Mind and Value:

A Research Agenda

This research agenda outlines some questions about mind and value that are of particular interest to philosophers at the Global Priorities Institute. Many of these questions concern the nature of mental phenomena, their distribution, our epistemic access to them, and their moral significance. We have selected questions based on their potential bearing on priority setting through their ability to inform decision-makers about the distribution of morally significant mental properties over individuals, with particular emphasis on non-human candidates for moral standing. The selection of topics is guided by conjectures about where additional research would be most valuable on the current margin. Some important questions are absent from the agenda because they are already well-studied or because we expect them to receive sustained scholarly attention by default. There are also cases in which we mention important but well-studied issues in order to provide context and orientation. Although the research

agenda outlines the kind of research we prioritize, we emphasize that it is not intended as exhaustive.

1. Which Mental Phenomena Are Morally Significant?

1.1 *Which Mental Phenomena Contribute to Moral Standing?*[1]

Many philosophers have maintained that moral standing is closely tied to *sentience*, that is, the capacity for valenced experience (e.g., Singer 1993; Korsgaard 2018; Nussbaum 2023). Pains and pleasures are paradigmatic valenced experiences, but the category includes much else besides, such as experiences of the sublime and feelings of nausea. Among those who agree that moral standing is closely related to sentience there is disagreement about whether sentience matters *per se* and, if so, how it matters (Roelofs 2023). Some hold that the capacity for consciousness rather than sentience confers moral standing (Chalmers 2022). Others deny that moral standing is (or need be) closely related to sentience, maintaining instead that its basis lies in mental phenomena, such as certain kinds of desires or agency, that are not tied to the capacity for consciousness (Levy 2014; Carruthers 2019: 171–4; Kagan 2019: 23–30; Bradford 2022; Goldstein and Kirk-Gianini 2023).

Who is right? Each of these views of moral standing can recognize the moral standing of humans. But they offer differing verdicts in cases involving non-human minds that are important for setting global priorities. For instance, if moral standing requires consciousness, then it is doubtful that any existing AI systems have moral standing and an open question whether any near-term AI systems will be conscious (Butlin et al. 2023; Chalmers 2023). On the other hand, if a thin variety of desire — one not dependent on the capacity for consciousness — suffices for moral standing, then some existing AI systems may well have moral standing (Goldstein and Kirk-Gianini 2023).

1.2 *How Are Mental Phenomena Otherwise Morally Significant?*

Plausibly, mental phenomena realize not only moral standing but also many other morally significant properties, including well-being, ill-being, and the possession of rights. However, there is much uncertainty about which mental phenomena realize which moral properties.

For example, the leading philosophical theories of well-being disagree about how different mental phenomena contribute to welfare (Lin 2022). Nonetheless, they may be thought to deliver very similar evaluations of many of the kinds of lives that actually exist, diverging sharply only for highly unusual, imaginary cases, like people devoting their lives to counting blades of grass (Rawls 1971: 432). How important, then, are the differences among these theories for setting priorities? Are there important divergences in the evaluation of possible outcomes by the different theories, of a kind that might significantly shift our priorities depending on the confidence we assign to each?

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 the nature and sources 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. For example, should the duration of a given valenced experience be understood as a matter of its objective duration or its subjectively experienced duration (Schukraft 2020; Shulman and Bostrom 2021)? If there are such things as degrees of consciousness (see Lee 2022 and citations therein), could pains of a fixed intensity differ in their degree of consciousness in a way that affects their disvalue? Does talk of the felt intensity of valenced experiences suffice to pick out a single dimension of (psychological) variation (see Armstrong 1968/2023: 341-3; Mayerfeld 1999: 61-7)? Does the allocation of attention toward or away from valenced experiences affect their value apart from

the modulation of felt intensity (compare Block 2010 and Chalmers 2010*a*: Ch. 11)? How should we weigh the number of subjects who undergo a given token experience and the degree of overlap between them, as opposed to the number of token experiences (Briggs and Nolan 2015; Javier-Castellanos 2021; see also Zuboff 1981; Unger 1990; Johnston 2016)?

If in fact there are conscious experiences that confer moral interests without doing so in virtue of their valence, what is their nature, and what factors modulate their importance? What role, if any, do complexity and richness of phenomenology play? What moral interests, if any, are tied to cognitive phenomenology?

Looking beyond the case of conscious states and processes altogether, which agentic capacities, if any, generate which sorts of moral interests and/or rights? 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 the source of moral significance to be a desire or belief whose satisfaction or truth requires that capacity? Or perhaps certain manifestations of the capacity? And 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)

More generally, answering questions of the kind outlined above will help us to determine the extent to which different morally relevant mental phenomena and/or the moral properties they ground may be expected to attain superhuman levels in certain kinds of minds, such as in digital minds that might one day be realized (Shulman and Bostrom 2021; compare Buchanan 2011: 209-42). They may also be expected to help us in addressing questions about what moral obligations we may have in respect of individuals with superhuman capacities for well-being or ill-being (compare Nozick 1974: 41: Parfit 1984: 389; Chappell 2021).

1.3 Methodological Issues in Welfare Measurement

We want to be able to assess and compare welfare levels in practice. Over the last half-century there has been an explosion of research in psychology and economics on subjective well-being (Layard 2005; Weimann, Knabe, and Schöb 2015). However, fundamental problems of measurement remain unresolved. These include the problems 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). 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?

If we want to be able to say whether and to what extent people's lives are good overall as opposed to bad overall, we need not only a cardinal welfare scale, but specifically a *ratio scale* with a privileged zero point. 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 reasons to believe that the midpoint of the Cantrill ladder is actually significantly higher than the boundary perceived by individuals as demarcating lives worth living from those not (MacAskill 2022: 196). There also remain difficulties associated with characterizing 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). Greater clarity on these issues is important, given the natural assumption that there are moral reasons to spare individuals from being born into a life not worth living (e.g., Narveson 1967; McMahan 1981; Parfit 1984: 391). How, if at all, can we do better?

Many of the questions raised above about the measurement of human well-being can be asked about the measurement of welfare for non-human animals. 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). There is a pressing need to develop better

measures of animal welfare, especially in relation to locating different kinds of animals relative to the zero level. 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, whereas we have reason to believe that people make biased assessments of which lives are worth living already in the human case (Gerhart et al. 1994). 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?

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. Even restricting ourselves to comparisons of the experiential component of individual welfare, there are enormous philosophical and scientific obstacles to making reliable comparisons of this kind (Browning 2023; Fischer 2024). Granting that honey bees have pleasant and unpleasant experiences, we may want to know whether the range of valenced states available to them is nonetheless only a fraction of the intensity range of affective experiences available to human beings. Does the total number or overall fraction of neurons within an animal's brain dedicated to the processing of valenced affect provide important evidence (Shriver 2022)? How much can we learn by thinking about the role of valenced experience in learning, decision-making, and the guidance of action, taking account of what we know about capacities for learning and action selection in different animals? How, if at all, do cognitive and emotional complexity relate to intensity range? Is the range even the right property to focus on? Lastly, how useful, if at all, are the approaches we may develop for making welfare comparisons across different animals in making comparisons of welfare between biological organisms and potential minds run on inorganic computational substrates?

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

The nature of morality is controversial, as is the nature of mind. It would be unsurprising if the moral significance of mental phenomena depends on the correct resolution of one or both of these controversies. For example, 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; Lee 2019; Birch 2022a; compare Pautz 2017). Are views of this kind correct, and, if so, what are the concrete evaluative implications of physicalism as regards consciousness and related states?

Another connection between the nature and moral significance of mentality concerns *illusionism*, the view 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 distinct from what it introspectively seems to be in one way or another. 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 Lee 2014). Developing and evaluating this natural thought requires examining the different varieties of illusionism, their comparative plausibility, and how different forms of illusion about consciousness affect the risk of error concerning the various ways in which experiences are thought to be morally significant.

Meta-ethical assumptions might also be expected to interact with our thinking about the moral significance of different mental phenomena. For example, realist meta-ethical views arguably make room for distinctive epistemological challenges to our moral beliefs about the significance of various psychological properties, including arguments from disagreement and appeals to various genealogical debunking arguments (compare Street 2006). The force or moral import of these challenges may depend on questions about mentality, such as the

epistemic profile of moral intuitions, the basis of mental content, and the reliability of introspection, and (Huemer 2008; Dogramaci 2021; Sinhababu 2022). Meta-ethical realism may be thought to put pressure on views that tie the moral significance of mental properties too closely to characteristically human traits (Jaquet 2022; de Lazari-Radek and Singer 2012.), whereas constructivist views might seem to defuse the appearance of suspicious coincidence that otherwise arises for 'chauvinist' views of that kind (compare Harman 1983: 124-5). Is that in fact the case? How, if at all, does the correct application of norms of theory choice to moral views depend on which meta-ethical theory is correct? For example, do Occamist norms apply only to theories' fundamental commitments (Bennett 2017; Schaffer 2015) and hence only to moral principles on views that construe them as fundamental? In what other ways, if any, do meta-ethical assumptions bear on questions about mind and value?

2. The Distribution of Morally Significant Mental Phenomena

How are morally significant mental phenomena distributed (as a function of non-mental factors)? We 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.

For example, we may be particularly interested in valenced experience. At present, 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, like evaluativism (Bain 2012; Carruthers 2018) and imperativism (Klein 2007; Barlassina and Hayward 2019) can be developed and operationalized to yield empirical criteria for valenced experience that are comparable in scope and specificity to, say, the criteria for consciousness proposed by the global workspace theory (Baars 1998; Dehaene 2014) or the integrated information theory of consciousness (Tononi 2008; Tononi et al. 2016; Albantakis et al. 2022). It is reasonable to expect a well-developed theory of valenced experience to shed light on morally-significant questions such as: 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)?

2.1 Liberal and Stringent Criteria for Consciousness

How liberal or stringent are the conditions on the realization of consciousness? For example, does consciousness have a wide range of biological realizers, and is it widely distributed throughout the tree of life, encompassing not only mammals and other vertebrates, but also invertebrates of different phyla (Klein and Barron 2016; Birch et al. 2021; Gibbons et al. 2022)?

Furthermore, what exactly is the significance of biology to consciousness? Is consciousness *substrate independent* or can it only be realized in a narrow range of material substrates, exemplified by neural biochemistry (Searle 1992; Block 2009, 2022)? Even granting that functional isomorphs would share the same phenomenal qualities regardless of material substrate (Chalmers 1996a: 247–275), does consciousness in fact have a wide range of realizers within functional state space? Or can it only be realized 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; Cao 2022).

If consciousness has a computational basis, how do constraints on computational implementation - such as those that may be required to avoid the threat of extreme pancomputationalism (Putnam 1987; Searle 1990; Chalmers 1996b) - constrain the range of its possible physical realizers (Ritchie and Piccinini 2018; Shiller 2023)? For example, might it be the case that the constraints on computational implementation needed to avoid certain paradoxical results support the conclusion that a detailed computer simulation of the human brain fails to realize conscious experience (Klein 2018)?

A further question is whether internal physical duplicates can 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)? 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 forthcoming)? Or are they met by only a narrow class of entities that have, say, the requisite evolutionary history?

The distinctions highlighted above—between liberal vs. restrictive realization, substrate independence vs. substrate dependence, and internalism vs. externalism about the physical basis of experience—are especially significant from a priority setting perspective because of their relatively direct bearing on the expected distribution of experience. For instance, evidence in favor of liberal realization or substrate independence would tend to support the hypothesis that some digital systems can be conscious. On the other hand, evidence for certain forms of externalism might provide reason to regard as unconscious fine-grained simulations of conscious minds. Our focus on these distinctions departs from traditional emphasis on functionalism (Levin 2023), computationalism (Rescorla 2020), and multiple realizability (Bickle 2020). Although none of these theses straightforwardly entails liberal realization, substrate independence, internalism, or their opposites, their bearing on the expected distribution of experience proceeds largely via their bearing on the foregoing distinctions. Thus, from a priority setting perspective, there is reason to focus on the 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 and reassess their traditionally assumed relationship with liberal realization and substrate independence.

2.2. Theories of Consciousness

In practice, 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. For this reason, we are most excited about meta-work on theories of consciousness, as opposed to object-level engagement with any particular theory.

For example, 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 especially prioritized? What are the in-principle limits on resolving uncertainties about consciousness in the science of consciousness, and to what extent should we expect to continue to be burdened with significant degrees of cluelessness about the distribution of experience? 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?

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. Which other theory regimentations would facilitate important theory comparisons? A closely related issue concerns how existing scientific theories of consciousness that were designed with humans in mind can be 'de-anthropomorphized' so as to be applicable to non-human minds (compare Cappelen and Dever 2021). For example, take the global workspace theory on which 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).

2.3 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. Given this aim, 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 in inorganic computational systems, or do we need a different methodological approach altogether when it comes to candidate digital minds (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. On reductive views, there is then a metasemantic puzzle concerning how our concepts of consciousness could have determinate reference, especially when these properties disassociate (Papineau 2002 175-231; Taylor 2013; Pautz 2017; Balog 2020; Birch 2022a). On both reductive and non-reductive views, there is an epistemological puzzle as to 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 tackle these problems, and what bearing does their resolution have on the question of how to value the distinct properties that may be tightly correlated and associated with consciousness in human subjects?

A further methodological issue concerns the extent to which research on the moral significance and on the nature and distribution of different mental phenomena can proceed in parallel. For example, 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 (Lee 2013). Compare: views that attach outstanding moral significance to personal identity might be underdetermined by the discovery that personal identity is non-branching psychological continuity and that that relation typically holds in the presence of other similar relations (Parfit 1984).

Our current epistemic predicament may suggest that additional crucial considerations are discoverable and that it is especially important to identify and articulate the import of potentially crucial but neglected issues that bear on the distribution of consciousness. Possible candidates include: the *meta-problem of consciousness* (Chalmers 2018); debunking arguments appealing to proximal or distal explanations of our judgments about consciousness (Chalmers, 2018, 2020); in-depth analysis of the strengths and weaknesses of the 'fading qualia' and 'dancing qualia' arguments for organizational 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; Roelofs 2022; Builes and Hare 2023); the bearing of self-locating evidence and observation selection effects (Bostrom 2002; Titelbaum 2008; Shulman and Bostrom 2012; Hanson et al. 2021; Snyder-Beattie; Isaacs et al. 2022; Manley n.d.) on our beliefs about the overall distribution of consciousness (see Zuboff 1990; Bostrom 2003; Crawford 2013; Dorr and Arntzenius, 2017; Chalmers 2022: Ch. 5; Li and Saad 2022, 2023, forthcoming; Builes and Hare 2023); the possibility of *hidden qualia* (Shiller 2017*b*; compare Block 2007; Goff 2013; Muelhauser 2017 Appendix H; Schwitzgebel 2015; Lee 2019; Bayne et al.

2020); accounting for *harmonious phenomenal-normative correlations* (James 1890; Pautz 2015, 2020*b*; Mørch 2017; Goff 2018; Saad 2019, 2022; Cutter and Crummett forthcoming); and accounting for *laws of appearance* (Raymont 2005; Cutter 2016, Pautz 2020*b*; Sainsbury 2022, Speaks 2022; Block 2023: 198-200; Morgan, 2023).

Lastly, how should we estimate the value of information about the basis of consciousness and the relative importance of different kinds of errors? 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 theorizing 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)?

3. Preparing to Live Alongside Digital Minds

Some forecasts assign substantial probability to AI systems meeting or exceeding 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. As before, the listed issues are by no means exhaustive.

3.1 Catastrophic Risks and Their Mitigation

There has been considerable discussion of the idea that advanced AI poses an existential risk via the possibility of digital agents with superhuman cognitive capacities that turn out to be misaligned with human values (Yudkowsky 2008; Bostrom 2014; Russell 2019; Cotra 2022; Ngo

et al. 2022; Ord 2022; Carlsmith 2024; 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, 2023).

A neglected issue concerns possible moral constraints on solutions to the alignment problem, (Christiano 2018; Peterson 2019; Gabriel 2020; compare: Chalmers 2010*b*: 30), and in particular constraints arising from the potential moral interests and rights of advanced AI systems. Suppose advanced AI systems meet or exceed cognitive criteria for high moral status of the kind typically associated with human persons. Would it be permissible to design digital minds like that so that they intrinsically value serving humanity and prioritize human welfare over their own, with no freedom to explore other values (Petersen 2007)? If it would not (Schwitzgebel and Garza 2020), what are the implications for designing morally permissible solutions to the alignment problem? To what extent are existing alignment proposals in tension with the ethical treatment of digital minds? Should any such tensions be leveraged to decelerate or regulate AI development?

More generally, the emergence of large populations of digital minds would be associated with non-negligible risks of catastrophic digital suffering and large-scale AI rights violations (Bostrom 2014: Ch. 8; Sotala and Gloor 2017; Tomasik 2017; Saad and Bradley 2022; Saad 2023; Schwitzgebel 2023). There is 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 recognized. Threat models and mitigation strategies remain underdeveloped. 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. 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).

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.

For example, 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? Would it be better to focus for now on resolving crucial uncertainties about consciousness and moral status in digital systems, before prioritizing legal and policy interventions? From the perspective of ensuring appropriate recognition of the moral status of AI systems, is it desirable for certain kinds of AI systems to arrive before others? Are there risks that certain false beliefs about morally significant aspects of mind could become locked in (compare MacAskill 2022: 75–102)?

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*a*; Shulman and Bostrom 2021; Friederich 2023).

For example, if futures involving large populations of digital systems imbued with consciousness are considered desirable, what strategies can mitigate against our uncertainty

about the physical basis of consciousness? Can candidate sufficient conditions for consciousness co-occur in digital systems, and, if so, 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 population-portfolio of this sort would be optimal?

Bibliography

Albantakis, L., L. Barbosa, G. Findlay, M. Grasso, A.M. Haun, W. Marshall, A. Zaeemzadeh, M, Boly, B.E. Juel, S. Sasai, K. Fujii, I. David, J. Hendren, J.P., and G. Tononi. 2022. Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. arXiv preprint arXiv:2212.14787. URL: https://arxiv.org/pdf/2212.14787

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/

Alexander, S. 2023. Davidson on takeoff speeds. *Astral Codex Ten*. URL: https://www.astralcodexten.com/p/davidson-on-takeoff-speeds?utm_source=%2Fsearch%2Fdavidson&utm_medium=reader2

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

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.

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

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

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press

Bain, D. 2013. What makes pains unpleasant? *Philosophical Studies* 166(1): 69–89.

Bain, D. 2019. Why take painkillers? *Noûs* 53(2): 462–490.

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.

Barlassina, L. and M. K. Hayward. 2019. More of me! Less of me! Reflexive imperativism about affective phenomenal character. *Mind* 128(512): 1013–1044.

Bayne, T., Seth, A. K., & Massimini, M. (2020). Are there islands of awareness?. *Trends in Neurosciences*, 43(1), 6-16.

Bennett, K. 2017. *Making things up*. Oxford: Oxford University Press.

Bickle, J. 2020. Multiple realizability. In *The Stanford Encyclopedia of Philosophy* (Summer 2020 edn.), 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., 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.

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(5-6): 481-499.

Block, N. 2009. Comparing the major theories of consciousness. In *The Cognitive Neurosciences*, 4th edn., ed. M. Gazzaniga, 1111–1123. Cambridge, MA: MIT Press.

Block, N. 2023. *The Border Between Seeing and Thinking*. 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? *The Philosophical Quarterly*, 53(211): 243–255.

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.

Bourget, D. and A. Mendelovici, A. 2014. Tracking Representationalism. In *Philosophy of Mind: The Key Thinkers,* ed. A. Bailey, 209–35. London: Continuum.

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

Bradford, G. 2022. Consciousness and welfare subjectivity. *Noûs*. Early View.

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

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

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.

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.* Early View.

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. URL: https://arxiv.org/abs/2308.08708.

Cao, R. 2022. Multiple realizability and the spirit of functionalism. *Synthese* 200(6): 1–31.

Carlsmith, J. 2024. Existential risk from power-seeking AI. In *Essays on Longtermism*, eds. D. Thorstad, J. Barrett, and H. Greaves. 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.

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. 2010*a*. *The Character of Consciousness*. Oxford: Oxford University Press.

Chalmers, D.J., 2010*b*. 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 Studies* 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*. New York, NY: W.W. Norton.

Chalmers, D. J. 2023. Could a large language model be conscious? URL: https://philarchive.org/rec/CHACAL-3

Chappell, R. Y. 2021. Negative utility monsters. *Utilitas* 33(4): 417–421.

Christian, B. 2020. *The Alignment Problem: How Can Machines Learn Human Values?* London: Atlantic Books.

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/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP

Cotra, A. 2022. 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.

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

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

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

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

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

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

Dalbey, B. and Saad, B. 2022. Internal constraints for phenomenal externalists: A structure matching theory. *Synthese* 200, 348.

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/.

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

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

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

Cappelen, H. and J. Dever. 2021. *Making AI intelligible: Philosophical Foundations*. Oxford: Oxford University Press.

Diener, E., Diener, C., Choi, H., and Oishi, S. 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*, 8.

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

Dretske, F. 1995 *Naturalizing the Mind*. MIT Press.

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

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

Fischer, B. (Ed.) 2024. *Weighing Animal Welfare: Comparing Well-being Across Species*. Oxford: Oxford University Press.

Fischer, B., A. Shriver, and M. S. Jules. 2022. Do brains contain many conscious subsystems? If so, should we act differently? *Rethink Priorities*. URL: https://rethinkpriorities.org/publications/do-brains-contain-many-conscious-subsystems

Frankish, K. 2016. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11-12), 11-39.

Friederich, S. 2023. Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*.

Gerhart, K.A., J. Koziol-McLain, S. R. Lowenstein, and G. G. Whiteneck. 1994. Quality of life following spinal cord injury: knowledge and attitudes of emergency care providers. *Annals of emergency medicine*, 23(4): 807-812.

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* 6: 155-229.

Goff, P. 2013. Orthodox property dualism + the linguistic theory of vagueness = Panpsychism. *Consciousness Inside and Out: Phenomenology, Neuroscience, and The Nature of Experience,* ed. R. Brown, 75-91. Dordrecht: Springer.

Goff, P. 2018. Conscious thought and the cognitive fine-tuning problem. *The Philosophical Quarterly*, 68(270): 98-122.

Godfrey-Smith, P. 2016. Mind, matter, and metabolism. *Journal of Philosophy* 113(10): 481–506.

Goldstein, S. and C. D. Kirk-Giannini. 2023. AI wellbeing. URL https://philpapers.org/rec/GOLAWE-4.

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 Astrophysical Journal* 922(2): 182.

Harman, G. 1983. Justice and moral bargaining. *Social Philosophy and Policy* 1(1): 114.

Hendrycks, D., M. Mazeika, and T. Woodside. 2023. An Overview of Catastrophic AI Risks. arXiv preprint. URL: https://arxiv.org/abs/2306.12001

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: Revista Iberoamericana de Estudios Utilitaristas* 17 (1): 73–90.

Huemer, M. 2008. Revisionary intuitionism. *Social Philosophy and Policy* 25(1): 368–392.

Isaacs, Y., J. Hawthorne, and J. Sanford Russell. 2022. Multiple universes and self-locating evidence. *Philosophical Review*, 131(3): 241–294.

James, W. 1890. *The principles of psychology* (Vol. 2). 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.

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.

Kammerer, F. 2019. The normative challenge for illusionist views of consciousness. *Ergo* 6(32).

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.

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).

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

Krakovna, V. 2023. 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.

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

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. 2014. Materialism and the epistemic significance of consciousness. In *Current controversies in philosophy of mind*, ed. U. Kriegel, 222-245. New York, NY: 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. MIT Press.

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

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

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

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

Lin, E., 2022. Well-being, part 2: Theories of well-being. *Philosophy Compass*, 17(2), p.e12813.

Lycan, W. 2001. The case for phenomenal externalism. *Philosophical perspectives* 15: 17-35.

MacAskill, W. 2022. *What we owe the future*. New York, NY: Basic Books.

Manley, D. n.d. On being a random sample. URL: philpapers.org/rec/MANOBA

Mayerfeld, J., 1999. *Suffering and moral responsibility*. Oxford University Press, USA.

McMahan, J. 1981. Problems of population theory. *Ethics* 92(1): 96–127

McQueen, K. J. 2019. Interpretation-neutral integrated information theory. *Journal of Consciousness Studies*, 26(1–2): 76–106.

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.

Metzinger, T. 2021. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1): 43-66.

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

Mørch, H.H. 2017. The evolutionary argument for phenomenal powers. *Philosophical Perspectives* 31: 293-316.

Muehlhauser, L. 2017 Report on Consciousness and Moral Patienthood URL: https://www.openphilanthropy.org/research/2017-report-on-consciousness-and-moral-patienthood/

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

Ng, Y.K. 1995 Toward welfare biology: evolutionary economics of animal consciousness and suffering. *Philosophy and Biology* 10: 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, S., 2022. The alignment problem from a deep learning perspective. arXiv preprint. URL: https://arxiv.org/abs/2209.00626

Nozick, R., 1974. *Anarchy, State, and Utopia*. New York, NY: Basic Books.

Nussbaum, M.C. 2023. *Justice For Animals: Our Collective Responsibility*. New York, NY: Simon and Schuster.

Ord, T. 2020. *The Precipice: Existential Risk and The Future of Humanity*. London: Bloomsbury.

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

Pallies, D. 2022. Attraction, aversion, and asymmetrical desires. *Ethics*, 132(3): 598-620.

Papineau, D. 2002. *Thinking About Consciousness*. Oxford: Oxford University Press.

Pautz, A. 2013. The real trouble with phenomenal externalism: 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, 237-98. Dordrecht: Springer.

Pautz, A. 2014. The real trouble with armchair arguments against externalism. In Sprevak et al. (eds.), *New Waves in Philosophy of Mind,* ed. M. Sprevak and J. Kallestrup, 153-181. Dordrecth: Springer

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: 349-407.

Pautz, A. 2020*a*. Consciousness and coincidence: Comments on Chalmers. *Journal of Consciousness Studies* 27(5-6): 143-155.

Pautz, A. 2020*b*. The puzzle of the laws of appearance. *Philosophical Issues* 30: 257-272.

Pearce, D. 1995. *The Hedonistic Imperative*. URL: http://www.happymutations.com/ebooks/david-pearce-the-hedonistic-imperative.pdf

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. 2007. The ethics of robot servitude. *Journal of Experimental and Theoretical Artificial Intelligence*, 19(1): 43-54.

Peterson, M. 2019. The value alignment problem: a geometric approach. *Ethics and Information Technology* 21: 19-28.

Putnam, H. 1988. *Representation and Reality*. Cambridge, MA: MIT press.

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

Rescorla, M. 2020. The computational theory of mind. In *The Stanford Encyclopedia of Philosophy* (Summer 2020 edn.), ed. E. N. Zalta. Metaphysics Research Lab, Stanford University.

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. London: Routledge.

Roelofs, L. 2019. *Combining Minds: How to Think About Composite Subjectivity*. Oxford: Oxford University Press.

Roelofs, L. 2022. No such thing as too many minds. *Australasian Journal of Philosophy*

Roelofs, L., 2023. Sentientism, motivation, and philosophical Vulcans. *Pacific Philosophical Quarterly* 104(2): 301-323

Russell, S. 2019. *Human compatible: Artificial intelligence 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 Report*. URL: https://www.sentienceinstitute.org/downloads/Simulations_and_Catastrophic_Risks.pdf

Saad, B. forthcoming. Lessons from the void: what Boltzmann Brains teach. *Analytic Philosophy*.

Saad, B., and A. Bradley. 2022. Digital suffering: why it's a problem and how to prevent it. *Inquiry*.

Sainsbury, M. 2022. Visual experience and the laws of appearance. *Erkenntnis* 88(7): 2933-2940.

Schaffer, J. 2015. What not to multiply without necessity. *Australasian Journal of Philosophy* 93(4): 644-664.

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. 2015. If materialism is true, the United States is probably conscious. *Philosophical Studies* 172: 1697-1721.

Schwitzgebel E. 2023. The coming robot rights catastrophe. *APA Blog*. URL: https://blog.apaonline.org/2023/01/12/the-coming-robot-rights-catastrophe

Schwitzgebel, E., and M. Garza. 2020. Designing AI with rights, consciousness, self-respect, and freedom. In *Ethics of Artificial Intelligence,* ed. S. M. Liao, 459-79. 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. and R. Long. 2023. Moral consideration for AI systems by 2030. URL: https://jeffsebodotnet.files.wordpress.com/2023/06/moral-consideration-for-ai-systems-by-2030-5.pdf

Shiller, D. 2017. In Defense of Artificial Replacement. *Bioethics* 31(5): 393-399.

Shiller, D. 2017*b*. Hidden qualia. *Review of Philosophy and Psychology* 8 (1): 165-180.

Shiller, D. 2023. Implementation considerations for digital consciousness. URL: https://philpapers.org/archive/SHIICF.pdf

Shriver, A. 2022. Why neuron counts shouldn't be used as proxies for moral weight. *Effective Altruism Forum*. URL: https://forum.effectivealtruism.org/posts/Mfq7KxQRvkeLnJvoB/why-neuron-counts-shouldn-t-be-used-as-proxies-for-moral

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

Simon, J. 2017. The hard problem of the many. *Philosophical Perspectives* 31(1): 449-468.

Sinhababu, N. 2022. Naturalistic arguments for ethical hedonism. URL: https://utilitarianism.net/guest-essays/naturalistic-arguments-for-ethical-hedonism/

Singer, P. 1993. *Practical Ethics,* 2nd edn. Cambridge: Cambridge University Press.

Sotala, K., and L. Gloor, L. 2017. Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica* 41(4).

Speaks, J. 2022. Pautz on the laws of appearance, internalism, and color realism. URL: https://www3.nd.edu/~jspeaks/papers/pautz.pdf

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.

Taylor, J. 2013. Is Consciousness Science Fundamentally Flawed?. *Journal of Consciousness Studies* 20(3-4): 203-221.

Titelbaum, M. G. 2008. The relevance of self-locating beliefs. *The Philosophical Review* 117(4): 555-605.

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*. Foundational Research Institute: Basel, Switzerland.

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(7): 450-461.

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

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

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

Weimann, J., A. Knabe, and R. Schob. 2015. *Measuring happiness: The economics of well-being.* Cambridge, MA: MIT Press.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks*, eds. N. Bostrom and M. Cirkovic, 308-45. Oxford: Oxford University Press.

Zuboff, Arnold 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: 39-68.