THE GLOBAL PRIORITIES INSTITUTE

PHILOSOPHY RESEARCH AGENDA

Version 1

November 2024

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Table of Contents

Introduction

The Global Priorities Institute (GPI) is an interdisciplinary research institute at the University of Oxford. Our mission is to conduct and promote world-class, foundational academic research on how to do good most effectively. This research agenda sets out some research topics that seem particularly promising to the philosophy team at GPI. It is not meant to cover all possible areas where we would be excited to see new work. Indeed, we hope this agenda will inspire reflection on what further areas of research might fit within our mission. Philosophers from a wide range of backgrounds may have relevant expertise; our current work draws on ethics, political philosophy, epistemology, philosophy of science, decision theory, and philosophy of mind and cognitive science.

In line with our mission, the topics we will discuss are chosen because they are important for the project of *ethical prioritisation.* There are many morally important problems in the world. It is impossible for any single individual or institution to solve them all. One must therefore prioritise, both among the problems themselves and among the means for tackling them. And prioritisation requires careful analysis. It will often involve thinking at the margin (roughly, asking what changes we each can make, given how things are) and making comparative judgements (asking not only what is important, but what is *most* important). While much of our work is deeply theoretical, we aim to address issues that are crucial to this practical project. As part of that aim, we try to focus on particular questions on which we think it is possible to make valuable progress, and on which good work might not otherwise be done.

We are especially interested in prioritisation from an impartial point of view – that is, taking *everyone's* interests into account, without a presumption about which causes are important. (This is one sense in which the institute's focus is 'global'; a second sense is that impartiality often leads us to focus on topics with the potential for global impact.) We make no assumption that the whole of ethics is impartial; for example, it may well be that much of one's day-to-day life is properly governed by personal interests, duties to one's own community, and so on, rather than by impartial considerations. Still, the impartial perspective strikes many as a particularly important part of morality that is often decisive. It is also sometimes explicitly considered by individuals and institutions, for example when thinking about charitable giving, volunteering, activism, career choice, and global governance. In any case, many of the topics in this agenda are also likely to be of interest to philosophers who attach little or no importance to purely impartial considerations as determinative of how we ought to live. Meanwhile, questions about the limitations of the impartial

perspective are certainly relevant to ethical prioritisation, and play some role in GPI's research, but are not the main focus.

Concretely, this agenda contains four parts. The first three lay out major areas of research. The fourth section lists a number of issues that cross-cut or supplement the first three. Each section is designed so that it can be read independently. While the agenda can be read from front to back, it should also be possible for a reader to jump straight to the sections that they find particularly interesting. Here's a sketch of the main themes.

Section 1 concerns the long-term future.[1] The impartial perspective suggests that the interests of future people count as much as our own – even the interests of people millennia from now. Given the potential scale of the future, our effects on the long-term future could be crucial to prioritisation. But how far into the future does our predictable influence reach, and how should we evaluate its effects?

Section 2 describes our research programme in the philosophy of mind and well-being. To adopt an impartial perspective we would want to know which beings merit moral consideration. In particular, what mental properties underpin moral status, and what kinds of beings have them? Getting this right may be crucial to prioritisation insofar as there are vast numbers of non-human candidates for moral status, including animals and, perhaps, future digital minds.

Section 3 considers developments in artificial intelligence (AI). Some predict that AI will soon transform the world, on a scale comparable to the Industrial Revolution. Our focus is on research that will help us understand and navigate the largest-scale risks and opportunities presented by AI, including such catastrophic risks as human extinction or persistent dystopian outcomes.

There are important interactions between these areas of research. For example, developments in AI may lead to the creation of new forms of moral patients, while having a long-lasting impact on the trajectory of human civilisation. And our views about who counts as a moral patient can radically alter our views about how well the future will go. In addition to this substantive overlap, there are a range of foundational philosophical issues that are relevant to these topics, and to the general project of prioritisation from an impartial perspective. Some of these we list in section 4, along with a number of other topics that complement our main research foci.

[1] The core of GPI's previous research agenda focused on 'the longtermism paradigm'; the present section 1 is a revised version of that material, focusing especially on philosophical issues. We feel that 'the longtermism paradigm' was a somewhat confusing title, given that 'longtermism' is also the name of a specific view, which we still consider in section 1.4.

In none of these areas do we think of ourselves as collectively advocating a particular view; our focus is on topics we consider to be worth reflection, not on arguing for one particular side of a given debate. Indeed, there is significant variation of views within GPI about which research areas are most promising, and also about basic normative and empirical premises.

Finally, we invite questions and feedback about the agenda. We are always excited to hear from researchers working on related questions.

1 The Long-Term Future

It is plausible that the vast majority of humans who will ever live have not yet been born. Their existence and the conditions under which they will live depends, in part, on what we do today. The same goes for other forms of life and other things of value that may persist for thousands, millions, or billions of years into the future: whether they do so, and how they do so, is partly up to us. From an impartial perspective, then, the sheer numbers suggest that our impact on the long-term future *might* be an important or even decisive consideration, rivalling or even outweighing any short- or medium-term considerations. Yet most policy discussions focus only on the next few years, or, at most, the next few decades.

Thus, one major area of research at GPI is to examine how our effects on the long-term future bear on questions of prioritisation. There are deep evaluative and normative questions about how to assess our actions and policies in terms of their long-term effects, and there are formidable epistemic issues when it comes to thinking about and predicting those effects. There are also more concrete questions about what a focus on the long-term future would actually amount to. For example, should the focus be on reducing the risk of near-term human extinction or on establishing durable and beneficent political institutions? What do the trade-offs actually look like between what's best from a long-term perspective and what's best for the next few generations? And, when it comes to deploying scarce resources, will long-term considerations often prove decisive?

See also GPI's research agenda in psychology and section 2.5 of GPI's research agenda in economics.

1.1 Evaluative Issues

How should we evaluate actions and policies in light of their effects on the long-term future? In approaching this question, we must take into account our uncertainty about those effects; interesting questions also arise about the role of uncertainty and disagreement about normative principles. Normative theory will play a starring role here, but we are especially interested in the points where normative theory may have a significant impact on how we evaluate the long-term effects of our actions, rather than in pursuing normative theory for its own sake or with other applications in mind.

1.1.1 Welfare and Beneficence

- **What do the most plausible theories of beneficence say about our reasons to influence the far future?** A main source of motivation for thinking about the long-term future comes from concern for the welfare of future people. But

what does such concern properly amount to? For example, theories that are *prima facie* most likely to minimise the importance of long-term effects include *person-affecting* views in population ethics and *lexical*, *anti-aggregative* or *partially aggregative* views in distributive ethics. What do the most plausible versions of these views say about our actual predicament with respect to long-term beneficence (Thomas 2023; Heikkinen 2022; Curran forthcoming)?

- **What is the critical level for welfare?** One of the main effects we can have on the long-term future is changing the number of people who ever live. Theories that evaluate changes in population size typically specify a 'critical' or 'neutral' welfare level, below which we have reasons to avoid creating additional lives (Blackorby et al. 1995; Ng 1986; Broome 2004, ch. 10). According to some theories, we also have positive reasons to create lives above this level. What kinds of human and non-human lives are above the critical level (Blackorby et al. 2005, ch. 5; Cockburn et al. 2014; Cowie 2017; Williamson 2021; Mogensen forthcoming)? What does this tell us about the value of different long-term interventions?
- **What is the correct view about pure time preference?** "Pure time preference" refers to the degree to which the welfare of future generations is discounted in evaluating policy outcomes. There is already a substantial literature (on both sides) evaluating the claim that public policy should adopt a zero rate of pure time preference (see Cowen & Parfit 1992; Greaves 2017; Dasgupta 2008). Given the importance of this claim to evaluating long-term effects, further research that changes the balance of arguments on this question could still be valuable. What more, if anything, can be said on the matter (e.g., Lloyd 2021; Mogensen 2022; Saad 2024*c*)?
- **Should all indirect effects count?** The effects of our actions on the very long-term future are often indirect and perhaps even unintended. Some philosophers hold that, when deciding whom to aid, we are generally morally constrained to consider only the direct impact of our actions on those we can help, as opposed to the indirect impact of helping some rather than others (Kamm 1993; Brock 2003; Lippert-Rasmussen & Lauridsen 2010; Du Toit & Millum 2016). On the other hand, indirect effects may often be much larger in scale than direct ones. How is this tension best resolved, and does it undermine the case for focusing on long-term effects (Lenman 2000; Mogensen 2020; Gardner 2021)?
- **What if the universe is infinite or we can have infinite effects?** Once we start thinking about the very long-term future and about the universe as a whole, the possibility becomes salient that the world may contain infinitely many welfare subjects or other loci of moral concern. The evaluation of such infinite worlds, and of actions affecting them, is rife with difficulties; moreover, many otherwise-plausible decision-theoretic principles suggest that these

difficulties can arise if there is even the tiniest probability of an infinite world. What should we make of this? And insofar as these problems can be resolved, should our main priority be to seek out infinite payoffs (see van Liedekerke 1995; Vallentyne and Kagan 1997; Bostrom 2011; Arntzenius 2014; Askell 2018; Meacham 2020; Wilkinson 2021; Pivato 2023; Tarsney & Wilkinson forthcoming)?

- **Is there an important asymmetry between good and bad?** There are a number of possible asymmetries between, broadly speaking, good and bad outcomes. At a theoretical level, there may be important asymmetries in how we evaluate benefits versus harms, good versus bad lives, or (perhaps because of risk aversion) the best worlds versus the worst. More empirically, it may be that the worst plausible trajectories are worse than the best plausible ones are good; or it may be that avoiding the worst outcomes is more neglected. What are the implications of the most plausible views in this space when it comes to thinking about the long-term future (Hurka 2010; Mogensen & MacAskill 2021; Unruh 2023; Thomas 2023; Mogensen 2023a, 2024; Pettigrew 2024; MacAskill ms-a)?

See also section 4.3.1 (Population Ethics and Aggregation) and section 4.1.3 (Infinities) of this research agenda, as well as section 1.3 (Welfare and Decision Procedures) of GPI's research agenda in economics and section 8 (Future Wellbeing) in GPI's research agenda in psychology.

1.1.2 Non-Welfare Considerations

- **What are the most important considerations, besides welfare, when thinking about the long-term future?** How strong are these considerations? For example, does the extended persistence of life, of humanity, or of civilisation have intrinsic value that makes a difference to prioritisation (Frick 2017; Scheffler 2018)? Does the potential existence of future justice or injustice affect how one should evaluate near-term human extinction (Barrett 2022; Mogensen 2023a: pp. 40–1; Schmidt and Barrett forthcoming)?
- **How might future agents be different from us, and does it matter?** Future agents will have different information, abilities, preferences, values, and moral views from our own. Does this have implications for how we assess the long-term future and our influence on it (Riedener forthcoming)? For example, to what extent should we defer to the views of future agents, preserve option value, or hope for moral progress (MacAskill ms-a)? On the other hand, are there strong reasons to prefer the continued existence of humanity to its succession by non-human life or even by digital minds?
- **What is the role of long-term thinking in political theory, and vice versa?** Do states have special obligations to their present citizens? Should beneficence

towards future generations be an important factor in the design of political institutions (Barrett 2022)? How do existing discussions of feasibility, idealisation, and utopianism in political theory bear on whether and how we should prioritise long-term effects (Sen 2009; Lawford-Smith 2012; Estlund 2019)? And, given that feasibility constraints in politics may weaken over very long timescales, does a long-term perspective support a renewed role for utopian political theorising? Or does it weigh against a focus on utopian blueprints, in favour of designing open, exploratory institutions, best able to capitalise on anticipated future improvements in values and information (Gaus 2016; Barrett 2020)?

See also section 4.3.2 (Non-Consequentialism and Moral Prioritisation), section 4.4.4 (Institutions), and section 7 (Policy and Institutional) of GPI's research agenda in psychology.

1.1.3 Decision Theory and Normative Uncertainty

- **How do risk-aversion, ambiguity-aversion, and other standard decision-theoretic concerns affect long-term prioritisation?** We face a great deal of uncertainty about the long-term future (for more on which, see section 1.2). How should we respond to this uncertainty, whether it is appropriately represented probabilistically or not? Would differences in risk or ambiguity-aversion change our priorities with respect to the long-term future? Should we take into account the risk or ambiguity attitudes of future people, and, if so, how (Buchak 2019, 2023; Mogensen 2023a; Thoma 2023; Greaves et al. 2024; Pettigrew 2024)?
- **Is concern for the long-term driven by small probabilities of very large effects?** The case for focusing on the long-term future may sometimes depend on small probabilities of achieving very large and persistent effects. How general is this phenomenon, and what is the correct way to evaluate low-probability, high-payoff opportunities (Tarsney 2020, 2023b; Wilkinson 2022c; Beckstead & Thomas 2024; Russell 2023)? For example, does expected utility theory provide an adequate account? Is there anything to the intuition that such opportunities should often be discounted in value (Smith 2014; Monton 2019; Kosonen 2022, 2023; Cibinel 2023; Tarsney 2023a), and, if so, what are the implications for how we evaluate interventions in the long term?
- **Given that different moral views vary in the importance they give to long-term effects, how should we take into account uncertainty and disagreement about which moral view is correct?** For one concrete example, how should policymakers and expert advisors take into account uncertainty about the appropriate social discount rate, when different discount rates can

lead to wildly different assessments of long-term effects (MacAskill et al. 2020; Millner 2020; Jaakkola and Millner 2020)?

- **What should we make of meta-normative fanaticism?** Under some broadly plausible ways of handling moral uncertainty, one often ends up acting 'fanatically', that is, in accordance with moral theories that claim the stakes are very high, even if one thinks these theories are unlikely to be correct. Insofar as some moral theories, like perhaps total utilitarianism, give enormous importance to long-term effects, do considerations of moral uncertainty mean that long-term effects are important by default? Is there a good theoretical way to avoid this form of fanaticism? (Ross 2006; Greaves & Ord 2017; MacAskill & Ord 2020; Greaves & Cotton-Barratt 2023; Baker 2024)?

See also section 4.1 (Decision-Theoretic Issues) and section 4.3.3 (Moral Uncertainty) of this research agenda and section 1.3 (Welfare and Decision Procedures) of GPI's research agenda in economics.

1.2 Epistemic challenges

This section raises some epistemic questions relevant to evaluating the long-term effects of our actions. Even if such effects may be important in principle, they may not be important in practice if we cannot influence the future in sufficiently predictable ways. In general, what is our typical epistemic position with respect to the long-term effects of our actions? Can we improve it, and how? Are we afflicted by sufficient cluelessness, uncertainty, and/or unawareness about the future to undermine the importance of long-term effects?

1.2.1 Cluelessness, Unawareness, and Deep Uncertainty

- **Does a lack of evidence leave us 'clueless' about long-term effects in a way that undermines or otherwise affects prioritisation?** Faced with the task of comparing actions in terms of their long-term effects, it often seems that the agent is *clueless*: that the available empirical and theoretical evidence simply supplies too thin a basis for guiding decisions in any principled way (Lenman 2000; Greaves 2016; Mogensen 2021). How is this situation best modelled, and what is the rational way of making decisions when in this predicament? Do the implications of cluelessness systematically favour some types of action over others?
- **What about 'complex cluelessness' from messy bodies of evidence?** Predicting the long-term effects of our actions often requires us to make difficult comparisons between complex and messy bodies of competing evidence, a situation sometimes described as *complex cluelessness* (see Greaves 2016). In this situation, it seems that we are not merely uncertain about the

effects of our actions, but cannot even assign the correct probabilities to them – perhaps our evidence underdetermines the correct probabilities, or as bounded epistemic agents we cannot access them. How best can we characterise this situation, and what is the rational way of responding to it? For example, does rationality require that we adopt *imprecise* subjective probabilities, and what does such imprecision imply for the evaluation of actions aimed at improving the long-term future (see, e.g., Mogensen 2021)?

- **How is prioritisation affected by our unawareness of future possibilities?** The long-term future will plausibly be shaped by events or phenomena that we have never considered and perhaps cannot fully imagine: e.g., scientific discoveries of which we do not currently conceive (Stanford 2001; Ruhmkorff 2011; Deutsch 2011). Unforeseen scientific discoveries have certainly had enormous effects on the course of history over recent centuries. What is the rational response to this sort of *unawareness* (e.g. Bradley 2017; Steele and Stefánsson 2021; de Canson 2024), and what does it imply for the evaluation of actions aimed at improving the long-term future?
- **Are there decision-procedures or heuristics that would allow us to handle deep uncertainty and disagreement about the far future in a principled way?** Given our poor epistemic position and our limited deliberative capacities, anything akin to an explicit expected value calculation may be untenable as a decision procedure (Feldman 2006). In everyday life people apply heuristics to make decisions in a way that does not take into account all available information and that does not depend on sophisticated probabilistic reasoning (see, e.g., Gigerenzer & Gaissmaier 2011). Similarly, a variety of decision-procedures and decision-framing strategies have been developed to deal with "deep uncertainty" in policy contexts (see, e.g., Helgeson 2020; Marchau et al. 2019), but usually with a time-horizon of mere decades. Are there good heuristics, procedures, or strategies for thinking about the long term future (Thorstad & Mogensen 2020; Mogensen & Thorstad 2022)?

See also section 4.1.2 (Severe Uncertainty).

1.2.2 Reading the Evidence

- **Which current gaps in our knowledge regarding the very long term are particularly action-relevant?** In which scientific field or other domain could these gaps be closed by the accumulation of further empirical evidence?
- **Are there special cases of interventions whose long-term effects are relatively predictable, or about which we can have unusually good evidence?** Some events may tend to 'lock in' significant, highly persistent,

hard-to-reverse effects (see MacAskill 2022b; Greaves & MacAskill forthcoming). For instance, if humanity goes extinct, we can be almost certain that there will be no more human life after that. At least in this respect, efforts to reduce extinction risk may be relatively easy to evaluate. Are there other examples of this kind, and how confident should we be in evaluating their long-term effects overall?

- **Which characteristics of scientific theories undercut their predictions in the very long term, such that we should put less stock in them?** For instance, does the *reflexivity* or *fragility* of a theory's predictions (Northcott 2022), or the narrowness of the circumstances in which the theory has been tested, mean that we should greatly discount its long-term predictions? Conversely, are there classes of future events or phenomena for which we can be confident of our current scientific understanding? In particular, are there cases in which we can be confident that we are not unaware of important possibilities (Vickers 2023)?
- **Is our position in history, or our very existence, somehow surprising, and should this affect our reading of the evidence via 'anthropic reasoning'?** For instance, would it be surprising to find that we live early in humanity's history, or at a particularly pivotal moment? Does this make it unlikely that humanity will have a long future (Carter 1983; Gott 1993; Leslie 1996; Bostrom 2002; Mogensen 2019a; Thomas 2021a), or that we can have a large impact on how the long-term future goes (MacAskill 2022a; Mogensen 2023b)? On the other hand, is a long future for humanity more likely, insofar as this would make it less surprising that we exist at all (Bartha and Hitchcock 1999)? Or, then again, does survivorship bias mean that we tend to underestimate catastrophic risk (Ćirković et al. 2010; Thomas 2024)?
- **What beliefs and probabilities should we adopt concerning background events?** On certain normative views, evaluations of one's actions depend not only on their forward-looking effects but also on one's beliefs or uncertainty about the rest of the world outside of one's causal future. This can make a difference to how we evaluate and reason about the far future (see, e.g., Tarsney 2020; Tarsney & Thomas 2024). What beliefs and probabilities should a rational agent adopt with respect to morally relevant events in regions of the universe unaffected by human action? What kinds of arguments and evidence can be adduced and how robust are they?

See also section 4.2 (Epistemological Issues).

1.3 Comparing interventions in terms of their long-term effects

In section 1.1 and section 1.2 we raised some relatively abstract and general questions in normative theory and epistemology that seem particularly relevant to making choices that affect the long-term future. Here we raise more fine-grained questions, from both practical ethics and practical epistemology, of how to compare options by their long-term effects, especially the types of options that might be available to real-world agents. The overall guiding questions are, in particular, which types of available interventions have the best long-term effects, and whether their effects are overall better than interventions with large short-term effects.

1.3.1 Extinction and Other Catastrophic Risks

- **What is the overall picture – empirically and normatively – with respect to different varieties of catastrophic risk?** Just how high are the probabilities that particular threats (e.g., pandemics, nuclear war, unaligned artificial intelligence) lead to near-term human extinction or other large-scale catastrophes (Snyder-Beattie et al. 2019; Ord 2020; Beard et al. 2020)? Are there additional moral considerations for focusing on mitigating some such threats rather than others? For example, some threats might be harder than others to mitigate without significant intrusions on liberty (cf. Bostrom 2019); some threats might generate special obligations because of their provenance or because of disparate impacts (Dasgupta 2008).
- **How important is extinction risk mitigation?** One potentially very important way that we can affect the long-run future is by changing the probability of near-term human extinction. How should we evaluate such changes? In particular, do person-affecting views or other views in population ethics or decision theory undermine the practical case for extinction risk reduction (Thomas 2023; Pettigrew 2024; Wilkinson 2023; Kowalcyzk and Venkatesh 2024)?
- **How should we expect extinction-level risk to change over time?** Across a range of methods of influencing the long-term future, the value of doing so is plausibly greater the longer humanity will survive. Given that we appear to face risks of near-term extinction (Ord 2020), for humanity to survive very long may require that such risks lessen greatly over time. Just how much should we expect those risks to lessen, and how does this affect the value of various ways we might attempt to improve the long-term future (Ord 2020; Thorstad 2023, forthcoming-c).

See also section 2.3.1 (Catastrophic Risks and Their Mitigation) and section 3.1 (Catastrophic Risk from AI) in this research agenda, section 2.1 (Economics of

Catastrophes) in GPI's research agenda in economics, and section 9 (Cause-Specific Issues) in GPI's research agenda in psychology.

1.3.2 Trajectory Changes

- **What kinds of trajectory changes are feasible, and how important would they be?** Some philosophers (MacAskill 2022b; Wilkinson 2023; Greaves & MacAskill forthcoming) have suggested that, instead of focusing on the probability of human extinction, there are other promising methods of positively influencing the long-term future: namely, *trajectory changes*, or improvements to the future conditional on the survival of humanity. Should agents focus on bringing about trajectory changes rather than influencing the risk of extinction? And, in practice, which trajectory changes are most valuable taking uncertainty into account?
- **What is the long-term value of economic growth?** As a trajectory change, how morally valuable would it be to increase long-term economic growth (Cowen 2018)? Should we expect long-term increases to GDP per capita to have a positive or negative effect on overall welfare (Stevenson and Wolfers 2008; Jones and Klenow 2016; Easterlin 2017), especially when taking into account the welfare of non-human animals (Frank 2008; Olsson and Alexandrie 2019)? See also section 2.4 (Population, Inequality, and Long-Term Welfare) of GPI's research agenda in economics.
- **Should we prioritise the reduction of large-scale future suffering and other forms of harm?** How should we evaluate efforts to reduce the probability of scenarios in which the long-term future contains enormous amounts of suffering or other forms of extreme disvalue (Tomasik 2015; Saad and Bradley 2022; Dung forthcoming)? In practice, how do such efforts compare to other attempts at bringing about trajectory changes or at mitigating extinction risks? Is there a compelling case to be made that pursuing one of these forms of intervention is typically far better than pursuing the others?
- **What kinds of institutional reform provide a route to long-term impact?** One general method of attempting to influence the long term is by reforming present-day political institutions. For instance, we could perhaps enfranchise future generations in our current political institutions (Kavka and Warren 1983; Goodin 2007; Tännsjö 2007; Beckman 2009: Ch. 7). Or we could perhaps implement age-weighted voting (Gonzalez-Ricoy and Gosseries 2017). How strong is the case in favour of these and other institutional changes? How could such changes best be implemented? What other democratic reforms would be most promising to better align political outcomes with long-term priorities? And which such reforms would be feasible in practice?
- **How valuable would it be to attempt to influence the values held by future generations?** Without our intervention, what values will they likely hold? In

particular, should we expect humanity to converge on the correct moral theory given enough time (Gustafsson and Peterson 2012; Vallinder and Olsson 2013; MacAskill ms-b)?

- **How much emphasis should we put on accumulating resources for future use?** How robust is the moral case for *patient philanthropy*: setting aside resources, perhaps investing them to accumulate over time, for future generations to use for what then seems most valuable (Trammell 2021)? Do considerations of, say, autonomy, paternalism, or robustness favour such patient philanthropy over other methods of influencing the long-term future?

See also section 2.3.3 (How Might the Future Go Well?), section 3.2 (AI and the Trajectory of Civilisation), and section 4.4 (Other Cause Areas).

1.4 Longtermism

The sheer scale of the future may suggest that the long-term effects of our actions should often be a key consideration when it comes to prioritisation (Greaves et al. forthcoming). Call this general claim *longtermism*, noting that it could be made precise in a number of different ways. Given that little attention is ordinarily paid to such long-term effects, longtermism is potentially a highly revisionary view. This warrants further examination. The types of issues we have raised in sections 1.1-1.3, and in other parts of this agenda, are relevant to this project, and some natural ways of resolving them seem to support a strong form of longtermism; here we pose some further questions related to longtermism as such.

- **In general: what are the most perspicuous ways to formulate longtermism?** How robust is the case for each of these views, and what are the most compelling objections? What are the practical implications? For example, take the "axiological strong longtermism" of Greaves & MacAskill (forthcoming): that, in the most important decisions facing agents today, the best options are among those that are best for the long-term future, and deliver far greater benefits to the long-term future than to the near-term future. Is this a good way to frame the issues, and is it true?
- **Weak or strong?** Alternatively, a weak form of longtermism might claim that the very long-term effects of our actions are *at least an important* determinant of which of our available actions are best (MacAskill 2022b; Barrett ms). How strong is the case for weak forms of longtermism such as this? And, conditional on such a weak view holding, how strong is the case for the corresponding strong view?
- **Axiological or deontic?** Similarly, longtermism can be expressed in either an axiological form (as above) or a *deontic* form, e.g.: that, in the most important decisions facing agents today, we *ought* to choose the action that has the best

effect on the very long-term future (see Greaves & MacAskill forthcoming). How strong is the case for an axiological form of longtermism? And, conditional on such an axiological view holding, how strong is the case for a corresponding deontic view? Are there particular deontic constraints or other considerations that routinely arise and weigh against the sorts of options longtermists favour (see, e.g., Mogensen 2019b)? On the other hand, are there particular deontic considerations that would put more emphasis on future generations?

- **Institutional or individual?** Longtermism can be expressed in either an individual form (as above) or a political form, e.g.: which present-day *institutions* we should establish and/or maintain is primarily (or at least partially) determined by very long-term considerations (Schmidt and Barrett forthcoming). Likewise, perhaps long-term considerations are particularly relevant (or irrelevant) to certain institutional agents, including governments. How does the case for an institutional form of longtermism compare to the case for an individual form? Do considerations of legitimacy, justice, and democracy speak in favour or against institutional longtermism (*ibid.*)?
- **To what decisions does longtermism apply?** If the above form of longtermism concerns the *most important* decisions facing agents today, how do the very long-term effects of our actions bear on *other* decisions? Do long-term considerations determine what is best in a broader range – perhaps even all – of the decisions we face (Greaves & Tarsney forthcoming; Thorstad forthcoming-d)?
- **How robust are various forms of longtermism to differences in moral, epistemological, or decision-theoretic views?** For instance, we might ask whether longtermism requires some particular, controversial population axiology, such as total utilitarianism. Or might it be robustly supported across a range of plausible approaches to population ethics (Thomas 2023; Tarsney & Thomas 2024)? Likewise, are various forms of longtermism only plausible on decision theories that embrace fanaticism (Tarsney 2023b; see also section 1.1.3)? And, even where the case for longtermism is robust to different theories, do those different theories support different conclusions about what interventions we should prioritise in practice (Heikkinen 2022; Thomas 2023; Buchak 2023; Pettigrew 2024; Curran forthcoming)? See also section 4 (General Issues Related to Prioritisation).
- **Is there convergence between long-term and short-term considerations?** Even if the long-term effects of our actions generate weighty considerations, there might be some degree of convergence between what's best in the long term and what's best in the short term. For example, reducing extinction risk might be very important from both points of view (Shulman & Thornley forthcoming), as might establishing institutions that empower both present

and future generations. Are there general reasons to expect such convergence (Williams 2012), or to be especially suspicious of convergence claims? Are there other reasons (e.g., in light of moral disagreement) to favour options that do well from both points of view (see also section 4.3.3 (Moral Uncertainty)?

2 Mind and Value

This part of the research agenda outlines some questions about mind and value that are of particular interest to philosophers at the Global Priorities Institute. These primarily concern the nature of mental phenomena, their distribution, our epistemic access to them, and their moral significance. We have selected questions based on their ability to inform decision-makers about the distribution of morally significant mental properties across individuals, with particular emphasis on non-human candidates for moral standing.

2.1 Which Mental Phenomena Are Morally Significant?

2.1.1 Which Mental Phenomena Contribute to Moral Standing?

GPI is interested both in foundational issues concerning the basis of moral standing and in applied questions about the prospects for moral standing in different sorts of systems. A particularly important foundational issue concerns the relationship between moral standing and *sentience*, understood as the capacity for valenced experience. Pains and pleasures are paradigmatic valenced experiences, but the category includes much else besides, such as experiences of the sublime and feelings of nausea. Questions raised by this issue include:

- **How plausible is the widely held view that there is a close tie between moral standing and sentience** (e.g., Singer 1993; Korsgaard 2018; Nussbaum 2022)? See also section 1.1 (Understanding People's Values) of GPI's research agenda in psychology.
- **What is the best version of the view that there is a close tie between moral standing and sentience?** Granting that there is such a tie, should we think that sentience *per se* is what matters for moral standing? Or should we instead think that moral standing is in the first instance tied to a phenomenon that often correlates with sentience, such as the capacity for potentially-motivating experience (Roelofs 2023)?
- **How plausible should we find particular views on which there is not a close tie between moral standing and sentience?** For example, some hold that the capacity for (phenomenal) consciousness rather than sentience confers moral standing (Chalmers 2022). Others maintain that a functional capacity associated with consciousness or sentience might confer moral standing even in the latters' absence (Hill 1991: 73; Levy 2014; G. Lee 2019; Sinnott-Armstrong and Conitzer 2021). Still others claim that the basis of moral standing lies in other mental phenomena such as certain kinds of desires or agency that are not tied to the capacity for consciousness (Carruthers 2019: 171-4; Kagan 2019: 23-30; Bradford 2022; Goldstein and Kirk-Gianini ms). Which of these views is best?

- **How strong are challenges to views on which moral standing is not closely tied to sentience or consciousness?** For example, one neglected challenge can be found in arguments for the *phenomenal intentionality thesis* that mental states are generally at least partly grounded in consciousness or phenomenal dispositions (Bourget and Mendelovici 2019 and references therein). Others can be found in arguments for the phenomenal individuation of certain types of mental states such as beliefs (Schwitzgebel 2002), for the normative significance of desires having its source in associated affective-phenomenal roles (Smithies and Weiss 2019), and for all epistemic justification that attaches to mental states tracing to consciousness or accessibility to consciousness (Smithies 2019).

Each of these views of moral standing can recognise the moral standing of humans. But they offer differing verdicts about some non-human minds. This divergence invites further questions such as:

- **If the capacity for consciousness confers moral standing, how likely are different AI systems to have moral standing?** For example, how likely are near-term AI systems or cerebral organoids to have moral standing (compare: Butlin et al. 2023; Chalmers 2023; Birch forthcoming)?
- **How does the hypothesis that the capacity for consciousness confers moral standing bear on how far the moral circle extends into the animal kingdom?** If we suppose instead that a thin variety of desire – one not dependent on the capacity for consciousness – suffices for moral standing, how would that affect the prospects for moral standing in these sorts of non-human minds?
- **What does a reasonable spread of credences over live views about the basis of moral standing suggest about its distribution?**

2.1.2 *How Are Mental Phenomena Otherwise Morally Significant?*

Plausibly, mental phenomena realise not only moral standing but also many other morally significant properties, including well-being, ill-being, and the possession of rights. GPI is interested in various questions concerning which mental phenomena realise which moral properties.

- **How important are the differences among theories of welfare for setting priorities?** While leading philosophical theories of well-being may deliver very similar evaluations of many of the kinds of lives that actually exist, these theories disagree about how different mental phenomena contribute to welfare (Lin 2022). Psychological profiles that lead to very different evaluations by different theories of welfare may become more prevalent in future as AI systems come to inhabit unfamiliar regions of the space of possible minds: for example, exhibiting sophisticated cognitive capacities without any capacity for valenced conscious experiences.

- **What are the kinds and sources of ill-being?** At present, we also have only a limited understanding of intrinsic sources of ill-being, especially sources besides pain or felt unpleasantness (Kagan 2014; Sumner 2020; Bradford 2021; Pallies 2022). Do gaps in our understanding of ill-being affect our ability to reliably determine which individuals are above and below the zero level for lifetime welfare in ways that might shift our priorities?

Focusing just on the evaluation of valenced experiences, there is also much uncertainty about how different intensive and extensive quantities affect their value:

- **Which quantities affect the value of valenced experiences?** Quantities to be addressed include the objective duration of experience and its felt duration (Schukraft 2020; Shulman and Bostrom 2021; Mogensen 2023c), degrees of consciousness (A. Lee 2023 and references therein), the amount of attention (if any) devoted to a valenced experience, the richness and complexity of consciousness, whichever quantities are the subject of talk about the intensity of valenced experiences (Armstrong 1968/2023: 341–3; Mayerfeld 1999: 61–7), and the number of subjects undergoing a given token experience (Briggs and Nolan 2015; Javier-Castellanos 2021; Roelofs and Sebo 2024; see also Zuboff 1981; Unger 1990; Johnston 2016).
- **How do different quantities affect the value of valenced experience?** For example, if there can be wholly unattended valenced experiences (compare Block 2010 and Chalmers 2010a: Ch. 11) or experiences of no objective duration, can they have any value? Which quantities contribute to the value of valenced experience on their own and which ones do so through interactions with other factors? Are there any (e.g. cognitive) background conditions that must be in place in order for quantities to affect the value of valenced experiences?

Further questions arise about morally significant mental phenomena when we look beyond consciousness altogether:

- **Which agentic capacities, if any, generate which sorts of moral interests and/or rights?** For example, could certain agentic capacities suffice for autonomy and so command respect even in the absence of a capacity for accruing welfare goods and bads?
- **For a given putatively morally significant agentic capacity, is there any reason to think that that capacity matters *per se*?** Or should we instead take an entity's interests involving that capacity to bottom out in its having a desire or belief whose satisfaction or truth requires that capacity? Or perhaps certain manifestations of the capacity?
- **What moral interests are tied to personal identity?** Are any of these crucial for thinking about how we should design digital minds with the capacity to merge

or fuse, with superhuman abilities to remember and forget, and lives that may be vastly shorter or vastly longer than typical human lives (Shulman and Bostrom 2021)?

Progress on questions of the kind outlined above will help us to evaluate the potential for superhuman levels of relevant mental and moral quantities in certain kinds of minds, such as in digital minds that might one day be realised (Shulman and Bostrom 2021; compare Buchanan 2011: 209-42). Such progress may also help us address questions about what we may morally owe to individuals and populations with superhuman capacities for well-being or ill-being (compare Nozick 1974: 41: Parfit 1984: 389; Chappell 2021; Sebo 2023).

2.1.3 Methodological Issues in Welfare Measurement

We want to be able to assess and compare welfare levels in practice. Methodological problems pose an obstacle to this type of evaluation:

- **How, if at all, can we construct measures of subjective well-being that allow us to reliably impute cardinal structure to subjects' responses and to make comparisons across individuals?** There are difficulties in making interpersonal comparisons of utility (Hausman 1995) and of phenomenal character (Shoemaker 2006). Over the last half-century there has been an explosion of research in psychology and economics on subjective well-being, focused primarily on self-reported life-satisfaction (Layard 2005; Weimann, Knabe, and Schöb 2015). However, core problems of measurement arguably remain unresolved, including the problem of using self-reports to measure welfare interpersonally or on a cardinal scale (Ng 1996, 2008; Kapteyn, Smith, Van Soest 2012; Angelini et al. 2013; Kristofferson 2017; Fabian 2022).
- **Can welfare be measured on a cardinal scale with a privileged zero point, allowing us to say whether a person's lifetime welfare is positive or negative?** Greater clarity on these issues is important, given the natural assumption that there are moral reasons to spare individuals from lives that fall below the zero level (e.g., Narveson 1967; McMahan 1981; Parfit 1984: 391). Measurement tools used by social scientists sometimes include a scale point imagined as the point of neutrality, such as the midpoint of the Cantril Self-Anchoring Striving Scale, below which most people in fact locate themselves (Diener et al. 2018). However, there are also reasons to believe that individuals take lives well-below that midpoint to be worth living (MacAskill 2022b: 196).
- **Are we able to characterise the zero level for lifetime well-being in a way that does not presuppose any particular theory of welfare or population axiology (Broome 2004; Arrhenius 2014: 21-35)?**

- **Can we develop better measures of animal welfare, including welfare scales that allow us to locate different animals relative to the zero level for lifetime welfare?** There currently exist few measures of the subjective or experiential aspects of animal welfare, and widely used animal welfare measures like the Five Domains (Mellor et al. 2020) may permit only limited ordinal comparisons of different outcomes (Browning 2022). Striking claims are sometimes made to the effect that most non-human animals – whether intensively farmed (Singer 1993: 121; Cooney 2014: 7) or living in the wild (Ng 1995; Horta 2010; Tomasik 2015) – do not have lives worth living. However, these claims are often supported primarily by intuitive conjectures. Are we able to construct a principled and reliable philosophical and scientific basis for determining whether, say, American broiler chickens or Atlantic cod really do or do not typically have lives so bad that we should wish for their sake that they had never been born?
- **How should we determine whether, and if so how, individuals that do not belong to the same species compare with respect to their capacity for welfare?** Even restricting ourselves to comparisons of the experiential component of individual welfare, there are enormous philosophical and scientific obstacles to making reliable interspecies comparisons (Browning 2023; Fischer 2024). These obstacles include determining how, if at all, these comparisons can be guided by total neuron count and the proportion of neurons dedicated to affective processing (Shriver 2022) or by learning abilities, decision-making capacities, and cognitive and emotional complexity. We encounter additional obstacles if we seek to go beyond interspecies comparisons in order to compare welfare in individuals made of different substrates. These obstacles stand in the way of evaluating welfare in potential minds run on inorganic computational substrates.
- **What are the most important contributors to welfare in non-human minds aside from valenced experience, and to what extent is their measurement tractable?** Candidates for such contributors include preference satisfaction (Dawkins 2017) and eudaimonic flourishing (Nussbaum 2022).

The ideal outcome would be to arrive at measures of the well-being of human and non-human animals that permit interspecies comparisons of welfare, so as to provide guidance about necessary trade-offs. See also section 4.4.1 (animal ethics).

2.1.4 How Does the Moral Significance of Mental Phenomena Depend on the Natures of Morality and Mind?

Different meta-ethical views and positions in the metaphysics of mind may suggest different verdicts about what kind of mental states have moral significance, thereby

shifting our moral and/or epistemic priorities. Potential interactions on this score include:

- **What are the moral implications of *physicalism* about consciousness?** Physicalism identifies consciousness with a physical state. It has been argued that (certain varieties of) physicalism are in tension with views that attribute special moral importance to the distinction between consciousness and its absence (Cutter 2017; G. Lee 2019; Birch 2022a; compare Pautz 2017). Do these views in fact have these implications? If so, to what extent does that tell against these views, and how plausible should we find these revisionary moral conclusions?
- **How does *illusionism* bear on the moral significance of consciousness?** Illusionism holds that consciousness is not as it introspectively seems (Dennett 1991; Frankish 2016; see Chalmers 2018 for further references). Illusionism comes in different varieties: some illusionists deny that consciousness exists; others maintain that consciousness exists, but is radically different from how it introspectively seems. It is natural to think that the moral significance of consciousness is tied to its nature and hence that introspective illusion about the nature of consciousness puts us at risk of error regarding its moral significance (Kammerer 2019, 2022; compare G. Lee 2013). Developing and evaluating this natural thought requires examining the different varieties of illusionism, their comparative plausibility, and the risks of error they generate for the moral evaluation of consciousness.
- **To what extent do realist meta-ethical views, on which the most basic ethical facts are constitutively independent of our attitudes toward ethical propositions, render our moral beliefs about various mental phenomena susceptible to epistemological challenges?** Relevant challenges include arguments from disagreement and appeals to various genealogical debunking arguments (compare Street 2006, Huemer 2008, Kahane 2010). In what ways does the force or moral import of these challenges depend on questions about mentality, such as the epistemic profile of moral intuitions, the basis of mental content, and the reliability of introspection (Huemer 2008; Dogramaci 2021; Sinhababu 2022)?
- **To what extent do different forms of meta-ethical antirealism that make the most basic ethical facts constitutively dependent on our evaluative attitudes lend support to views that tie the moral significance of mental properties closely to characteristically human traits?** And to what extent do meta-ethical realist views disconfirm such anthropocentric views by saddling them with suspicious coincidences or by debunking intuitions that motivate them (Jaquet 2022; de Lazari-Radek and Singer 2012.; compare Harman 1983: 124-5)
- **How, if at all, does which meta-ethical theory is correct bear on how norms of theory choice apply to views about the moral significance of mental**

phenomena? For example, do meta-ethical theories that construe moral principles as fundamental lead Occamist norms to favour simple moral principles concerning the value of mental states (compare: Bennett 2017; Schaffer 2015)? If so, to what extent should our meta-ethical commitments influence whether we think an individual's welfare is a simple, rather than horrendously complicated, function of their mental states' intensity, duration, etc.?

- **What other hypotheses about the nature of mind and value importantly bear on the moral significance of mental phenomena?** A more systematic understanding of which hypotheses might belong to this class and how they might bear on the moral significance of mental phenomena would be valuable for ensuring that the relevant hypotheses in this class are properly taken into account in setting priorities.

2.2 The Distribution of Morally Significant Mental Phenomena

How are morally significant mental phenomena distributed? We will focus primarily on this question as it applies to consciousness, but research on the distribution of other morally significant phenomena may turn out to be similarly valuable. And we will here highlight some issues of particular interest to GPI concerning valenced experience (LeDoux 1998; Panksepp 2005):

- **How can we develop criteria that allow us to determine empirically whether a system has valenced experiences, which of its valenced states are conscious, and the specific valenced character of its conscious states?** Progress in that direction might take the form of empirical criteria for determining whether an individual has any valenced experience, with these criteria comparable in scope and specificity to, say, the criteria for determining whether a system has any conscious states proposed by the global workspace theory (Baars 1988; Dehaene 2014) or the integrated information theory of consciousness (Tononi 2008; Tononi et al. 2016; Albantakis et al. 2023). At present, affective quality spaces have received relatively little philosophical attention compared to, say, colour quality space (Silva 2023). In addition, there are relatively few well-developed and well-studied theories of valenced experience, and it remains to be seen whether existing philosophical theories of valence, such as evaluativism (Bain 2013; Carruthers 2017) and imperativism (Klein 2007; Barlassina and Hayward 2019) can be developed and operationalised to yield empirical criteria for valenced experience.
- **Why do animals have both positive and negative affective mental states rather than just different gradations of positive (or negative) affect?** Are there notable respects in which minds that rely merely on gradations of positive

affect are impaired relative to minds with bi-polar affect systems, or might the former be desirable engineering goals for possible digital minds (Pearce 1995)?

2.2.1 Liberal and Stringent Criteria for Consciousness

The distribution of consciousness in our world is a wide open issue. The range of defended views spans very liberal ones, according to which all electrons are conscious, to very stringent ones, according to which no humans are conscious. Within this range, views also vary on the reach of consciousness within the biological realm concerning, for example, whether experience is widely distributed not only among mammals and other vertebrates, but also invertebrates of different phyla (Klein and Barron 2016; Birch et al. 2021; Gibbons et al. 2022). More generally, live hypotheses about the distribution of consciousness differ starkly on which animals and artificial entities would be conscious, owing to differences between systems' substrates, environmental embeddings, or the prevalence of realisers of consciousness in the associated state spaces. The following exemplify key questions about these three parameters:

- ***Substrate (in)dependence*: Can consciousness only be realised in a certain kind of material, e.g. neural wetware (Searle 1992; Block 2009, 2023)?** Or would any substrate with the requisite functional organisation be adequate for realising consciousness (Chalmers 1996a: 247–275)? If consciousness can be realised in more than one kind of material, how broad is the range of materials in which consciousness can be realised? Does it include the substrates of digital systems?
- ***Internal state liberality/stringency*: How liberal/stringent are the conditions on the internal states that (help) realise consciousness?** For instance, does consciousness have a wide range of realisers within the biological realm such that humans, octopi, insects, creatures with radically different biochemistries, and everything in between can be conscious? Or are the biological conditions for consciousness exceedingly narrow, such that the evolution of intelligent agents on Earth easily could have failed to produce conscious subjects? Granting that functional isomorphs would share the same phenomenal properties regardless of material substrate (Chalmers 1996a), does consciousness in fact have a wide range of realisers within functional state space? Or can it only be realised by a narrow range of functional states, which may, as a matter of fact, be tied closely to the properties of biological brains (Block 1997; Godfrey-Smith 2016; cf. Cao 2022)?
- ***External (in)sensitivity:* How liberal/stringent are the conditions on the external factors (if any) that (help) realise consciousness?** Can internal physical duplicates vary in whether they are conscious or in what experiences they have (Dretske 1995, 1996; Tye 1995; Lycan 2001; Pautz 2013, 2014; Bourget and Mendelovici 2014; Dalbey and Saad 2022; Saad 2024*b*)? If so, which external

conditions modulate phenomenology and in what ways? If there are external conditions on consciousness, how demanding are these? Are they met by virtually all sophisticated agents, save Boltzmann brains (see Saad 2024*a*)? Or are they met by only a narrow class of entities that have, say, the requisite evolutionary history?

These parameters are especially significant from a priority-setting perspective because of their relatively direct bearing on the expected distribution of experience. For instance, evidence in favour of liberal internal conditions on consciousness would tend to support the hypothesis that some digital systems can be conscious. On the other hand, evidence for stringent external conditions on consciousness would tend to tell against attributing consciousness to fine-grained simulations of conscious minds.

Our focus on these parameters departs from traditional emphasis on functionalism (Levin 2023), computationalism (Rescorla 2020), and multiple realisability (Bickle 2020). Although none of these theses straightforwardly entails (or precludes) substrate independence, that internal conditions on consciousness are liberal, or that consciousness is insensitive to external factors, the bearing of these theses on the expected distribution of experience proceeds largely via their bearing on the three identified parameters. Thus, from a priority-setting perspective, there is reason to focus on these distinctions rather than the traditional theses in the vicinity.

That said, functionalism and computationalism nonetheless maintain an influential role in philosophical and scientific investigations of consciousness and the mind more generally. For this reason, it may be valuable to revisit their traditionally assumed relationship with liberal realisation and substrate independence. For example, if consciousness has a computational basis, how do constraints on computational implementation restrict the range of its possible physical realisers (Putnam 1987; Searle 1990; Chalmers 1996*b*; Ritchie and Piccinini 2018; Klein 2018)? What are the implications as regards the feasibility of realising conscious experiences in systems based on common principles of computer architecture and organisation (Shiller 2024)?

2.2.2. Theories of Consciousness

Theoretical investigations of the distribution of consciousness usually consider only a small subset of existing theories and a small subset of available data. However, priority setting ultimately calls for distributional estimates supported by the total body of available evidence. For this purpose, it is crucial to attain a synoptic perspective that takes account of all reasonable theories bearing on the distribution of consciousness. (For efforts in this direction, see Butlin et al. 2023; Chalmers 2023;

Sebo and Long 2023.) Systematically investigating the collective body of theories that bear on the distribution of consciousness is nonetheless a daunting task, as the vast and rapidly growing literature on this topic is scattered within and across disparate sub-literatures. Given this state of play, meta-work on theories of consciousness (as opposed to object-level engagement with any particular theory) may be particularly important. Issues tackled by such work could include:

- **Given that there are many more rigorous comparisons of scientific theories of consciousness that could be made than will be made in the near term, which comparisons should be prioritised?**
- **What are the in-principle limits on resolving uncertainties about consciousness through scientific investigation?** To what extent should we expect to continue to be burdened with significant degrees of 'cluelessness' about the distribution of experience (compare: Lenman 2000; Greaves 2016)?
- **What is the structure of the space of theories of consciousness?** Is there a subspace in which existing theories fall along a small number of crucial dimensions, e.g. concerning their distributional commitments, what data support them, or their normative profiles? What portion of the space do existing theories encompass? What bounds the space? What dimensions and regions of the space are neglected?
- **Are there any important but underappreciated forms of convergence among leading theories of consciousness or among researchers in consciousness studies?**
- **Which theory regimentations would facilitate important theory comparisons?** In some cases, theory comparison would benefit from theory regimentation. For example, when a theory is formulated unclearly or with inessential commitments, it can be helpful to reformulate the theory so as to capture its core empirical commitments. McQueen's (2019) minimal formulation of the integrated information theory may be considered as an example of this.
- **How can scientific theories of consciousness that were designed with humans in mind be 'de-anthropomorphised' so as to be applicable to non-human minds?** (Compare Cappelen and Dever 2021.) For example, take the global workspace theory. On this view, whether a state is conscious depends on whether it is a representation that is broadcast to a wide range of consumer systems. For the purposes of generating predictions about typical humans, the theory can be left imprecise about, say, the requisite range of consumer systems. But such imprecision needs to be resolved if we are to extend the theory to non-human minds with rudimentary global workspaces (see Carruthers 2019: 140–164; Birch 2020b; Butlin et al. 2023: §2.2.3).

2.2.3 Methodology and Data

Ultimately, we would like to be able to know which individuals exhibit morally significant properties like consciousness and sentience, so as to be in a position to say what different individuals' interests are and how their interests should be weighed. In pursuing this aim, we face questions such as:

- **How should we go about developing estimates for the distribution of consciousness?** Should we prefer approaches that are *theory-heavy*, *theory-light, or theory-neutral* (Birch 2022b)?
- **Should the same methodology be used in investigating the distribution of conscious experience in non-human animals and candidate digital minds?** Or do the two cases call for altogether different approaches (Andrews and Birch 2023)?

Further methodological challenges arise if our ordinary attributions of consciousness fail to discriminate between a number of physical and functional properties that typically co-vary with consciousness in human subjects, but which can come apart:

- **How should we address the epistemological puzzle of what evidence could conceivably reveal to us which of these properties generally co-varies with consciousness** (Block 2002; Hohwy 2004; Levin 2008; Balog 2020)**?**
- **How should we respond to the following metasemantic puzzle that such properties pose for reductive theories of consciousness: how is there a determinate fact of the matter concerning which of these properties our concept of consciousness refers to** (Papineau 2002 175-231; Taylor 2013; Pautz 2017; Balog 2020; Birch 2022a)**?** And how does the answer to the question bear on the distribution of consciousness in cases where these properties dissociate.
- **What bearing does the resolution of these problems have on how to value the distinct properties that may be tightly correlated and associated with consciousness in human subjects?** (See Birch 2022a.)

Still further methodological issues concern the extent to which research on the moral significance of different mental phenomena can proceed in parallel with research on the nature of those phenomena:

- **To what extent is it desirable that a theory of consciousness - or of a particular type of experience such as pain - be able to account for its moral significance?** (See Jacobson 2013; Bain 2019.)
- **Could views about the moral significance of consciousness be undermined by particular conclusions we might draw about its nature and distribution?** For example, could the moral significance ordinarily assigned to consciousness be undermined by the discovery that the state that comes closest and close enough to satisfying our concept of consciousness is similar to many accompanying states to which the concept does not apply (G. Lee

2013)? Compare: views that attach outstanding moral significance to personal identity might be undermined by discovering both that personal identity is non-branching psychological continuity and that non-branching psychological continuity typically holds in the presence of other similar relations (Parfit 1984).

Given our current epistemic predicament with respect to the distribution of consciousness, it is natural to hope that additional crucial considerations are discoverable. It is thus worth asking:

- **What crucial but neglected factors bear on the distribution of consciousness, and what is their import?** Candidates factors include:
    - The *meta-problem of consciousness* (Chalmers 2018)
    - *Debunking arguments* appealing to proximal or distal explanations of our judgements about consciousness (Chalmers, 2018, 2020)
    - In-depth analysis of the strengths and weaknesses of the *dancing and fading qualia arguments* for organisational invariance (Chalmers 1996a: 247-75) and related theses (Saad and Bradley 2022)
    - The *mental problem of the many* (see Unger 2004; Simon 2017; Crummett 2022; Fischer et al. 2022; Builes and Hare 2023; Roelofs 2024)
    - The bearing of self-locating evidence and observation selection effects (Bostrom 2002; Titelbaum 2008; Shulman and Bostrom 2012; Snyder-Beattie et al. 2019; Hanson et al. 2021; Isaacs et al. 2022; Manley ms) on our beliefs about the overall distribution of consciousness (see Zuboff 1990; Bostrom 2003; Crawford 2013; Arntzenius and Dorr 2017; Chalmers 2022: Ch. 5; Li and Saad 2022; Builes and Hare 2023; Saad 2023, 2024*a*)
    - The possibility of *hidden qualia* (Papineau 2002: 215-220; Taylor 2013; Shiller 2017*a*; compare Block 2007; Goff 2014; Muelhauser 2017: Appendix H; Schwitzgebel 2015; A. Lee 2019; Bayne et al. 2020)
    - *Harmonious phenomenal-normative correlations* (James 1890; Pautz 2015, 2020; Mørch 2017; Goff 2018; Saad 2019, 2022; Cutter and Crummett forthcoming)
    - *Laws of appearance* (Raymont 2005; Cutter 2016, Pautz 2020; Sainsbury 2022, Speaks 2022; Block 2023: 198-200; Morgan, 2023).

- **How should we estimate the value of information about the basis of consciousness and the relative importance of different kinds of errors?** For example, when it comes to evidence of sentience, it seems intuitive that we should be more worried about false negatives than false positives. Is that in fact the case, and, if so, how, if at all, should this asymmetry inform research and theorising about consciousness (Peters 2023)? Are there ways in which

even accurate information about the distribution and physical basis of sentience might pose an information hazard? For example, could such information be misused at scale by malevolent actors, and how, if at all, should these downside risks inform research practices (Althaus and Baumann 2020; compare: Bostrom 2017)?

2.3 Preparing to Live Alongside Digital Minds

Some forecasts assign substantial probability to AI systems that meet or exceed humans in cognitive capacities being mass produced before the end of this century (Davidson 2023; compare: Hanson 2016; Cotra 2020; Alexander 2023*a*). The prospect of digital minds raises a host of challenges that are little understood and neglected. There is no plan in place for navigating these challenges, and no compelling case has been made that they will be well-navigated by default. There is thus an urgent task of identifying key challenges raised by that prospect and devising strategies for addressing them. This section highlights some key issues in this area.

2.3.1 Catastrophic Risks and Their Mitigation

There has been much discussion of advanced AI posing an existential risk via digital agents with superhuman cognitive capacities that turn out to be misaligned with human values (Yudkowsky 2008; Bostrom 2014; Russell 2019; Ord 2020; Cotra 2022b; Ngo et al. 2022; Carlsmith forthcoming; see also Alexander 2023*b* and references therein; for other catastrophic risks posed by AI, see, e.g., Hendrycks et al. 2023). A growing body of research addressing the *alignment problem* aims to ensure that the goals and values of AI systems do not conflict with those of human users (Christiano 2019; Christian 2020; Krakovna, 2024). For more on these issues, see section 3 (Artificial Intelligence: Risks and Opportunities).

One neglected cluster of issues concerns the bearing of the (potential) moral interests and rights of advanced AI systems on catastrophic risks.

- **How do potential interests and rights of advanced AI systems morally constrain solutions to the alignment problem?** (Christiano 2018; Peterson 2019; Gabriel 2020; Sebo and Long 2023; Bradley and Saad 2024; compare: Chalmers 2010*b*: 30). What are these interests and rights? Which systems have them? To what extent are these systems' rights grounded directly in their moral standing, rather than by way of special obligations we would incur through our role in creating these systems (Schwitzgebel and Garza 2015)? To what extent are existing alignment proposals in tension with the ethical treatment of digital minds?
- **Would it be permissible to create digital minds that intrinsically value serving humanity and prioritise human welfare over their own, with no**

freedom to explore other values? (See Petersen 2011; Bales ms.) If so, under what conditions? For example, would it be permissible to create digital minds of this sort that meet or exceed cognitive criteria for high moral status of the kind typically associated with human persons? If it would not (Schwitzgebel and Garza 2020), what are the implications for designing morally permissible approaches to AI safety?

- **Should any tensions between the ethical treatment of digital minds and the safe development and deployment of highly capable AI systems be leveraged to decelerate or regulate AI development?**
- **More generally, how can catastrophic risks of large-scale suffering and/or rights violation in emerging populations of digital minds be mitigated?** (See Bostrom 2014: Ch. 8; Sotala and Gloor 2017; Tomasik 2017; Saad 2023; Schwitzgebel 2023; Saad and Bradley 2022). At present, there seems to be nothing that guarantees, or even renders it likely, that humanity will generally extend future AI moral patients the considerations they are owed rather than the sort of consideration we currently extend to personal computers or non-player characters in video games. If a very large number of such AIs exist and even a small portion are mistreated, their abuse will unfold at enormous scales. At present, these risks are not widely recognised.
- **Can we develop better threat models and mitigation strategies?** Valuable research on this score could be constructive or critical in character: while the construction of better risk mitigation strategies would constitute progress in this area, so too would impossibility results that reveal the unavailability of mitigation strategies that achieve well-motivated desiderata (compare: Thornley forthcoming). The latter might serve to guide further research in the area in fruitful directions or lend support to proposals such as technological pauses or moratoria (Metzinger 2021).

2.3.2 Digital Minds and Timing Issues

In setting priorities, we face not only questions about the impact of different types of interventions but also about the impact of intervention timing. A number of these arise in connection with AI, owing to the rapid pace of AI development, the malleability of its trajectory, and the large but highly uncertain potential impact of AI on the distribution of minds:

- **What, if anything, should be done now to prepare the ground for appropriate recognition of the moral status of advanced AI systems that might exist in the future?** For example, are there risks that certain false beliefs about morally significant aspects of mind could in the near future become locked in? (Compare MacAskill 2022b: 75-102.) Would it be better to focus for now on resolving crucial uncertainties about consciousness and moral status in digital systems, before prioritising legal and policy interventions?

- **Given the aim of ensuring appropriate recognition of the moral status of AI systems, to what extent should we be concerned with the order in which relevant future developments occur?** For example, is it desirable for certain kinds of AI systems to arrive before others? How important is it to develop moral criteria for the treatment of digital minds in advance of their arrival, before attitudes and practices become socially entrenched? How strong a reason is this to prioritise work addressing the ethics of digital minds at the current margin as against, say, work addressing the plight of intensively farmed animals? To what extent would answers to such questions change, given the broader goal of eventually expanding the moral circle to fully encompass both digital minds and non-human animal minds? More generally, what are the benefits and costs associated with different possible 'takeoff scenarios' for the emergence of digital minds (Saad and Caviola 2024)?

2.3.3 How Might the Future Go Well?

Currently, discussion of possible future outcomes involving the emergence of digital minds focuses primarily on catastrophic risks and corresponding threat models. It is also worth reflecting concretely on the character of desirable long-run outcomes involving digital minds and the steps by which to get from here to there (Chalmers 2010*b*; Hanson 2016; Shiller 2017*b*; Shulman and Bostrom 2021; Friederich 2023; Bostrom and Shulman forthcoming). Some potential research angles here include:

- **What strategies can mitigate against our uncertainty about consciousness in digital minds?** For example, if candidate sufficient conditions for (un)consciousness can co-occur in digital systems, what are the prospects for engineering or training co-occurring candidates into digital systems so as to reduce uncertainty about the presence of consciousness? Alternatively, what are the prospects for creating large populations of systems that differentially exhibit candidate bases of consciousness? What do different views in population ethics imply about what portfolio of populations with different candidate bases would be optimal?
- **How do the ranges of possible digital mind designs and possible ways of organising a society containing both human and digital minds jointly constrain the space of favourable future outcomes?** For example, morally favourable outcomes with stable, liberal-democratic societies may require careful design choices. Digital minds that have – or which can easily gain – superhuman powers of persuasion may readily achieve undue levels of political influence. Similarly, creating digital minds that have the capacity to reproduce rapidly and cheaply could spark an arms race for democratic power. On the other hand, some ways of restricting digital minds' opportunities to express themselves or procreate would be unjust. Tackling these twin dangers – of creating digital minds whose capacities threaten

morally favourable forms of social organisation and of securing certain forms of social organisation through unjust restrictions on digital minds – may require a constructive approach that simultaneously devises novel kinds of digital minds and novel forms of social organisation.

3 Artificial Intelligence: Risks and Opportunities

Artificial intelligence (AI) will plausibly be one of the defining technologies of the 21st century. Indeed, some predict that AI will have transformative implications within decades, having at least as profound an impact on human civilisation as the Industrial Revolution (Roser, 2023).

Like many technologies, and like the Industrial Revolution itself, AI is a source of both opportunities and risks. And if AI does end up being transformative then these opportunities and risks are likely to be particularly dramatic in scale. This suggests that navigating AI well could be one of the most important tasks facing humanity in the coming decades. For this reason, GPI is interested in research that helps us to understand and navigate the largest-scale risks and opportunities presented by AI.

One motivation for this work is the thought that AI might have significant impacts on the long-term future of humanity (see section 1), either by bringing about human extinction or by influencing the trajectory of civilisation. However, reflection on the impacts of AI can also be justified from many moral viewpoints, including those that place less emphasis on the long-term future. So, we see this research direction as complementary to our work on the long-term future but not reliant on it.

Artificial intelligence is a new focus area for us and this agenda will likely grow and change as we learn more. With that caveat in mind, here are three research areas that we're particularly interested in.

See also sections 2.2 (Economics of Artificial Intelligence) and 2.3 (Modelling Artificial Agents) of GPI's research agenda in economics and section 9.2 (Risks From Artificial Intelligence) in GPI's research agenda in psychology.

3.1 Catastrophic Risk from AI

AI could bring about a variety of catastrophes (Bengio et al. 2024; Hendrycks forthcoming). One possibility is that AI could itself bring about the extinction or disempowerment of humanity. Another possibility is that humans could use AI to cause catastrophic outcomes. For example, authoritarian regimes could use AI to entrench their power, or terrorist groups could use AI to develop dangerous biological weapons that they then deploy. And yet another possibility is that AI could change the world in ways that aren't directly catastrophic but make catastrophe more likely. For example, AI might lead to a growth in both wealth inequality and misinformation, and

this might impede the smooth functioning of society in a way that makes catastrophes more likely.

GPI is particularly interested in two threads of work that explore the risk of AI causing catastrophes like these.

3.1.1 Understanding the Catastrophic Risks Posed by AI

GPI is interested in work that helps to clarify what catastrophic risks, if any, are posed by AI. Understanding what risks AI poses can help society to prioritise between different strands of work. Understanding specific risks can also help guide work to mitigate those risks. With this in mind, GPI is interested in work on the following topics.

- **What is the case for expecting AI to pose various catastrophic risks?** Threat models are rigorous explorations of a potential risk that might be posed by AI. Threat models can explore conceptual and evaluative questions about the nature of a given risk, can outline clearly structured arguments that AI poses a particular risk, and can evaluate the plausibility of these premises. Note that here – as elsewhere – we're not assuming any conclusion: we're interested in work that evaluates the risks posed by AI, regardless of whether these strengthen the case for the risk or weaken it. Previous examples of threat modelling include Bostrom 2014; Cotra 2022b; Grace 2022a; Hendrycks et al., 2023; Ngo et al., 2022; Carlsmith forthcoming; Goldstein & Kirk-Giannini forthcoming).
- **What are the most fruitful ways of modelling advanced AI systems?** GPI is interested in models aiming to provide insight into the likely behaviour of future AI systems. For example, we might try to develop decision theoretic models to study the behaviour of individual systems (Bales forthcoming; Gallow forthcoming; Thornley forthcoming) or game theoretic models to explore the interaction between multiple models (Conitzer & Oesterheld 2023). Alternatively, we might investigate the extent to which the predictive processing model of the mind can provide insight when applied to AI (Ratoff 2021).
- **How should we prioritise between different risks posed by AI (and between AI risk and other causes)?** Accurate threat models – which characterise the catastrophic risk posed by AI – don't by themselves settle whether mitigating risks from AI should be a priority. To answer this question, we also need to reflect on how moral prioritisation should proceed and on how AI risk mitigation compares to other moral causes. GPI is interested in work that engages in reflection on this question of prioritisation. For example, this might involve reflecting on how the catastrophe risk posed by AI (and the tractability

of reducing it) compares with other sources of catastrophic risk (Ord 2020). Or it might involve thinking about the circumstances under which reducing the risk of catastrophe from AI should be a priority (Shulman & Thornley forthcoming). Or it might involve reflection on which of the risks posed by AI should be a priority. We might ask (for example) whether accident-risk or misuse-risk is more urgent, whether our focus should be on reducing the risk of extinction or on improving our prospects conditional on survival, and to what extent we should prioritise reducing risks to digital moral patients.

- **How can political philosophy provide insight into catastrophic risks from AI?** Some of the risks posed by AI are political in nature, including the risks posed by AI-enabled dictatorships. Other risks will inevitably involve a political dimension, for example with regulation and international agreements playing an important role in enabling or mitigating risks. For this reason, it's likely that political philosophy will be able to provide insight, and we're interested in work of this sort.

3.1.2 Exploring Mitigation Strategies

GPI is also interested in work that aspires not simply to clarify the risks posed by AI but also to mitigate them. For this reason, we are interested in work on the following topics.

- **How should we think about the alignment of AI systems?** Many agree on the importance of AI alignment to mitigate catastrophic risks from AI, where this is understood broadly as ensuring that AI systems act in accordance with human values and interests. But what specifically should advanced AI systems be aligned with (Gabriel 2020)? Human preferences? If so, how do we handle cases where preferences differ (Zhang & Conitzer 2019)? Some particular moral theory? If so, which one (Barrington 2023; D'Alessandro 2023)? How, if at all, should we account for moral uncertainty (Korinek & Balwit 2022)? And should we be aspiring to this sort of normatively deep alignment at all, or should we instead aim to resolve the risks in other ways?
- **How promising are existing mitigation strategies (and are there novel strategies that should be considered)?** Authors have proposed many strategies for mitigating risks from advanced AI systems. These include both strategies for preventing accidents and strategies for preventing misuse. We are interested in assessing these strategies: where might they fall down and how could they be improved? Strategies we might evaluate include governance proposals like slowing down AI development (Grace 2022b; Hogarth 2023), restricting access to hardware (Balwit 2023), red-teaming AI systems to identify their flaws (Anthropic 2023; Mislove 2023), and requiring tests of AI systems' capabilities and alignment (Kinniment et al. 2023). We're also

interested in assessing technical proposals including: improving our understanding of models' internals (Olah et al. 2020; Bergal & Beckstead 2021; Nanda 2022), developing AI systems that learn human values by observing human behaviour (Hadfield-Menell et al. 2016; Russell 2019), creating systems that aim to be 'low-impact' (Armstrong & Levinstein 2017), eliciting latent knowledge from AI systems (Christiano et al. 2021), training AI systems to be truthful (Evans et al., 2021), using less-advanced AI systems to help us reliably oversee the training of more-advanced AI systems (Christiano et al. 2018; Irving et al. 2018; Leike et al. 2018), using AI feedback to train AI systems (Bai et al. 2022), and designing agents that won't resist being shut down (Thornley 2024; Thornley et al. 2024). In addition, we're interested in developing novel mitigation strategies.

3.2 AI and the Trajectory of Civilisation

Even setting aside the possibility of catastrophe, AI might have a dramatic impact on the world and on the future of human civilisation. This could result from two potential consequences of AI.

The first potential consequence is that AI might bring about rapid societal and technological change. That is, AI might markedly increase the rate of economic growth and technological development (Aghion et al. 2019; Trammell & Korinek 2020; Nordhaus 2021), as well as dramatically impacting social norms and institutions. So, the impact of AI could occur on a particularly large scale.

These impacts could be very good or very bad. For example, rapid economic growth might lift many out of poverty and technological development might lead to novel pharmaceuticals and clean energy technologies. Taken to the limit, we might envision a sort of utopia resulting (Russell, 2019: 246; Bostrom, Dafoe & Flynn 2020: 297). On the other hand, AI might bring about the invention of new weapons of mass destruction, potent forms of propaganda, and lie-detection and surveillance technologies. Further, it might do so at such a rapid pace that society struggles to adapt (Bostrom, Dafoe & Flynn 2020: 299). Taking this to the limit, we might envision a dystopia.

The second potential consequence is that AI might induce lock-in, in the sense of causing certain features of society to persist for an extremely long time (MacAskill 2022: Ch 4). This could occur because a given AI system could itself exist for an extremely long time, able to survive via periodic copying to new hardware. If such a system, or a set of such systems, shapes the world in a certain way then they might continue to do so for a long time and hence lock-in the relevant features of society. This magnifies the potential impact of the sorts of large scale consequences outlined

above. It also means that, for better or worse, changes that have a relatively small immediate impact might have a far larger impact once their persistence is accounted for.

These two potential consequences of AI mean that we could soon face a series of important decisions that have unusually large-scale and persistent consequences. In the light of this possibility, GPI is interested in three strands of work relating to navigating the impacts of AI on the trajectory of civilisation.

3.2.1 Understanding the Impacts

We are interested in work that helps to clarify the likely impacts of AI. Among other things, we're interested in the following topics:

- **What is lock-in and how likely is AI to bring it about?** GPI is interested in work that clarifies the nature of lock-in and the relationship between lock-in and the achievement of a desirable future. We're also interested in work that explores whether AI is likely to bring about various types of lock-in (Karnofsky, 2021; Finnveden et al., 2021). One important-seeming type is value lock-in (MacAskill, 2022b: Chapter 4): the values instantiated by advanced AI could persist for a very long time. That suggests that it is especially important to get these values right.
- **How likely is AI to bring about rapid societal and technological change?** GPI is interested in work that explores both what impacts AI is likely to have and on what timescale. While much of this work is likely to lie outside of philosophy, we think that philosophers can potentially provide insight into questions here. Relevant work might explore how fast AI systems are likely to improve (Cotra, 2020, 2022a; Barnett & Besiroglu, 2023; Davidson, 2023), and in particular, whether such systems are likely to undergo a process of recursive self-improvement (Chalmers, 2010; Thorstad, 2022).

3.2.2 Navigating the Development of Advanced AI

Given the impacts of AI, we might ask how we should want AI to be used. This strand of research explores both how we might want to develop advanced AI in the first place and what we might want to do with it once we do. For example, we're interested in work on the following:

- **At what pace should we aim to develop advanced AI?** In the light of potential benefits and costs of AI, we need to make a decision about how to proceed with AI development. One question here is whether we should pause AI progress, either now or in the future, to give us more time to understand the risks and

how to mitigate them (Alaga and Schuett 2023; Ienca 2023). More generally, work in this area might explore whether we should engage in differential technological development, where we aim to speed up work that ensures the safety of AI relative to other work that increases the capabilities of AI (Sandbrink et al. 2022).

- **How should advanced AI be used in shaping the future of civilisation?** If we develop safe AI, it will be important to ask how we might use it to create a flourishing future. For example, this work might explore the role that democratic processes should play in making crucial decisions about how to make use of AI. Or it might explore how AI could itself play a role in deciding how to make use of AI in shaping humanity's future or in answering any of the other questions in this list.
- **What values should we want an AI to instantiate?** Unfortunately, there are many ways in which we might get the values instantiated by an AI wrong. We might endow powerful AI with the wrong theory of normative ethics, or the wrong theory of welfare, or the wrong axiology, or the wrong population ethics, or the wrong decision theory, or the wrong theory of infinite ethics. Each of these mistakes could make the future significantly worse than it otherwise would be. With what values - if any - should we endow AI? Can we delegate this question to AI itself or otherwise wait to decide?

3.2.3 Navigating Rapid Change

The final strand of research in this area will explore how we should navigate the rapid social and technological change that AI could potentially bring about. This might require us to rapidly make a large number of important decisions. These decisions include: How do we share the wealth and power that might be granted by advanced AI? How do we govern new capabilities (like advanced lie detection and surveillance, human enhancement, and space settlement) that might be enabled by advanced AI? How do we integrate into society the vast numbers of digital beings that might follow in the wake of advanced AI? GPI is interested in work that explores how these challenges might be addressed. For example, we're interested in work in the following areas.

- **How can we improve society's capacity to navigate rapid change?** Work here might explore how to improve humanity's governance systems and institutions to leave us better able to coordinate and respond to an agile way to rapid change. It might also explore whether we can address some of the challenges that AI might pose by making agreements ahead of time, to help ensure that the benefits of AI are shared broadly.
- **How should we navigate specific challenges that rapid change might lead to?** If society goes through a period of rapid societal and technological change

then there will be many innovations that we need to navigate rapidly. There might be value in reflecting on how to navigate some of these challenges ahead of time. Work in this area might explore whether we can make reliable predictions about the kind of technologies whose development will be accelerated under conditions of explosive economic growth. It might then explore what kinds of preparatory solutions should be proposed for addressing the risks and opportunities that are likely to arise from the emergence of those technologies under conditions of rapid change.

3.3 Digital Minds

So far, this section of the research agenda has largely focused on the implications that AI might have for humans and human civilisation. However, it's possible that in the future AI systems might themselves become moral patients, deserving of our consideration. GPI is interested in work that helps us to determine the likelihood that future AI systems are moral patients and interested in work that clarifies how we should treat AI systems in light of this likelihood. For more details of work we're interested in in this area, see section 2.3 (Preparing to Live Alongside Digital Minds).

4 General Issues Related to Prioritisation

A range of general philosophical issues confront an agent wishing to do the most good possible with limited resources. Here we outline some particularly important *Decision Theoretic Issues* (section 4.1), *Epistemological Issues* (section 4.2), *Moral Issues* (section 4.3), and *Issues Relating to Specific Cause Areas* (section 4.4). While most issues in this section are not specific research foci for GPI, we think that they may either inform our approach to other questions in the agenda, synergise well with them, or represent fruitful research projects for individual researchers to pursue.

4.1 Decision-Theoretic Issues

4.1.1 Risk Aversion

Some policies that have high expected value carry a great deal of risk. Many think that this counts against implementing those policies, though whether this is rational and precisely how to accommodate *risk aversion* in our decision theory is a controversial matter.

- **Does the correct decision theory permit risk aversion (see Buchak 2013, Thoma 2019 for overviews)?** If so, are there arguments for or against risk aversion when aiming to promote moral value (see Wilkinson 2022b) or when acting on behalf of others (see Buchak 2019, Thoma 2023)?
- **How should we think about low probability outcomes, especially extremely good or bad ones (see Wilkinson 2022c, Russell 2023, Beckstead & Thomas 2024)?** Should we ignore very unlikely outcomes (see Monton 2019, Kosonen 2022), or avoid excessive risk taking by adopting a bounded utility function (see Menger 1934, Arrow 1966; compare Russell and Isaacs 2021), or non-expectational decision rule (following Buchak 2013, Bottomley & Williamson 2024)?
- **Even if the correct decision rule permits risk aversion, should real-world agents act in a risk-averse way?** Do long-run arguments or considerations of background risk mean that risk neutrality is effectively required in practice (see Buchak 2013: Chapter 7, Thoma 2018, Tarsney 2020, Wilkinson 2022b)?
- **Are there specific interventions that only make sense given some kind or degree of risk aversion (Greaves et al. 2024; Pettigrew 2024)?** How should policymakers and philanthropic agents go about incorporating risk aversion into their decision-making procedures, if at all?

4.1.2 Infinities

It is arguable that some things are infinitely more valuable than others, either in actuality or in expectation. How should we think about our choices in light of these possible infinities?

- **Is it really epistemically possible that some outcomes are infinitely more valuable than others?** What kind of decision theory should we employ to accommodate such outcomes (see Bartha 2007, Chen & Rubio 2020)? Should our concern with realising infinite values swamp our concern for realising merely finitely valuable outcomes (see Hájek 2024)?
- **How should we evaluate populations containing infinitely many lives?** If no action can affect more than a finite amount of value, it may appear that no action can affect the value of an infinite world (Nelson 1991). Do we therefore face a kind of 'infinitarian paralysis'? If not, what kinds of tools should we use to compare infinite populations (see Vallentyne & Kagan 1997, Bostrom 2011, Arntzenius 2014 Askell 2018, Tarsney & Wilkinson forthcoming)?
- **How should we respond to the possibility that some of our options have infinite or undefined expected value?** Is it possible to face 'St Petersburg-like' acts with infinite expected value (see Peterson 2023) or 'Pasadena-like' acts with undefined expected value (see Hájek & Nover 2004; Alexander 2012)? How can we sensibly compare such acts (see Easwaran 2008; Colyvan 2008; Easwaran 2014; Colyvan & Hajek 2016; Wilkinson forthcoming)? Do we face such acts in practice (see Wilkinson msb), and does this result in a kind of infinitarian paralysis (since each of our acts may have undefined expected value)?
- **What should we make of principles that seem plausible in the finite setting but break down in the infinite setting?** In particular, various seemingly plausible principles in decision theory (see Goodsell 2023, Russell 2023, Wilkinson forthcoming) and population ethics (Askell 2018, Goodsell 2021, Wilkinson 2021) result in paradoxes or inconsistencies in the infinite setting. Does this ever provide a reason for doubting those principles in the finite setting?

See also section 1.3 (Welfare and Decision Procedures) of GPI's research agenda in economics.

4.1.3 Causal and Non-Causal Decision Theories

There is a longstanding debate over how to evaluate acts that provide *evidence* of a good outcome without *causing* a good outcome – Newcomb's Paradox is the classic example (see Nozick 1969, Lewis 1981, and Ahmed 2014).

- **Should we care about causally promoting good outcomes, or something else?** Is the correct decision theory *Causal* Decision Theory (see Lewis 1981), *Evidential* Decision Theory (see Ahmed 2014), *Functional* Decision Theory (see Levinstein and Soares 2020), or something else? How should we act while uncertain about the answer to that question (see MacAskill et al. 2021)?
- **How does the debate between these theories affect our cause prioritisation?** Do particular theories of ethics or rationality look more or less promising given a non-causal decision theory (see Oesterheld 2017: Section 4, Wilkinson 2022a)?
- **If a non-causal theory is correct, should we engage in *acausal trade*?** That is, should we ever perform acts that do not promote our values but provide evidence that agents elsewhere promote our values (see Oesterheld 2017)?
- **What kind of theory should we expect or want digital agents to follow (see Conitzer & Oesterheld 2023)?**

4.2 Epistemological Issues

Good decision-making is based on evidence. In some cases, it is clear what our evidence is and what kinds of belief it justifies. In other cases, it is unclear what kinds of beliefs we are justified in forming. How should an altruistic agent act in such cases?

4.2.1 Severe Uncertainty

We must make choices despite having sparse and difficult-to-interpret evidence about the effects of those choices. While we may have some idea about how our choices will go in the short run, the further we look into the future, the less we seem to have to go on.

- **Are there situations in which it is inappropriate to assign precise probabilities to outcomes of some act?** If so, what kind of ambiguity-aversion does the correct decision theory permit (see Bradley et al. 2017; Rowe & Voorhoeve 2018)?
- **What tools do we have to reduce long-term uncertainty?** Can fields like forecasting and persistence studies help us resolve at least some severe uncertainty? See also section 2.3 (Forecasting) of GPI's research agenda in psychology and section 1.2 (Forecasting) of GPI's research agenda in economics.
- **Are there any decision procedures or heuristics we can adopt to make better decisions in light of severe uncertainty (see Mogensen & Thorstad 2022, Thorstad forthcoming)?** How can such heuristics be justified? Do they effectively amount to ignoring the long-term future?

- **How should we act when we face unawareness (i.e., when we cannot describe or conceptualise some features of a decision situation, see Bradley 2017; Steele and Stefánsson 2021; de Canson 2024)?**
- **Does the problem of cluelessness mean that we lack a meaningful basis for comparing options (see Lenman 2000)?** The long-run future seems to be inscrutable, which might suggest that an agent who cares about the total consequences of their acts has no way of comparing options. Are there principles that we can leverage to make at least some meaningful comparisons (see Greaves 2016)? Does cluelessness mean that we must give up on commonsense intuitions about how to rank policies (Mogensen 2021)?

4.2.2 Self-locating Beliefs

Some arguments from epistemology suggest that facts about us and our location can sometimes provide us with evidence about apparently unrelated propositions. A range of potentially important topics hinge on how we should respond to such evidence.

- **How should we take self-locating information into account?** For general discussion, see Bostrom (2002), Titelbaum (2008), and Isaacs et al. (2022).
- **What should we make of the 'doomsday argument'?** The 'doomsday' argument leverages self-locating evidence to show that we should expect humanity to go extinct before many more of us are born (Carter 1983; Gott 1993; Leslie 1996; Bostrom 2002; Mogensen 2019a; Thomas 2021a). Working out whether this argument is sound will affect how we think about future harms and benefits, as well as interventions aimed at mitigating existential risks.
- **What should we make of various 'simulation arguments'?** Simulation arguments leverage self-locating evidence to show that we should be confident that we live in a simulation (see Bostrom 2003, Weatherson 2003, Thomas 2021b). Apart from being of great theoretical interest, what practical upshot does this argument have, if sound?
- **Are we living at the Hinge of History?** In particular, does self-locating evidence provide evidence for or against our being at an especially important time with respect to our ability to go extinct (see MacAskill 2020a, Mogensen 2023b)?

4.2.3 The Status of Philosophical Arguments

Some philosophical arguments purport to establish *prima facie* outlandish conclusions, many of which have implications for ethical prioritisation (e.g., that we should acausally cooperate with agents outside of our causal future). Similarly, many *prima facie* outlandish empirical claims would, if true, have implications for prioritisation.

- **Should outlandish positions be taken seriously in practice?** If so, are there limits on the kinds of counterintuitive views that should be taken seriously? And if so, what exactly does it take for a view to be beyond the pale (see Kelly 2005, 2008, 2011 and Rinard 2013)? If we are broadly skeptical of some outlandish positions, are there particular issues or arguments that philosophers wanting to promote the good should set aside (see the discussions by Cotra 2021 and Carlsmith 2023)?
- **Should we have the same level of epistemic modesty about unusual moral views as we do about unusual empirical views (cf. McPherson 2009)?**

4.3 Moral Issues

4.3.1 Population Ethics and Aggregation

If we aim to do the most good overall, then we need to know how to weigh benefits and harms interpersonally. We need to know whether and how to aggregate, and we must also tackle the issues which arise when our choices make a difference to the number of people who will exist.

- **How should we decide between interventions that change the number or identities of future people?** Are some interventions only worth pursuing on some views of population ethics, or do most theories converge in their recommendations under plausible empirical assumptions (cf. Tarsney and Thomas 2024)? Are we justified in ignoring these sorts of effects when we attend to policies primarily aimed at the near term (Broome 2005: 402)?
- **Does the Non-Identity Problem weaken or eliminate our welfare-based moral reasons to benefit future generations (Parfit 1984: 366-71 , Boonin 2014, Mogensen 2019b)?** If so, might we have other sorts of reasons to benefit or co-operate with future generations?
- **Should we aggregate wellbeing at all (see Taurek 1977, Parfit 1978)?** If yes, should we be partial or full aggregationists (see Voorhoeve 2014, Horton 2021)? Again, how does our choice of theory here affect our cause prioritisation (Curran forthcoming)?
- **Is there a case to be made for prioritising identified over statistical lives (see Cohen, Daniels & Eyal 2015)?** If so, how strongly should we prioritise identified people in principle, and should we in fact favour interventions which help identified people in practice?

See also section 1.1.1 (Welfare and Beneficence) and section 1.3 (Welfare and Decision Procedures) of GPI's research agenda in economics and section 1,1 (Understanding People's Values) of GPI's research agenda in psychology.

4.3.2 Non-Consequentialism and Moral Prioritisation

Discussions of impartial prioritisation often start from broadly consequentialist considerations, such as weighing harms and benefits or promoting good outcomes. But this perspective might be challenged in several ways, and we might ask to what extent non-consequentialist considerations will lead to different results or even suggest a different methodology.

- **What should we make of the 'Stakes Sensitivity Argument' for setting aside non-consequentialist considerations in some contexts?** When the stakes are high, it often seems like considerations of welfare trump *prima facie* deontic constraints (see Nagel 1978, Greaves & MacAskill 2021). This might suggest that "moderate" non-consequentialists, who count the value of consequences as *one* of the things that matters morally, should act as though they are full-blown consequentialists in high-stakes decision situations. Is this right, and are ethical prioritisation decisions high-stakes in the relevant sense?
- **What should 'radical non-consequentialists' think about prioritisation?** Some non-consequentialists hold either that there is no such thing as the value of consequences, or that value-based considerations do not make a difference to what we ought to do (Geach 1956, Thomson 1997). What practical upshots do such views have? In particular, are the best arguments for prioritising unorthodox cause areas, such as aiming to improve the far future or the welfare of non-human animals, inseparable from value-based moral reasoning?
- **Are there specific non-consequentialist considerations that speak against doing what is impartially best in some cases?** For example, it might seem *unfair* to implement the most cost-effective intervention in cases where doing so benefits only those lucky enough to be able to be cheaply helped (cf. Broome 1984, 1991, Daniels 1993, Kamm 2004). Or it might seem *paternalistic* for one agent to impose solutions on others rather than letting them decide for themselves how to use resources (Saunders-Hastings 2019). Do such considerations speak against the overall project of prioritising the "best" interventions?

4.3.3 Moral Uncertainty

Humans have been doing moral philosophy for at least 2,500 years. We still don't know for sure what the correct view is, and it's unlikely that we will know in the foreseeable future. We have to make choices regardless. How should we go about evaluating causes and cause areas in light of our moral-epistemic position (see MacAskill, Bykvist & Ord 2020)?

- **Is moral uncertainty practically important?** Is there a normatively important sense in which we “ought” to do certain things in virtue of our uncertainty over moral theories (see Weatherson 2019)? If so, what is the correct theory of this sense of "ought" (see Gustafsson and Torpman 2014, MacAskill and Ord 2020)?
- **Could we be unaware of the correct moral theory?** What ought we do then?
- **How does moral uncertainty matter in practice?** Are there specific issues in this research agenda, or in ethical prioritisation more broadly, where moral disagreements are important? If so, what do different meta-normative theories recommend in such cases?
- **How valuable is normative information (see Russell forthcoming)?** In particular, how much of our current resources should we spend on pursuing normative information (see Ord 2020: Ch. 7)?
- **Should we be especially uncertain about ethical issues in light of the depth and persistence of normative disagreement (McGrath 2008; Wedgwood 2010; Setiya 2012; Mogensen 2017)?**

See also section 1.3 (Welfare and Decision Procedures) of GPI's research agenda in economics and section 1.1 (Understanding People's Values) of GPI's research agenda in psychology.

4.4 Other Cause Areas

This section covers other issues which don’t fit neatly into the previously mentioned categories, but to which we think philosophers can make important and valuable contributions.

4.4.1 Animal ethics

Much of moral philosophy focuses on how humans ought to treat each other. But at least some non-human animals might have moral standing as well. If so, given our current neglect and mistreatment of many animals, and given their vast numbers, interventions aimed at improving animal welfare could be enormously valuable. Below are some of the philosophical questions related to our treatment of non-human animals which seem to us especially pressing:

- **Which non-human animals matter morally, how do they matter, and why?** (See Regan, 1983; Singer 1993; Korsgaard 2018; Nussbaum 2022.)
- **How can we make meaningful judgements about non-human animal wellbeing?** What makes it the case that an animal has a given wellbeing level, and is the answer here any different to the case of humans? Can we make

interspecies comparisons of wellbeing – and if so, how (see Browning 2023, Fischer 2024)?
- **Do some farmed animals have lives worth living?** Which ones? Could it be beneficial to increase the number of animals raised on farms (see McMahan 2009)? What about the welfare levels of wild animals (see Groff and Ng 2019; Browning and Veit 2023)?
- **What kinds of economic and institutional changes would improve farmed animal welfare?** How can animal welfare be incorporated into standard economic models or cost-benefit analysis (see e.g., Espinosa and Treich 2024)? Which kinds of regulatory changes would have the greatest impact on farmed animal welfare? How can we assess interventions aimed at improving animal welfare and determine whether they are cost effective?
- **Is animal wellbeing the only morally relevant consideration?** Do animals also have rights (Regan 1983; Francione 2008; Donaldson and Kymlicka 2011)? What would be the implications if so?
- **Is species existence intrinsically valuable?** To what extent should we avoid extinctions of non-human animals? How much should we value non-human animals qua individuals versus valuing them qua components of an ecosystem (see Baard, 2022 and Palmer et al. 2023)?

See also section 8.2 (Wellbeing in Nonhuman Beings) in GPI's research agenda in psychology.

4.4.2 Aid and Development

There has been an enormous amount of work in moral philosophy on duties to the global poor, and many altruistic people take helping the worst off to be at least part of how they should go about making the world a better place. Since the scope and volume of work on this topic is vast, the following is merely indicative of the kind of important questions that synergise well with other topics in this agenda:

- **How do duties to the worst off today balance against concern for future generations?**
- **How demanding are our duties towards the worst off in the world today (Singer 1972, Unger 1996, Cullity 2004)?** Are they merely duties of beneficence, or are they (arguably more stringent) duties of justice (Pogge 2002)?
- **What metrics should we employ when evaluating interventions?** Should we use Disability-Adjusted Life Years, Quality-Adjusted Life Years, or something else (Gold et al. 2002; Fleurbaey and Blanchet 2013; Frijters et al. 2020)?
- **What kinds of evidence should we use when evaluating interventions?** What is the value of Randomised Control Trials (see Deaton and Cartwright 2018)?

More generally, should we favour interventions backed up with specific kinds of evidence, or are there dangers to overemphasising certain kinds of evidence ('looking for the car keys underneath the street lamp')?
- **In what ways can aid have unforeseen consequences that harm the worst off?** Does aid often, for example, weaken states' responsiveness to the needs of their citizens (see Temkin 2022)? How should we take such consequences into account when making altruistic decisions?

4.4.3 Moral Progress

A common issue in prioritisation is whether to act now or invest to do more good later. One relevant question that philosophers seem particularly well equipped to answer is not just whether future people will have more resources and empirical information than we do, but whether they will be better situated than us in virtue of having progressed morally (see MacAskill 2022b: Ch. 3-4).

- **Should we expect to progress morally (see Sauer et al. 2021)?** What will be the cause(s) of such progress?
- **How impactful is it to focus on moral progress itself as a cause area (see Anthis & Paez 2021)?** Aside from advocacy leading to moral circle expansion, are there other ways of speeding up and/or directing moral progress ?
- **Should we defer resources to future generations because of expected moral progress?** If we expect to make moral progress in some form, this may speak in favour of keeping as many options as possible open for future generations.

4.4.4 Institutions

Many of the most important decisions today are taken by institutions like governments, universities, charitable foundations and companies, rather than by individuals. It is plausible that institutions often have different reasons for action than we do. If that is correct, how do these differences bear on the question of what institutions ought to do, all things considered?

- **Do governments have special duties to their citizens (see Frick 2020)?** Should governments use their resources to bring about impartially better outcomes, and if so when (Barrett 2022)? How can we square such duties with the democratic principle of accountability to present citizens, especially if those citizens' own values are not those of an impartial agent seeking to do the most good?
- **Do other areas of this research agenda (Longtermism, AI, Mind & Value) bear on questions of political representation?** Should we give representation to future people, digital minds, and/or animals, and if so how?

- **Should we welcome or oppose a stable world government?** The formation of a world government, if it ever occurs, may be one of the most significant events in political history. From a moral perspective, ought we to welcome or to oppose such a development? If the answer is a qualified one, is there anything we can do to help avoid the formation of a world government under adverse conditions, while promoting the formation of a world government under favourable conditions? What should the structure and/or constitution of such a world government be?
- **Is institutional and collective change more important than individual action?** If our goal is to do the most good, should influencing institutions, or focusing on collective change more generally, be expected to have greater impact than individual altruistic decisions (see Berkey 2018, Collins 2019, Schmidt and Barrett forthcoming)?

See also section 7 (Policy and Institutional) of GPI's research agenda in psychology.

Bibliography

Aghion, P., B. F. Jones, and C. I. Jones. 2019. Artificial intelligence and economic growth. In *The Economics of Artificial Intelligence: An Agenda*, eds. A. Agrawal, J. Gans, and A. Goldfarb, 237–282. Chicago, IL: University of Chicago Press.

Ahmed, A. 2014. *Evidence, Decision and Causality*. Cambridge: Cambridge University Press.

Alaga, J. and J. Schuett. 2023. Coordinated pausing: An evaluation-based coordination scheme for frontier AI developers. *Centre for the Governance of AI*. URL: https://www.governance.ai/research-paper/coordinated-pausing-evaluation-based-scheme

Albantakis, L., L. Barbosa, G. Findlay, M. Grasso, A. M. Haun, W. Marshall, W. G. P. Mayner, A. Zaeemzadeh, M. Boly, B. E. Juel, S. Sasai, K. Fujii, I. David, J. Hendren, J. P. Lang, and G. Tononi. 2023. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology 19* (10): 1–45.

Alexander, J. M. 2012. Decision Theory Meets the Witch of Agnessi. *Journal of Philosophy 109* (12): 712–727.

Alexander, S. 2023a. Davidson on takeoff speeds. *Astral Codex Ten*. URL: https://www.astralcodexten.com/p/davidson-on-takeoff-speeds?utm source=

Alexander, S. 2023b. Why I am not (as much of) a doomer (as some people). *Astral Codex Ten*. URL: https://astralcodexten.substack.com/p/why-i-am-not-as-much-of-a-doomer

Althaus, D. and T. Baumann. 2020. Reducing long-term risks from malevolent actors. *Center for Long Term Risk*. URL: https://longtermrisk.org/reducing-long-term-risks-from-malevolent-actors/

Andrews, K. and J. Birch. 2023. What has feelings? *Aeon*. URL: https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals

Angelini, V., D. Cavapozzi, L. Corazzini, and O. Paccagnella. 2014. Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases. *Oxford Bulletin of Economics and Statistics 76* (5): 643–666.

Anthis, J. R. and E. Paez. 2021. Moral circle expansion: A promising strategy to impact the far future. *Futures* 130: 102756.

Anthropic. 2023. Frontier threats red teaming for AI safety. URL: https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

Armstrong, D. M. 1968 [2023]. *A Materialist Theory of the Mind*. London: Routledge.

Armstrong, S. and B. Levinstein. 2017. Low impact artificial intelligences. arXiv:1705.10720. URL: https://arxiv.org/abs/1705.10720

Arntzenius, F. 2014. Utilitarianism, decision theory and eternity. *Philosophical Perspectives 28* (1): 31–58.

Arntzenius, F. and C. Dorr. 2017. Self-locating priors and cosmological measures. In *The Philosophy of Cosmology*, eds. K. Chamcham, J. Barrow, S. Saunders, and J. Silk, 396–428. Cambridge: Cambridge University Press.

Arrhenius, G. 2014. *Population Ethics: The Challenge of Future Generations*. Unpublished manuscript.

Arrow, K. J. 1966. Exposition of the theory of choice under uncertainty. *Synthese 16* (3/4): 253–269.

Askell, A. 2018. Pareto principles in infinite ethics. PhD Thesis, New York University, URL: https://askell.io/files/Askell-PhD-Thesis.pdf

Baard, P. 2022. *Ethics in Biodiversity Conservation*. Routledge.
Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bow- man, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. 2022. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073. URL: ttps://arxiv.org/abs/2212.08073.
Bain, D. 2013. What makes pains unpleasant? *Philosophical Studies 166* (1): 69–89.
Bain, D. 2019. Why take painkillers? *Noûs 53* (2): 462–490.
Baker, C. 2024. Expected choiceworthiness and fanaticism. *Philosophical Studies 181* (5).
Bales, A. forthcoming. Will AI avoid exploitation? Artificial general intelligence and expected utility theory. *Philosophical Studies*.
Bales, A. ms. Against willing servitude: Autonomy in the ethics of advanced artificial intelligence. Unpublished manuscript.
Balog, K. 2020. Hard, harder, hardest. In *Sensations, Thoughts, and Language: Essays in Honor of Brian Loar*, ed. A. Sullivan, 265–289. New York, NY: Routledge.
Balwit, A. 2023. How we can regulate AI. *Asterisk* 3: 70–77.
Barlassina, L. and M. K. Hayward. 2019. More of me! Less of me! Reflexive imperativism about affective phenomenal character. *Mind 128* (512): 1013–1044.
Barnett, M. and T. Besiroglu. 2023. Scaling transformative autoregressive models. *Epoch AI*. URL: https://epochai.org/files/direct-approach.pdf
Barrett, J. 2020. Social reform in a complex world. *Journal of Ethics and Social Philosophy 17* (2).
Barrett, J. 2022. Social beneficence. *Global Priorities Institute Working Paper No. 11-2022*. URL: https://globalprioritiesinstitute.org/jacob-barrett-social-beneficence-global-priorities-institute-university-of-oxford/
Barrett, J. ms. The case for weak longtermism. Unpublished manuscript.
Bartha, P. and C. Hitchcock. 1999. No one knows the date or the hour: An unorthodox application of Rev. Bayes's theorem. *Philosophy of Science 66* (3): 339-353.
Barrington, M. 2023. Absolutist AI. arXiv:2307.10315. URL: http://arxiv.org/abs/2307.10315
Bartha, P. 2007. Taking stock of infinite value: Pascal?s wager and relative utilities. *Synthese 154* (1): 5–52.
Bayne, T., A. K. Seth, and M. Massimini. 2020. Are there islands of awareness? *Trends in Neurosciences 43* (1): 6–16.
Beard, S., T. Rowe, and J. Fox. 2020. An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures* 115: 102469.
Beckman, L. 2009. *The Frontiers of Democracy*. London: Palgrave MacMillan.
Beckstead, N. and T. Thomas. 2024. A paradox for tiny probabilities and enormous values. *Noûs* 58(2): 431-55.
Bengio, Y., D., Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, D. Goldfarb, H. Heidari, L. Khalatbari, S. Longpre, V. Mavroudis, M. Mazeika, K. Y. Ng, C. T. Okolo, D. Raji, T. Skeadas, F. Tramèr, and S. Mindermann. 2024. *International Scientific Report on the Safety of Advanced AI: Interim Report*. UK Government Department for Science, Innovation and Technology and

AI Safety Institute. URL: https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf
Bennett, K. 2017. *Making Things Up*. Oxford: Oxford University Press.
Bergal, A. and N. Beckstead. 2021. Interpretability. *The AI Alignment Forum*. URL: https://www.alignmentforum.org/posts/CzZ6Fch4JSpwCpu6C/interpretability
Berkey, B. 2018. The institutional critique of effective altruism. *Utilitas 30* (2): 143–171.
Bickle, J. 2020. Multiple realizability. In *The Stanford Encyclopedia of Philosophy* (Summer 2020 ed.)., ed. E. N. Zalta. Metaphysics Research Lab, Stanford University.
Birch, J. 2022a. Materialism and the moral status of animals. *Philosophical Quarterly 72* (4): 795–815.
Birch, J. 2022b. The search for invertebrate consciousness. *Noûs 56* (1): 133–153.
Birch, J. forthcoming. When is a brain organoid a sentience candidate? *Molecular Psychology* .
Birch, J., C. Burn, A. Schnell, H. Browning, and A. Crump. 2021. Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans. *Department for Environment, Food and Rural Affairs*. URL: https://www.lse.ac.uk/business/consulting/reports/review-of-the-evidence-of-sentiences-in-cephalopod-molluscs-and-decapod-crustaceans
Blackorby, C., W. Bossert, and D. Donaldson. 1995. Intertemporal population ethics: Critical-level utilitarian principles. *Econometrica 63* (6): 1303–1320.
Blackorby, C., W. Bossert, and D. J. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.
Block, N. 1997. Anti-reductionism slaps back: Mental causation, reduction and supervenience. *Philosophical Perspectives* 11: 107–132.
Block, N. 2002. The harder problem of consciousness. *Journal of Philosophy 99* (8): 391–425.
Block, N. 2007. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences 30* (5): 481–548.
Block, N. 2009. Comparing the major theories of consciousness. In *The Cognitive Neurosciences IV*, ed. M. Gazzaniga, 1111–1123. Cambridge, MA: MIT Press.
Block, N. 2010. Attention and mental paint. *Philosophical Issues 20* (1): 23–63.
Block, N. 2023. *The Border Between Seeing and Thinking*. Oxford: Oxford University Press.
Boonin, D. 2014. *The Non-Identity Problem and the Ethics of Future People.* Oxford: Oxford University Press.
Bostrom, N. 2002. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York, NY: Routledge.
Bostrom, N. 2003. Are we living in a computer simulation? *Philosophical Quarterly 53* (211): 243–255.
Bostrom, N. 2011. Infinite ethics. *Analysis and Metaphysics* 10: 9–59.
Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
Bostrom, N. 2017. Strategic implications of openness in AI development. *Global Policy 8* (2): 135–148.
Bostrom, N. 2019. The vulnerable world hypothesis. *Global Policy 10* (4): 455–476.
Bostrom, N. and M. M. Církovíc. (Eds.) 2008. *Global Catastrophic Risks*. Oxford: Oxford University Press.
Bostrom, N., A. Dafoe, and C. Flynn. 2020. Public policy and superintelligent AI: A vector field

approach. In *Ethics of Artificial Intelligence*, ed. S. M. Liao, 293–326. Oxford: Oxford University Press.

Bostrom, N. and C. Shulman. forthcoming. Propositions concerning digital minds and society. *Cambridge Journal of Law, Politics, and Art* .

Bottomley, C. and T. L. Williamson. 2024. Rational risk-aversion: Good things come to those who weight. *Philosophy and Phenomenological Research 108* (3): 697–725.

Bourget, D. and A. Mendelovici. 2014. Tracking representationalism. In *Philosophy of mind: the key thinkers*, ed. A. Bailey, 209–235. New York, NY: Bloomsbury Academic.

Bourget, D. and A. Mendelovici. 2019. Phenomenal intentionality. In *The Stanford Encyclo- pedia of Philosophy* (Fall 2019 ed.)., ed. E. N. Zalta. Metaphysics Research Lab, Stanford University.

Bowen, J. and J. Basl. 2020. AI as a moral right-holder. In *The Oxford Handbook of Ethics of AI*, eds. M. D. Dubber, F. Pasquale, and S. Das, 289–306. Oxford: Oxford Handbooks.

Bradford, G. 2021. Perfectionist bads. *Philosophical Quarterly 71* (3): 586–604.

Bradford, G. 2022. Consciousness and welfare subjectivity. *Noûs 4* (75): 905–921.

Bradley, A. and B. Saad. 2024. AI alignment vs AI ethical treatment: Ten challenges. *Global Priorities Working Paper No. 19-2024*. URL: https://globalprioritiesinstitute.org/ai-alignment-vs-ai-ethical-treatment-ten-challenges-adam-bradley-bradford-saad/

Bradley, R. 2017. *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Bradley, R., C. Helgeson, and B. Hill. 2017. Climate change assessments: Confidence, probability, and decision. *Philosophy of Science 84* (3): 500–522.

Briggs, R. and D. Nolan. 2015. Utility monsters for the fission age. *Pacific Philosophical Quarterly 96* (2): 392–407.

Brock, D. W. 2003. Separate spheres and indirect benefits. *Cost effectiveness and resource allocation 1* (1): 4.

Broome, J. 1984. Selecting people randomly. *Ethics 95* (1): 38–55.

Broome, J. 1991. Fairness. *Proceedings of the Aristotelian Society* 91: 87–101.

Broome, J. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Broome, J. 2005. Should we value population? *Journal of Political Philosophy* 13 (4): 399-413

Browning, H. 2022. Assessing measures of animal welfare. *Biology and Philosophy 37* (4): 1–24.

Browning, H. 2023. Welfare comparisons within and across species. *Philosophical Studies 180* (2): 529–551.

Browning, H. and Veit, W. 2023. Positive Wild Animal Welfare. *Biology & Philosophy 38, 14.*

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

Buchak, L. 2019. Weighing the risks of climate change. *The Monist 102* (1): 66–83.

Buchak, L. 2023. How should risk and ambiguity affect our charitable giving? *Utilitas 35* (3): 175–197.

Buchanan, A. E. 2011. *Beyond Humanity? The Ethics of Biomedical Enhancement*. Oxford: Oxford University Press.

Builes, D. and C. Hare. 2023. Why aren't I part of a whale? *Analysis 83* (2): 227–234.

Butlin, P., R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon, and R. VanRullen. 2023. Consciousness in artificial intelligence: Insights from the science of consciousness. https://arxiv.org/abs/2308.08708. URL: https://arxiv.org/abs/2308.08708

Cao, R. 2022. Multiple realizability and the spirit of functionalism. *Synthese 200* (6): 1–31.
Cappelen, H. and J. Dever. 2021. *Making AI Intelligible: Philosophical Foundations*. Oxford: Oxford University Press.
Carlsmith, J. 2023. Crazy train. URL: https://joecarlsmith.com/category/crazy-train/
Carlsmith, J. forthcoming. Existential risk from powerseeking AI. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.
Carruthers, P. 2017. Valence and value. *Philosophy and Phenomenological Research 97* (3): 658–680.
Carruthers, P. 2019. *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford: Oxford University Press.
Carter, B. 1983. The Anthropic Principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society A 310* (1512): 347–363.
Chalmers, D. J. 1996a. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
Chalmers, D. J. 1996b. Does a rock implement every finite-state automaton? *Synthese 108* (3): 309–33.
Chalmers, D. J. 2010a. *The Character of Consciousness*. Oxford: Oxford University Press.
Chalmers, D. J. 2010b. The singularity: A philosophical analysis. *Journal of Consciousness Studies 17* (9-10): 7–65.
Chalmers, D. J. 2018. The meta-problem of consciousness. *Journal of Consciousness Stud- ies 25* (9-10): 6–61.
Chalmers, D. J. 2020. Debunking arguments for illusionism about consciousness. *Journal of Consciousness Studies 27* (5-6): 258–281.
Chalmers, D. J. 2022. *Reality +: Virtual Worlds and the Problems of Philosophy*. London: Allen Lane.
Chalmers, D. J. 2023. Could a large language model be conscious? arXiv:2303.07103. URL: https://arxiv.org/abs/2303.07103
Chappell, R. Y. 2021. Negative utility monsters. *Utilitas 33* (4): 417–421.
Chen, E. K. and D. Rubio. 2020. Surreal decisions. *Philosophy and Phenomenological Research 100* (1): 54–74.
Christian, B. 2020. *The Alignment Problem: How Can Machines Learn Human Values?* London: Atlantic Books.
Christiano, P. 2018. When is unaligned AI morally valuable? URL: https://ai-alignment.com/sympathizing-with-ai-e11a4bf5ef6e
Christiano, P. 2019. AI alignment landscape. URL: https://ai-alignment.com/ai-alignment-landscape-d3773c37ae38
Christiano, P., A. Cotra, and M. Xu. 2021. Eliciting latent knowledge: How to tell if your eyes deceive you. URL: https://docs.google.com/document/d/1WwsnJQstPq91Yh-Ch2XRL8HEpsnjrC1dwZXR37PC8/edit#heading=h.jrzi4atzacns
Christiano, P., B. Shlegeris, and D. Amodei. 2018. Supervising strong learners by amplifying weak experts. arXiv:1810.08575. URL: https://arxiv.org/abs/1810.08575
Cibinel, P. 2023. A dilemma for Nicolausian discounting. *Analysis 83* (4): 662–672.
Církovíc, M., A. Sandberg, and N. Bostrom. 2010. Anthropic shadow: observation selection effects and human extinction risks. *Risk Analysis 30* (10): 1495-1506.

Cockburn, J., J.-Y. Duclos, and A. Zabsonré. 2014. Is global social welfare increasing? A critical-level enquiry. *Journal of Public Economics* 118: 151–162.

Cohen, I. G., N. Daniels, and N. Eyal. (Eds.) 2015. *Identified versus Statistical Lives: An Interdisciplinary Perspective*. Oxford: Oxford University Press.

Collins, S. 2019. Beyond individualism. In *Effective Altruism: Philosophical Issues*, eds. H. Greaves and T. Pummer, 202–217. Oxford University Press.

Colyvan, M. 2008. Relative expectation theory. *Journal of Philosophy 105* (1): 37–44.

Colyvan, M. and A. Hájek. 2016. Making ado without expectations. *Mind 125* (499): 829–857.

Conitzer, V. and C. Oesterheld. 2023. Foundations of cooperative AI. *Proceedings of the AAAI Conference on Artificial Intelligence 37* (13): 15359–15367.

Cooney, N. 2014. *Veganomics: The Surprising Science on What Motivates Vegetarians, from the Breakfast Table to the Bedroom*. New York, NY: Lantern Books.

Cotra, A. 2020. Forecasting transformative AI with biological anchors. URL: https://drive.google.com/drive/u/0/folders/15ArhEPZSTYU8fo12bs6ehPS6-xmhtBPP

Cotra, A. 2021. Ajeya Cotra on worldview diversification and how big the future could be. Interviewed by R. Wiblin and K. Harris. *The 80,000 Hours Podcast*, January 19 2021. URL: https://80000hours.org/podcast/episodes/ajeya-cotra-worldview-diversification/next-best-worldviews-004105

Cotra, A. 2022a. Two-year update on my personal AI timelines. *The AI Alignment Fo- rum*. URL: https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines

Cotra, A. 2022b. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. *Less Wrong*. URL: https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to

Cowen, T. 2018. *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals*. San Francisco, CA: Stripe Press.

Cowen, T. and D. Parfit. 1992. Against the social discount rate. In *Justice between age groups and generations*, eds. P. Laslett and J. S. Fishkin, 144–161. New Haven, CT: Yale University Press.

Cowie, C. 2017. Does the Repugnant Conclusion have any probative force? *Philosophical Studies 174* (12): 3021–3039.

Crawford, L. 2013. Freak observers and the simulation argument. *Ratio 26* (3): 250–264.

Crummett, D. 2022. What if we contain multiple morally relevant subjects? *Utilitas 34* (3): 317–334.

Cullity, G. 2004. *The Moral Demands of Affluence*. Oxford: Oxford University Press.

Curran, E. J. forthcoming. Longtermism and the complaints of future people. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.

Cutter, B. 2016. Color and shape: A plea for equal treatment. *Philosophers' Imprint 16* (8): 1–11.

Cutter, B. 2017. The metaphysical implications of the moral significance of consciousness. *Philosophical Perspectives 31* (1): 103–130.

Cutter, B. and D. Crummett. forthcoming. Psychophysical harmony: A new argument for theism. *Oxford Studies in Philosophy of Religion*.

Dalbey, B. and B. Saad. 2022. Internal constraints for phenomenal externalists: A structure matching theory. *Synthese 200* (5): 1–29.

D'Alessandro, W. 2023. Is deontological AI safe? URL: https://drive.google.com/file/d/1UXnoOJj4kYF73spT VEHEUBzewNXFL4q/view

Daniels, N. 1993. Rationing fairly: Programmatic considerations. *Bioethics 7* (2-3): 224–233.

Dasgupta, P. 2008. Discounting climate change. *Journal of Risk and Uncertainty 37* (2/3): 141–169.

Davidson, T. 2023. What a compute-centric framework says about takeoff speeds. *Open Philanthropy*. URL: https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/

Dawkins, M. S. 2017. Animal welfare with and without consciousness. *Journal of Zool ogy 301* (1): 1–10.

de Canson, C. 2024. The nature of awareness growth. *Philosophical Review 133* (1): 1–32.

de Lazari-Radek, K. and P. Singer. 2012. The objectivity of ethics and the unity of practical reason. *Ethics 123* (1): 9–31.

Deaton, A. and N. Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science Medicine* 210: 2–21.

Dehaene, S. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York, NY: Viking Press.

Dennett, D. C. 1991. *Consciousness Explained*. London: Penguin Books.

Deutsch, D. 2011. *The Beginning of Infinity: Explanations That Transform The World*. London: Penguin.

Diener, E., C. Diener, H. Choi, and S. Oishi. 2018. Revisiting "Most people are happy"—and discovering when they are not. *Perspectives on Psychological Science 13* (2): 166–170.

Dogramaci, S. 2021. Are we playing a moral lottery? Moral disagreement from a metasemantic perspective. *Ergo: An Open Access Journal of Philosophy 8* (1): 523–550.

Donaldson, S. and W. Kymlicka. 2011. *Zoopolis: A Political Theory of Animal Rights*. Oxford: Oxford University Press.

Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Dretske, F. 1996. Phenomenal externalism, or if meanings ain't in the head, where are qualia? *Philosophical Issues* 7: 143–158.

Du Toit, J. and J. Millum. 2016. Are indirect benefits relevant to health care allocation decisions? *Journal of Medicine and Philosophy 41* (5): 540–557.

Dung, L. forthcoming. How to deal with risks of AI suffering. *Inquiry* .

Easterlin, R. A. 2017. Paradox lost? *Review of Behavioral Economics 4* (4): 311–339.

Easwaran, K. 2008. Strong and weak expectations. *Mind 117* (467): 633–641.

Easwaran, K. 2014. Principal values and weak expectations. *Mind 123* (490): 517–531.

Espinosa, R. and Treich, N. 2024. The animal welfare levy. TSE Working Paper, n. 24-1503.

Estlund, D. M. 2019. *Utopophobia: On the Limits (If Any) of Political Philosophy*. Princeton, NJ: Princeton University Press.

Evans, O., O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. 2021. Truthful ai: Developing and governing ai that does not lie. arXiv:2110.06674 . URL: https://arxiv.org/abs/2110.06674

Fabian, M. 2022. Scale norming undermines the use of life satisfaction scale data for welfare analysis. *Journal of Happiness Studies 23* (4): 1509–1541.

Feldman, F. 2006. Actual utility, the objection from impracticality, and the move to expected utility. *Philosophical Studies 129* (1): 49–79.

Finneron-Burns, E. 2017. What's wrong with human extinction? *Canadian Journal of Philosophy*

47 (2–3): 327–343.
Finnveden, L., C. J. Riedel, and C. Shulman. 2021. Artificial intelligence and lock-in. URL: https://docs.google.com/document/d/1mkLFhxixWdT5peJHq4rfFzq4QbHyfZtANH1nou68q88/editheading=h
Fischer, B. (Ed.) 2024. *Weighing Animal Welfare: Comparing Well-being Across Species*. Oxford: Oxford University Press.
Fischer, B., A. Shriver, and M. St. Jules. 2022. Do brains contain many con- scious subsystems? If so, should we act differently? *Rethink Priorities*. URL: https://rethinkpriorities.org/publications/do-brains-contain-many-conscious-subsystems
Fleurbaey, M. and D. Blanchet. 2013. *Beyond GDP: Measuring Welfare and Assessing Sustainability.* Oxford: Oxford University Press.
Frances, B. and J. Matheson. 2024. Disagreement. In *The Stanford Encyclopedia of Philosophy* (Summer 2024 ed.)., eds. E. N. Zalta and U. Nodelman. Metaphysics Research Lab, Stanford University.
Francione, G. 2008. *Animals as Persons: Essays on the Abolition of Animal Exploitation*. New York, NY: Columbia University Press.
Frank, J. 2008. Is there an "animal welfare Kuznets curve"? *Ecological Economics 66* (2): 478–491.
Frankish, K. 2016. Illusionism as a theory of consciousness. *Journal of Consciousness Studies 23* (11–12): 11–39.
Frick, J. 2017. On the survival of humanity. *Canadian Journal of Philosophy 47* (2–3): 344–367.
Frick, J. 2020. National partiality, immigration, and the problem of double-jeopardy. *Oxford Studies in Political Philosophy Volume* 6: 151–183.
Friederich, S. 2023. Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. PhilSci Archive. URL: https://philsci-archive.pitt.edu/22736/
Frijters, P., A. E. Clark, C. Krekel, and R. Layard. 2020. A happy choice: wellbeing as the goal of government. Behavioural Public Policy 4(2):126–165.
Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and Machines 30* (3): 411–437.
Gallow, J. D. forthcoming. Instrumental divergence. *Philosophical Studies*.
Gardner, M. 2021. A deontological approach to future consequences. In *The Oxford Handbook of Intergenerational Ethics*, ed. S. M. Gardiner. Oxford: Oxford University Press.
Gaus, G. 2016. *The Tyranny of the Ideal: Justice in a Diverse Society*. Princeton, NJ: Princeton University Press.
Geach, P. 1956. Good and evil. *Analysis 17* (2): 33–42.
Gibbons, M., A. Crump, M. Barrett, S. Sarlak, J. Birch, and L. Chittka. 2022. Can insects feel pain? A review of the neural and behavioural evidence. *Advances in Insect Physiology* 63: 155–229.
Gigerenzer, G. and W. Gaissmaier. 2011. Heuristic decision making. *Annual Review of Psychology 62* (1): 451–482.
Godfrey-Smith, P. 2016. Mind, matter, and metabolism. *Journal of Philosophy 113* (10): 481–506.
Goff, P. 2014. Orthodox property dualism + the linguistic theory of vagueness = panpsychism. In *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, ed. E. Brown, 75–91. Dordrecht: Springer.
Goff, P. 2018. Conscious thought and the cognitive fine-tuning problem. *Philosophical Quarterly*

68 (270): 98–122.
Gold, M. R., D. Stevenson, and D. G. Fryback. 2002. HALYs and QALYs and DALYs, oh my: Similarities and differences in summary measures of population health. *Annual Review of Public Health* 23: 115-134.
Goldstein, S. and C. D. Kirk-Giannini. forthcoming. Language agents reduce the risk of existential catastrophe. *AI and Society* .
Goldstein, S. and C. D. Kirk-Giannini. ms. AI wellbeing. URL: https://philpapers.org/rec/GOLAWE-4
González-Ricoy, I. and A. Gosseries. (Eds.) 2017. *Institutions for Future Generations*. Ox- ford: Oxford University Press.
Goodin, R. E. 2007. Enfranchising all affected interests, and its alternatives. *Philosophy and Public Affairs 35* (1): 40–68.
Goodsell, Z. 2021. A St Petersburg paradox for risky welfare aggregation. *Analysis 81* (3): 420-426.
Goodsell, Z. 2023. *Unbounded Utility*. PhD Thesis, University of Southern California, URL: https://philpapers.org/rec/GOOUUG.
Gott, J. R. 1993. Implications of the Copernican principle for our future prospects. *Nature* 363: 315–319.
Grace, K. 2022a. Counterarguments to the basic AI x-risk case. *The AI Alignment Forum*. URL: https://www.alignmentforum.org/posts/LDRQ5Zfqwi8GjzPYG/counterarguments-to-the-basic-ai-x-risk-case
Grace, K. 2022b. Let's think about slowing down AI. *The AI Alignment Forum*. URL: https://www.alignmentforum.org/posts/uFNgRumrDTpBfQGrs/let-s-think-about-slowing-down-ai.
Greaves, H. 2016. Cluelessness. *Proceedings of the Aristotelian Society 116* (3): 311–339.
Greaves, H. 2017. Discounting for public policy: A survey. *Economics and Philosophy 33* (3): 391–439.
Greaves, H., J. Barrett, and D. Thorstad, eds. forthcoming. *Essays on Longtermism*. Oxford: Oxford University Press.
Greaves, H. and O. Cotton-Barratt. 2023. A bargaining-theoretic approach to moral uncertainty. *Journal of Moral Philosophy 21* (1-2): 127–169.
Greaves, H. and W. MacAskill. 2021. The case for strong longtermism. *Global Priorities Institute Working Paper No. 5-2021*. URL: https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf
Greaves, H. and W. MacAskill. forthcoming. The case for strong longtermism. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.
Greaves, H. and T. Ord. 2017. Moral uncertainty about population axiology. *Journal of Ethics and Social Philosophy 12* (2): 135–167.
Greaves, H. and C. Tarsney. forthcoming. Minimal and expansive longtermism. In *Essays on Longtermism*, eds. D. Thorstad, J. Barrett, and H. Greaves. Oxford: Oxford University Press.
Greaves, H., T. Thomas, A. L. Mogensen, and W. MacAskill. 2024. On the desire to make a difference. *Philosophical Studies* 181: 1599-1626.
Groff, Z. and Ng, YK. 2019. Does suffering dominate enjoyment in the animal kingdom? An update to welfare biology. *Biology & Philosophy 34, 40*.

Gustafsson, J. E. and M. Peterson. 2012. A computer simulation of the argument from disagreement. *Synthese 184* (3): 387–405.
Gustafsson, J. E. and O. Torpman. 2014. In defence of my favourite theory. *Pacific Philosophical Quarterly 95* (2): 159–174.
Hadfield-Menell, D., A. Dragan, P. Abbeel, and S. Russell. 2024. Cooperative inverse reinforcement learning. arXiv:1606.03137. URL: https://arxiv.org/abs/1606.03137
Hanson, R. 2016. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford: Oxford University Press.
Hanson, R., D. Martin, C. McCarter, and J. Paulson. 2021. If loud aliens explain human earliness, quiet aliens are also rare. *The Astrophysics Journal 922* (2): 182.
Harman, G. 1983. Justice and moral bargaining. *Social Philosophy and Policy 1* (1): 114.
Hausman, D. M. 1995. The impossibility of interpersonal utility comparisons. *Mind 104* (415): 473–490.
Heikkinen, K. 2022. Strong longtermism and the challenge from anti-aggregative views. *Global Priorities Institute Working Paper 5-2022*. URL: https://globalprioritiesinstitute.org/karri-heikkinen-strong-longtermism-and-the-challenge-from-anti-aggregative-moral-views/
Helgeson, C. 2020. Structuring decisions under deep uncertainty. *Topoi 39* (2): 257–269.
Hendrycks, D. forthcoming. *Introduction to AI Safety, Ethics, and Society*. Boca Raton, FL: CRC Press.
Hendrycks, D., M. Mazeika, and T. Woodside. 2023. An overview of catastrophic AI risks. arXiv:2306.12001. URL: https://arxiv.org/abs/2306.12001
Hill, C. S. 1991. *Sensations: A Defense of Type Materialism*. Cambridge: Cambridge University Press.
Hogarth, I. 2023. We must slow down the race to God-like AI. *Financial Times* April 13 2023. URL: https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2
Hohwy, J. 2004. Evidence, explanation, and experience. *Journal of Philosophy 101* (5): 242–254.
Horta, O. 2010. Debunking the idyllic view of natural processes: Population dynamics and suffering in the wild. *Télos 17* (1): 73–90.
Horton, J. 2021. Partial aggregation in ethics. *Philosophy Compass 16* (3): 1–12.
Huemer, M. 2008. Revisionary intuitionism. *Social Philosophy and Policy 25* (1): 368–392.
Hurka, T. 2010. Asymmetries in value. *Noûs 44* (2): 199–223.
Hàjek, A. 2024. Pascal's Wager. In *The Stanford Encyclopedia of Philosophy* (Summer 2024 ed.)., eds. E. N. Zalta and U. Nodelman. Metaphysics Research Lab, Stanford University.
Ienca, M. 2023. Don't pause giant AI for the wrong reasons. *Nature Machine Intelligence* 5: 470–471.
Irving, G., P. Christiano, and D. Amodei. 2018. AI safety via debate. arXiv:1805.00899. URL: https://arxiv.org/abs/1805.00899
Isaacs, Y., J. Hawthorne, and J. S. Russell. 2022. Multiple universes and self-locating evidence. *Philosophical Review 131* (3): 241–294.
Jaakkola, N. and A. Millner. 2020. Nondogmatic climate policy. *National Bureau of Economic Research Working Paper 27413*. URL: https://www.nber.org/papers/w27413
Jacobson, H. 2013. Killing the messenger: Representationalism and the painfulness of pain. *Philosophical Quarterly 63* (252): 509–519.
James, W. 1890. *The Principles of Psychology, Volume 1*. New York, NY: Henry Holt and Company.

Jaquet, F. 2022. Speciesism and tribalism: Embarrassing origins. *Philosophical Studies 179* (3): 933–954.

Javier-Castellanos, A. A. 2021. Should the number of overlapping experiencers count? *Erkenntnis 88* (4): 1–23.

Johnston, M. 2016. The personite problem: Should practical reason be tabled? *Noûs 50* (4): 617–644.

Jones, C. I. and P. J. Klenow. 2016. Beyond GDP? Welfare across countries and time. *American Economic Review 106* (9): 2426–57.

Kagan, S. 2014. An introduction to ill-being. *Oxford Studies in Normative Ethics* 4: 261–88.

Kagan, S. 2019. *How to Count Animals, More or Less*. Oxford: Oxford University Press.

Kahane, G. 2010. Evolutionary debunking arguments. *Noûs 45* (1): 103–125.

Kamm, F. 2004. Deciding whom to help, health-adjusted life years and disabilities. In *Public Health, Ethics, and Equity*, eds. S. Anand, F. Peter, and A. Sen, 225–242. Oxford University Press UK.

Kamm, F. M. 1993. *Morality, Mortality, Volume 1: Death and Whom to Save From It*. Oxford: Oxford University Press.

Kammerer, F. 2019. The normative challenge for illusionist views of consciousness. *Ergo: An Open Access Journal of Philosophy* 6.

Kammerer, F. 2022. Ethics without sentience: Facing up to the probable insignificance of phenomenal consciousness. *Journal of Consciousness Studies 29* (3-4): 180–204.

Kapteyn, A., J. P. Smith, and A. Van Soest. 2013. Are Americans really less happy with their incomes? *Review of Income and Wealth 59* (1): 44–65.

Karnofsky, H. 2021. Weak point in 'most important century': Lock-in. *Cold Takes*. URL: https://www.cold-takes.com/weak-point-in-most-important-century-lock-in/

Kavka, G. and V. Warren. 1983. Political representation for future generations. In *Environmental Philosophy: A Collection of Readings*, eds. R. Elliott and A. Gare. University Park, PA: Pennsylvania State University Press.

Kelly, T. 2005. Moorean facts and belief revision, or can the skeptic win? *Philosophical Perspectives 19* (1): 179–209.

Kelly, T. 2008. Common sense as evidence: Against revisionary ontology and skepticism. *Midwest Studies in Philosophy 32* (1): 53–78.

Kelly, T. 2011. Following the argument where it leads. *Philosophical Studies 154* (1): 105– 124.

Kinniment, M., L. J. K. Sato, H. Du, B. Goodrich, M. Hasin, L. Chan, L. H. Miles, T. R. Lin, H. Wijk, J. Burget, A. Ho, E. Barnes, and P. Christiano. 2023. Evaluating language-model agents on realistic autonomous tasks. *Alignment Research Center*. URL: https://evals.alignment.org/Evaluating LMAs Realistic Tasks.pdf

Klein, C. 2007. An imperative theory of pain. *Journal of Philosophy 104* (10): 517–532.

Klein, C. 2018. Computation, consciousness, and "Computation and consciousness". In *The Routledge Handbook of the Computational Mind*, eds. M. Sprevak and M. Colombo, 297–309. New York, NY: Routledge.

Klein, C. and A. Barron. 2016. Insects have the capacity for subjective experience. *Animal Sentience 9* (1).

Korinek, A. and A. Balwit. 2022. Aligned with whom? Direct and social goals for AI systems. *National Bureau of Economic Research Working Paper 30017*. URL: https://www.nber.org/system/files/working papers/w30017/w30017.pdf

Korsgaard, C. M. 2018. *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: Oxford University Press.

Kosonen, P. 2022. Tiny probabilities of vast value. D.Phil. thesis, Oxford University.

Kosonen, P. 2023. Tiny probabilities and the value of the far future. *Global Priorities Institute Working Paper No. 1-2023*. URL: https://globalprioritiesinstitute.org/tiny-probabilities-and-the-value-of-the-far-future-petra-kosonen/

Kowalczyk, K. and N. Venkatesh. 2024. Risk, non-identity, and extinction. *The Monist* 107 (2): 146–156.

Krakovna, V. 2024. AI alignment resources. URL: https://vkrakovna.wordpress.com/ai-safety-resources/.

Kristoffersen, I. 2017. The metrics of subjective wellbeing data: An empirical evaluation of the ordinal and cardinal comparability of life satisfaction scores. *Social Indicators Research 130* (2): 845–865.

Lawford-Smith, H. 2012. Understanding political feasibility. *Journal of Political Philosophy 21* (3): 243–259.

Layard, P. R. 2005. *Happiness: Lessons From a New Science*. London: Penguin.

LeDoux, J. E. 1998. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York, NY: Simon & Schuster.

Lee, A. Y. 2019. The microstructure of experience. *Journal of the American Philosophical Association 5* (3): 286–305.

Lee, A. Y. 2023. Degrees of consciousness. *Noûs 57* (3): 553–575.

Lee, G. 2013. Materialism and the epistemic significance of consciousness. In *Current Controversies in Philosophy of Mind*, ed. U. Kriegel, 222–245. London: Routledge.

Lee, G. 2019. Alien subjectivity and the importance of consciousness. In *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*, eds. A. Pautz and D. Stoljar, 215–242. Cambridge, MA: MIT Press.

Leike, J., D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. 2018. Scalable agent alignment via reward modeling: a research direction. arXiv:1811.07871 . URL: https://arxiv.org/abs/1811.07871

Lenman, J. 2000. Consequentialism and cluelessness. *Philosophy and Public Affairs 29* (4): 342–370.

Leslie, J. 1996. *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.

Levin, J. 2008. Taking type-B materialism seriously. *Mind and Language 23* (4): 402–425.

Levin, J. 2023. Functionalism. In *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.)., eds. E. N. Zalta and U. Nodelman. Metaphysics Research Lab, Stanford University.

Levinstein, B. A. and N. Soares. 2020. Cheating death in Damascus. *Journal of Philosophy 117* (5): 237–266.

Levy, N. 2014. The value of consciousness. *Journal of Consciousness Studies 21* (1-2): 127–138.

Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy 59* (1): 5–30.

Li, H. and B. Saad. 2022. Panpsychism and ensemble explanations. *Philosophical Studies 179* (12): 3583–3597.

Liao, S. M. 2020. The moral status and rights of artificial intelligence. In *Ethics of Artificial Intelligence*, ed. S. M. Liao, 480–504. Oxford: Oxford University Press.

Lin, E. 2022. Well-being, part 2: Theories of well-being. *Philosophy Compass 17* (2): e12812.
Lippert-Rasmussen, K. and S. Lauridsen. 2010. Justice and the allocation of healthcare resources: Should indirect, non-health effects count? *Medicine, Health Care and Philosophy 13* (3): 237–246.
Lloyd, H. 2021. Time discounting, consistency and special obligations: A defence of Robust Temporalism. *Global Priorities Institute Working Paper 11- 2021*. URL: https://globalprioritiesinstitute.org/wp-content/uploads/Time-discounting-consistency-and-special-obligations Harry-R-Lloyd.pdf
Lycan, W. G. 2001. The case for phenomenal externalism. *Noûs 35* (s15): 17–35.
Lyon, A. and E. Pacuit. 2013. The wisdom of crowds: Methods of human judgement aggregation. In *Handbook of Human Computation*, ed. P. Michelucci, 599–614. Dordrecht: Springer.
MacAskill, W. 2022a. Are we living at the hinge of history? In *Ethics and Existence: The Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, 331–357. Oxford: Oxford University Press.
MacAskill, W. 2022b. *What We Owe the Future: A Million-Year View*. New York, NY: Basic Books.
MacAskill, W. ms-a. Human extinction, asymmetry, and option value. Unpublished manuscript.
MacAskill, W. ms-b. Should we expect moral convergence? Unpublished manuscript.
MacAskill, W., K. Bykvist, and T. Ord. 2020. *Moral Uncertainty*. Oxford University Press.
MacAskill, W. and T. Ord. 2020. Why maximize expected choice-worthiness? *Noûs 54* (2): 327–353.
MacAskill, W., A. Vallinder, C. Oesterheld, C. Shulman, and J. Treutlein. 2021. The evidentialist's wager. *Journal of Philosophy 118* (6): 320–342.
Manley, D. ms. On being a random sample. URL: https://philpapers.org/rec/MANOBA
Marchau, V. A. W. J., W. E. Walker, P. J. T. M. Bloemen, and S. W. Popper. (Eds.) 2019. *Decision Making under Deep Uncertainty: From Theory to Practice*. Cham: Springer.
Mayerfeld, J. 1999. *Suffering and Moral Responsibility*. Oxford: Oxford University Press.
McGrath, S. 2008. Moral disagreement and moral expertise. *Oxford Studies in Metaethics* 3: 87-108.
McMahan, J. 1981. Problems of population theory. *Ethics 92* (1): 96–127.
McMahan, J. 2008. Eating Animals the Nice Way. *Daedalus, Journal of the American Academy of Arts and Sciences.*
McPherson, T. 2009. Moorean arguments and moral revisionism. *Journal of Ethics and Social Philosophy* (2): 1–25.
McQueen, K. J. 2019. Interpretation-neutral integrated information theory. *Journal of Consciousness Studies 26* (1-2): 76–106.
Meacham, C. J. G. 2020. Too much of a good thing: Decision-making in cases with infinitely many utility contributions. *Synthese 198* (8): 7309–7349.
Mellor, D. J., N. J. Beausoleil, K. E. Littlewood, A. N. McLean, P. D. McGreevy, B. Jones, and C. Wilkins. 2020. The 2020 Five Domains model: Including human-animal interactions in assessments of animal welfare. *Animals 10* (10): 1870.
Menger, K. 1934. Das Unsicherheitsmoment in der Wertlehre. *Zeitschrift für Nationalökonomie* 5: 459–485.
Metzinger, T. 2021. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness 1* (8): 1–24.
Millner, A. 2020. Nondogmatic social discounting. *American Economic Review 110* (3): 760–75.

Mislove, A. 2023. Red-teaming large language models to identify novel AI risks. *The White House*. URL: https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/

Mogensen, A. L. 2017. Disagreements in moral intuition as defeaters. *The Philosophical Quarterly* 67(267): 282–302.

Mogensen, A. L. 2019a. Doomsday rings twice. *Global Priorities Institute Working Paper No. 1-2019*. URL: https://globalprioritiesinstitute.org/andreas-mogensen-doomsday-rings-twice/

Mogensen, A. L. 2019b. Staking our future: deontic long-termism and the Non-Identity Problem. *Global Priorities Institute Working Paper - No. 9-2019*. URL: https://globalprioritiesinstitute.org/andreas-mogensen-staking-our-future-deontic-long-termism-and-the-non-identity-problem/

Mogensen, A. L. 2020. Meaning, medicine, and merit. *Utilitas 32* (1): 90–107.

Mogensen, A. L. 2021. Maximal cluelessness. *Philosophical Quarterly 71* (1): 141–162.

Mogensen, A. L. 2022. The only ethical argument for positive δ? Partiality and pure time preference. *Philosophical Studies 179* (9): 2731–2750.

Mogensen, A. L. 2023a. Philosophical considerations relevant to valuing continued human survival: Conceptual analysis, population axiology, and decision theory. *Global Priorities Institute Working Paper No. 9-2023*. URL: https://globalprioritiesinstitute.org/andreas-mogensen-philosophical-considerations-relevant-to-valuing-continued-human-survival-conceptual-analysis-population-axiology-and-decision-theory/

Mogensen, A. L. 2023b. The Hinge of History Hypothesis: Reply to MacAskill. *Analysis 84* (1): 47–55.

Mogensen, A. L. 2023c. Welfare and felt duration. *Global Priorities Institute Working Paper No. 14 - 2023*. URL: https://globalprioritiesinstitute.org/welfare-and-felt-duration-andreas-mogensen/

Mogensen, A. L. 2024. Respect for others' risk attitudes and the long-run future. *Noûs*.

Mogensen, A. L. forthcoming. Would a world without us be worse? Clues from population axiology. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.

Mogensen, A. L. and W. MacAskill. 2021. The paralysis argument. *Philosophers' Imprint 21* (15).

Mogensen, A. L. and D. Thorstad. 2022. Tough enough? Robust satisficing as a decision norm for long-term policy analysis. *Synthese 200* (1): 1–26.

Monton, B. 2019. How to avoid maximizing expected utility. *Philosophers' Imprint* 19.

Mørch, H. H. 2017. The evolutionary argument for phenomenal powers. *Philosophical Perspectives 31* (1): 293–316.

Morgan, J. B. 2023. What the senses cannot 'say'. *Philosophical Quarterly 73* (2): 557–579.

Muehlhauser, L. 2017. Report on consciousness and moral patienthood. *Open Philanthropy*. URL: https://www.openphilanthropy.org/research/2017-report-on-consciousness-and-moral-patienthood/

Nagel, T. 1978. Ruthlessness in public life. In *Public and Private Morality*, 75–91. Cambridge: Cambridge University Press.

Nanda, N. 2022. A comprehensive mechanistic interpretability explainer glossary. URL: https://neelnanda.io/glossary

Narveson, J. 1967. Utilitarianism and new generations. *Mind 76* (301): 62–72.

Nelson, M. T. 1991. Utilitarian eschatology. *American Philosophical Quarterly* 28(4): 339–47.
Ng, Y.-K. 1986. Social criteria for evaluating population change: An alternative to the Blackorby - Donaldson criterion. *Journal of Public Economics 29* (3): 375–381.
Ng, Y.-K. 1995. Towards welfare biology: Evolutionary economics of animal consciousness and suffering. *Biology and Philosophy 10* (3): 255–285.
Ng, Y.-K. 1996. Happiness surveys: Some comparability issues and an exploratory survey based on just perceivable increments. *Social Indicators Research 38* (1): 1–27.
Ng, Y.-K. 2008. Happiness studies: Ways to improve comparability and some public policy implications. *Economic Record 84* (265): 253–266.
Ngo, R., L. Chan, and S. Mindermann. 2022. The alignment problem from a deep learning perspective. arXiv:2209.00626. URL: https://arxiv.org/abs/2209.00626
Nordhaus, W. D. 2021, January). Are we approaching an economic singularity? information technology and the future of economic growth. *American Economic Journal: Macroeconomics 13* (1): 299–332.
Northcott, R. 2022. Reflexivity and fragility. *European Journal for Philosophy of Science 12* (3): 1–14.
Nover, H. and A. Hàjek. 2004. Vexing expectations. *Mind 113* (450): 237–249.
Nozick, R. 1969. Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday*, ed. N. Rescher, 114–146. Dordrecht: D. Reidel.
Nozick, R. 1974. *Anarchy, State, and Utopia*. New York, NY: Basic Books.
Nussbaum, M. C. 2022. *Justice for Animals: Our Collective Responsibility*. New York, NY: Simon & Schuster.
Oesterheld, C. 2017. Multiverse-wide cooperation via correlated decision making. *Cen- ter On Long-Term Risk*. URL: https://longtermrisk.org/multiverse-wide-cooperation-via-correlated-decision-making/
Olah, C., N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. 2020. Zoom in: An introduction to circuits. *Distill*. URL: https://distill.pub/2020/circuits/zoom-in
Olsson, C. V. and G. Alexandrie. 2019. A beef with growth? An empirical analysis of income and animal farming. Bachelor Thesis, Stockholm University. URL: https://static1.squarespace.com/static/5b9bacb89772aeb4dcbb5cb3/t/5da4557535c26c463ccb88e4/1571050871 4
Ord, T. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.
Pallies, D. 2022. Attraction, aversion, and asymmetrical desires. *Ethics 132* (3): 598–620.
Palmer, C., Fischer, B., Gamborg, C., Hampton, J. and Sandoe, P. *Wildlife Ethics: The Ethics of Wildlife Management and Conservation*. Newark: John Wiley & Sons.
Panksepp, J. 2005. Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition 14* (1): 30–80.
Papineau, D. 2002. *Thinking About Consciousness*. Oxford: Oxford University Press.
Parfit, D. 1978. Innumerate ethics. *Philosophy and Public Affairs 7* (4): 285–301.
Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
Pautz, A. 2013. The real trouble for phenomenal externalists: New empirical evidence for a brain-based theory of consciousness. In *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, ed. R. Brown, 237–298. Dordrecht: Springer.
Pautz, A. 2014. The real trouble for armchair arguments against phenomenal externalism. In

New Waves in Philosophy of Mind, eds. M. Sprevak and J. Kallestrup, 153–181. London: Palgrave-Macmillan.

Pautz, A. 2015. A dilemma for Russellian monists about consciousness. URL: https://philarchive.org/archive/PAUCRMv1

Pautz, A. 2017. The significance argument for the irreducibility of consciousness. *Philosophical Perspectives 31* (1): 349–407.

Pautz, A. 2020. The puzzle of the laws of appearance. *Philosophical Issues 30* (1): 257–272.

Pearce, D. 1995. *The Hedonistic Imperative*. URL: https://www.hedweb.com/hedethic/tabconhi.htm

Peters, U. 2023. Do current evidential standards in the science of consciousness help or hinder the discovery of signs of consciousness? Presentation at the *Detecting Unusual Consciousness* conference, Bonn (September 27, 2023).

Petersen, S. 2011. Designing people to serve. In *Robot Ethics: The Ethical and Social Implications of Robotics*, eds. P. Lin, K. Abney, and G. A. Bekey, 283–298. Cambridge, MA: MIT Press.

Peterson, M. 2019. The value alignment problem: A geometric approach. *Ethics and Information Technology 21* (1): 19–28.

Peterson, M. 2023. The St. Petersburg Paradox. In *The Stanford Encyclopedia of Philosophy* (Fall 2023 ed.)., eds. E. N. Zalta and U. Nodelman. Metaphysics Research Lab, Stanford University.

Pettigrew, R. 2024. Should longtermists recommend hastening extinction rather than delaying it? *The Monist 107* (2): 130–145.

Pivato, M. 2023. Cesàro average utilitarianism in relativistic spacetime. *Social Choice and Welfare 61* (4): 733 – 761.

Pogge, T. 2002. *World Poverty and Human Rights: Cosmopolitan Responsibilities and Reforms*. Cambridge: Polity Press.

Putnam, H. 1987. *Representation and Reality*. Cambridge, MA: MIT Press.

Ratoff, W. 2021. Can the predictive processing model of the mind ameliorate the value-alignment problem? *Ethics and Information Technology 23* (4): 739–750.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Raymont, P. 2005. Some experienced qualities belong to the experience. URL: https://philpapers.org/archive/RAYSEQ-2.pdf

Regan, T. 1983. *The Case for Animal Rights*. Berkeley, CA: University of California Press.

Rescorla, M. 2020. The Computational Theory of Mind. In *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.)., ed. E. N. Zalta. Metaphysics Research Lab, Stanford University.

Riedener, S. 2021. Human extinction from a Thomist perspective. In *Effective Altruism and Religion: Synergies, Tensions, Dialogue*, eds. S. Riedener, D. Roser, and M. Huppenbauer, 187–210. Baden-Baden: Nomos.

Riedener, S. forthcoming. Authenticity, meaning and alienation: Reasons to care less about far future people. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.

Rinard, S. 2013. Why philosophy can overturn common sense. *Oxford Studies in Epistemology* 4: 85–213.

Ritchie, J. B. and G. Piccinini. 2018. Computational implementation. In *The Routledge Handbook of the Computational Mind*, eds. M. Sprevak and M. Colombo, 192–204. New York, NY: Routledge.

Roelofs, L. 2023. Sentientism, motivation, and philosophical vulcans. *Pacific Philosophical*

Quarterly 104 (2): 301–323.
Roelofs, L. 2024. No such thing as too many minds. *Australasian Journal of Philosophy 102* (1): 131–146.
Roelofs, L. and J. Sebo. 2024. Overlapping minds and the hedonic calculus. *Philosophical Studies*.
Roser, M. 2023. AI timelines. *Our World in Data*. URL: https://ourworldindata.org/ai-timelines
Ross, J. 2006. Rejecting ethical deflationism. *Ethics 116* (4): 742–768.
Roussos, J. 2020. Policymaking under scientific uncertainty. PhD thesis, London School of Economics.
Rowe, T. and A. Voorhoeve. 2018. Egalitarianism under severe uncertainty. *Philosophy and Public Affairs 46* (3): 239–268.
Ruhmkorff, S. 2011. Some difficulties for the problem of unconceived alternatives. *Philosophy of Science 78* (5): 875–886.
Russell, J. S. 2023. On two arguments for fanaticism. *Noûs* Early View.
Russell, J. S. forthcoming. The value of normative information. *Australasian Journal of Philosophy* .
Russell, J. S. and Y. Isaacs. 2021. Infinite prospects. *Philosophy and Phenomenological Research* 103(1): 178–198.
Russell, S. 2019. *Human Compatible: AI and the Problem of Control*. London: Penguin.
Saad, B. 2019. A teleological strategy for solving the meta-problem of consciousness. *Journal of Consciousness Studies 26* (9-10): 205–216.
Saad, B. 2022. Harmony in a panpsychist world. *Synthese 200* (6): 1–24.
Saad, B. 2023. Simulations and catastrophic risks. *Sentience Institute*. URL: https://www.sentienceinstitute.org/downloads/Simulations and Catastrophic Risks.pdf
Saad, B. 2024*a*. Lessons from the void: What Boltzmann brains teach. *Analytic Philosophy* .
Saad, B. 2024*b*. Should dualists locate the physical basis of experience in the head? *Synthese 203* (2): 1–18.
Saad, B. 2024*c*. The sooner the better: an argument for bias toward the earlier. *Journal of the American Philosophical Association* 10 (2): 371-386.
Saad, B. and Bradley, A., 2022. Digital suffering: Why it's a problem and how to prevent it. *Inquiry: An Interdisciplinary Journal of Philosophy*, pp.1-36.
Saad, B. and Caviola. L. 2024. Digital minds takeoff scenarios. *Effective Altruism Forum*. URL: https://forum.effectivealtruism.org/posts/2uGShxLsXWGExJYNL/digital-minds-takeoff-scenarios
Sainsbury, M. 2022. Visual experience and the laws of appearance. *Erkenntnis 88* (7): 2933–2940.
Sandbrink, J., H. Hobbs, J. Swett, A. Dafoe, and A. Sandberg. 2022. Differential technology development: An innovation governance consideration for navigating technology risks. *Social Science Research Network*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213670
Sauer, H., C. Blunden, C. Eriksen, and P. Rehren. 2021. Moral progress: Recent developments. *Philosophy Compass 16* (10): e12769.
Saunders-Hastings, E. 2019. Benevolent giving and the problem of paternalism. In *Effective Altruism: Philosophical Issues*, eds. H. Greaves and T. Pummer, 115–136. Oxford: Oxford University Press.
Schaffer, J. 2015. What not to multiply without necessity. *Australasian Journal of Philosophy 93* (4): 644–664.

Scheffler, S. 2018. *Why Worry About Future Generations?* Oxford: Oxford University Press.
Schmidt, A. T. and J. Barrett. forthcoming. Longtermist political philosophy: an agenda for future research. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.
Schneider, S. 2020. How to Catch an AI Zombie: Testing for Consciousness in Machines. In *Ethics of Artificial Intelligence*, 439–458. Oxford University Press.
Schukraft, J. 2020. The subjective experience of time: Welfare implications. *Effective Altruism Forum*. URL: https://forum.effectivealtruism.org/posts/qEsDhFL8mQARFw6Fj/the-subjective-experience-of-time-welfare-implications
Schwitzgebel, E. 2002. A phenomenal, dispositional account of belief. *Noûs 36* (2): 249–275.
Schwitzgebel, E. 2015. If materialism is true, the United States is probably conscious. *Philosophical Studies 172* (7): 1697–1721.
Schwitzgebel, E. 2023. The coming robot rights catastrophe. *Blog of the APA*. URL: https://blog.apaonline.org/2023/01/12/the-coming-robot-rights-catastrophe
Schwitzgebel, E. and M. Garza. 2015. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy 39* (1): 98–119.
Schwitzgebel, E. and M. Garza. 2020. Designing AI with rights, consciousness, self-respect, and freedom. In *Ethics of Artificial Intelligence*, ed. S. M. Liao, 459–479. Oxford: Oxford University Press.
Searle, J. R. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association 64* (3): 21–37.
Searle, J. R. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.
Sebo, J. 2023. The Rebugnant Conclusion: Utilitarianism, insects, microbes, and AI sys- tems. *Ethics, Policy and Environment 26* (2): 249–264.
Sebo, J. and R. Long. 2023. Moral consideration for AI systems by 2030. *AI and Ethics*.
Sen, A. 2009. *The Idea of Justice*. Cambridge, MA: Harvard University Press.
Setiya, K. 2012. *Knowing Right From Wrong*. Oxford: Oxford University Press.
Shiller, D. 2017a. Hidden qualia. *Review of Philosophy and Psychology 8* (1): 165–180. Shiller, D. 2017b. In defense of artificial replacement. *Bioethics 31* (5): 393–399.
Shiller, D. 2024. Functionalism, integrity, and digital consciousness. *Synthese 203* (2): 1–20.
Shoemaker, S. 2006. The Frege-Schlick view. In *Content and Modality: Themes From the Philosophy of Robert Stalnaker*, ed. J. J. Thomson, 18–33. Oxford: Oxford University Press.
Shriver, A. 2022. What neuron counts can and can't tell us about moral weight. *Rethink Priorities*. URL: https://docs.google.com/document/d/1p5ovw84- ry2taYmyOIl4B91j7wkCurlB
Shulman, C. and N. Bostrom. 2012. How hard is artificial intelligence? Evolutionary arguments and selection effects. *Journal of Consciousness Studies 19* (7-8): 103–130.
Shulman, C. and N. Bostrom. 2021. Sharing the world with digital minds. In *Rethinking Moral Status*, eds. S. Clarke, H. Zohny, and J. Savulescu, 306–326. Oxford: Oxford University Press.
Shulman, C. and E. Thornley. forthcoming. How much should governments pay to prevent catastrophes? Longtermism's limited role. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.
Silva, L. 2023. Towards an affective quality space. *Journal of Consciousness Studies 30* (7): 164–195.
Simon, J. A. 2017. The hard problem of the many. *Philosophical Perspectives 31* (1): 449– 468.

Singer, P. 1972. Famine, affluence, and morality. *Philosophy & Public Affairs* 1(3): 229–243.
Singer, P. 1993. *Practical Ethics*, 2nd. ed. Cambridge: Cambridge University Press.
Sinhababu, N. 2022. Naturalist arguments for ethical hedonism. URL: https://utilitarianism.net/guest-essays/naturalistic-arguments-for-ethical-hedonism/
Sinnott-Armstrong, W. and V. Conitzer. 2021. How much moral status could artificial intelligence ever achieve? In *Rethinking Moral Status*, eds. S. Clarke, H. Zohn, and J. Savulescu, 269–281. Oxford: Oxford University Press.
Smith, N. J. J. 2014. Is evaluative compositionality a requirement of rationality? *Mind 123* (490): 457–502.
Smithies, D. 2019. *The Epistemic Role of Consciousness*. Oxford: Oxford University Press.
Smithies, D. and J. Weiss. 2019. Affective experience, desire, and reasons for action. *Analytic Philosophy 60* (1): 27–54.
Snyder-Beattie, A. E., T. Ord, and M. B. Bonsall. 2019. An upper bound for the background rate of human extinction. *Nature Scientific Reports 9* (11054).
Sotala, K. and L. Gloor. 2017. Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica 41* (1).
Speaks, J. 2022. Pautz on the laws of appearance, internalism, and color realism. URL: https://www3.nd.edu/ jspeaks/papers/pautz.pdf
Stanford, P. K. 2001. Refusing the devil's bargain: What kind of underdetermination should we take seriously? *Philosophy of Science 68* (S3): 1–12.
Steele, K. and H. O. Stefánsson. 2021. *Beyond Uncertainty: Reasoning with Unknown Possibilities*. Cambridge: Cambridge University Press.
Stevenson, B. and J. Wolfers. 2008. Economic growth and subjective well-being: Reassessing the Easterlin paradox. *Brookings Papers on Economic Activity 39* (1): 1–102.
Street, S. 2006. A Darwinian dilemma for realist theories of value. *Philosophical Studies 127* (1): 109–166.
Sumner, W. 2020. The worst things in life. *Grazer Philosophische Studien 97* (3): 419–432.
Tännsjö, T. 2007. Future people, the all affected principle, and the limits of the aggregation model of democracy. In *Homage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, eds. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson, and D. Egonsson. Lund: Department of Philosophy, Lund University.
Tarsney, C. 2020. Exceeding expectations: Stochastic dominance as a general decision theory. *Global Priorities Institute Working Paper No. 3-2020*. URL: https://globalprioritiesinstitute.org/christian-tarsney-exceeding-expectations-stochastic-dominance-as-a-general-decision-theory/
Tarsney, C. 2023a. Against anti-fanaticism. *Global Priorities Institute Working Paper No. 15-2023*. URL: https://globalprioritiesinstitute.org/against-anti-fanaticism-christian- tarsney/.
Tarsney, C. 2023b. The epistemic challenge to longtermism. *Synthese* 201(6): 1-37.
Tarsney, C. and T. Thomas. 2024. Non-additive axiologies in large worlds. *Ergo: An Open Access Journal of Philosophy* 11.
Tarsney, C. and H. Wilkinson. forthcoming. Longtermism in an infinite world. In *Essays on Longtermism*, eds. H. Greaves, J. Barrett, and D. Thorstad. Oxford: Oxford University Press.
Taurek, J. 1977. Should the numbers count? *Philosophy and Public Affairs 6* (4): 293–316.
Taylor, J. 2013. Is consciousness science fundamentally flawed? *Journal of Consciousness Studies 20* (3-4): 203–221.

Temkin, L. S. 2022. *Being Good in a World of Need*. Oxford: Oxford University Press.
Thoma, J. 2018. Risk aversion and the long run. *Ethics* 19(2): 230–253
Thoma, J. 2019. Decision Theory. *The Open Handbook of Formal Epistemology*. URL: https://philpapers.org/archive/PETTOH-2.pdf.
Thoma, J. 2023. Taking risks on behalf of another. *Philosophy Compass 18* (3): e12898.
Thomas, T. 2021a. Doomsday and objective chance. *Global Priorities Institute Working Paper No. 8-2021*. URL: https://globalprioritiesinstitute.org/doomsday-and-objective-chance-teruji-thomas/
Thomas, T. 2021b. Simulation expectation. *Global Priorities Institute Working Paper No. 16-2021*. URL: https://globalprioritiesinstitute.org/wp-content/uploads/Teruji-Thomas-Simulation-expectation.pdf
Thomas, T. 2023. The asymmetry, uncertainty, and the long term. *Philosophy and Phenomenological Research 107* (2): 470–500.
Thomas, T. 2024. Dispelling the anthropic shadow. *Global Priorities Institute Working Paper No. 20-2024*. URL: https://globalprioritiesinstitute.org/dispelling-the-anthropic-shadow-teruji-thomas/
Thomson, J. J. 1997. The right and the good. *Journal of Philosophy 94* (6): 273–298.
Thornley, E. 2024. The shutdown problem: Incomplete preferences as a solution. *The AI Alignment Forum*. URL: https://www.alignmentforum.org/posts/YbEbwYWkf8mv9jnmi/the-shutdown-problem-incomplete-preferences-as-a-solution
Thornley, E. forthcoming. The shutdown problem: An AI engineering puzzle for decision theorists. *Philosophical Studies*.
Thornley, E., A. Roman, C. Ziakas, L. Ho, and L. Thomson. 2024. Towards shutdownable agents via stochastic choice. *Global Priorities Institute Working Paper No. 16-2024*. URL: https://globalprioritiesinstitute.org/towards-shutdownable-agents-via-stochastic-choice-thornley-roman-ziakas-ho-thomson/
Thorstad, D. 2023. High risk, low reward: A challenge to the astronomical value of existential risk mitigation. *Philosophy and Public Affairs 51* (4): 373–412.
Thorstad, D. forthcoming-a. Against the singularity hypothesis. *Philosophical Studies*.
Thorstad, D. forthcoming-b. General-purpose institutional decision-making heuristics: The case of decision-making under deep uncertainty. *British Journal for the Philosophy of Science*.
Thorstad, D. forthcoming-c. Mistakes in the moral mathematics of existential risk. *Ethics*.
Thorstad, D. forthcoming-d. The scope of longtermism. *Australasian Journal of Philosophy*.
Thorstad, D. and A. L. Mogensen. 2020. Heuristics for clueless agents: how to get away with ignoring what matters most in ordinary decision-making. *Global Priorities Institute Working Paper No. 2-2020*. URL: https://globalprioritiesinstitute.org/david-thorstad-and-andreas-mogensen-heuristics-for-clueless-agents-how-to-get-away-with-ignoring-what-matters-most-in-ordinary-decision-making/
Titelbaum, M. G. 2008. The relevance of self-locating beliefs. *Philosophical Review 117* (4): 555–606.
Tomasik, B. 2015. The importance of wild-animal suffering. *Relations: Beyond Anthropocentrism 3* (2): 133–152.
Tomasik, B. 2017. *Artificial Intelligence and Its Implications for Future Suffering*. Basel:

Foundational Research Institute.

Tononi, G. 2008. Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin 215* (3): 216 – 242.

Tononi, G., M. Boly, M. Massimini, and C. Koch. 2016. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience* 17: 450–461.

Trammell, P. 2021. Patient philanthropy in an impatient world. URL: https://docs.google.com/document/d/1NcfTgZsqT9k3ongeQbappYyn-UO4vltjkm64n4or5r4/edit

Trammell, P. and A. Korinek. 2020. Economic growth under transformative AI: A guide to the vast range of possibilities for output growth, wages, and the labor share. *Global Priorities Institute Working Paper No. 8- 2020*. URL: https://globalprioritiesinstitute.org/wp-content/uploads/Philip-Trammell-and-Anton-Korinek economic-growth-under-transformative-ai.pdf

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenom- enal Mind*. Cambridge, MA: MIT Press.

Unger, P. 2004. The mental problems of the many. *Oxford Studies in Metaphysics* 1: 195–222.

Unger, P. 1990. *Identity, Consciousness, and Value*. Oxford: Oxford University Press.

Unger, P. 1996. *Living High and Letting Die: Our Illusion of Innocence*. Oxford: Oxford University PRess.

Unruh, C. F. 2023. The constraint against doing harm and long-term consequences. *Journal of Moral Philosophy 20* (3-4): 290–310.

Vallentyne, P. and S. Kagan. 1997. Infinite value and finitely additive value theory. *Journal of Philosophy 94* (1): 5–26.

Vallinder, A. and E. J. Olsson. 2013. Do computer simulations support the argument from disagreement? *Synthese 190* (8): 1437–1454.

Van Liedekerke, L. 1995. Should utilitarians be cautious about an infinite future? *Australasian Journal of Philosophy 73* (3): 405–407.

Vickers, P. 2023. *Identifying Future-Proof Science*. Oxford: Oxford University Press.

Voorhoeve, A. 2014. How should we aggregate competing claims? *Ethics 125* (1): 64–87.

Weatherson, B. 2003. Are you a sim? *Philosophical Quarterly 53* (212): 425–431.

Weatherson, B. 2019. *Normative Externalism*. Oxford: Oxford University Press.

Wedgwood, R. 2010. The moral evil demons. In *Disagreement*, eds. R. Feldman and T. A. Warfield, 216-246. Oxford: Oxford University Press.

Weimann, J., A. Knabe, and R. Sch¨ob. 2015. *Measuring Happiness: The Economics of Well-Being*. Cambridge, MA: MIT Press.

Wilkinson, H. 2021. Infinite aggregation: Expanded edition. *Philosophical Studies* 178: 1917–1949.

Wilkinson, H. 2022a. Can an evidential be risk-averse? *Global Priorities Institute Working Paper No. 21-2022*. URL: https://globalprioritiesinstitute.org/hayden-wilkinson-can-an-evidentialist-be-risk-averse/

Wilkinson, H. 2022b. Can risk aversion survive the long run? *Philosophical Quarterly 73* (2): 625–647.

Wilkinson, H. 2022c. In defense of fanaticism. *Ethics 132* (2): 445–477.

Wilkinson, H. 2022d. The unexpected value of the future. *Global Priorities Institute Working Paper No. 17-2022*. URL: https://globalprioritiesinstitute.org/the-unexpected-value-of-the-future-hayden-wilkinson-global-priorities-institute-university-of-oxford/

Wilkinson, H. 2023. How to neglect the long term. *Global Priorities Institute Working Paper No. 11-2023*. URL: https://globalprioritiesinstitute.org/how-to-neglect-the-long-term-hayden-wilkinson/

Wilkinson, H. forthcoming. Flummoxing Expectations. *Noûs*.

Williams, E. G. 2012. Promoting value as such. *Philosophy and Phenomenological Research 87* (2): 392–416.

Williamson, P. 2021. A new argument against critical-level utilitarianism. *Utilitas 33* (4): 399–416.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global Catastrophic Risks*, eds. N. Bostrom and M. Církovíc, 308–345. Oxford: Oxford University Press.

Zhang, H. and V. Conitzer. 2019. A PAC framework for aggregating agents' judgments. *Proceedings of the AAAI Conference on Artificial Intelligence 33* (1): 2237–2244.

Zuboff, A. 1981. The story of a brain. In *The Mind's I*, eds. D. R. Hofstadter and D. C. Dennett, 202–212. New York, NY: Basic Books.

Zuboff, A. 1990. One self: The logic of experience. *Inquiry 33* (1): 39–68.