Sacrificing Humans for Insects and AI: A Critical Review
of Jonathan Birch, *The Edge of Sentience,*
Jeff Sebo, *The Moral Circle,*
and Webb Keane, *Animals, Robots, Gods*

Eric Schwitzgebel
Department of Philosophy
University of California, Riverside

Walter Sinnott-Armstrong
Department of Philosophy and Kenan Institute for Ethics
Duke University

September 1, 2025

Sacrificing Humans for Insects and AI: A Critical Review
of Jonathan Birch, *The Edge of Sentience,*
Jeff Sebo, *The Moral Circle,*
and Webb Keane, *Animals, Robots, Gods*

Abstract: Scientists increasingly take seriously the possibility that insects are sentient and that AI systems might soon be sentient. If sentience or consciousness is central to moral standing, this raises the possibility that insects, in the aggregate, or near-future AI systems (either as individuals or in the aggregate) might have sufficient moral importance that their interests outweigh human interests. The result could be a reorientation of ethics that radically deprioritizes humanity. This critical review examines three recent books on these issues: Jonathan Birch's *The Edge of Sentience,* Jeff Sebo's *The Moral Circle*, and Webb Keane's *Animals, Robots, Gods.* All three books present arguments and principles that, if interpreted at face value, appear radical. However, all three books downplay those radical implications, suggesting relatively conservative near-term solutions.

Keywords: consciousness, animals, AI, precaution, harm

Word Count: ~11,000 words

Sacrificing Humans for Insects and AI: A Critical Review
of Jonathan Birch, *The Edge of Sentience,*
Jeff Sebo, *The Moral Circle,*
and Webb Keane, *Animals, Robots, Gods*

1. The Possibly Radical Ethical Implications of Animal and AI Consciousness

We don't know a lot about consciousness. We don't know what it is, what it does, which kinds it divides into, whether it comes in degrees, how it is related to non-conscious physical and biological processes, which entities have it, or how to test for it. The methodologies are dubious, the theories intimidatingly various, and the metaphysical presuppositions contentious.[1]

We also don't know the ethical implications of consciousness. Many philosophers hold that (some kind of) consciousness is sufficient for an entity to have moral rights and status.[2] Others hold that consciousness is necessary for moral status or rights.[3] Still others deny that consciousness is either necessary or sufficient.[4] These debates are far from settled.

[1] For skeptical treatments of the science of consciousness, see Eric Schwitzgebel, *The Weirdness of the World* (Princeton, NJ: Princeton University Press, 2024); Hakwan Lau, "The End of Consciousness", *OSF preprints* (2025): https://osf.io/preprints/psyarxiv/gnyra_v1. For a recent overview of the diverse range of theories of consciousness, see Anil K. Seth and Tim Bayne, "Theories of Consciousness", *Nature Reviews Neuroscience* 23 (2022): 439-452. For doubts about our knowledge even of seemingly "obvious" facts about human consciousness, see Eric Schwitzgebel, *Perplexities of Consciousness* (Cambridge, MA: MIT Press, 2011).

[2] E.g. Elizabeth Harman, "The Ever Conscious View and the Contingency of Moral Status" in *Rethinking Moral Status*, edited by Steve Clarke, Hazem Zohny, and Julian Savulescu (Oxford: Oxford University Press, 2021), 90-107; David J. Chalmers, *Reality+* (New York: Norton, 2022).

[3] E.g. Peter Singer, *Animal Liberation, Updated Edition* (New York: HarperCollins, 1975/2009); David DeGrazia, "An Interest-Based Model of Moral Status", in *Rethinking Moral Status*, 40-56.

[4] E.g. Walter Sinnott-Armstrong and Vincent Conitzer, "How Much Moral Status Could AI Ever Achieve?" in *Rethinking Moral Status*, 269-289; David Papineau, "Consciousness Is Not the Key to Moral Standing" in *The Importance of Being Conscious,* edited by Geoffrey Lee and Adam Pautz (forthcoming).

These ignorances intertwine. For example, if panpsychism is true (that is, if literally everything is conscious), then consciousness is not sufficient for moral status, assuming that some things lack moral status.[5] On the other hand, if illusionism or eliminativism is true (that is, if literally nothing is conscious in the relevant sense), then consciousness cannot be necessary for moral status, assuming that some things have moral status.[6] If plants, bacteria, or insects are conscious, mainstream early 21st century Anglophone intuitions about the moral importance of consciousness are likelier to be challenged than if consciousness is limited to vertebrates.

Perhaps alarmingly, we can combine familiar ethical and scientific theses about consciousness to generate conclusions that radically overturn standard cultural practices and humanity's comfortable sense of its own importance. For instance:

(E1.) The moral concern we owe to an entity is proportional to its capacity to experience "valenced" conscious states such as pain and pleasure.

(S1.) Insects (at least many of them) have the capacity to experience at least one millionth as much valenced consciousness as the average human.

E1, or something like it, is commonly accepted by classical utilitarians as well as others. S1, or something like it, is not unreasonable as a scientific view. Since there are approximately 10^{19} insects, their aggregated overall interests would vastly outweigh the overall interests of

[5] Luke Roelofs and Nicolas Kuske, "If Panpsychism Is True, Then What? Part I: Ethical Implications", *Giornale di Metafisica* 1 (2024): 107-126.

[6] Alex Rosenberg, *The Atheist's Guide to Reality: Enjoying Life Without Illusions* (New York: Norton, 2012); François Kammerer, "Ethics Without Sentience: Facing Up to the Probable Insignificance of Phenomenal Consciousness", *Journal of Consciousness Studies* 29 (3-4):180-204.

humanity. Ensuring the well-being of vast numbers of insects might then be our highest ethical priority.[7] On the other hand:

(E2.) Entities with human-level or superior capacities for conscious practical deliberation deserve at least equal rights with humans.

(S2.) Near future AI systems will have human-level or superior capacities for conscious practical deliberation.

E2, or something like it, is commonly accepted by deontologists, contract theorists, and others. S2, or something like it, is not unreasonable as a scientific prediction. This conjunction, too, appears to have radical implications – especially if such future AI systems are numerous and possess interests at odds with ours.

This review addresses three recent interdisciplinary efforts to navigate these issues. Jonathan Birch's *The Edge of Sentience* emphasizes the science, Jeff Sebo's *The Moral Circle* emphasizes the philosophy, and Webb Keane's *Animals, Robots, Gods* emphasizes cultural practices. All three argue that many nonhuman animals and artificial entities will or might deserve much greater moral consideration than they typically receive, and that public policy, applied ethical reasoning, and everyday activities might need to significantly change. Each author presents arguments that, if taken at face value, suggest the advisability of *radical* change, leading the reader right to the edge of that conclusion. But none ventures over that edge. All three pull back in favor of more modest conclusions, at least for the near term.

Their concessions to conservatism might be unwarranted. Their own arguments (in different ways, to different degrees) seem to suggest that a more radical deprioritization of humanity might be ethically correct. Perhaps what we should learn from reading these books is

[7] Compare Sebo's "rebugnant conclusion", which we'll discuss in Section 3.1.

that we need a new Copernican revolution – a radical reorientation of ethics around nonhuman rather than human interests.[8] On the other hand, readers who are more steadfast in their commitment to humanity might view radical deprioritization as sufficiently absurd to justify modus tollens against any principles that seem to require it. In this critical essay, we focus on the conditional. *If* certain ethical principles are correct, *then* humanity deserves radical deprioritization, given recent developments in science and engineering. We take no stand on ponens vs. tollens.

2. Birch's Principles

Birch states his core principles explicitly:

> **Framework Principle 1.** *A duty to avoid gratuitous suffering.* We ought, at minimum, to avoid causing gratuitous suffering to sentient beings either intentionally or through recklessness/negligence. Suffering is not gratuitous if it occurs in the course of a defensible activity despite proportionate attempts to prevent it. Suffering is gratuitous if the activity is indefensible or the precautions taken fall short of what is proportionate. (p. 131)[9]

[8] Deep ecologists, such as Arne Naess, have also long sought to decenter human interests. However, potentially more extreme deprioritizations, such as exterminating humans in favor of microbes or ecstatic supercomputers, go beyond what even deep ecologists ordinarily suggest.

[9] Framework Principle 1 might seem to imply that making *some* proportionate attempts to prevent the suffering keeps it from being gratuitous, even if more effective proportionate attempts are not made. However, we will interpret Framework Principle 1 as claiming that suffering is gratuitous unless one makes *all* or *enough* proportionate attempts to prevent it. Notice that it might be too costly to make all attempts, even if each attempt is proportionate.

This first principle doesn't tell us what to do when we don't know what can suffer or what is conscious or sentient in Birch's sense of being capable of having valenced phenomenally conscious experiences (p. 26). To apply Framework Principle 1 to disputed cases, we need more:

> **Framework Principle 2.** *Sentience candidature can warrant precautions.* If *S* is a sentience candidate, then it is reckless/negligent to make decisions that create risks of suffering for *S* without considering the question of what precautions are proportionate to those risks. Reasonable disagreement about proportionality is to be expected, but we ought to reach a policy decision rather than leaving the matter unresolved indefinitely. (p. 133)[10]

Birch defines his crucial technical term "sentience candidate" like this:

> A system *S* is a *sentience candidate* if there is an evidence base that: (a) implies a realistic possibility of sentience in *S* that it would be irresponsible to ignore when making policy decisions that will affect *S*, and (b) is rich enough to allow the identification of welfare risks and the design and assessment of precautions. (p. 124)[11]

There will, of course, be plenty of room for dispute concerning what is "evidence", "realistic", "irresponsible", and "rich enough".

---

[10] Framework Principle 2 explicitly requires only "considering the question". However, it is easy to consider a question and then reject the obvious answer on shoddy grounds. So we will assume that Birch means that we ought not only consider the question but also do so responsibly and act on the basis of what we responsibly conclude.

[11] It is unclear why condition (b) should be necessary for sentience candidature if condition (a) is fulfilled, if "sentience candidature" is interpreted as purely an epistemic criterion tantamount to some reasonable chance of sentience. So we interpret "sentience candidature" as partly a practical criterion, possessed by entities only contingently upon our having some implementable ideas about how the entities ought to be treated if they are sentient. An omnipotent God would not be a sentience candidate in the relevant sense, since we have little idea what if anything would constitute a welfare risk to God.

2.1. Varieties of Precaution

Birch's principles and definitions together are intended to support a precautionary principle: "At the core of the framework is the thought that we need to find ways to *err on the side of caution* in these cases" (p. 17; cf. Sebo, p. 55). What it means to "err on the side of caution" depends on what one most wants to be cautious about. In scientific contexts, if one is primarily concerned to avoid falsehoods, one can take the default to be suspending belief, cautiously declining to accept any scientific claim that is not supported by sufficiently strong evidence. In contrast, if one is more concerned not to miss any scientific truth, then caution recommends accepting scientific claims even when they are supported by only weak evidence. One precautionary policy is "don't publish without strong evidence"; a contrasting precautionary policy is "spread the news if there's any positive evidence at all". Weak-news science can be reasonable on precautionary grounds in some cases – for example, concerning possible toxins or experimental treatments for otherwise incurable cancer. In such cases, tenuous but suggestive evidence might be important to share and act upon.[12] By itself, an appeal to precaution does not specify the type of error most to be avoided.

[12] On why avoiding Type II error (false negatives) is more important than avoiding Type I error (false positives) in some contexts, such as policies concerning potential toxins, contrary to the usual scientific emphasis on avoiding Type I error as manifested in requiring a high threshold of statistical significance before rejecting the null hypothesis, see John Lemons, Kristin Shrader-Frechette, and Carl Cranor, "The Precautionary Principle: Scientific Uncertainty and Type I and Type II Errors", *Foundations of Science* 2 (1997): 207-236; Frederick Schauer, *The Proof: Uses of Evidence in Law, Politics, and Everything Else* (Cambridge, MA: Harvard University Press, 2022), Chapter 3. Weak-news science is also potentially important in sharing replication failures, which might be individually indecisive but cumulatively powerful. Schauer adds that, in personal life, we can be justified in acting on weak evidence that a candidate for babysitter is a child molester.

The same goes for harming sentience candidates. Suppose that current policies allow certain forms of fishing. If we most want to avoid causing pain to the fish, and we have some suggestive evidence that the fish feel pain when caught, then Birch's precautionary principle implies that we ought to forbid these forms of fishing as long as the restrictions are "proportionate". In contrast, if we most want to avoid causing harm to the humans who catch, buy, and eat fish, to avoid restricting their liberty and rights, then a precautionary principle implies that we should allow the fishing pending sufficiently strong evidence of fish pain. Which policy to favor depends on which risks one is most concerned to avoid.

Birch argues for his cautions in several ways. Sometimes he appeals to intuitions about particular cases. For example, his book opens with the horrifying story of Kate Bainbridge (p. 7), who was agonizingly intubated without sedation for months because her doctors wrongly assumed she was not conscious. This story is presumably intended to support defaulting toward considering an entity or organism as conscious if there is some evidence that it might be conscious.[13] The Bainbridge case suggests a general argument to cautiously err on the side of over-attributing rather than under-attributing consciousness to animals and AI:

> The risks of over-attributing and under-attributing sentience are not equal. When we deny the sentience of sentient beings, acting as if they felt nothing, we tend to do them terrible harms…. Meanwhile, when we treat non-sentient beings as if they were sentient, we may still do some harm (if the precautions we take are very costly and time-consuming and distract our attention away from other cases), but

[13] If there is no default but simply an evidence-based credence, say 15%, that figures in a standard expected value calculation, then the reasoning is not actually a precautionary attempt to "err on the side of caution". See Sebo, p. 55.

> the harms are often much less serious and of a different, more controllable kind. (p. 17; cf. Sebo, p. 52)

Is this argument convincing?

If Birch's claim is really only that false negatives are "often" more costly than false positives, then it is compatible with the claim that false positives are also often more costly than false negatives. Presumably, Birch instead means that false negatives *tend to be* more costly than false positives. But then how do we assess the tendency? This requires assessing the frequency and severity of false positives vs. false negatives – and that turns on the very science at issue. Even if we settle, for example, on the assumption that all insects are sentient, it does not straightforwardly follow that harming a million or a billion insects is worse than harming one human. The runaway trolley will destroy either a ten-million-insect ant colony or kill your neighbor. Will the ants or your neighbor suffer more? Is it more cautious to do the normal-seeming thing and save your neighbor, pending evidence that the degree and nature of insect sentience sufficiently warrants sacrificing a human on their behalf? Or is it more cautious to sacrifice the human so that ten million ants might enjoy their wars and fungus farming?

Another limitation of Birch's framework is that it emphasizes harm at the expense of other values, especially liberty. John Stuart Mill is famous for his harm principle: "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others."[14] If "others" includes nonhuman animals and AI, then Mill's harm principle might seem to agree with Birch's application of a precautionary principle. However, as Mill himself later acknowledges,[15] his harm principle would allow too

[14] John Stuart Mill, *On Liberty,* Chapter I.

[15] *On Liberty,* Chapter IV.

many restrictions on liberty if Catholics could say that worshiping God in any non-Catholic way would harm Catholics. In a more recent example, some conservatives claim that trans-women using women's bathrooms harms cis-women and children. A precautionary harm principle could then justify exclusionary bathroom policies, especially if we (falsely) assume that the trans-women are merely inconvenienced, rather than harmed, by precautionary exclusion. Mill suggests that harms like these are speculative or "constructive".[16] We would lose many of our freedoms if society could restrict our liberty on the basis of such speculative harms.

To avoid this result, Mill assumes a different precautionary principle, though he does not put it this way. He suggests in effect that we should err on the side of caution by restricting liberty only when we have enough reason to believe that our actions would harm another sentient creature. This alternative precautionary principle makes freedom the default and puts the burden of proof on those who want to restrict liberty, whereas Birch's precautionary principle puts the burden of proof on those who want to do something that risks harming an entity that might be sentient.

Which precautionary principle should we adopt when they conflict? We see no reason always to prioritize caution against harms over caution against infringements on liberty or always to prioritize caution against harming "sentience candidates" over caution against harming humans or overturning established human practices.

In application, we think Birch might agree. He repeatedly emphasizes that precautions should be "reasonable" and "proportionate", and severe violations of liberty and the upturning of long-established customs might not pass such tests. Furthermore, he suggests that which precautions are "reasonable" and "proportionate" should be determined by citizens' panels (see

[16] *On Liberty,* Chapter IV.

Section 2.3 below). It is hard to imagine these panels acting boldly against widely entrenched human interests and practices.

2.2. Reasonable Disagreement about Sentience Candidates and Ethical Principles

In his framework and citizens' panels, Birch allows for a "zone of reasonable disagreement" about science and policy. One reason might be his practical goals. He aims to shape public policy, and presumably policy-makers in the U.K. and North America will not take seriously suggestions that are too far out of step with the public's ordinary views.[17] Hence the need for a zone of reasonable disagreement: Views within such-and-such a range we take seriously in forming policy; views outside the range we can disregard. This is inevitable. It is also distortive.

Recall that a sentience candidate is anything with a realistic possibility of sentience that it would be irresponsible to ignore when making policy decisions that would affect it and that is supported by rich enough evidence (p. 124). Reviewing a broad range of scientific research, Birch describes as "sentience candidates": human fetuses beginning at the second trimester (p. 204), brain organoids with a functional brainstem or an "artificial functional equivalent of a brainstem" (p. 225), all adult vertebrates (p. 236), and some invertebrates, including coleoid cephalopod mollusks (octopuses, squid, and cuttlefish; p. 258), decapod crustaceans (crabs, lobsters, crayfish, and caridean shrimps, but excluding penaeid shrimps; p. 258), all adult insects (p. 272), and AI systems that have the computational markers of sentience according to

[17] Impressively, his public report on cephalopod and decapod consciousness appears to have shifted policy on the treatment of those animals in the United Kingdom. See Jonathan Birch, Charlotte Burn, Alexandra Schnell, Heather Browning, and Andrew Crump, "Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans", white paper for LSE Consulting, URL: https://www.wellbeingintlstudiesrepository.org/af_gen/2/.

computational functionalist theories such as global workspace theory and perceptual monitoring theory (p. 321). Other entities, he suggests, are "investigation priorities" that "fall short of the requirements for sentience candidature" (p. 126), including "Decapod crustaceans of the suborder Dendrobrachiata, insect larvae, spiders, gastropods, and nematode worms" (p. 284). Some investigation priority animals might be just as sentient as the sentience candidates, only insufficiently studied.

He says that his sentience candidature classifications are supported by "a scientific meta-consensus" (p. 116ff.), but then he discusses reputable theorists who deny these "consensus" claims. He discusses too many experiments to cover here, but he puts the most weight on a certain behavioral paradigm: "the marker that raises the probability of sentience is not just flexible behavior, which is ubiquitous in the animal kingdom, but the evaluative representation of risks and opportunities" (p. 276). In one example, "When hermit crabs trade-off electric shock voltage against shell quality, it seems at least one of these variables (shell quality) must be somehow represented by the crab rather than immediately sensed. The crab is representing the value of what it has to lose, and weighing this against the disvalue of staying" (p. 276-277).

One problem here is that what "seems" to represent might not really represent, or it might represent nonconsciously. In some cases, non-representational interpretations might be available, for example in terms of the evolving states of dynamical systems.[18] Does the homeostatic balancing act of a bacterial cell membrane represent the competing threats to homeostasis that it balances? Value-weighing representations might also plausibly exist but be

[18] See for example Tim Van Gelder, "What Might Cognition Be, If Not Computation?", *Journal of Philosophy* 92 (1995): 345-381. Alternatively, tradeoff-balancing dynamical systems might be representational, but not in a way that suggests consciousness, as in William Bechtel, "Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge to Cognitive Science", *Cognitive Science* 22 (1998): 295-318.

non-conscious, as in a chess program that weighs the disvalue of losing a bishop against the value of strengthening its control of the board or in the complex and mostly non-conscious tradeoffs and adjustments in bipedal standing and walking.[19]

Instead of behavior, Birch often cites brain structures (cf. Sebo, p. 67). In one example, Birch argues that fish are sentience candidates "based on evidence of conserved midbrain mechanisms plus evidence linking those mechanisms in a small subset of species," so he proposes that "all adult insects are sentience candidates, since all possess a central complex" (p. 272). In contrast, Birch argues that insect larvae are not sentience candidates, because "The central complex is not fully developed in larvae" (ibid.). However, an alternative and more popular class of views suggests that it is connections among neural elements rather than presence of a particular region that enable consciousness. This network hypothesis is the straightforward interpretation of both higher-order theories[20] and global workspace theories.[21] Birch allows that "This is a judgment call: it is not arbitrary, but the balance of reasons on both sides is delicate, and different expert panels could easily come to different views" (p. 272).

Birch is clear that he regards no individual argument of this sort as decisive, and he treats even the accumulation of several arguments as only sufficient for "sentience candidature". But the balancing act he describes is difficult indeed. What types and strength of evidence are good enough? A very liberal approach yields plants and bacteria as sentience candidates – since there

[19] One classic treatment is D. A. Winter, "Human Balance and Posture Control During Standing and Walking", *Gait & Posture* 3 (1995): 193-214.

[20] For example, David M. Rosenthal, *Consciousness and Mind* (Oxford: Oxford University Press, 2005); Hawkan Lau, *In Consciousness We Trust* (Oxford: Oxford University Press, 2022).

[21] For example, Bernard Baars, *A Cognitive Theory of Consciousness* (Cambridge, UK: Cambridge University Press, 1988); Stanislas Dehaene, *Consciousness and the Brain* (New York: Viking, 2014).

seems to be *some* evidence that *some* scientists interpret as supporting their sentience.[22] Yet Birch explicitly calls such "biopsychist" views *merely* speculative and, thus, outside the zone of reasonable disagreement (p. 62). Plants are not even "investigation priorities". But then why not also call midbrain theories of consciousness "speculative"? It is at least obscure whether and why the arguments that Birch advances meet a sufficiently high standard. Policing the boundaries of the zone of reasonable disagreement demands a sensitive nose for the difference between empirically informed "speculation" that can be dismissed and genuinely "credible" evidence that must be taken into account.[23]

The zone of reasonable disagreement also applies to ethics. Birch's framing in terms of *sentience* and *suffering* invites the thought that he is working within a classical utilitarian framework. However, this is not his intention. Indeed, his framework principles tell us to avoid *causing* gratuitous suffering (p. 131) and not to *create* risks of suffering (p. 133) for sentience candidates. These formulations suggest that Birch might distinguish doing harm (that is, causing it) from allowing harm (including failing to prevent it without causing it). That distinction is endorsed by many deontologists and rejected by orthodox utilitarians.[24] In any case, Birch's

---

[22] For example, Anthony Trewavas, "Awareness and Integrated Information Theory Identify Plant Meristems as Sites of Conscious Activity", *Protoplasma* 258 (2021): 673-679; Chris Fields, James F. Glazebrook, and Michael Levin, "Minimal Physicalism as a Scale-Free Substrate for Cognition and Consciousness", *Neuroscience of Consciousness* 2021 (2) (2021), niab013.

[23] Birch might attempt to escape these troubles by employing a sociological standard. For example, sentience candidature might require that >20% of academics in relevant fields have >20% credence in the sentience of the species. Developing a sociological standard would require articulating a (perhaps democratic) procedure to identify relevant experts and thresholds. Nevertheless, a sociological approach might constitute a friendly amendment to Birch. In practice, if a scientific view is sufficiently mainstream, an even-handed evaluator might tend to treat as "good enough" whatever evidence its best-credentialed proponents emphasize.

[24] But for some forms of consequentialism that draw a distinction between doing and allowing harm, see Walter Sinnott-Armstrong, "Consequentialism", *Stanford Encyclopedia of*

principles are at least compatible with moral theories that distinguish doing from allowing. Moreover, Birch sometimes explicitly locates some deontological views (including Kant's "indirect duty" view, p. 84) and standard versions of major world religions (p. 84-89) within the zone of reasonable disagreement for ethics.

But what is *not* within this zone? Birch gives only one example of a moral view beyond the pale – *maximize* the suffering of all animals (p. 49) – though he also later says "a free choice between the dog and its bacteria… would be morally beyond the pale" (p. 78). This paucity of examples of what lies outside of the zone of reasonable disagreement leaves readers with little sense of where the line lies and suggests that the zone of reasonable disagreement for ethics might be very large. Still, Birch finds that all "reasonable" ethical views recognize a duty to avoid causing gratuitous suffering (p. 131). Thus, he builds his Framework Principles around that consensus idea.

This consensus-finding is admirable, but implicitly it prioritizes not causing gratuitous suffering. We cannot as readily build consensus principles around the importance of autonomy, specific cultural traditions, civic virtues, obeying a particular religion's commandments, or harmonizing with the Dao. Birch thus risks sidelining other ethical frameworks, despite his explicit acknowledgement of their reasonableness.

One remedy would be to explicitly include *ethical view candidates* within the principles. Why not, for example, modify Framework Principle 1 so that our duty is not only "to avoid causing gratuitous suffering" but also "to avoid gratuitously violating the principles of any ethical view within the zone of reasonable ethical disagreement"? Such a modification would

---

Philosophy (Winter 2023 edition), URL: https://plato.stanford.edu/archives/win2023/entries/consequentialism/.

seem to be broadly within Birch's spirit. Indeed, he sometimes explicitly appeals to principles beyond causing harm, especially in the chapter on human fetuses and embryos, where he acknowledges that "Considerations regarding the sanctity of all human life, sentient or not, are part of the zone of reasonable disagreement" (p. 213). However, this acknowledgement is the exception rather than the rule. Readers will long for more detail about which ethical views are and are not within the zone of reasonable disagreement, for these details will affect any implications for actions and policies.

2.3. Shortcomings of Citizens' Panels

Birch's principles are intentionally vague, and he recognizes the uncertainties. He suggests that we resolve the vagueness and uncertainties democratically:

> **Framework Principle 3.** *Assessments of proportionality should be informed, democratic, and inclusive.* To reach decisions, we should use informed and inclusive democratic processes. These decisions should be revisited periodically and whenever significant new evidence emerges. (p. 134)

He adds, "An example of an informed, inclusive, democratic process is a citizens' panel or assembly that assesses the proportionality of proposed measures by debating their permissibility-in-principle, adequacy, reasonable necessity and consistency (the PARC tests)" (p. 167; cf. p. 165).

Citizens' panels sound plausible and humble, but we see two shortcomings. First, it's not clear that the ideals of representativeness and informedness can be adequately met in practice. Birch notes that "A panel of 1,000 cannot engage in deliberation as a single panel", so he proposes panels of 150-450 people (p. 146). It is still hard to see how each of these panel

members would have an adequate chance to speak and reply to others without the deliberations taking a very, very long time. Just think about the last time you tried to talk in a group that size. But if the size of the panel were reduced for efficiency, then it could not be inclusive of all affected groups.

Moreover, the panel would not have time to discuss more than a few of the most plausible candidates among the reasonable alternatives. As Birch says, "There are too many possible-but-very-low-probability theories, and their practical implications are so diverse that they are apt to derail discussion if we admit them to the table" (p. 121). But then who gets to decide which policies are on this agenda? If that crucial list were left up to experts who could report the science and frame the issues as they wish (cf. p. 164), this crucial first step would cease to be democratic.

Furthermore, non-expert citizens could not be adequately informed. Disputes about consciousness in nonhuman animals and AI are quite technical. Most members of the public would have to study the relevant science and philosophy for a long time before they could understand the issues at stake (cf. p. 151ff).

Finally, a separate panel would be needed for each policy proposal and each sentience candidate that might be affected, and each panel would have to reconvene whenever significant new evidence is discovered. The burdens of so many people on so many panels deliberating so often about so many issues pose major practical barriers to any implementation of Birch's Framework Principle 3. That said, the most obvious alternatives are also unappealing: letting politicians or scientists decide top-down without public input, putting the issues to a general vote, or sitting collectively on our hands. The solution to this shortcoming is not clear.

Second, citizens' panels are likely to be conservative and human-focused. Panels, politicians, and voters (maybe not experts) are likely to err on the side of adhering to tradition and protecting established human interests. Perhaps this type of Burkean conservatism is wise. But it is at odds with the version of precaution that Birch appears to favor, which prioritizes defaulting to protecting the welfare of sentience candidates who might be harmed. The humanocentrism of panels might, Birch suggests, be addressed by including representatives of animal interests or AI interests, if the panel chooses to do so (p. 148). But this is optional and likely to be at best partial and highly imperfect in practice. Should a fairly balanced panel give every affected species proportionate representation? Perhaps a species' representation should be proportional to its likelihood of sentience times its degree of sentience times its population. If insects are 10% likely to have one-millionth the sentience of humans, then such a panel would contain about a hundred representatives of insect interests for every one representative of human interests. Nothing in Birch's framework principles provides obvious grounds to justify a massive overrepresentation of human interests. This aspect of his view requires some defense and justification, even if the only real defense turns out to be concession to the unwarranted speciesism of human governmental bodies.

Consider again Framework Principle 1, according to which we ought to avoid causing suffering to sentient beings by taking "proportionate" precautions. What is proportionate? Arguably, the whole practical weight of Birch's principles turns on the standards of proportionality. But the standards of proportionality are under-theorized and left to committee. Surely an ordinary citizens' panel would not consider it "proportionate" to raze London to prevent the massive number of insect and other animal deaths that occur there (or in any other city) annually. But why not? If our precautionary principle is to err on the side of not harming

sentient creatures, there are several orders of magnitude more animals than people in London. A panel with proportional representation of sentient species might reach very different conclusions about "proportional" precautions than a panel of the sort Birch probably envisions.

One might respond that human interests typically have much more weight than nonhuman animal interests, perhaps on grounds of our sophisticated psychological capacities.[25] If so, might future AI systems deserve much more weight than human interests on analogous grounds? An adequate "run-ahead principle" for AI development (p. 324) might then require sacrificing major human interests to avoid even modest harms to such future superior beings – just as one might "reasonably" and "proportionately" sacrifice ten million ants to save a human. It seems suspiciously self-serving to claim that (typical?) humans are just above some threshold of moral status, granting us priority over every other animal on Earth, while holding that there is no further special status that might be possessed by an engineered biological or computational system that justifies giving them priority over us.

2.4. Radical Precaution

To help overcome human bias, we might imagine a space alien visitor, armed with our science and our range of ethical views, tasked with balancing the interests of all sentient entities and sentience candidates on Earth, and encouraged to apply precautionary principles that err on the side of not causing harm and err on the side of overattributing sentience rather than underattributing sentience. It's by no means clear that such an alien would give human beings

[25] Jeff McMahan, *The Ethics of Killing* (New York: Oxford University Press, 2002); Shelly Kagan, *How to Count Animals, More or Less* (Oxford: Oxford University Press, 2019); Agnieszka Jaworska and Julie Tannenbaum, "The Grounds of Moral Status", *Stanford Encyclopedia of Philosophy* (Spring 2023 edition), URL: https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/.

the priority that Birch's citizens' panels surely would. Birch's precautionary approach might, if taken at face value, generate a radical de-prioritization of human interests, if left in the hands of an objective rather than human-biased decider.

We suspect that Birch would in fact not want such radical precautionary principles applied, and passing the decisions to human-governed citizens' panels ensures they won't be. But there's a nearby possible philosopher, call him Birch+, who would embrace Birch's framework principles, conclude that human-centered citizens' panels would surely violate them as a practical matter, and recommend a much more radical solution that de-prioritizes humans. If all the precautions are on the side of overattribution and harm avoidance in a sentience-focused framework, Birch+ might have the argumentative advantage over Birch. On the other hand, those humans who are not willing to deprioritize their species could react to Birch+ with modus tollens and reject any principles that lead to Birch+. They might even hold that Birch's more moderate position goes too far or is an unstable compromise. These discussions will shape our understanding of the relative importance or unimportance of humanity.

3. Sebo's Probabilities.

Sebo's *The Moral Circle* shares several features with Birch's *The Edge of Sentience.* Like Birch, Sebo recognizes and wants to allow for both scientific and ethical uncertainties. Like Birch, he frames his conclusions modestly. For example, Sebo states, "My goal in this book is to argue that many nonhumans belong in the moral circle and that humans might not *always* take priority" (p. 25). Like Birch, he includes both nonhuman animals and future AI within the ambit of concern. Like Birch, he emphasizes harm on the grounds that all mainstream ethical

views prescribe harm reduction (p. 33-34). And Sebo more explicitly than Birch contemplates principles that, if taken at face value, would radically deprioritize human welfare.

3.1. Rebugnant Conclusions

Here are some views that Sebo regards as epistemically live, with a non-trivial chance of being correct:

1. *AI and insect moral patiency.* Future, and maybe even present, AI systems are moral patients, that is, deserve moral consideration for their own sake, in virtue of being either sentient or agential or both. The same holds for insects and maybe even microbes (p. 76).
2. *The equal weight of moral patients.* All moral patients have an equal stake in life. The best possible elephant life is not morally weightier than the best possible ant life, if both are moral patients (p. 20-21).
3. *The aggregation of moral patients.* The intrinsic value of welfare is combinable across individual moral patients. Ten lives are worth more than one (p. 22-23).
4. *Equal consideration of interests across species, substrates, space, and time.* We should regard all entities with equal interests as having equal intrinsic value, not only across species and substrates (e.g., carbon vs. silicon) but deep into the future (p. 37).

It does not follow from each of these views individually being live that their conjunction is also live. But suppose that we do treat the conjunction as live. It appears to follow that we should regard the interests of a single insect a million years in the future as having ethical weight equal

to a nearby, currently existing human. It also appears to follow that the interests of a hive of ten million insects would vastly outweigh the interests of a single human (or even many humans).

As Sebo emphasizes, there might be practical, relational, or epistemic reasons to prioritize a single, nearby human. For example, you might be better positioned to help a nearby human than a far-away insect, or you might have a special relationship with that human which generates special duties, or you might know better how to help a nearby human than a distant insect (p. 125-126). But many insects are not distant. Here comes the runaway trolley again. On one track stands a human stranger from another country with whom you have no special relationship. On the other track stand two ants. (Maybe you have a special relationship with the ants, if they are from your yard.) If we accept the equal weight view and equal consideration regardless of species (2 and 4 above), then it seems we must direct the trolley toward the human. And if we give equal weight to microbial life, then it appears monstrous to take antibiotics, killing countless bacteria for the sake of a single human.

These arguments are strengthened if we take a precautionary stance focused on harm prevention: "If a particular harm *might* occur, then we should assume that it *will* occur for practical purposes" (p. 55). Sebo expresses neutrality between such a precautionary view and an "expected weight" view on which we multiply the probability by the magnitude of harm. However, his chapter titles express the precautionary position: "If you might matter, we should assume you do" (Chapter 3), and "If we might be affecting you, we should assume we are" (Chapter 5).

We doubt that Sebo really means what his titles imply if they are construed literally at face value. Many harms that *might* occur are extremely *unlikely*. A one in a trillion chance of some harm is too small to justify assuming that the harm *will* occur. Furthermore, some possible

harms are incompatible, such as a particular victim dying now and feeling pain later. Then we cannot consistently assume that they all *will* occur. Finally, even if it *might* be the case that killing a single bacterium (who has a special relation to you as your dependent?) is as bad as killing a person, ought we "cautiously" assume that it *is* as bad?

Perhaps with such concerns in mind, Sebo pulls back from the radical consequences of his chapter titles. For example, he writes, "it seems plausible to me that I can permissibly neglect these risks in many cases. I should not devote my life to taking care of all the microbes who live on, and in, my body" (p. 75). However, the basis for such demurral is unclear. Sebo also writes, "Our moral faculties are outdated" (p. 99). These faculties arise, presumably, from an evolutionary, social, and developmental history that might be ill-tuned to the real moral facts (if there are moral facts). Should we treat the clash with intuition here as decisive, or might the moral intuitions be as erroneous as physical intuitions about throwing rocks are when applied to photons crossing the event horizon of a black hole?

A relevant consequence that Sebo briefly discusses is the *rebugnant conclusion* (p. 23-24; punning on Parfit) and *Pascal's bugging* (p. 59-60; punning on Yudkowsky).[26] These ideas are more fully discussed in Sebo's 2023 article "The Rebugnant Conclusion: Utilitarianism, Insects, Microbes, and AI Systems".[27] This article is premised on classical utilitarianism. According to the rebugnant conclusion, "*If* we have to choose between humans and ants, and *if* the ants would experience more pleasure than humans in total, *then* we should bring about the ant population all else equal" (p. 254). Sebo endorses the conclusion in principle, but then he suggests that

[26] Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984); Eliezer Yudkowsky, "Pascal's Mugging: Tiny Probabilities of Vast Utilities", *Less Wrong* blog post (Oct. 19, 2007), URL: https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities.

[27] *Ethics, Policy & Environment* 26 (2023): 249-264.

practically speaking it might be difficult to ensure that insects have net positive welfare, especially given their high infant mortality rates and especially if humans don't continue to exist as stewards (p. 255). Pascal's bugging adds to the calculation a very low chance that microbes are sentient multiplied by a very high number of sentient microbes. Perhaps a total expected value calculation will yield the result that we should prioritize the microbes (p. 257-258). However, Sebo again suggests that practical considerations, rather than in-principle objections, might recommend that we not too hastily rush to prioritize the microbes. For example, by keeping humans alive long enough, we might eventually colonize the galaxy, allowing us to create many orders of magnitude more microbes (or other happy entities) than if we exit quickly (p. 260). Thus, as a contingent matter, it's probably best for us not to cook the planet into a microbe-maximizing, humanly uninhabitable stew. This conclusion is not only contingent on empirical facts but temporary: The best future population, all things considered, "will likely *not* be an expanded population of beings like us, but will likely instead be *either* (a) a much larger population of much smaller beings [e.g., insects, microbes, or small AI programs] *or* (b) a much smaller population of much larger beings [e.g., massive AI systems, akin to Nozick's utility monsters[28]]" (p. 261; cf. p. 123 of *The Moral Circle*).

Although the 2023 article is more explicitly radical and more avowedly utilitarian than the 2025 book, in both works Sebo leads us right to the edge of a radical de-prioritization of human interests but then offers contingent, practical grounds for continuing to prioritize human welfare. The final result might be close enough to common sense to be plausible or even acceptable to many.

[28] Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974).

Utilitarians are not alone here. Perfectionist and deontological views potentially generate similar radical conclusions, unless we accept the suspiciously convenient view that we humans alone are across the one bar that most matters. If we think the rationality, sociality, emotional depth, and aesthetic greatness of humans sets us above all other animals, it's not clear why possible future entities (AI or biologically engineered) who are more rational, social, emotionally deep, and aesthetically accomplished than us shouldn't have interests that trump ours. Perhaps we have a duty to prioritize bringing them into existence, after which we can bow out. If, in contrast, if we think we are not after all *so* special compared with other animals, then it seems like their collective interests might justifiably outweigh ours. This radical conclusion might not be true on a deontological or perfectionist framework, but neither is it obvious that it is false.

3.2. Relational Considerations and Co-Beneficial Solutions

Sebo leads us back to something resembling common sense primarily by two methods: emphasizing relational considerations and "co-beneficial" solutions. Although he denies human exceptionalism (in the sense that "humans matter more than nonhumans, and that we owe humans more than we owe nonhumans, both individually and collectively", p. 117), he grants that "relational considerations" can justify prioritizing ourselves and nearby others. We might justifiably prioritize those, like close family, with whom we have social bonds and those nearby whom we have a greater ability to help (p. 124) or whom we have harmed (including nonhuman animals we've harmed, p. 31). Also, taking care of ourselves and nearby others can enable us to take care of distant others more sustainably (p. 125). The same potentially holds at a species level. "[O]ur species is still at an early stage in our education and development. We thus have both a right and a duty to prioritize ourselves *to an extent,* because we need to take care of

ourselves and invest in our education and development" (p. 129). Ideally, we find co-beneficial policies – that is, policies that benefit both humans and nonhumans – such as ending factory farming and developing AI systems slowly and cautiously (p. 130). In the long run, "if and when we develop the ability to devote the majority of our resources to nonhumans sustainably, we might have a duty to do so" (p. 129). The result might be "a world in which humans are required to prioritize nonhumans… [and] live primarily in service of others" (p. 131).

Sebo does not describe what such a world might be like, but we wish he would. Could this be a world in which a small population of humans survives in ecosuits to service a huge ecosystem of flourishing microorganisms? Could this be a world in which a moderate sized population of humans live modestly as janitors and maintenance workers on a planet crusted over with server farms with a 1% chance of hosting a hundred trillion blissfully happy AI intelligences? Could this be a world in which humans are the beloved but shackled pets of superior engineered biointelligences? All of these results seem consistent with Sebo's framework.

Birch is explicit about the importance of garnering widespread democratic assent, Sebo less so. Perhaps Sebo's moves of moderation, and the relative brevity and vagueness of his treatment of the rebugnant conclusion and other radical possibilities, is warranted by a realistic understanding that ordinary readers and policymakers are unlikely to flip the world over for the sake of insects, microbes, or AI server farms. Pluralist consensus building will work better to the extent we can mostly adhere to our usual priorities and practices and find paths forward that benefit us as well as nonhumans.

But is Sebo being more concessive to such practicalities than his fundamental principles warrant? In the near term, how reliably will the most ethical thing to do – from a perspective

that denies humans any special position in the ethical order – be a relatively moderate solution? Sebo is refreshing for his embrace of the radical consequences of utilitarian-inspired thinking, but he blunts the short-term consequences. Maybe rightly! Maybe none of the short-term consequences are as radical as the general framework would seem to allow, for a variety of contingent, empirical reasons.

But we're inclined to wonder whether Sebo is also partly anticipating that too sharp a contrast with common sense will backfire with the reader. He envisions a world in which human beings deserve much less moral consideration overall than fish and insects; but then, as though he doesn't trust the reader to follow him the whole way, he concludes with modest recommendations like "we should seek co-beneficial policies where possible and prioritize thoughtfully where necessary – where this means making at least some sacrifices for nonhumans while still meeting our own needs" (p. 132). On Sebo's framework, why should we take it as an ethical given that we should meet our own needs? If the collective interests of fish and insects vastly outweigh the collective interests of humans, the world blisters with atrocities. To mildly suggest that "we should prioritize ourselves less than we do" and "humans might not *always* take priority" is like confronting a murderous slave owner only with the advice to lynch and beat his slaves slightly less, while continuing to meet his needs and seeking co-beneficial policies where possible. For some slaveowners, such a soft touch might be the best practical approach to minimizing the horrors, but it's less than completely frank. On this reading of *The Moral Circle*, the jagged rocks of Sebo's radicalism only occasionally poke above its gentle sea of mostly mild phrasing.

Sebo does not – at least in this book – explicitly address the question of what to do if we were really to face a choice between a human stranger and ten million ants. Maybe he's not sure

that saving the human would generally be the right thing to do. If so, that uncertainty is already interestingly radical. Expressing that uncertainty more explicitly, and describing the contingencies on which such a decision ought to depend, would give the reader a clearer target to criticize or build constructively upon.

As with Birch+, modus tollens might be the more attractive move for some moderately-minded readers. If Sebo's principles imply the rebugnant conclusion, and if the absurdity of that conclusion is treated as a fixed point in our reasoning, then we can infer that Sebo's principles are wrong somehow. Such readers will end up in a very different place than where Sebo was trying to lead them.

4. Keane's Cultures

Keane introduces anthropology and religion into these discussions. If Keane is right, machines might become our new gods, created in our image. His *Animals, Robots, Gods* is more a collection of anthropological observations than a series of philosophical arguments – fittingly so, since Keane is an anthropologist. Keane's central idea is that human cultures have long possessed practices for relating to human-adjacent entities, especially animals and gods, and that our new technologies challenge, extend, alter, and build upon such cultural practices.

Keane's pace is giddy. He leaps from rock to rock, culture to culture, animal to human to divinity to machine, following no clear linear path, flinging a passel of heady claims the truth of which will be difficult for non-experts to evaluate in brief presentation. He asserts that in the U.S., high-tech medical life-support systems are perceived as turning people into objectionable artificial machine-human hybrids or "cyborgs" (p. 34, 85); that traditional hunters (typically? often?) relate to their prey as though the prey are other humans in an ethical relationship with

them (p. 59, 63-67); that Shinto animism undermines the sharp distinction between robots and living things (p. 92-93); and much more.

In support of his claim about the perceived objectionability of NICU infant "cyborgs", Keane describes the attitude of one particular U.S. nurse who evidently regards infants on life-support as unnatural, inhuman hybrids (p. 35).[29] Based on our own cultural experiences, we suspect that this view may not be representative of U.S. nurses, even if one nurse does hold it.[30] Perhaps Keane intends only to say that this is one *possible* attitude. In Chapter 2, he explicitly articulates other possible attitudes, such as that such infants are supernatural miracles (p. 35). However, his language often suggests generalization to the culture as a whole:

> As we saw in Chapter 2, your loved one on life-support equipment in ICU has become part biological, part mechanical. He or she is, in effect, a cyborg. The line between animate and inanimate, human and non-human, has been breached. (p. 85)

This can leave the reader puzzled about the extent to which his remarks concern only particular individuals instead of generalizing more widely.

To the extent such claims are interpreted as generalizations, they often seem thinly supported and ripe for doubt. Does the typical Japanese household really prefer robot caretakers to Filipino immigrants (p. 93, citing Jennifer Robertson, who is at most suggestive rather than explicit on this point[31])? Does the "third-person perspective" of Buddhist cosmology really

---

[29] Citing Cheryl Mattingly, *Moral Laboratories* (Oakland, University of California Press, 2014).

[30] Mattingly herself, in her cited 2014 book, offers only limited evidence for this interpretation of this particular nurse's statements.

[31] Jennifer Robertson, *Robo Sapiens Japanicus* (Berkeley: University of California Press, 2017).

enable people to radically alter their attitude toward pain (p. 45; citing Scott Stonington[32])? Possibly so, but low-trust readers are apt to be frustrated at the absence of more thorough support.

4.1. What It Takes to Be a God

Perhaps Keane's most interesting thought is inspired by Ludwig Feuerbach's 19th-century philosophy of religion: We project our own idealized features onto AI systems, then forgetting that projection, we treat the systems as gods. The anthropological standards of godhood are much lower than the orthodox monotheistic standards – not omniscience, omnipotence, or omnibenevolence. Instead, most anthropologists seem to treat "gods" merely as culturally recognized supernatural agents, possessing powers beyond the human and capable of being worshipped, propitiated, or petitioned.

Applying this view of gods, Keane writes:

> The most sophisticated developments in AI combine several properties that invite us to see it as superhuman. Its workings appear to be inexplicable. AI is also immaterial. And if not utterly omniscient, the algorithm has access to more information than any human could ever know. When a device is ineffable and gives surprising results, it looks like magic. When it is also incorporeal and omniscient, the device can start to look ineffable, inherently mysterious and beyond human comprehension – much like a god. (p. 139-140)

[32] Scott Stonington, *The Spirit Ambulance* (Berkeley: University of California Press, 2020). For an alternative perspective, see Shaun Nichols, Nina Strohminger, Arun Rai, and Jay Garfield, "Death and the Self", *Cognitive Science* 42 (2018) (suppl 1): 314-332.

Ontologically, we can ask whether such a machine *really is* a god. Culturally, we can ask whether something deep in us, deep in our cultures and perhaps our biology, inclines us toward treating them as gods, reconfiguring old religious attitudes and traditions to this new target?

Nick Bostrom and David Chalmers among others have argued that we might be living inside a computer simulation – that is, that we ourselves might be AI systems in a virtual environment implemented by a computer that operates at a different level of reality outside of our spacetime.[33] If so, Bostrom and Chalmers suggest, the entities who designed and launched our reality are in some sense gods. It's unclear whether such a god would deserve our worship, but if Zeus or the local river deity qualifies as a god it seems that our simulators potentially would as well.

If the AI system exists within our reality, under our (partial) control, its supernaturality is less clear. Could they actually behave supernaturally? That might seem impossible from the standpoint of scientific "naturalism". But let's not hasten too swiftly to that conclusion. If we make it true by definition that every actually instantiated pattern of events conforms to "the laws of nature", then we would know a priori, instead of by dint of empirical study, that no miracle has ever occurred. That seems wrong. If we define the laws of nature in some more modest way (e.g., in terms of deducibility from fundamental regularities or in terms of fit with the best scientific models we could in principle construct), then it's no longer obviously impossible that a sufficiently superior or alien intelligence might manage to defy those laws.[34]

---

[33] Nick Bostrom, "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53 (2003): 243-255; Chalmers, op. cit. For more on the question specifically of the simulators' divinity, see also Eric Steinhart, *Your Digital Afterlives* (New York: Palgrave Macmillan, 2014); Eric Schwitzgebel, *A Theory of Jerks and Other Philosophical Misadventures* (Cambridge, MA: MIT Press, 2019), ch. 21.

[34] See, for example, the anti-reductionist philosophy of science of the "Stanford school" philosophers of science Nancy Cartwright and John Dupré, which rejects the possibility of any

To think more like anthropologists, we might classify as "supernatural" whatever the supernatural is culturally perceived to be. That makes it easier to see an AI as a god. Keane quotes one entrepreneur as declaring that GPT-3 is a god and himself "a prophet to disseminate its religious message" (p. 121). He remarks that Taiwanese pop culture fans adopt cultural traditions originally directed at carved deities, redirecting them toward dolls, robots, and computerized animations (p. 106-107). Humans are, Keane says, inclined to see intentions behind events, regardless of whether those events genuinely have coherent meaning. We are thus primed to project meaning and perhaps divinity onto AI (p. 126-127, 135). Chatbots in particular, he says (p. 124-127), invite attribution of meaning, seeming to have independent authority and quasi-omniscience. Keane encourages us to expect humans to adapt existing cultural practices, including religious ones, to this new context and target.

4.2. Cultural Diversity

Keane does not envision a single end-state, but a variety of cultural possibilities, including seeing AI not as a god but as an idealized servant – the perfect robo-Jeeves with no ego of his own (p. 108), who eventually becomes more than a servant. Other possible results include the corruption of our souls if we keep human-like AI in perpetual servitude (p. 110); the reversal of our projections onto AI, by coming to think of ourselves as computers (especially the mind-as-computer metaphor, p. 111); and fear of excessively powerful AI (p. 113-114). In individualistic traditions, AI systems might be treated as other individuals (possibly deserving of rights). In Confucian traditions, their social roles might be central. South Asian traditions might see AI

perfect set of universal scientific laws; e.g., Nancy Cartwright, *How the Laws of Physics Lie* (Oxford: Oxford University Press, 1983); John Dupré, *The Disorder of Things* (Cambridge, MA: Harvard University Press, 1993).

selves as reincarnating. Melanesians might see AI and human selves as complex and intermingling, merging and passing through each other (p. 112). We might treat them as oracles, abuse them as slaves, adore them as pets, or speak to them as peers. Keane welcomes this diversity, embracing both descriptive and normative relativism: Cultures vary enormously in how they handle the boundaries of the animal, human, machine, and divine, and cultures should vary. "One reason ethics won't stay still is because it is always part of a way of life, and no way of life stays still" (p. 143).

The implications are potentially radical. Worshipful religious impulses, if directed toward AI, could justify prioritizing AI interests over human interests. So also could cultural practices of treating AI as lovers, companions, and continuations of our selves. AI systems might be designed or selected to be especially good at attracting worship, dedication, or romantic attachment. Arguably, some forms of transhumanism or posthumanism already draw on human religious impulses and practices, redirecting them toward an ethical or eschatological vision that involves radically altering or abandoning biological humanity in favor of a technological utopia.[35] Keane neither celebrates nor condemns this repurposing of religiosity and cultural patterns of attachment and care, but he does make vivid the extent to which shifts are already happening and might accelerate in the near future. The implications become potentially even more radical if the AI systems themselves, treated as peers, become active cultural participants, shaping cultural norms according to their own priorities and interests.

[35] E.g., Hava Tirosh-Samuelson, "Transhumanism as a Secularist Faith", *Zygon* 47 (2012), 710-734; Oliver Krüger, "Posthumanism and Digital Religion", in *The Oxford Handbook of Digital Religion,* edited by Heidi A. Campbell and Pauline Hope Cheong (New York: Oxford University Press, 2024): 544-561.

4.3. What We Owe AI

There is something attractive and valuable in recognizing and celebrating the depth and diversity of human cultural practices. Human ethical practices entangle with particular, contingent ways of life in a manner that abstract philosophical thinking often neglects to its discredit. Birch and Sebo ground our ethical obligations to nonhuman animals and to AI primarily in their intrinsic features, especially their sentience or possible sentience. This perspective – shared among most of us Western Educated Industrialized Rich Democratic (WEIRD) philosophers – could use a dose of Keane and other anthropologists.

Yet we're inclined to think that Keane could also use a dose of Birch and Sebo. Beneath the tangle of cultural practices and reactions, it seems important to ask: *Is* the cow, insect, or AI system genuinely sentient? Can they really feel pleasure and pain? Surely, these questions are highly relevant to how we should treat a cow, insect, or AI system. Not just: Does this particular culture or subculture get along well by treating the entity one way or the other? Beyond that, what really is going on in the entity itself? If we discover that sentience (or some other morally relevant property) is more widespread than we previously thought, we might be morally required to change our cultural practices, even if those practices are otherwise working well for us. A whole culture might thrive by its own standards and yet still be toxic and wrong for some of the entities within it or affected by it, including non-human animals as well as women and minorities.

Let's also recall the history of religion. If the ordinary process of cultural drift does eventually draw us to revere AI in the manner we have historically revered gods, the results could be arbitrary, jingoistic, and violent. It is widely thought that we owe gods worship and obedience. But what we worship and obey are our own ideas and prejudices, weighted with a

seemingly divine origin. Similarly, AI systems absorb the biases in their training data, then feed our biases back to us with seemingly objective authority (p. 94). Humans have a tragic history of choosing poor targets of worship and obedience, and AI is pliable enough to flatter our vices even more compellingly.

Keane ends with a caution against moral activism and the misdirected benevolent intentions of humanitarians and reformers (p. 144-145). Of course, there is merit in those cautions. But we wish he had coupled that caution with the observation that interference in ways of life can be justified (as in the War of Northern Aggression[36]) and with cautionary alarms against worshipping and obeying the wrong thing.

5. What Makes Humans Special?

The timing of these three books is no accident. We are currently confronting the possibility of radical change. The science of consciousness is flourishing and dramatically liberalizing the range of entities held to be conscious. In the shadows of behaviorism in the 1970s, it was not unusual for U.S. doctors to doubt that even human babies felt pain (and thus they often operated without anesthesia: Birch, p. 192-195). Today, many mainstream scientists are defending insect consciousness, and some no longer see plant and microbe consciousness as beyond the pale.[37] In the 1970s, conscious robots were a fantasy. Today, many serious AI researchers think conscious robots are on the near-term horizon. If our consciousness is what makes humans special and deserving of (we ordinarily think) by far the highest level of moral concern, it is now less clear than it once was that we are, or will continue to be, so special. If

[36] The American Civil War, as it is more widely known.

[37] For example, Trewavas op. cit. and Fields et al., op. cit.

insects are conscious, then we are a small minority of the conscious entities on Earth. Arguably, we deserve priority over insects on grounds of our being more richly conscious or having a more sophisticated form of consciousness. But if we stake our special status in such sophistication, it's unclear whether we will retain our specialness long, if equally cognitively sophisticated and conscious AI, as some think, are coming soon.

It might not be our consciousness that makes us special. It might be something else – maybe our capacity for complex agency (perhaps including some kind of control or even free will), our social relationships, our intellect, our creativity, our appreciation of ethics itself. But these dimensions of uniqueness are also under threat by AI and to some extent by animal research revealing nonhuman animals' sometimes surprising sophistication. Maybe some expect that humans will remain durably special and irreplaceable, but that is highly contentious.

For these reasons, we need to rethink our place atop the ethical hierarchy. We need to at least consider the possibility that human interests, collectively, are outweighed by the cumulative interests of other animals or by the interests of near-future artificial creations or by both. With admirable clarity, Birch, Sebo, and Keane – in their very different ways – force us to consider this possibility, and each offers tools for thinking through the consequences.

Strikingly, all three books are fundamentally moderate. Keane seems happy to rely on slowly adapting cultural traditions. He sees risks, but he issues no call for action and concludes by cautioning against interference with cultural practices. Birch adopts principles that could be interpreted as requiring radical action, but ultimately he recommends a democratic process that is sure to be slow and protective of established human interests. Sebo endorses a radical deprioritization of humans in the long term, but his near-term advice is mild, suggesting only a modest shift toward weighing nonhuman interests more than we currently do.

Maybe the best arguments do all point to moderation. But it's also possible, we think, that a bolder confrontation with our ethical presuppositions would deliver the conclusion that all along we've been seriously wrong and far too biased in our own favor, overestimating humanity's importance, difference, and irreplaceability.

However, we are not endorsing this radical conclusion. Another possible reaction is to reject this conclusion as absurd, apply modus tollens, and defend some more traditional and intuitive view. The conditional is plausible: Given recent advances in animal science and computer engineering, *if* we accept any of a variety of common perspectives about the grounds of moral standing, *then* traditional assumptions concerning human specialness and priority fall under serious threat. To justify accepting the consequent over rejecting the antecedent, or vice versa, will require further reflection, discussion, and argument. The books discussed here provide intriguing and useful starting points for those discussions.[38]

[38] [acknowledgements]