interdependent Utilities and Pareto Optimality'. *Quarterly Journal of Economics* 86, no. 1 (1972): 19–24.
- Sharpe, William F. 'Capital asset prices: A theory of market equilibrium under conditions of risk'. *Journal of Finance* 19, no. 3 (1964): 425–42.
- Snowden, James. 'Should We Give to More Than One Charity?' In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer. Oxford: Oxford University Press, forthcoming.
- Sobel, Joel. 'Interdependent Preferences and Reciprocity'. *Journal of Economic Literature* 43, no. 2 (2005): 392–436.
- Stern, Nicholas. *The Economics of Climate Change*. Cambridge, UK: Cambridge University Press, 2006.
- Tarsney, Christian. 'Does A Discount Rate Measure The Costs Of Climate Change?' *Economics & Philosophy* 33, no. 03 (2017): 337–65.
- ———. 'Exceeding Expectations: Stochastic Dominance as a General Decision Theory'. Working paper, 2018.
- Tengs, Tammy O., et al. 'Five-Hundred Life-Saving Interventions and Their Cost-Effectiveness'. *Risk Analysis* 15, no. 3 (1994): 369–90.
- Toit, Jessica du, and Franklin Miller. 'The Ethics of Continued Life-Sustaining Treatment for Those Diagnosed as Brain-Dead'. *Bioethics* 30, no. 3 (2016): 151–58.
- Tran, Brigitte Roth. 'Divest, Disregard, or Double Down?' *Finance and Economics Discussion Series* 2017, no. 042 (2017).
- Vallentyne, Peter, and Shelly Kagan. 'Infinite Value and Finitely Additive Value Theory'. *The Journal of Philosophy* 94, no. 1 (1997): 5–26.
- Vivalt, Eva. 'Heterogeneous Treatment Effects in Impact Evaluation'. *American Economic Review, Papers and Proceedings* 105, no. 5 (2015): 467–70.
- ———. 'How Much Can We Generalize from Impact Evaluations?' Working paper, 2019.
- Weitzman, Martin L. 'Why the Far Distant Future Should Be Discounted at Its Lowest Possible Rate'. *Journal of Environmental Economics and Management* 36, no. 3 (1998): 201–8.

- Weitzman, Martin L. 'On Modeling and Interpreting the Economics of Catastrophic Climate Change'. *Review of Economics and Statistics* 91, no. 1 (2009): 1–19.
- Wilson, Alastair. 'Disagreement, Equal Weight, and Commutativity'. *Philosophical Studies* 149 (2010): 321–6.
- Zame, William R. 'Can intergenerational equity be operationalized?' *Theoretical Economics* 2, no. 2 (2007): 187–202.

Appendix A. Research areas for future engagement

This appendix indicates additional areas of possible research that would further GPI's mission, but that GPI itself is not working on now or for the immediately foreseeable future, for reasons of capacity and focus.

A.1 Animal welfare

Given the vast numbers of animals (both wild and farmed) that exist, together with the fact that many animals live in conditions far worse than those faced by the typical human, it is natural to suspect that promoting animal welfare may be among the most cost-effective ways of doing good. Assessing this idea raises a number of interesting and unresolved theoretical questions, including about the ways in which we can improve the world vis-a-vis animal welfare and how we ought to prioritise between interventions that improve human lives and interventions that improve non-human animal lives. These questions are currently particularly neglected within academia.

Potential research projects:

- What sorts of entities have the capacity for sentience? Humans, presumably. But what about non-human animals? Insects? The natural environment? Some forms of artificial intelligence? And how can we make welfare comparisons across them? (This questioning of course takes us beyond specifically animal welfare issues, but the case of non-human animals is a natural place to start this investigation.) How should we make interspecies comparisons of welfare? Is brain size a reasonable proxy? If not, how can we do better?

  PHIL - PHILOSOPHY OF MIND, NEUROETHICS **OTHER** - WELFARE BIOLOGY

- Where is the 'zero level' for wellbeing? Which farm animals have lives that are net positive vs net negative? On balance, do wild animals have lives that are net positive or net negative? What are the implications of different population axiologies for this question?

  PHIL - POPULATION ETHICS **OTHER** - WELFARE BIOLOGY

- There are consequence-based reasons to promote the consumption of farm animals with higher welfare over those with lower welfare, because such consumption spurs the creation of animals in similar conditions. For hunted meat (most notably wild fish), on the other hand, consumption simply shortens animal lives and/or reduces populations, so it may be better to eat hunted meat of lower welfare. More generally, what are the ethical implications of farmed vs hunted meat consumption?

  PHIL - POPULATION ETHICS **ECON** - AGRICULTURAL ECONOMICS, ECOLOGICAL ECONOMICS **OTHER** - WELFARE BIOLOGY

- Economic models typically represent animal welfare, if at all, only to the extent that it is represented in human preferences. Can we develop a rigourous 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?

ECON - AGRICULTURAL ECONOMICS, WELFARE ECONOMICS

- To what extent, and on what scales, do various shocks to supply and demand (e.g. increased levels of vegetarianism/veganism, bans on battery cages) affect the number of animals farmed for food (in total and/or under given welfare conditions)?

  ECON - AGRICULTURAL ECONOMICS, INDUSTRIAL ORGANISATION

- To what extent would changes to the farm production of one animal affect the numbers of other (farmed and wild) animals born?

  ECON - AGRICULTURAL ECONOMICS, INDUSTRIAL ORGANISATION **OTHER** - ECOLOGY

- Consider the 'meat eater' problem: interventions that save human lives and/or boost economic growth have obvious direct benefits, but both lead to increases in the consumption and production of animal products. To what extent could this significantly decrease the net positive impact of such interventions, or even imply that they have net negative value? To what extent do positive indirect effects of economic growth push in the other direction? Which of these sets of considerations is larger, when all indirect effects are counted? How, in general, should we think about the impact on animals of improving human lives?

  ECON - AGRICULTURAL ECONOMICS, GROWTH

- What is the case for the claim that improving the living conditions of non-human animals in the wild is among the most cost-effective causes? What tractable activities are there aimed at promoting such improvements?

  ECON - ECOLOGICAL ECONOMICS **OTHER** - WELFARE BIOLOGY

Existing academic literature:

- Cowen, Tyler. 'Policing Nature'. *Environmental Ethics* 25, no. 2 (2003): 169–82.
- Delon, Nicolas, and Duncan Purves. 'Wild Animal Suffering is Intractable'. *Journal of Agricultural and Environmental Ethics* 31, no. 2 (2018): 239–60.
- Lusk, J. L., and F. B. Norwood. 'Animal Welfare Economics'. *Applied Economic Perspectives and Policy* 33, no. 4 (2011): 463–83.
- ———. 'Speciesism, Altruism and the Economics of Animal Welfare'. *European Review of Agricultural Economics* 39, no. 2 (2012): 189–212.
- Malone, Trey, and Jayson Lusk. 'Putting the Chicken Before the Egg Price: An Ex Post Analysis of California's Battery Cage Ban'. *Journal of Agricultural & Resource Economics* 41, no. 3 (2016): 518–32.
- Matheny, Gaverick. 'Least Harm: A Defense of Vegetarianism from Steven Davis's Omnivorous Proposal'. *Journal of Agricultural and Environmental Ethics* 16, no. 5 (2003): 505–11.

- Matheny, Gaverick, and Kai M. A. Chan. 'Human Diets and Animal Welfare: The Illogic of the Larder'. *Journal of Agricultural and Environmental Ethics* 18, no. 6 (2005): 579–94.
- McMahan, Jeff. 'The Moral Problem of Predation'. In *Philosophy Comes to Dinner: Arguments About the Ethics of Eating*, edited by Andrew Chignell, Terence Cuneo, and Matthew C. Halteman. New York: Routledge, 2015. 268–94.
- Ng, Yew-Kwang. 'Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering'. *Biology & Philosophy* 10, no. 3 (1995): 255–85.
- Norwood, F. Bailey, and Jayson Lusk. *Compassion, by the Pound: The Economics of Farm Animal Welfare*. New York: Oxford University Press, 2011.
- Singer, Peter. 'Speciesism and Moral Status'. *Metaphilosophy* 40, no. 3–4 (2009): 567–81.
- Višak, Tatjana, and Robert Garner, eds. *The Ethics of Killing Animals*. New York: Oxford University Press, 2015.

Existing informal discussion:

- Rossa O'Keeffe-O'Donovan and Eva Vivalt, Animal Agriculture and Economics
- 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, 2017 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

A.2 The scope of welfare maximisation

This topic concerns whether impartial welfare maximisation 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.

Potential research projects:

- If it's the case that the long-run effects of one's actions are much larger in impact than the short-run effects, does this strengthen the case for there being strong duties of beneficence, simply because altruistic actions do so much more good than we might have thought?

  PHIL - DUTIES OF BENEFICENCE

- 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'?

  PHIL - DUTIES OF BENEFICENCE, DEONTOLOGICAL CONSTRAINTS

- If there is an obligation to engage in impartial welfare maximisation, what is the nature of that obligation? Should all our resources be spent in whichever way would do the most good? For example, is it that there is an obligation to maximise the effectiveness of whatever sacrifices one makes, but (at least beyond a certain point) no obligation to make the sacrifices?

  PHIL - DUTIES OF BENEFICENCE, CONDITIONAL OBLIGATION

- Do obligations of beneficence require cause-neutrality?

  PHIL - DUTIES OF BENEFICENCE

- 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?

  PHIL - DUTIES OF BENEFICENCE

- It is often claimed that all plausible moral theories recognise a pro tanto reason to promote the impartial good. To what extent does this claim justify the further claim that the project of impartial benevolence, and the associated research questions (as described in this research agenda), are important by the lights of all plausible moral theories?

  PHIL - MORAL UNCERTAINTY, DUTIES OF BENEFICENCE

Existing academic literature:

- Horton, Joe. 'The All or Nothing Problem'. *The Journal of Philosophy* 114, no. 2 (2017): 94–104.
- Mogensen, Andreas L. 'Should We Prevent Optimific Wrongs?' *Utilitas* 28, no. 2 (2016): 215–26.

- Pummer, Theron. 'Whether and Where to Give'. *Philosophy & Public Affairs* 44, no. 1 (2016): 77–95.
- Singer, Peter. *The Life You Can Save: Acting Now to End World Poverty*. New York: Random House, 2009.
- ———. *The Most Good You Can Do: How Effective Altruism Is Changing Ideas about Living Ethically*. New Haven: Yale University Press, 2015.

Existing informal discussion:

- Theron Pummer, People and charitable causes are importantly different things, 1 October 2014

Appendix B. Closely related areas of existing academic research

Here we indicate areas of existing academic literature that serve as particularly relevant background for the topics on this research agenda. Interested researchers who also have background expertise in one or more of these areas are likely to be particularly good fits to GPI's research agenda.

B.1 Methodology of cost-benefit analysis and cost-effectiveness analysis

Cost-benefit analysis and cost-effectiveness analysis are standard tools for evaluating projects. 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 pure time discount rate in trading off costs/benefits incurred earlier against those incurred later, and the use or not of 'distributional weights' (e.g. to account for the fact that a marginal dollar is worth more to a poor person than to a rich person).

Examples of relevant literature:

- Adler, Matthew D. 'Benefit-Cost Analysis and Distributional Weights: An Overview'. *Review of Environmental Economics and Policy* 10, no. 2 (2016): 264–85.
- Bronsteen, John, Christopher Buccafusco and Jonathan Masur. 'Well-Being Analysis vs. Cost-Benefit Analysis'. *Duke Law Journal* 62 (2013): 1603–89.
- Sen, Amartya. 'The Discipline of Cost-Benefit Analysis'. *The Journal of Legal Studies* 29, no. S2 (2000): 931–52.
- J-PAL. Conducting Cost-Effectiveness Analysis.
- HM Treasury. *The Green Book: Appraisal and Evaluation in Central Government*. London: TSO, 2013.

B.2 Multidimensional economic indices

A number of efforts have been made in the last decade or so to come up with macroeconomic measures that capture more than GDP. Some, for example, incorporate 'environmental capital', or value biodiversity loss, in addition to accounting for the resources already under human ownership and in productive use. Relatedly, a literature in development economics focuses on constructing 'multidimensional poverty indices', which define poverty in terms not only of income or consumption, but also other factors for which income may serve as an incomplete proxy: factors such as years of schooling, quality of housing, longevity, or literacy. In general, multidimensional indices are useful for accounting for the full impacts of any set

of interventions, but they are particularly important to the project of comparing interventions across very different causes.

Examples of relevant literature:

- Indices involving environmental capital, etc:
    - Daly, Herman E., John B. Cobb Jr and John B. Cobb. *For the Common Good: Redirecting the Economy toward Community, the Environment, and a Sustainable Future*. Boston: Beacon Press, 1994.
    - Anielski, Mark. 'Measuring the Sustainability of Nations: The Genuine Progress Indicator System of Sustainable Well Being Accounts'. *The Fourth Biennial Conference of the Canadian Society for Ecological Economics: Ecological Sustainability of the Global Marketplace*. 2001.
- The Multidimensional Poverty Index:
    - Alkire, Sabina, and James Foster. 'Counting and Multidimensional Poverty Measurement'. *Journal of Public Economics* 95, no. 7 (2011): 476–87.
- Examples of prior approaches to generating multidimensional poverty indices:
    - Axiomatic:
        - Chakravarty, Satya R., Diganta Mukherjee, and Ravindra R. Ranade. 'On the Family of Subgroup and Factor Decomposable Measures of Multidimensional Poverty'. *Research on Economic Inequality* 8 (1998): 175–94.
    - Information-theoretic:
        - Maasoumi, Esfandiar, and Maria Ana Lugo. 'The Information Basis of Multivariate Poverty Assessments'. In *Quantitative Approaches to Multidimensional Poverty Measurement*, edited by Nanak Kakwani and Jacques Silber. London: Palgrave Macmillan, 2008. 1–29.
    - Fuzzy set:
        - Cerioli, Andrea, and Sergio Zani. 'A Fuzzy Approach to the Measurement of Poverty'. In *Income and Wealth Distribution, Inequality and Poverty*, edited by Camilo Dagum and Michele Zenga. Berlin; Heidelberg: Springer-Verlag, 1990. 272–84.
    - Latent variable:
        - Kakwani, Nanak, and Jacques Silber, eds. *Quantitative Approaches to Multidimensional Poverty Measurement*. London: Palgrave Macmillan, 2008.
- 'Capability approach':
    - Nussbaum, Martha C. *Creating Capabilities*. Cambridge, MA: Harvard University Press, 2011.

- Sen, Amartya Kumar. *Commodities and Capabilities*. Amsterdam: North-Holland, 1985.

B.3 Infinite ethics and intergenerational equity

It is conceivable, and in fact implied by some contemporary cosmological theories, that the universe contains an infinite number of potentially value-bearing entities, such as happy and sad people, and therefore an infinite amount of positive and/or negative value. If no action can affect more than a finite amount of value, it follows in standard cardinal arithmetic that no action can affect the value of the world. This raises the question of how such 'infinitarian paralysis' can be avoided. Alternatively, if some of our actions may have consequences of infinite value, and if we do not render them finite by discounting—that is, if we act on some principle of 'intergenerational equity'—we face the question of how to compare such consequences, or probabilities of such consequences.

Examples of relevant literature:

- Arntzenius, Frank. 'Utilitarianism, Decision Theory, and Eternity'. *Philosophical Perspectives* 28, no. 1 (2014): 31–58.
- Asheim, Geir B. 'Intergenerational Equity'. *Annual Review of Economics* 2, no. 1 (2010): 197–222.
- Basu, Kaushik, and Tapan Mitra. 'Aggregating Infinite Utility Streams with Intergenerational Equity: The Impossibility of Being Paretian'. *Econometrica* 71, no. 5 (2003): 1557–63.
- Bostrom, Nick. 'Infinite Ethics'. *Analysis and Metaphysics*, no. 10 (2011): 9–59.
- Lauwers, Luc, and Peter Vallentyne. 'Infinite Utilitarianism: More Is Always Better'. *Economics & Philosophy* 20, no. 2 (2004): 307–30.
- Vallentyne, Peter, and Shelly Kagan. 'Infinite Value and Finitely Additive Value Theory'. *The Journal of Philosophy* 94, no. 1 (1997): 5–26.
- Zame, William R. 'Can intergenerational equity be operationalized?' *Theoretical Economics* 2, no. 2 (2007): 187–202.

B.4 Epistemology of disagreement

Given our state of uncertainty, many topics within global priorities research will inevitably be subject to disagreement among intelligent and well-informed people. As a result, we must often deal with the question of how to act in the face of disagreement among 'epistemic peers': those of roughly equal competence with respect to the question at hand. This question has been studied extensively both in the abstract and with explicit reference to contentious issues central to global prioritisation, such as the social discount rate.

Examples of relevant literature:

- Christensen, D. 'Epistemology of Disagreement: The Good News'. *Philosophical Review* 116, no. 2 (2007): 187–217.
- Christensen, David. 'Disagreement as Evidence: The Epistemology of Controversy'. *Philosophy Compass* 4, no. 5 (2009): 756–67.
- Christensen, David and Jennifer Lackey, eds. *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press, 2013.
- Elga, Adam. 'Reflection and Disagreement'. *Noûs* 41, no. 3 (2007): 478–502.
- Feldman, Richard, and Ted A. Warfield, eds. *Disagreement*. Oxford: Oxford University Press, 2010.
- Freeman, Mark C., and Ben Groom. 'Positively Gamma Discounting: Combining the Opinions of Experts on the Social Discount Rate'. *The Economic Journal* 125, no. 585 (2015): 1015–24.
- Jouini, Elyes, Jean-Michel Marin and Clotilde Napp. 'Discounting and Divergence of Opinion'. *Journal of Economic Theory* 145, no. 2 (2010): 830–59.
- Wilson, Alastair. 'Disagreement, Equal Weight, and Commutativity'. *Philosophical Studies* 149 (2010): 321–6.

B.5 Demandingness

Maximising consequentialism is sometimes objected to on the grounds that it is overly demanding. For example, going out for dinner at a mid-range restaurant is seen as a permissible option by 'common-sense morality', but such an action is unlikely to have the best consequences impartially considered, and is therefore judged impermissible by maximising consequentialism. Research into the scope of individuals' and institutions' moral obligations toward global welfare maximisation must therefore contend with such demandingness objections.

Examples of relevant literature:

- Scheffler, Samuel. *The Rejection of Consequentialism*. Oxford: Clarendon Press, 1982.
- Kagan, Shelly. *The Limits of Morality*. Oxford: Clarendon Press, 1989.
- Unger, Peter. *Living High and Letting Die*. New York: Oxford University Press, 1996.

B.6 Forecasting

It is difficult to estimate the consequences of some projects empirically, for example investments to reduce risks of low frequency events, or investments in developing new technologies. To evaluate such projects, it is important to use the most reliable forecasting

techniques available, and to understand how to compare the evaluations these techniques produce with our evaluations of projects regarding which there is more direct empirical evidence.

Examples of relevant literature:

- Arrow, Kenneth J., et al. 'The Promise of Prediction Markets'. *Science* 320, no. 5878 (2008): 877–8.
- Hanson, Robin. 'Shall We Vote on Values, But Bet on Beliefs?' *Journal of Political Philosophy* 21, no. 2 (2013): 151–78.
- Helmer, Olaf. *Analysis of the future: The Delphi Method*. No. RAND-P-3558. Santa Monica, CA: RAND Corp, 1967.
- Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. London: Random House, 2016.

B.7 Population ethics

Our relative evaluations of projects across many cause areas depends to a large extent on our understanding of how to compare outcomes in which different groups of individuals may exist. Answers to questions in population ethics appear particularly important regarding questions about the value of extinction risk reduction, about the value of farm animal welfare efforts, and about whether to save or improve lives.

Examples of relevant literature:

- Arrhenius, Gustaf. 'Population Ethics: The Challenge of Future Generations'. Manuscript in preparation.
- Greaves, Hilary. 'Population axiology'. *Philosophy Compass* 12, no. 11 (2017): e12442.
- Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984. Part 4.

B.8 Risk aversion and ambiguity aversion

Our uncertainty about activities' long-term consequences can differ widely by cause area. Risk aversion can therefore substantially affect the decision of whether, for example, to prioritise reductions in existential risk or in near-term suffering. Because the *precision* of our beliefs about long-term consequences can also differ widely, ambiguity aversion can affect our prioritisation decisions similarly. The question of global prioritisation therefore relies heavily on the question of whether, and to what extent, we ought to avoid risk and ambiguity.

Examples of relevant literature:

- On risk aversion:

- Surveys with some empirical emphasis:
    - Chandler, Jake. 'Descriptive Decision Theory'. *The Stanford Encyclopaedia of Philosophy (Winter 2017 Edition)*. *Metaphysics Research Lab, Stanford University* (2017).
    - Fox, Craig R., Carsten Erner, and Daniel J. Walters. 'Decision Under Risk: From the Field to the Laboratory and Back'. In *The Wiley Blackwell Handbook of Judgment and Decision Making*, edited by Gideon Keren and George Wu. Chichester, UK: John Wiley & Sons, 2015. 41–88.
- Recent normative theories and discussions:
    - Buchak, Lara. *Risk and Rationality*. Oxford: Oxford University Press, 2013.
    - Stefánsson, H. Orri, and Richard Bradley. 'What is Risk Aversion?' *British Journal for the Philosophy of Science*, forthcoming.
- Other discussions related to risk aversion in expected utility theory:
    - Allais, Maurice. 'The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School'. *Econometrica* 21, no. 4 (1953): 503–46.
    - Hansson, Bengt. 'Risk aversion as a Problem of Conjoint Measurement'. In *Decision, Probability and Utility: Selected Readings*, edited by Peter Gärdenfors and Nils-Eric Sahlin. Cambridge, UK: Cambridge University Press, 1988. 136–58.
    - Rabin, Matthew. 'Risk Aversion and Expected-Utility Theory: A Calibration Theorem'. *Econometrica* 68, no. 5 (2000): 1281–92.
    - Rothschild, Michael, and Joseph E. Stiglitz. 'Increasing Risk: I. A Definition'. *Journal of Economic Theory* 2, no. 3 (1970): 225–43.
- Cumulative prospect theory and related descriptive theories:
    - Wakker, Peter P. *Prospect Theory: For Risk and Ambiguity*. Cambridge, UK: Cambridge University Press, 2010.

- On ambiguity aversion:
    - Ellsberg, Daniel. 'Risk, Ambiguity, and the Savage Axioms'. *Quarterly Journal of Economics* 75, no. 4 (1961): 643–69.
    - Models of ambiguity aversion:
        - Etner, Johanna, Meglena Jeleva and Jean-Marc Tallon. 'Decision Theory under Ambiguity'. *Journal of Economic Surveys* 26, no. 2 (2012): 234–70.
    - Normative discussion of ambiguity aversion:

- Al-Najjar, Nabil I., and Jonathan Weinstein. 'The Ambiguity Aversion Literature: A Critical Assessment'. *Economics & Philosophy* 25, no. 3 (2009): 249–84.
- Bradley, Richard. *Decision Theory with a Human Face.* Cambridge, UK: Cambridge University Press, 2017.
- Raiffa, Howard. 'Risk, Ambiguity, and the Savage Axioms: Comment'. *Quarterly Journal of Economics* 75, no. 4 (1961): 690–4.
- Siniscalchi, Marciano. 'Two Out Of Three Ain't Bad: A comment on "the ambiguity aversion literature: A critical assessment"'. *Economics & Philosophy* 25, no. 3 (2009): 335–56.
- Voorhoeve, Alex, and Thomas Rowe. 'Egalitarianism under Ambiguity'. Manuscript in preparation.

- Experimental literature (descriptive rather than normative):
  - Trautmann, Stefan, and Gijs van de Kuilen. 'Ambiguity Attitudes'. In *The Wiley-Blackwell Handbook of Judgment and Decision Making*, edited by Gideon Keren and George Wu. Chichester, UK: John Wiley & Sons, 2015. 89–116.

B.9 Moral uncertainty

Attempts to compare the importance of different problems or the effectiveness of different interventions, for example in programme evaluation research in economics, often default to using a utilitarian framework. But, even if one is sympathetic to utilitarianism, it would clearly be overconfident to be *certain* in that moral theory. So, plausibly, we should try to incorporate moral uncertainty into our reasoning when we prioritise among problems. This raises the general question of what form appropriate action under moral uncertainty takes. A framework for action under moral uncertainty is ultimately necessary for resolving questions regarding which causes are most important, given said moral uncertainty; regarding whether and in what way it is permissible to cause harm in the course of doing good; and regarding the extent of individuals' and institutions' obligations toward impartial benevolence (including, for example, benevolence toward individuals in the distant future).

Examples of relevant literature:

- Cotton-Barratt, Owen, William MacAskill and Toby Ord. 'Normative Uncertainty, Intertheoretic Comparisons, and Variance Normalisation'. Manuscript in preparation.
- Greaves, Hilary, and Toby Ord. 'Moral Uncertainty About Population Axiology'. *Journal of Ethics & Social Philosophy* 12, no. 2 (2017): 135–67.
- Gustafsson, Johan E., and Olle Torpman. 'In Defence of My Favourite Theory'. *Pacific Philosophical Quarterly* 95, no. 2 (2014): 159–74.

- Lockhart, Ted. *Moral Uncertainty and Its Consequences*. Oxford: Oxford University Press, 2000.
- Harman, Elizabeth. 'Does Moral Ignorance Exculpate?' *Ratio* 24, no. 4 (2011): 443–68.
- MacAskill, William, Krister Bykvist and Toby Ord. *Moral Uncertainty*. Oxford: Oxford University Press, forthcoming.
- MacAskill, William. 'The Infectiousness of Nihilism'. *Ethics* 123, no. 3 (2013): 508–20.
- Mason, Elinor. 'Moral Ignorance and Blameworthiness'. *Philosophical Studies* 172, no. 11 (2015): 3037–57.
- Ross, Jacob. 'Rejecting Ethical Deflationism'. *Ethics* 116, no. 4 (2006): 742–68.
- Sepielli, Andrew. 'Along an Imperfectly-Lighted Path'. PhD dissertation. New Brunswick: Rutgers University, 2010.
- ———. 'Moral Uncertainty and the Principle of Equity among Moral Theories'. *Philosophy and Phenomenological Research* 86, no. 3 (2013): 580–9.
- Weatherson, Brian. 'Running Risks Morally'. *Philosophical Studies* 167, no. 1 (2014): 141–63.

B.10 Value of information

In situations of uncertainty, information can greatly increase our chances of choosing better actions. The timelines on which we expect to acquire information, the costs of acquiring it, and the extent to which we expect that it will be action-guiding can all affect our decisions concerning, for example, whether to commit resources sooner or later. More generally, considerations regarding the value of information inform the importance we place on the 'option value' of delaying any irreversible development of unknown value, such as human extinction, until after more information has been acquired.

Examples of relevant literature:

- Arrow, Kenneth J., and Anthony C. Fisher. 'Environmental Preservation, Uncertainty, and Irreversibility'. *Quarterly Journal of Economics* 88, no. 2 (1974): 312–9.
- Bishop, Richard C. 'Option Value: An Exposition and Extension'. *Land Economics* 58, no. 1 (1982): 1–15.
- Dixit, Avinash, and Robert S. Pindyck. *Investment Under Uncertainty*. Princeton: Princeton University Press, 1994.
- Eeckhoudt, Louis, and Philippe Godfroid. 'Risk Aversion and the Value of Information'. *The Journal of Economic Education* 31, no. 4 (2000): 382–8.

- Good, Irving John. 'On the Principle of Total Evidence'. *The British Journal for the Philosophy of Science* 17, no. 4 (1967): 319–21.

B.11 Harnessing and combining evidence

When choosing among approaches to working on a particular problem or cause area, individuals and organisations should use empirical evidence to estimate which approach will be most effective. In some fields, for example in development economics, there has been a large increase in the availability of high-quality studies, including randomised controlled trials, estimating the effect of different interventions or programmes. However, it is often not clear how to combine information from different studies, particularly when they were undertaken in different settings or use different empirical methods, even if they are evaluating essentially the same intervention. For other questions of interest, it is inherently more difficult (and sometimes impossible) to run randomised trials, and we must use information from other sources, including theoretical models and other types of empirical evidence, to make informed judgements. General research into how best to harness and combine the available sources of evidence therefore has broad relevance to the enterprise of global prioritisation.

Examples of relevant literature:

- Deaton, Angus, and Nancy Cartwright. 'Understanding and Misunderstanding Randomized Controlled Trials'. *Social Science & Medicine* 210 (2018): 2–21.
- External validity:
    - Bold, Tessa, et al. 'Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education'. *Center for Global Development* Working paper 321, 2013.
    - Dehejia, Rajeev, Cristian Pop-Eleches and Cyrus Samii. 'From Local to Global: External Validity in a Fertility Natural Experiment'. *IZA Discussion Papers* 9300, 2015.
    - Vivalt, Eva. 'How Much Can We Generalize from Impact Evaluations?' Working paper, 2019.
- New approaches to drawing inferences out of sample:
    - Chassang, Sylvain, Padró I. Miquel and Erik Snowberg. 'Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments'. *American Economic Review* 102, no. 4 (2012): 1279–309.
    - Gechter, Michael. 'Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India'. Manuscript in preparation.
    - Gechter, Michael, Cyrus Samii, Rajeev Dehejia and Cristian Pop-Eleches. 'Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference'. Manuscript in preparation.

    - Kowalski, Todd J., Shanu N. Kothari, Michelle A. Mathiason and Andrew J. Borgert. 'Impact of Hair Removal on Surgical Site Infection Rates: A Prospective Randomized Noninferiority Trial'. *Journal of the American College of Surgeons* 223, no. 5 (2016): 704–11.
- Structural modelling:
    - Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 'Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate Progress'. *The Review of Economic Studies* 79, no. 1 (2011): 37–66.
    - Keane, Michael P. 'Structural vs. Atheoretic Approaches to Econometrics'. *Journal of Econometrics* 156, no. 1 (2010): 3–20.
    - Low, Hamish, and Costas Meghir. 'The Use of Structural Models in Econometrics'. *Journal of Economic Perspectives* 31, no. 2 (2017): 33–58.
- Qualitative evidence:
    - Bennett, Andrew, and Jeffrey T. Checkel. 'Process Tracing: From Philosophical Roots to Best Practices'. In *Process Tracing: From Metaphor to Analytic Tool*, edited by Andrew Bennett and Jeffrey T. Checkel. Cambridge, UK: Cambridge University Press, 2014. 3–37.
    - Freedman, David A. 'On Types of Scientific Enquiry: The Role of Qualitative Reasoning'. In *Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady and David Collier. Oxford: Oxford University Press, 2008. 300–18.
    - Goertz, Gary, and James Mahoney. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton: Princeton University Press, 2012.
    - Seawright, Jason. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge, UK: Cambridge University Press, 2016.

B.12 The psychology of altruistic decision-making

Various apparently altruistic and reasonable behaviours seem puzzling on closer inspection, if we assume that the agent is attempting to maximise the expected impact of their actions. These behaviours include (a) donating to more than one charity and (b) avoiding supporting work on mitigating existential risks on the grounds of 'risk aversion'. The same behaviours might make 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 and certain forms of risk aversion).

A better understanding of the variety of psychological mechanisms underlying altruistic behavior might aid efforts to work around behavioural limitations, and maximise the good done by imperfectly altruistic agents.

Examples of relevant literature:

- Andreoni, James. 'Privately Provided Public Goods in a Large Economy: The Limits of Altruism'. *Journal of Public Economics* 35 (1988): 57–73.
- ———. 'Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence'. *Journal of Political Economy* 97 (1989): 1447–58.
- ———. 'Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving'. *The Economic Journal* 100 (1990): 464–77.
- Ashraf, Nava, and Oriana Bandiera. 'Altruistic Capital'. *American Economic Review* 107, no. 5 (2017): 70–5.
- Batson, Charles Daniel. *Altruism in Humans*. New York: Oxford University Press, 2011.
- Batson, C. Daniel, Nadia Ahmad, David A. Lishner and Jo-Ann Tsang. 'Empathy and Altruism'. In *Oxford Handbook of Positive Psychology*, edited by C. R. Snyder and Shane J. Lopez. Oxford: Oxford University Press, 2009. 417–27.
- Bloom, Paul. *Against Empathy: The Case for Rational Compassion*. New York: Harper Collins, 2016.
- Caviola, Lucius, Nadira Faulmüller, Jim AC Everett, Julian Savulescu, and Guy Kahane. 'The Evaluability Bias in Charitable Giving: Saving Administration Costs or Saving Lives?' *Judgment and Decision Making* 9, no. 4 (2014): 303–15.
- Elster, Jon. 'The Valmont Effect: The Warm-Glow Theory of Philanthropy'. In *Giving Well: The Ethics of Philanthropy*, edited by Patricia M. L. Illingworth, Thomas Pogge and Leif Wenar. New York: Oxford University Press, 2011. 67–83.
- Fong, Christina M., and Felix Oberholzer-Gee. 'Truth in Giving: Experimental Evidence on the Welfare Effects of Informed Giving to the Poor'. *Journal of Public Economics* 95, no. 5–6 (2011): 436–44.
- Karlan, Dean, John A. List and Eldar Shafir. 'Small Matches and Charitable Giving: Evidence from a Natural Field Experiment'. *Journal of Public Economics* 95, no. 5–6 (2011): 344–50.
- Null, Clair. 'Warm Glow, Information, and Inefficient Charitable Giving'. *Journal of Public Economics* 95, no. 5–6 (2011): 455–65.
- Oakley, Barbara, Ariel Knafo, Guruprasad Madhavan and David Sloan Wilson, eds. *Pathological Altruism*. Oxford: Oxford University Press, 2011.
- Povey, Richard. 'The Limits to Altruism - A Survey'. Working paper, 2014.
- ———. 'The Socially Optimal Level of Altruism'. Working paper, 2015.

- Simler, Kevin, and Robin Hanson. *The Elephant in the Brain: Hidden Motives in Everyday Life*. Oxford: Oxford University Press, 2017. Ch. 12.
- Small, Deborah A., and George Loewenstein. 'Helping *a* Victim or Helping *the* Victim: Altruism and Identifiability'. *Journal of Risk and Uncertainty* 26, no. 1 (2003): 5–16.
- Smith, Sarah, Frank Windmeijer and Edmund Wright. 'Peer Effects in Charitable Giving: Evidence from the (Running) Field'. *The Economic Journal* 125, no. 585 (2014): 1053–71.

Appendix C. Additional informal discussion

This appendix contains links to additional informal discussion of the themes discussed in this research agenda.

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 effective altruism community are as follows:

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

Finally, here is an incomplete list of some of the most important articles and blog posts from the effective altruism 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 in Global Health
- Carl Shulman, Are pain and pleasure equally energy efficient?
- Carl Shulman and 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