A RESEARCH AGENDA FOR THE GLOBAL PRIORITIES INSTITUTE

Hilary Greaves, Will MacAskill, Rossa O'Keeffe-O'Donovan & Phil Trammell

December 2018

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Table of Contents

Executive Summary

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.

The Global Priorities Institute (GPI) exists to develop and promote rigorous, 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.

Within the very broad class of research questions that this perspective could in principle motivate, GPI's own particular interest and research focus concerns 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.
The first section of our research agenda covers research questions whose motivation is quite specific to longtermism. A second section covers general issues in global prioritisation. This covers issues that are not specific to a longtermist point of view, but that arise more generally for agents engaged in an exercise of global prioritisation, and inter alia are important for longtermists.

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 rigorous foundational research, many of the ideas in question are not yet mainstream in academic circles.

The Global Priorities Institute exists to develop and promote rigorous, 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:

Our Vision
A world in which global priorities are set by using evidence and reason to determine what will do the

most good.

Our 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.

We divide our research agenda as follows.

The first section 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.

The second section is on 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.
The intended audience for this research agenda 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.

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. This is the main focus of GPI's current research.

Articulation and evaluation of longtermism

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. How is this tension best resolved?

  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. 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 rigorous 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. How sensitive is this argument to variations in other evaluative assumptions over which there is reasonable disagreement?

  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
- Should altruists in general be moved primarily by explicit considerations of long-run impact, or are such efforts intractable?

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

Sign of the value of continued existence of humanity

Potential research projects:

- Assess the expected value of the continued existence of the human race. Might this expected value be negative, or just unclear? How do our answers to these questions vary if we (i) assume utilitarianism; (ii) assume a non-utilitarian axiology; (iii) fully take axiological uncertainty into account?

  PHIL - MORAL UNCERTAINTY, POPULATION ETHICS ECON - MEASUREMENT, MODEL UNCERTAINTY
- What is the 'zero level' of human well-being?

  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? 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 - WELFARE ECONOMICS
- 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)? If so, what activities would be best?

  PHIL - MORAL UNCERTAINTY, DECISION THEORY ECON - CATASTROPHIC RISK
- How does moral uncertainty change our assessment of the likely value of the long-run future?

  PHIL - MORAL UNCERTAINTY ECON - DISCOUNTING, MODEL UNCERTAINTY, INTERGENERATIONAL EQUITY

Mitigating catastrophic risk

Potential research projects:

- Is there a fruitful notion of 'existential' risk that is broader than the notion of extinction risk? What is the most fruitful such generalisation?

  ECON - CATASTROPHIC RISK

- Does longtermism lead to the conclusion that existential risk reduction should be the highest priority? Does it further lead to the stronger conclusion that reducing extinction risk should be the highest priority? 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?

  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?

  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? Does any plausible alternative lead away from the idea that the opportunities in question are among the best from an ex ante evaluative standpoint?

  PHIL - DECISION THEORY ECON - DECISION THEORY

- A catastrophic risk can be called 'existential' to the extent that it carries, in expectation, a truly 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?

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

- How, concretely, should we adapt (endogenous) growth models to account for the risks and benefits that growth may pose for the long term?

  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? How do the risks of climate change and the benefits from mitigating them compare with more neglected risks?

  ECON - CATASTROPHIC RISK, ENVIRONMENTAL ECONOMICS

Other ways of leveraging the size of the future

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?
- 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, brain emulation, and nanotechnology) 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?

PHIL - MORAL UNCERTAINTY 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?

  PHIL - ETHICS OF CHAOS ECON - TIME SERIES ECONOMETRICS, MACROECONOMIC THEORY
- Let finitism be the claim that we ought to try to bring about an astronomically large finite amount of value in the future, but not an infinitely large amount of value. 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 can we develop a framework for determining what might be the best activities to pursue?

  PHIL - DECISION THEORY, INFINITE ETHICS ECON - INTERGENERATIONAL EQUITY

Intergenerational governance

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. 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. In delaying an important social decision intergenerationally, however, we may worry that future generations' values and preferences will differ from our own. Facing this tradeoff, under what circumstances should we 'principals' defer irreversible decisions to better-informed future 'agents'? For example, does option value give us good reason to prevent human extinction even if we're unsure about the sign of the value of the future?

  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?

  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. 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. Does this idea make sense? If so, what are the conditions that we

should try to bake into the long reflection?

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)?

Economic indices for longtermists

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, and growth in general may have other negative effects (for instance, on political stability and/or on climate). How radically do these drawbacks render growth an imperfect proxy for expected long-term wellbeing? Is the correlation between consumption growth and long-term well-being even positive, given the current drivers of growth, from a geographical, sectoral and technological perspective?

  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 other promising proxies for long-term well-being? If so, what would be the policy implications of using those proxies in place of economic growth?

  ECON - GROWTH, MACROECONOMIC MEASUREMENT

Moral uncertainty for longtermists

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?

  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?

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

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

Potential research projects:

- Is there any motivation for prioritising these interventions that is respectable from a longtermist perspective?
- To what extent should a worry of 'suspicious convergence' 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 or improving the conditions of caged hens? 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

General issues in global prioritisation

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. If so, it is at least psychologically natural to default to a kind of decision paralysis, under which one does not take any action that has altruistic motivation but some private cost. Is this cluelessness-induced inaction rational? If it is rational, what is the theory of rationality that describes it? If it is not rational, why does the phenomenon occur?

  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?

  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 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?

  PHIL - DECISION THEORY ECON - DECISION THEORY
- What are the implications of ambiguity aversion (whether rational or not) for the project of doing good?

  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?

PHIL - DECISION THEORY, EPISTEMOLOGY

Epistemological issues

The issues in this section can broadly be grouped into two categories. The first concerns the fact that thinking about global prioritisation, particularly (although not only) within the longtermist paradigm, tends to rely on heavily philosophical considerations and to reach some unusual and/or counterintuitive conclusions. To what extent, if at all, should this undermine our confidence in the conclusions in question?

The second cluster of issues concerns the implications of the fact that various forms of progress might render our distant descendants significantly better placed to answer questions of global prioritisation than we are ourselves.

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?

  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
- To what extent should those with unusual views be exceptionally epistemically modest? Should disagreement among peers lead one to decrease one's belief in an unusual view, even if one would otherwise have held it strongly? In such settings, which mechanisms can induce individuals to report their moral views honestly to each other? Should one have the same levels of epistemic modesty about unusual moral views as one should about unusual empirical views?

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

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 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 she 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:

- How should we discount efforts to do good at different points in time, if at all? What is the justification for discounting: a 'preference-based' discount rate, discounting based on uncertainty about the future, forecasted changes in the marginal utility of consumption, or something else? Should we use exponential discounting to value the impact of monetary donations, or some other discount schedule? Is there reason to think that the present is an unusual time with respect to how quickly we ought to discount future donations?

  PHIL - ETHICS OF DISCOUNTING ECON - DISCOUNTING
- Positive returns on investment, and increasing information about where to give, constitute important reasons to consider giving later; rising global output, and accordingly declining opportunities for cost-effective philanthropy, constitute important reasons to consider giving earlier. More thoroughly, what are the considerations that are relevant to the question of when to donate? Can we build an economic quantitative model to represent these considerations?

  ECON - DISCOUNTING, VALUE OF INFORMATION, FORECASTING
- How might 'search theory', in which individuals have to decide whether to commit to taking some opportunity (for example accept a job offer) or hold out for a better opportunity, shed light on the question of philanthropic discounting and when to do good?

  ECON - DISCOUNTING, VALUE OF INFORMATION, SEARCH MODELS
- 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?

  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 perpe-

tuity?

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 an intertemporally neutral perspective? 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

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 hedg-

ing 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 strategies, if any, are successful for individuals? For a large foundation? For the worldwide collection of altruistic actors as a whole?

  ECON - MODERN PORTFOLIO THEORY, 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'? 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?

  ECON - APPLIED MICROECONOMICS
- Philanthropists face uncertainty about the rate at which doing good will grow more costly. 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' 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 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, 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?

  ECON - FINANCIAL ECONOMICS, DIVERSIFICATION

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 rigorous investigation of the properties of the relevant cost-effectiveness distributions.

Potential research projects:

- Establish more rigorously 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.
  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? How does variation of cost-effectiveness within a cause compare to variation of cost-effectiveness between causes? 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?
  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?
  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

Modelling altruism

Economic theory normally 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? What are the characteristics of these settings? 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? Under what assumptions do key results, such as the Fundamental Theorems of Welfare Economics, still hold? 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? 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. Which characteristics of individuals or the choices that they face are associated with different types of apparently altruistic acts?

  ECON - BEHAVIOURAL ECONOMICS

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?

  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.

  ECON - GAME THEORY
- How can results from the mechanism design literature help altruistic individuals and organisations to coordinate in a more effective manner? 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. Which mechanisms can induce individuals to report their beliefs about charitable interventions (approximately) truthfully?

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 (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?

  PHIL - MORAL UNCERTAINTY ECON - GAME THEORY, MECHANISM DESIGN

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

  ECON - MECHANISM DESIGN

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. and governments. 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, lumpy 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?

  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? 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?

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 or approval voting to have on social welfare?

  ECON - POLITICAL ECONOMY, SOCIAL CHOICE THEORY