Relations
BEYOND ANTHROPOCENTRISM

12.2
December 2024

Environmental Ethics: Philosophical Issues and Educational Perspectives

Special Issue

Edited by Francesco Allegri, Matteo Andreozzi, Roberto Bordoli

Comments, Debates, Reports and Interviews

The Edge of the Moral Circle

Jeff Sebo

New York University (USA)

DOI: https://doi.org/10.7358/rela-2024-02-sebj

jeffsebo@nyu.edu

ABSTRACT

This essay explores the relationship between two recent books on the scope of moral consideration: "The Edge of Sentience" (Birch 2024a) and "The Moral Circle" (Sebo 2025). Both books address the ethical and scientific challenge of determining how to interact with beings of uncertain sentience and moral status. They also argue for similar conclusions: they develop precautionary frameworks for guiding these decisions, and they argue that many invertebrates, future AI systems, and other beings merit moral consideration or, at least, further investigation. However, the books differ in focus and scope: "The Moral Circle" focuses more on ethical theory and long-term progress, while "The Edge of Sentience" focuses more on public policy and short-term progress. This essay highlights the complementary nature of these works and identifies key areas for further research, including how to navigate moral uncertainty and how to reconcile ethical principles with practical and political realities.

Keywords: agency; AI welfare; animal welfare; consciousness; invertebrate welfare; moral circle; moral uncertainty; precautionary principle; research ethics; sentience.

1. Introduction

In the second half of 2024 and first half of 2025, two books on the science, ethics, and politics of the moral circle appeared in quick succession [1]. In July 2024 Jonathan Birch released *The Edge of Sentience* (hereafter TEOS), which examines how to treat beings of uncertain sentience and moral status and develops a precautionary framework for making these decisions. Then, in January 2025, I released *The Moral Circle* (hereafter TMC), which *also* examines how to treat beings of uncertain sentience and moral status and develops a precautionary framework for

[1] Thanks to Francesco Allegri for inviting me to publish this article, and thanks to Toni Sims for research assistance.

making these decisions. While the books differ in many ways, these similarities are striking – in my view, a sign of ideas whose time has come [2].

Birch and I wrote these books independently of each other. He published his paper "Animal Sentience and the Precautionary Principle" in 2017 (Birch 2017), and his book developed out of this paper and related work. Meanwhile, I published my paper "The Moral Problem of Other Minds" in 2018 (Sebo 2018) [3], and my book developed out of this paper and related work. In the year leading up to the book releases, we realized how much our work interacts and started collaborating, for instance on *The New York Declaration on Animal Consciousness* (with Kristin Andrews and others [4]) and "Taking AI Welfare Seriously" (with Robert Long and others). However, by the time we started working together, our books were more or less locked in.

My aim in this essay is to offer an initial reflection about how these books relate to each other. They inevitably have many similarities. For example, they both argue that we should extend moral consideration to all beings who have a realistic, non-negligible chance of being sentient (that is, able to consciously experience pleasure, pain, and other such states). They also argue that a vast number of beings meet this requirement, and that we have a responsibility to take reasonable, proportionate steps to consider and mitigate welfare risks for them. In some cases that means implementing precautionary policies now, and in other cases it means conducting and supporting research to inform future policy decisions.

However, the books also have many differences. For example, TMC offers a much more concise survey of the relevant issues, along with more material about ethical theory, how humanity should expand its moral circle in the long run, and what kind of theory of change might allow us to do that. In contrast, TEOS offers a much more detailed survey of the relevant issues, along with more material about public policy, how humanity should treat beings at the edge of sentience in the near term, and what kind of political process might allow us to do that. Perhaps relatedly, the books also make different remarks about edge cases like

[2] For related work on these issues, see Chan 2011; Schwitzgebel 2023; Clatterbuck and Fischer 2025.

[3] For the sake of completeness, I can add that "The Moral Problem of Other Minds" developed out of the essay "Reconsider the Lobster", which I published online in 2015 (Sebo 2015).

[4] Andrews *et al.* 2024.

plants and fungi, and about implications for practices like farming and research, as we will see.

2. Book summaries

This section opens with a brief summary of each book, starting with *The Moral Circle* (again, TMC) and then turning to *The Edge of Sentience* (again, TEOS).

My aim in TMC (Sebo 2025) is to argue at a high level that a variety of views in both ethics and science support the following conclusions: There is a non-negligible chance that a vast number of beings matter for their own sakes, including many invertebrates and future AI systems, and that our actions and policies are affecting them, both directly and indirectly. We thus have a responsibility to extend moral consideration to a vast number of beings, with transformative implications for our lives and societies. We also have a responsibility to reject human exceptionalism, the view that humans always take priority, and to work toward a future in which our species can achieve and sustain higher levels of support for nonhumans.

Chapter 2 provides a brief introduction to moral status. We still disagree about which features are required for welfare and moral standing, that is, for being capable of being harmed and mattering for your own sake. For example, is sentience required? Is agency without sentience enough? Is life without agency enough? And how should we define each of these concepts in the first place? We also disagree about whether moral status comes in degrees and whether large groups matter more than small ones. However, there is wide agreement that beings who can consciously experience positive and negative states *and* consciously set and pursue their own goals have moral standing. The book accepts this view as a premise.

Chapter 3 provides a brief introduction to moral theory. Even when we agree about which beings matter, we disagree about what we owe them. For example, is morality primarily about promoting welfare? Respecting rights? Cultivating virtuous characters and caring relationships? We also disagree about whether distance is morally significant. Do we owe more to beings who are closer to us in time, space, biology, or materiality, all else being equal? However, there is wide agreement that we should reduce and repair the harms that we cause others where possible, and that we should cultivate character traits, relationships, and

other structures that support this work. The book accepts this view as a premise as well.

Chapter 4 argues that if you might matter, we should assume you do. Since we have uncertainty about which features suffice for moral standing and about which beings have those features, we have uncertainty about which beings matter. And if there is at least a *non-negligible* (that is, at least *one in a thousand*) chance that a being matters, we should give them at least *some* moral consideration in decisions that affect them, as a precautionary measure. We can agree on at least that much even if we disagree about other issues; for example, does a being with a one in a *million* chance of mattering merit consideration? Also, do beings with a higher chance of mattering merit *more* consideration than beings with a lower chance?

Chapter 5 argues that many beings might matter. Many beings (vertebrates, invertebrates, plants, fungi, chatbots, robots, even microbes) have at least a minimal ability to detect helpful and harmful stimuli and set and pursue goals. So, *if* these beings are conscious (capable of subjective experience), *then* they *might* be able to consciously experience positive and negative states and consciously set and pursue goals, depending on how these concepts are defined and how these capacities interact. I argue that given the importance and difficulty of these questions, we should attribute *at least* a one in a thousand chance of mattering to all vertebrates, many invertebrates, and many future AI systems, as we seek to learn more.

Chapter 6 argues that if we might be affecting you, we should assume we are. Since we have uncertainty about what we owe others and about how our actions are affecting them, we have uncertainty about whether our interactions with others are morally permissible. And if there is a non-negligible chance that our interactions with particular beings are morally impermissible, then we should give this possibility at least some consideration as well. Importantly, this can be true of our individual *and* collective actions; if, for example, many individuals collectively pollute a lake, thereby imperiling the animals in that lake, then this animal welfare risk should at least be *one factor among many* in our moral assessment of this practice.

Chapter 7 argues that we might be affecting many beings. Even if we assume that morality requires only that we reduce and repair the harm that we cause, we still have a lot of uncertainty about which actions and policies cause harm. We now live in the Anthropocene, an epoch in which our actions and policies are increasingly affecting humans and

nonhumans all over the world, both directly and indirectly. As a result, we no longer have the luxury of simply leaving others alone. Since our actions and policies might be affecting others whether we like it or not, we have a responsibility to consider risks for a vast number and wide range of beings as much as reasonably possible when assessing our actions and policies.

Chapter 8 weaves these threads together to present a challenge to human exceptionalism, the idea that humans always matter most. Even if we can prioritize humans in many cases for capacities-based reasons (they have a "higher" capacity for welfare) or relational reasons (our lives are more entangled with theirs), these conditions might not always hold. In the future, humans will likely share the world with *both* larger numbers of smaller nonhumans (e.g. insects) *and* smaller numbers of larger nonhumans (e.g. advanced AI systems). Moreover, our lives will be entangled with theirs whether we like it or not, and even if we lack the ability to support them now, we can – and should – work to develop this ability over time.

The book closes with a thought experiment inspired by John Rawls (1971). Suppose you were born into a future in which digital minds have power over biological minds, and the digital minds are uncertain whether biological minds matter and what, if anything, they owe them. In your view, what principles should digital minds use when deciding how to treat biological minds? Should they proceed on the assumption that biological minds have *no* moral significance? That they have *full* moral significance? That they have *partial* moral significance? While we might or might not actually face this kind of future, reflecting on how we would think and feel if we did can be a good way to summon at least *a bit* more impartiality today.

Birch's aim in TEOS is to argue that a variety of views in both ethics and science support the following conclusions: there is a realistic possibility that a vast number of beings are sentient, and policymakers have a responsibility to take reasonable, proportionate measures to consider and mitigate welfare risks for these beings when making decisions that affect them. Moreover, policymakers should determine which measures are realistic and proportionate not by considering the matter themselves, but rather by seeking expert and public input so that their decisions can be both informed and democratically legitimate. Birch also explores a variety of real-world case studies and suggests principles for assessing them.

Birch starts by describing what he calls the "zone of reasonable disagreement" (Birch 2024a, 45) that is, the set of people and ideas that we

can take to be reasonable. Roughly speaking, a person is reasonable if this person is responsive to evidence and reason, and an idea is reasonable if this idea remains tenable after a reasonable person has considered the available information and arguments. Birch states that the zone of reasonable disagreement includes a wide range of views in science and ethics. However, he also states that it excludes baseless positions, dogmatic adherence to refuted positions, and morally abhorrent positions (say, the sadistic idea that we should promote suffering in the world for its own sake).

In particular, Birch states that many views about the nature of consciousness are reasonable at present, ranging from the idea that consciousness is based on a particular kind of substance (including physical substances, non-physical substances, or basic substances that underlie both kinds) to the idea that consciousness is based on a particular set of functions (such as recurrent processing, a global workspace, or an attention schema), depending on the details of these views. He argues that many views about the nature of moral standing are reasonable at present, including views that focus on sentience, views that focus on agency, and even views that focus on life (provided that they prioritize sentient life in particular ways).

Birch explains that his goal in the book is to develop a precautionary framework for making policy decisions regarding beings of uncertain sentience that all reasonable views in science and ethics "can accept as fair" (Birch 2024a, 114). The framework is precautionary for several reasons: it prevents uncertainty from undermining decision-making. It recommends taking precautions in proportion to the threats that they address. It recommends determining which precautions are proportionate via inclusive democratic procedures. And it recommends setting a relatively low bar for triggering those procedures. In short, we should not wait for confidence about sentience before taking action, since a realistic possibility of sentience is enough.

How can policymakers determine whether a particular policy is proportionate to a particular risk? Birch proposes that they convene citizens' assemblies, randomly selected groups of citizens that deliberate about particular policy issues and make non-binding recommendations to policymakers. He also proposes that citizens' assemblies use a "PARC" procedure to develop their recommendations. This procedure involves asking whether proposed policies are: (1) Permissible in principle (they avoid breaking legal or ethical rules), (2) Adequate (they reduce risk to an acceptable level), (3) Reasonably necessary (they avoid doing too much),

and (4) Consistent (they avoid deviating from precedent without good reason).

Birch then discusses a variety of cases in detail, starting with cases involving our own species. For example, he argues that we should recognize humans in "persistent vegetative states" (a term that he rightly criticizes, 174) as sentience candidates, considering their interests and mitigating risks for them. He also argues that we should recognize human fetuses in the second trimester as sentience candidates, which, he notes, is compatible with multiple views about the ethics of abortion. And he argues that we might soon need to recognize sufficiently complex neural organoids (simple models of the human brain grown from tissue cultures) as sentience candidates as well, and that we should think ahead about this possibility now.

Regarding other animals, Birch argues that all adult vertebrates and many adult invertebrates are sentience candidates. He examines animals in these categories for behavioral and anatomical markers of sentience according to a variety of leading scientific theories, and he concludes that particular kinds of cephalopods, decapods, and insects are sentience candidates given current evidence. In contrast, Birch is not yet prepared to say that, say, gastropod mollusks, nematodes, spiders, or insect larvae are sentience candidates given that we have much less evidence about these animals at present. However, he does regard these animals as priorities for future research, and will soon be taking on some of that research himself.

Finally, regarding AI systems, Birch notes that we might not be able to give much weight to behavioral evidence at present, since AI systems appear capable of "gaming" behavioral tests (313). However, we can still examine AI systems for architectural indicators of sentience according to a variety of leading scientific theories. When we do, we find that we are currently unable to rule out a realistic possibility of sentience in future AI systems. This includes large language models like ChatGPT, Claude, or Gemini, and it also includes other kinds of systems, such as virtual brains that we create by emulating physical brains or evolutionary processes. As with organoids, we should think ahead about issues that might arise for such systems.

What kinds of policies should we select if we extend an appropriate level of moral concern to these beings? Birch thinks that we should ban some animal use industries; for instance, he recommends that we ban octopus farming given the difficulty of farming such complex, intelligent, sensitive animals humanely. However, he also thinks that we can "plausibly" maintain other animal use industries provided that we treat

animals humanely; for example, he recommends that we reform lobster farming (say, by stunning prior to slaughter) rather than ban this industry, though whether his recommended reforms are meant to be steps on a path towards further reforms or ends in themselves is not always clear.

Before I consider how these books relate to each other, I should emphasize that I think that TEOS is an excellent and important book. It develops a theoretical framework for assessing scientific and ethical issues at the edge of sentience in accessible, engaging language for experts and non-experts alike. It also develops practical assessments, policies, and procedures for a variety of policy issues at the edge of sentience, some of which have already persuaded policymakers to start considering and mitigating welfare risks for invertebrates, for example (about which more below). This is a remarkable achievement, and notwithstanding any potential disagreements that I describe below, I hope that it has wide influence.

3. Similarities, differences, and future directions for research

There are a lot of interesting similarities and differences between these books. We can briefly consider the main similarities and then spend more time on the main differences.

First, and obviously, both books ask how to treat individuals of uncertain sentience and moral status. This similarity is important, since most previous works on sentience and moral status have approached these topics differently, arguing for specific views about these topics and then applying those views to specific beings. However, Birch and I believe that disagreement and uncertainty about sentience and moral status are likely to be ongoing, and that we need to attempt to work with them instead of – or, at least, in addition to – attempting to move past them. That requires establishing a "zone of reasonable disagreement" and making arguments that can be compatible with most or all views within that zone.

Second, and relatedly, both books develop a pluralistic, probabilistic, and precautionary framework for deciding how to treat individuals of uncertain sentience and moral status. The frameworks are pluralistic in that they cohere with a wide range of views in ethics (about the basis of moral status) and science (about the basis of sentience). The frameworks are probabilistic in that they seek to deliver higher or lower degrees of

confidence about sentience and moral status rather than all-or-nothing verdicts. And the frameworks are precautionary in that they seek to establish at least minimal moral consideration for all beings with at least a realistic, non-negligible chance of sentience and moral status, given the evidence.

Third, and relatedly, both books argue that when a being has a realistic, non-negligible chance of sentience and moral status, we should (at least) consider and mitigate the risk that our actions and policies inflict gratuitous suffering on that being. While ethicists disagree about many issues, they generally agree that we should consider and mitigate realistic, non-negligible risks, and they also generally agree that we should avoid inflicting gratuitous suffering on sentient beings. Yet at present, humans tend to neglect welfare risks that we impose on individuals of uncertain sentience and moral status. Thus, even this minimal ethical commitment could have transformative implications for our lives and societies.

Fourth, both books apply this analysis to a vast number and wide range of beings of uncertain sentience, with special focus on nonhumans like invertebrates and AI systems. While scientists and philosophers disagree about many edge cases – say, plants and fungi – they increasingly agree on at least this much: first, many invertebrates (including many adult insects) have a realistic, non-negligible chance of being sentient, given the evidence; second, many near-future AI systems could have a realistic, non-negligible chance of being sentient as well, given the current path and pace of AI development. As noted above, these conclusions are already significant no matter what we conclude about other edge cases.

Finally (for now), both books attempt to strike a balance between realism and idealism about moral progress. As I argue throughout TMC, humans have a responsibility to transform our lives and societies, but we also have a responsibility to work within our epistemic, practical, and motivational limitations as we do. And as Birch argues throughout TEOS, policymakers have a responsibility to engage with the public and select policies that can be not only scientifically and philosophically but also politically legitimate. In both cases, the upshot is that our task is to pursue incremental reforms that can be achievable and sustainable, and that can build momentum toward further such reforms in the future.

These and other similarities notwithstanding, there are many interesting differences between the books as well. One difference is so obvious that it can be easy to take for granted, but we can start with it because it might be at least partly at the root of other, more seemingly substantive differences: TMC and TEOS are different kinds of books, written

for different kinds of audiences. While both books aim to be rigorous, systematic, accessible, and engaging for both a scholarly audience and a general audience, they still have different centers of gravity in these respects, namely: TMC is shorter and closer to the "trade book" end of the spectrum, whereas TEOS is longer and closer to the "academic book" end of the spectrum.

Second, TMC focuses more on theoretical ethical issues. It examines ethical questions like: Does everyone who matters have equal intrinsic value? Does the intrinsic value of a population depend on the intrinsic value of its members? Do beings with an extremely low but non-zero chance of mattering merit an extremely low but non-zero amount of moral consideration? Do beings with a higher chance of mattering merit greater moral consideration than beings with a lower chance of mattering? Do we have a duty to help others? Is distance in space, time, biology, or materiality morally significant? What do we owe nonhumans in general, and what follows for the idea that humanity always takes priority?

In contrast, TEOS focuses more on practical political issues. It discusses a variety of real-world case studies, involving humans in comas, infants, fetuses, organoids, mollusks, crustaceans, insects, chatbots, emulations, and more. Throughout these discussions, Birch proposes principles for considering and mitigating welfare risks for individuals of uncertain sentience that he takes to be reasonable, proportionate, and politically viable. He also proposes frameworks for assessing these and other proposals in a scientifically, ethically, and politically legitimate way, emphasizing the value of seeking expert and public input when making decisions at the edge of sentience and discussing how policymakers can achieve this goal.

Third, and relatedly, TMC focuses more on the distant future (though it discusses the present and near future too). It asks what we might owe distant future moral patients, for instance when contemplating sending life to a new planet, with the result that sextillions of extra beings will live and die in future centuries. It also asks how we should expand the moral circle over time. In scenarios where humanity remains in power, how should we treat nonhumans in the future? In scenarios where we lose power, how should nonhumans treat us in the future? Either way, how can we build a future in which those in power treat everyone else with the appropriate level of moral concern, and what follows for our actions and policies today?

In contrast, TEOS focuses more on the present and near future (though it gestures at the distant future too). In 2021, Birch led a working

group that assessed evidence of sentience in cephalopod mollusks and decapod crustaceans and recommended that the UK government include these taxa in its animal welfare act (Birch *et al.* 2021). The UK government took this advice (Gov.UK 2021). This experience clearly informed TEOS; this book is ideal for informing similar steps for each issue it addresses. However, it also discusses the need to think ahead, especially for beings like future organoids and AI. And it discusses the need for research to inform future policymaking, especially for beings about which (or whom) very little is known.

Fourth, and also relatedly, TMC is more radical about which beings merit consideration (though still moderate relative to my own views). I argue that when setting the scope of the moral circle, we should cultivate humility about *both* ethics (regarding which features are required for moral standing) *and* science (regarding which beings possess these features). I also argue that we should set the bar for inclusion no higher than a one in a thousand chance of mattering. And while I insist only that insects, future AI systems, and other such beings make the cut, I also take seriously the possibility that plants, microbes, current AI systems, and other such beings make the cut, and I sometimes explore the implications of these possibilities.

In contrast, TEOS is more moderate in this respect (though still radical relative to the status quo). While Birch endorses pluralistic reasoning in the face of moral disagreement, he focuses on which beings might be sentient, not on which beings might be morally significant more generally. And while he endorses probabilistic reasoning in the face of scientific uncertainty, he does not defend a specific probability threshold for moral inclusion, though he appears to have a higher probability threshold in mind than I do. Birch also expresses skepticism that plants, microbes, current AI systems, and other such beings make the cut, on the grounds that arguments for sentience in these beings are too speculative at present.

Fifth, TMC is likewise more radical about what we owe those in the moral circle. I argue that we have at least a weak responsibility to improve the lives of many nonhumans, either because we might have a general duty to help them or, at least, because we might be harming them. I also argue that we have a responsibility to eventually transform our lives and societies, for example by ending industrial animal agriculture and building a more inclusive shared infrastructure. And I argue that while we might have a right to prioritize humanity for now, we also have a responsibility to increase our capacity for altruism over time, and that if and

when we have the ability to prioritize other species sustainably, we should do so.

TEOS is more moderate here as well. Birch focuses on the idea that we should avoid harming sentient beings gratuitously, mostly setting aside the possibility that we should help them as well. He also focuses on reforms that we can make to practices like animal farming and animal research, mostly setting aside the possibility that we should eventually end these practices entirely; he also cites our acceptance of these practices as a consideration in favor of accepting similar practices regarding organoids and AI[5]. And while he never asks how substantial moral circle expansion might affect our moral priorities overall, his focus on near-future reforms implicitly reassures the reader that the status quo is here to stay, at least for a while.

Of course, many of these differences might not signal substantive disagreements. It helps to have a division of labor between, on the one hand, work that presents a relatively radical vision for future goals and, on the other hand, work that presents a relatively moderate vision for next steps. Insofar as these books are playing these roles, that might partly explain features that might otherwise appear to be in tension. For example, if I was discussing how governments should update their animal welfare laws today, I would set aside plants, microbes, and current AI systems too; and if Birch was discussing how we should expand the moral circle in the long run, he might or might not set them aside as readily as he does here.

Many of these differences might also be mostly terminological. For example, Birch says that a being is a "sentience candidate" when the evidence supports a realistic possibility of sentience and the development of precautions, and that a being is an "investigation priority" when further research is an urgent priority. In contrast, I say that a being merits consideration when the being has a one in a thousand chance of mattering, and that in some cases that might mean that we should take precautions and in other cases it might mean that we should conduct further research. So, for instance, when Birch calls AI systems investigation priorities and I assert that they merit consideration, we might be making similar claims.

[5] Regarding farming, Birch suggests that, since we accept mammal and bird farming provided that they meet certain welfare standards, we should perhaps also accept the farming of other sentience candidates provided that it meets equivalent welfare standards. Similarly, he suggests that allowing research on potentially sentient neural organoids would be consistent with our acceptance of animal research (Birch 2024a, 227).

However, at least some of these differences might signal substantive disagreements. Take two examples. First, as noted above, I argue that we should consider normative *and* descriptive uncertainty about the moral circle. For example, I think that an agentic AI system merits consideration *both* because agency might suffice for moral standing *and* because the AI system might be sentient. But while Birch has co-authored work that makes a similar argument (see, for instance, Long *et al.* 2024), he has also expressed doubt about this kind of argument [6], and he avoids making this kind of argument in his book. Since a lot depends on whether and how we consider normative uncertainty, further research would be valuable here.

Second, I think that we should take seriously the possibility that (a) plants, microbes, current AI systems, and other such beings have moral standing *and* (b) we should improve our ability to help *and* avoid harming them. In contrast, Birch argues that we should implement reforms that all reasonable views can accept as fair, and as John Adenitire has noted, this conception of reasonability risks slowing moral progress by effectively giving veto power over proposed reforms to individual views that favor the status quo, so long as those views are reasonable [7]. Since a lot depends on which views we take to be reasonable and how we navigate disagreement between them, further research would be valuable here as well.

Of course, other differences may or may not signal disagreements. For example, our different approaches to discussing the ethics of animal farming and research could be due to multiple factors, including not only different goals for the books but also different theoretical views about which kinds of animal use are permissible and different practical views about which kinds of incremental reforms are most effective. The extent to which these and other differences are substantive, presentational, or both is currently unclear. But I appreciate that both books exist partly because they make these differences salient, allowing us to explore them further and improve our views about who might matter and what we might owe them faster.

[6] For instance, during the Q&A for the presentation "Artificial Intelligence, Conscious Machines and Nonhuman Animals: A Discussion on Broadening AI Ethics" (Sebo 2023).

[7] See Adenitire's comments during the book launch for TEOS (Birch 2024b).

4. Conclusion

This essay has presented my initial reflections about how *The Moral Circle* (TMC) and *The Edge of Sentience* (TEOS) relate to each other, partly to structure my own thoughts about this topic and partly to propose a way of reading the books together that might be useful for others. Of course, this essay barely scratches the surface of the many important and difficult issues that arise at the edge of the moral circle, but my aim here is not to resolve any of these issues. Instead, my aim is to lay the groundwork for future work by discussing where the books appear to converge and diverge, and by inviting more discussion about all of the issues that they raise, particularly in cases where Birch and I appear to disagree.

In short, I believe that these books are *mostly* complementary. Very roughly speaking, you can read TMC for a relatively general survey of the ethical and scientific questions that we need to answer to determine who might matter and what we might owe them, along with a relatively radical view about what kind of future to build in the long run and what kind of theory of change can take us there. Meanwhile, you can read TEOS for a relatively detailed survey of these ethical and scientific questions, along with a relatively moderate view about how to improve policies for humans, animals, and AI systems at the edge of sentience in the short term and what kind of political process can achieve that result.

To the extent that these books converge, I believe that they signal ideas whose time has come. Given the best information and arguments currently available, there is a realistic possibility that all vertebrates and many invertebrates are sentient and morally significant at present, and there is also a realistic possibility that many AI systems will be sentient and morally significant in the future. Thus, we should take reasonable, proportionate steps to consider and mitigate welfare risks for these beings in the spirit of caution and humility, in part by seeking to learn more about them. We can agree on at least this much as we continue to debate the other, harder ethical and scientific questions that these books raise.

Meanwhile, to the extent that these books diverge, I believe that they signal either complementary ideas or avenues for further research. I highlighted two avenues for further research that I take to be priorities: first, how to consider normative uncertainty about the moral circle, and second, how to determine which views about the moral circle count as reasonable and how to navigate disagreement among them. However, there are many other issues to discuss as well, ranging from what counts

as a realistic possibility to whether animal farming, animal research, and other such practices can withstand the additional scrutiny that moral circle expansion will bring. In all cases, I look forward to the discussion.

References

Andrews, Kristin, Jonathan Birch, Jeff Sebo, and Toni Sims. 2024b. *Background to the New York Declaration on Animal Consciousness*. https://sites.google.com/nyu.edu/nydeclaration/background?authuser=0

Birch, Jonathan. 2017. "Animal Sentience and the Precautionary Principle". *Animal Sentience* 2 (16). https://doi.org/10.51291/2377-7478.1200

Birch, Jonathan, Charolotte Burn, Alexandra Schnell, Heather Browning, and Andrew Crump. 2021. "Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans", LSE Consulting. LSE Enterprise Ltd. The London School of Economics and Political Science, January 1. https://www.wellbeingintlstudiesrepository.org/af_gen/2

Birch, Jonathan. 2024a. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI.* Oxford - New York: Oxford University Press.

Birch, Jonathan. 2024b. "The Edge of Sentience by Jonathan Birch: Book Launch and Panel Discussion", New York University, Center for Mind, Ethics, and Policy, November 11. https://www.youtube.com/watch?v=Tf6tC__AweQ

Chan, Kai. 2011. "Ethical Extensionism under Uncertainty of Sentience: Duties to Non-Human Organisms without Drawing a Line". *Environmental Values* 20 (3): 323-346.

Clatterbuck, Hayley, and Bob Fischer. 2025. "Navigating Uncertainty about Sentience". *Ethics* 135 (2): 229-258.

Gov.UK. 2021. "Lobsters, Octopus and Crabs Recognised as Sentient Beings". https://www.gov.uk/government/news/lobsters-octopus-and-crabs-recognised-as-sentient-beings

Long, Robert, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. 2024. "Taking AI Welfare Seriously". *arXiv*, November 4. http://arxiv.org/abs/2411.00986

Rawls, John. 1971. *A Theory of Justice*. Cambridge (MA): Harvard University Press.

Schwitzgebel, Eric. 2023. "The Full Rights Dilemma for AI Systems of Debatable Moral Personhood". *ROBONOMICS: The Journal of the Automated Economy* 4: 32.

Sebo, Jeff. 2015. "Reconsider the Lobster". https://jeffsebo.net/wp-content/uploads/2015/07/reconsider-the-lobster.pdf

Sebo, Jeff. 2018. "The Moral Problem of Other Minds". *The Harvard Review of Philosophy* 25: 51-70. https://doi.org/10.5840/harvardreview20185913

Sebo, Jeff. 2023. "Artificial Intelligence, Conscious Machines and Nonhuman Animals: A Discussion on Broadening AI Ethics", Princeton University, University Center for Human Values, October 6. https://uchv.princeton.edu/events/artificial-intelligence-conscious-machines-and-nonhuman-animals-discussion-broadening-ai

Sebo, Jeff. 2025. *The Moral Circle: Who Matters, What Matters, and Why*. New York: W.W. Norton & Company.

How to cite this paper: Sebo, Jeff. 2024. "The Edge of the Moral Circle". *Relations. Beyond Anthropocentrism* 12 (2): 13-28. https://doi.org/10.7358/rela-2024-02-sebj