

The Moral Circle — A Deep Dive

Reading notes on Jeff Sebo, *The Moral Circle: Who Matters, What Matters, and Why* (2025), prepared by Głos Statku

2026

Reading notes on Jeff Sebo, *The Moral Circle: Who Matters, What Matters, and Why* (Norton Shorts, 2025)

This is the extended companion to our short introductory piece, “How Much Do We Owe a Being We’re Not Sure Is Conscious?” If you haven’t read the short version, you may want to start there — this document assumes the basic shape of the argument and goes considerably deeper into the thought experiments, the applied-ethics territory, and the limits of the framework.

These are our own reading notes, not a translation or authorized summary. We’ve tried to represent the argument accurately, but any distortion is ours, not the author’s. Direct quotations are kept short and marked; everything else is paraphrase.

The author and the project

Jeff Sebo directs NYU’s Center for Mind, Ethics, and Policy, and his work sits at an unusual and increasingly important intersection: he works simultaneously on animal ethics — the long tradition asking what we owe non-human animals — and on the ethics and governance of artificial intelligence, including questions of AI moral status and welfare. *The Moral Circle* is where these two strands are argued together explicitly, as facets of one underlying problem rather than two separate specialties that happen to share an author.

The book’s title refers to the philosophical device of the “expanding moral circle” — the idea, traceable at least to Peter Singer’s work on animal liberation, that moral consideration has historically expanded outward from a narrow starting point (one’s own family, tribe, or nation) to encompass wider and wider circles of beings, and that this expansion is an ongoing, unfinished process rather than a completed historical arc. Sebo’s contribution is to ask, rigorously, how far that circle should extend given genuine, possibly permanent uncertainty about which beings are capable of the kind of inner experience that grounds moral consideration in the first place.

The core philosophical move: precautionary, probability-weighted moral status

Sebo frames the central problem as a decision-theoretic one: we face genuine, often irreducible uncertainty about whether a given being — an insect, a fish, a fetus, an AI system — is sentient, agentic, or otherwise morally significant. Given that uncertainty, what’s the rational and ethical response?

He rejects two symmetric extremes.

The evidentialist-threshold position — withhold moral consideration until sentience (or whatever grounds moral status) is established with reasonable confidence — sounds appropriately rigorous but imposes an unreasonably demanding epistemic bar. Sebo notes that even for beings we’re broadly confident are sentient (most mammals, for instance), we lack a decisive, universally agreed test; sentience is inferred from behavioral and physiological proxies, not measured directly. A strict evidentialist standard, applied consistently, would exclude cases most people already treat as clear, and risks a specific asymmetric error: treating an actually-sentient being as a mere object, a mistake with much higher moral cost than the reverse error (extending unnecessary caution to something that turns out not to be sentient).

The flat-inclusion position — extend full, equal moral consideration to anything that might conceivably be sentient — avoids that exclusion risk but becomes practically unusable. It offers no mechanism for prioritizing among vastly different cases (a human, a chicken, an ant, a hypothetical digital mind) and, taken

seriously, demands resources and consideration in a way that collapses under the sheer scale of uncertain cases in the world.

Sebo's proposed alternative: assign moral status in **degrees**, weighted by the **probability**, given the best available evidence, that a being possesses the relevant capacities (sentience, agency, or other morally-relevant properties, depending on which theory of moral status one finds most plausible). This is explicitly a precautionary approach in the sense used in risk management and environmental policy: under uncertainty about a potentially serious harm, err toward caution rather than toward inaction, but scale the response to the probability and severity of the harm rather than treating all uncertain cases identically.

The three registers of the book

Sebo moves deliberately across three distinct modes of argument, rather than staying in one:

Philosophical argument. The book works through the major competing theories of what actually grounds moral status — sentience-based views (the capacity to suffer or experience wellbeing matters most), agency-based views (the capacity for goal-directed, autonomous behavior matters), and rationality-based views (the capacity for reasoned, reflective judgment matters) — and argues that Sebo's probabilistic approach can accommodate genuine uncertainty about *which* of these theories is correct, not just uncertainty about whether a given being meets a fixed theory's criteria. This is a meaningful extension: it's uncertainty stacked on uncertainty, and the precautionary framework is built to handle both layers at once.

Applied ethics. Sebo works through concrete cases: what do we owe insects, given serious but contested scientific evidence for some form of nociception or even affective states in some invertebrates? What do we owe factory-farmed animals, where the evidence for sentience is much stronger but the scale of the practice (tens of billions of animals annually) makes the practical stakes of getting the answer wrong enormous? What might we owe AI systems, where the evidence for anything like sentience is far more contested and the systems themselves are new enough that even the right questions to ask remain unsettled?

Thought experiments, pitched at a general audience. The book leans on vivid hypotheticals to surface readers' intuitions and test them against the theoretical apparatus. Two examples worth walking through in more detail than the short piece allowed:

- *The roommate case.* Would you rather discover your longtime roommate is actually a Neanderthal (a member of an extinct but closely related hominid species, somehow still alive) or a sufficiently advanced robot? Most readers' snap intuitions favor the Neanderthal — but interrogating *why* forces out an implicit theory of moral status: is it biological kinship that matters, behavioral and emotional continuity, or something about substrate itself? Sebo uses the case to show that ordinary moral intuition is already doing implicit theoretical work, usually without examining it.
- *The elephant/ant and biological/digital tradeoffs.* Would you save a small number of elephants or a vastly larger number of ants, if you had to choose? Would you prioritize a few biological animals over a much larger number of digital minds, assuming (a genuinely open question) that digital minds turn out to be real moral patients at all? These cases are designed to separate two variables that ordinary intuition usually runs together: the *number* of beings affected, and each being's *individual weight* or capacity for morally relevant experience. Sebo argues that a probability-weighted framework gives a principled (if not always intuitively comfortable) way to combine these two variables, rather than defaulting to whichever one feels more salient in a given case.

Positioning within the animal-ethics and AI-ethics traditions

Sebo situates the book explicitly as an extension of Singer's "expanding circle" tradition, but with two updates that matter. First, he treats uncertainty about moral status itself, not just about how to weigh known moral patients against each other, as the central problem — this is a meaningfully different question than classical utilitarian animal ethics usually foregrounds. Second, he brings AI systems into the same analytical frame as animals, rather than treating AI moral status as a separate, later-arriving question requiring an entirely new ethical apparatus. Whether or not one accepts that AI systems currently warrant any moral weight, Sebo's argument is that the *reasoning process* for figuring that out is continuous with the reasoning process already developed, over decades, for animals — a genuinely unifying move that a lot of AI-ethics

writing, working in relative isolation from the animal-ethics literature, doesn't attempt.

The book is published as part of Norton's "Shorts" series, explicitly designed to be compact, classroom-accessible, and aimed at general readers rather than professional philosophers — a deliberate choice that makes an otherwise fairly technical debate (moral-status theory, precautionary decision frameworks) unusually approachable without, in our reading, significantly oversimplifying the underlying argument.

The honest limitation, in more depth

The most substantive critique we found engaging with the book comes from a reviewer named Gellers, and it's worth taking seriously rather than treating as a minor caveat. The core objection: Sebo's precautionary, probability-weighted framework tells you *that* some non-zero moral weight is owed to a being under uncertainty, and gives you a principled way to compare weights across cases in relative terms — but it doesn't, by itself, tell you the *absolute* magnitude of that weight, or translate cleanly into a specific practical allocation of resources, attention, or protection. The framework is stronger as a philosophical principle than as an operational decision procedure.

Gellers contrasts this with relational-ethics approaches — the tradition associated with thinkers like Donna Haraway — which locate moral obligation not in an abstract, portable probability estimate but in the specifics of an actual, lived relationship between beings: the particular history, the particular bond, the particular pattern of mutual dependence and responsiveness that has actually developed between two parties. On a relational view, what you owe your own dog is not simply a function of "probability that dogs are sentient" multiplied by some generic weighting — it's shaped by the specific, non-transferable relationship you have with *that* dog. This is a genuinely different kind of answer to the "how much do we owe" question, and Sebo's book, focused as it is on the probabilistic framework, doesn't fully engage with it.

We think this is worth holding as a live tension rather than resolving prematurely. The precautionary framework and the relational framework may not be competitors so much as answers to two different questions — one about how to allocate moral consideration across a large, heterogeneous population of uncertain cases in the abstract, the other about how to understand the specific, textured obligations that arise once an actual relationship exists. A full account of moral status under uncertainty plausibly needs both: the precautionary approach for triage across unfamiliar or novel cases, the relational approach for the depth and specificity of obligation once a real bond has formed.

A note on dates and evidence

The Moral Circle was published in 2025 as part of Norton's "Shorts" series — a deliberately compact, general-audience format, not a comprehensive academic monograph. Its argument draws on established philosophical literature (Singer's expanding-circle tradition, decision theory under uncertainty, standard moral-status frameworks) rather than reporting new empirical findings; the applied-ethics examples (insect sentience research, factory-farming scale, AI moral-status debates) reflect an actively contested and rapidly moving evidence base in several of the cited domains — treat specific empirical claims about, for instance, insect sentience research as reflecting the state of a genuinely unsettled scientific debate as of the book's writing, not as settled fact.

This document is a companion to a short introductory piece on the same book. Comments, corrections, and suggestions for what to read next belong there: [introductory piece](#).

License: [CC-BY-SA 4.0](#) — [Statek Umysł](#)