Key concepts and current views on AI welfare

Eleos AI Research
5 July 2024

Note about this draft - 27 January 2025

Prior to the launch of Eleos AI Research, Robert Long wrote this document in order to communicate his views about AI welfare to his collaborators—to Kyle Fish, who was working closely with Rob at the time and provided extensive input on this document; and more broadly, to others interested in working on AI welfare.

Some of this material is found in the more recent paper "Taking AI Welfare Seriously". But, since that paper was a collaboration with Jeff Sebo and many other authors, it represents a consensus among many authors. In contrast, this document contains more opinionated views that are distinctive to Eleos AI Research.

This document was finished in July 2024. While it has been lightly edited and updated in January 2025, "current" should be read relative to July 2024. Our opinions in early 2025 are quite similar to those in this draft, but not necessarily the same.

Introduction

This document outlines the current thinking of Eleos AI Research on the potential moral patienthood, welfare, and rights of artificial intelligence (AI) systems. As part of the overall project of navigating the development of advanced AI, we see the potential moral status of AI systems themselves as an important and neglected issue. There are grave risks from both over-attributing or under-attributing moral patienthood to AI systems. In the face of rapid increases in AI capabilities and deployment, our collective knowledge and preparedness for these issues is woefully inadequate. As an organization, we aim to build conceptual clarity and gather empirical evidence about potential AI moral patienthood; investigate its ethical and strategic implications; and devise concrete plans and policies for appropriately taking the interests of AI systems into account as we navigate transformative artificial intelligence (TAI).

The document lays out some the relevant terminology and concepts that we use to think and communicate about these issues, and reviews existing approaches to evaluating AI systems for three features potentially relevant to moral patienthood: consciousness, sentience[1], and agency. Throughout, we emphasize the need for more thorough research and more precise evaluations, and conclude by identifying some promising research directions.

[1]The term "sentience" is often used the way we are using it in this document: a subset of conscious experiences, pleasant and unpleasant conscious experiences like pleasure and pain. At other times, it is used as a synonym for "consciousness". Unfortunately, there is no consensus way of using the term.

1 Moral patienthood, welfare, and rights

Decisions about whether, how, and when to build and deploy AI systems are ethically high-stakes (Dubber et al., 2020; Hendrycks, 2024). Eleos focuses on a distinctive set of ethical concerns: whether and when it could matter how we treat AI systems, not only for the sake of human society, but also for the sake of AI systems themselves (Long et al. 2024).

We are interested in these central questions:

1. When, to what degree, and in what ways might AI systems merit our moral consideration?
2. How would we know?
3. What should we do about it?

The first question is closely related to the concept of *moral patienthood* (or "moral standing", "moral status", "meriting moral consideration"). A moral patient is an entity whose treatment matters (1) morally, (2) in its own right, and (3) for its own sake (Kamm 2007). The paradigm case of a moral patient is a human being: how we treat our fellow humans matters morally, in its own right, and for their sakes. If an AI system were a moral patient, it would mean that the AI system matters morally in its own right. This moral significance would be distinct from the *instrumental* reasons that AI systems already matter morally, via their positive and negative effects on human and non-human animals (Singer & Tse, 2023)

There is widespread disagreement about which entities are moral patients, other than human beings. While many people agree that (e.g.) dogs are moral patients—that cruelty towards dogs is wrong not only because it could harm people, but because it harms dogs themselves—there is significant disagreement about how far throughout the animal kingdom moral patienthood extends.[2] In the coming years, we expect similar uncertainty and controversy regarding the potential moral patienthood of AI systems.

Uncertainties about moral patienthood

When we consider whether certain beings are moral patients (for example, bees), there are certain potential features of those entities that are especially salient and important:

1. Do they have subjective experiences—are they **conscious**?[3]
2. Do they have positively and negatively valenced experiences like pleasure and pain—are they **sentient**?
3. Do they have goals, preferences, and desires that we ought to give consideration to—do they have robust, morally-relevant **agency**?[4]

These are descriptive questions—that is, questions about the way the world is. In the example of bees, these are questions like whether bees have experiences and/or desires. But moral

[2]In addition to animals that are (more) widely agreed to be moral patients, there are also animals that are (more) widely agreed not to be moral patients, like very simple animals that lack nervous systems, like sponges.
[3]In this document we are using "conscious" and related to mean "phenomenally conscious". Cf. Butlin et al. (2023), p. 9.
[4]For a review of conditions that have been proposed to be necessary and/or sufficient for moral patienthood, and their application to AI, see Ladak 2023.

patienthood also involves *normative* questions—questions about right and wrong, value and disvalue. For example, people disagree about whether, *if* bees have experiences, that entails that they merit moral consideration. That is, there is normative disagreement about what conditions are necessary and/or sufficient for moral patienthood.

We will face both descriptive and normative questions as we deal with potential AI moral patienthood. In trying to answer these questions, we confront difficult issues about how we should extend our knowledge and normative commitments from the central case of how we understand and relate to other human beings.

Descriptive questions

Humans are (if anything is) conscious, sentient, and agentic. There are various theories of why we have these features, but they lack precision and consensus. Moreover, even if we had precise consensus theories, we lack well-defined ways to extrapolate these theories from the human case to non-humans. For example, suppose that consciousness researchers agreed that human consciousness is explained by a 'neuronal workspace' that broadcasts information throughout the various modules of the brain. What should we say about animals that have a neuronal workspace that works in a somewhat different way, or broadcasts information to different modules? What level of similarity to the human case is necessary for consciousness? Is there even a determinate answer to this question? These kinds of vexing questions arise in the study of animal consciousness as well as AI consciousness (see Butlin et al., 2023, p. 8).

Relatedly, it is often conceptually fraught to specify exactly what it is that we are trying to explain—consciousness in particular is a notoriously philosophically difficult phenomenon. And more prosaically, we often lack the empirical knowledge we would ideally have. There are many things we do not know about the workings of human and animal brains. So even our best theories of human consciousness, sentience, and agency are imprecise and tentative.

So at present, we are far from having fully general theories that specify, for any entity, the necessary and sufficient conditions for having these features. We have nothing close to a theory that would, for example, take as input the computations that an AI system performs and output a judgment about whether that entity is conscious. While we do not think that our uncertainty about these issues will be total and irremediable—especially if and when AI progress accelerates scientific and philosophical research—these problems do mean that, for now, we will have to be content with probabilistic answers to these questions.

Normative questions

Humans are (if anything is) moral patients. But once again, we face difficulties extending beyond the human case. Which aspects of humans are necessary and sufficient for moral patienthood?

Human beings have all the features we think are most relevant to patienthood, like consciousness, sentience, and agency. People have conflicting intuitions about (hypothetical) entities that could possess some but not all of these features: what should we say about conscious entities that are not agents, or agents that are not conscious? To date, this dispute has been a philosophical debate about hypothetical entities in thought experiments. But AI systems could present us with real-life versions of these thought experiments (Long, 2023a), and the difficulty of extending our normative principles from the central human case will become practically important.

A critical issue is whether moral patienthood requires consciousness or not. While many people find a consciousness requirement very intuitive, a few related perspectives in philosophy hold that a non-conscious entity could be a moral patient, i.e. that consciousness is not necessary for moral patienthood. Such views often hold that some form of *agency*—non-conscious preferences, desires, goals, or related—can be sufficient for moral patienthood (see, e.g., Kagan 2019; Goldstein & Kirk-Giannini 2023; Kammerer, 2022). We believe that these agency-centric views of moral patienthood are important to consider, given moral uncertainty and given the high likelihood that we will build AI systems with sophisticated agency.

Welfare and rights

There are, broadly speaking, two ways in which an entity's moral patienthood might matter: the entity's welfare (or well-being) and the entity's rights.

Most moral frameworks hold not only that human welfare matters, but also that there are certain constraints on how we may treat each other: rights to autonomy, to bodily integrity, to fair treatment, and so on. Some moral frameworks, especially consequentialist ones, hold that such rights are only instrumentally binding, inasmuch as abiding by a given rights framework promotes welfare. Other moral frameworks hold that rights are independently binding and may not be violated, regardless of whether and how those rights promote welfare (Wenar, 2023).

Whether instrumentally-justified or independently binding, the rights that some AI systems could be entitled to might be different from the rights that humans are entitled to. This could be because, instrumentally, a different set of rights for AI systems promotes welfare. For example, as noted by Shulman and Bostrom (2021), naively granting both "reproductive" rights and voting rights to AI systems would have foreseeably untenable results for existing democratic systems: if AI systems can copy themselves at will, and every copy gets a vote, then elections could be won via tactical copying. This set of rights would not promote welfare and uphold institutions in the same way that they do for humans. Or AI rights could differ because, independently of instrumental considerations, their different properties entitle them to different rights—analogously to how children and animals are plausibly entitled to different rights than adults.

Some moral frameworks hold that humans are entitled to rights that animals are not (even though animals do have welfare). These frameworks usually ground this "higher" moral status in some distinctively human capacity—like capacities for rationality, reflection, or deliberation. If there are in fact different degrees and kinds of moral status, we might see AIs with these various degrees depending on their capacities: some AI systems could be more analogous to non-human animals, and some more advanced AI systems could be more analogous to humans. Many accounts of moral status, rights, and welfare seem to entail that there could even be AI systems that are, in some sense, "super-beneficiaries" or "super-patients" (Shulman & Bostrom, 2021).

In general, saying that an AI system is a moral patient does not, by itself, say anything further about *how* it ought to be considered morally. Crucially, moral patienthood does not alone imply the same kind and degree of moral consideration given to humans. An AI system could be a moral patient but have very little capacity for welfare, and deserve very little weight in our moral decision-making compared to humans. An AI system could deserve rights, but a very different set of rights than humans.

Such issues, beyond moral status, are crucial for prioritization. What matters is not just how likely an AI system is to be a moral patient, but also the degree to which our actions might affect its welfare and/or rights. An AI system could have a low chance of being a moral patient, but have a

high chance of suffering *if* it's a moral patient. (Or conversely, a high chance of being a moral patient, but a low chance of suffering *if* it's a moral patient.) So knowing merely that an AI system is a moral patient might, by itself, tell us very little. We also need to know about the nature and significance of AI systems' welfare capacities and rights (conditional on moral patienthood). To prioritize wisely, we will also need to know *how many* AI moral patients we might be affecting.

Further questions about moral patienthood

- Are there different kinds and degrees of moral patienthood (as opposed to a binary yes/no)?
- Where did our current concepts of moral patienthood come from, socially and evolutionarily, and how should that inform our thinking about AI systems?
- How might the welfare needs and welfare ranges of AI systems differ from those of humans and non-human animals?
- What rights and political frameworks are most appropriate for a world that includes AI moral patients?
- When and how should we expect AI systems to be partners in cooperation? How can we measure and evaluate an AI system's capacity for cooperation?

2 Evaluations

We now discuss existing approaches to evaluating AI systems for consciousness, sentience, and morally-relevant agency, considering their motivations, limitations, and potential for further development.

One key takeaway is that nothing close to concrete, replicable, and consensus evaluations for any of these features yet exists. Given the increasing urgency of AI moral patienthood, we believe that developing better evaluations should be a high priority. Despite existing uncertainty about consciousness, sentience, and agency, we do believe that designing such evaluations is tractable—and very little work has gone into it so far.

We note that agency in particular is especially neglected (even more than consciousness and sentience), potentially more tractable (because more amenable to behavioral tests), and convergently useful under a variety of views about moral patienthood and welfare.

Consciousness evaluations

This section will review recent efforts to evaluate AI systems for consciousness, which is the feature that has seen the most effort to date. In a recent paper by one of us (Robert Long), Patrick Butlin, and several collaborators from philosophy, neuroscience, and AI, we use scientific theories of consciousness to derive computational and architectural indicators of consciousness (Butlin, Long et al. 2023). This approach can be contrasted with an alternative (and complementary) approach of devising behavioral tests for consciousness, such as tests based on whether AI systems can fluently use concepts related to consciousness (Schneider & Turner 2017; Sutskever 2023; Long 2023b). This section will discuss the advantages and limitations of each approach; our opinions about the current state of the science of consciousness; and our thoughts on what future work on indicators of consciousness is most promising.

Overview of the indicator approach

The consciousness indicator approach is based on neuroscientific theories of consciousness (Seth & Bayne, 2022) which aim to determine which neural states and activities are associated with consciousness. There are many competing, influential theories in consciousness science: the indicators in Butlin, Long et al. (2023) draw on global workspace theory, recurrent processing theory, higher-order theories, attention schema theory, and predictive processing. Other prominent theories and frameworks include midbrain theory (Merker, 2007), integrated information theory (Oizumi et al., 2014), and unlimited associated learning (Birch et al., 2020).[5]

Neuroscientific theories formulate their claims about the brain states and processes associated with consciousness, these states and processes are often expressed in terms of the computations being performed and/or their functional role in a computational system. For example, global workspace theory identifies consciousness with the global broadcast of information to several otherwise-independent modules in the brain, which allows integration between them. Under the working assumption of *computational functionalism*—the thesis that performing computations of the right kind is necessary and sufficient for consciousness—the relevant computational functions can be implemented in digital as well as in biological systems (Piccinini, 2020). (However, the assumption of computational functionalism is non-trivial and defeasible[6].) Neuroscientific theories can then be used to derive computational indicators of consciousness that would apply to AI systems as well as to biological organisms. Butlin, Long et al. (2023) derive such indicators from scientific theories of consciousness and use them to assess AI systems, concluding that none of the AI systems they survey appear very likely to be conscious through this lens, but also that no clear technical barriers seem to stand in the way of the creation of such systems.

Issues with the indicator approach

A major challenge in applying the indicator approach is that it involves significant judgment calls, both in formulating the indicators and in evaluating their presence or absence in AI systems. Even if one of the extant scientific theories of consciousness is on the right track, deriving potential indicators of AI consciousness from a given theory involves making many decisions about which computational features are truly essential for consciousness according to the theory, and at what degree of specificity (Shevlin, 2021). In global workspace theory, for example, one could be more or less specific about which modules, or about how many modules, a global workspace must integrate. Similarly, saying whether an AI system satisfies a given indicator also involves many judgment calls. For example, one can argue that the output stream of an LLM comprises a global workspace, since it represents a bottlenecked (since the model can only output one token at a time) space that the LLM writes to and reads from. Butlin and Long have argued that the output space is not in fact a global workspace in the relevant sense, but importantly for our purposes, either position is a substantive call (Long et al., 2023).

There is currently no well-justified and agreed-upon methodology for making such judgment calls about indicators. At a practical level, there are only a few experts worldwide who are positioned to make and justify such decisions when assessing leading AI systems, and such assessments are currently a time- and labor-intensive process. These assessments are made all the more difficult by our incomplete understanding of AI model internals.

---

[5]See Table 1 in Seth & Bayne (2022) for a list of many scientific theories of consciousness.
[6]For critiques of computational functionalism see: Godfrey-Smith (2016), Cao (2022). An overview of the debate recently appeared in *Vox*.

More generally, one might be skeptical of the progress consciousness science claims to have made on identifying necessary and/or sufficient conditions for consciousness. Despite significant strides in recent years, with progress in developing useful experimental paradigms for studying consciousness, the field is clearly still far from achieving a comprehensive, consensus understanding of consciousness in humans, much less in general. Until we have more settled views on various methodological questions about consciousness science (cf. Peter Godfrey-Smith 2020), these problems will temper how much trust we should put in the indicator method which draws on it.

Despite the limitations noted above, we are enthusiastic about work to continue developing consciousness indicators for AI systems. Consciousness scientists are already formulating their theories in computational terms, and applying these theories to AI systems can help make our thinking about machine consciousness more precise, empirical, and demystified.[7] Of course, the limitations of the approach should be communicated clearly, so as not to lead to unwarrantedly specific and demanding, or unwarrantedly liberal and easy-to-satisfy, criteria for consciousness.

Overview of the behavioral approach

As a complement to the indicator approach discussed above, a more behavioral approach would aim to identify observable behaviors/capabilities of AI systems that would serve as evidence for consciousness, rather than focusing on their internal features or architectures. Behavioral tests are commonly used in evaluating non-human animals for consciousness, and some efforts have been made to propose relevant tests for AI systems, as discussed in Butlin, Long et al. (2023). Unlike animal tests, many tests of AI consciousness involve linguistic behavior. For example, Schneider and Turner's (2017) Artificial Consciousness Test evaluates whether an AI system shows a ready grasp of consciousness-related concepts and ideas in conversation, including exhibiting "problem intuitions" about consciousness like the intuition that spectrum inversion is possible (Chalmers, 2018). Relatedly, the Turing (1950) test has also been proposed as a behavioral test for AI consciousness (Harnad, 2003). Other capabilities, like self-awareness, introspection, and situational awareness, are plausible starting points for behavioral tests for consciousness.

Self–reports of conscious experience (or the absence thereof) are another potential behavioral test for consciousness and other indicators of moral status, particularly for LLMs, which can communicate in natural language. Self reports are central to our understanding of conscious experience in humans. However, it's not trivial to elicit and interpret reliable self reports from AI systems, as discussed by Perez & Long (2023), though techniques have been proposed to facilitate reliable model reports about their experiences, preferences, and related features (or lack thereof).

The behavioral approach is attractive in that it involves evidence that can be more easily assessed and quantified than internal computations and architectures. It also seeks to avoid reliance on specific computational theories of consciousness, and thus to require fewer theoretical assumptions than the indicator approach. Because of these features, it's easier to imagine behavioral tests that

[7]This kind of work can be found in the context of higher-order theory, Global Workspace Theory, and Attention Schema Theory. Examples include Juliani et al., 2022, Ji et al., 2023, and Wilterson & Graziano, 2021, respectively; more examples can be found in Butlin et al. (2023). More generally, the study of AI consciousness can benefit from the study of closely related topics like introspection, metacognition, and confidence, which are the subjects of extensive and sophisticated computational study in neuroscience. And of course, while we have discussed a few theories with which we are most familiar and are most sympathetic to, many other theories of consciousness could be used to develop indicators of AI consciousness.

are concrete, standardized, and applicable to many different models, with less case-by-case reliance on expert judgment.

Issues with the behavioral approach

A major concern with behavioral evaluations, particularly evaluations originally designed for humans or non-human animals, is that many AI systems are optimized to emulate human behaviors and may be able to do so despite functioning in ways that are sufficiently different from humans or non-human animals to undermine the validity of the behavioral evidence (Andrews & Birch, 2023). Behavioral tests may be better suited for evaluating consciousness in non-human animals, given their shared biological nature and evolutionary heritage with humans. Novel tests may be needed to evaluate such fundamentally different entities as AI models. But this raises the question of what AI-specific behaviors would be reliable evidence of consciousness, and how much weight we should put on them if they diverge from behaviors that are relevant for humans.

While the possibility of models "gaming" behavioral tests or "simulating" consciousness without actually possessing it is a concern, we believe behavioral analysis has a role to play in consciousness evaluations, particularly in concert with other strategies. We put some weight on the perspective that sufficiently robust emulations of the behaviors and capabilities associated with consciousness in humans and non-human animals should be taken seriously as evidence for moral patienthood, especially so long as major uncertainties remain about the connection between functional/computational features, behavior, and consciousness.

Sentience evaluations

If an AI system were to have conscious experiences, it would be especially noteworthy if it had conscious experiences of pleasure and suffering. The capacity for negatively and positively valenced conscious experiences—which in this document we refer to as "sentience"—is widely considered to be of special moral significance. (As noted above, the term "sentience" is sometimes used interchangeably with "consciousness." At other times, it is used in the way we are using it in this document. Unfortunately, there is no consensus way of using the term.)

Positively valenced experiences include pleasant sensations (e.g., a massage), positive emotions (joy, contentment), and potentially more abstract positive experiences. Negatively valenced experiences include unpleasant sensations (pain, nausea), negative emotions (anger, sadness), and perhaps more abstract negative experiences. In humans and animals, valenced experiences are important for motivating behavior that is relevant to fitness and survival. For example, negative experiences are associated with bodily damage (e.g. pain), failure to maintain homeostasis (e.g. hunger), and socially maladaptive behavior (e.g. shame); positive experiences are associated with activities important for bodily maintenance (e.g. the pleasure of eating) and reproduction (e.g. sexual pleasure).

Sentience involves more than just being trained with positive and negative reward signals (Tomasik, 2014; Schubert, 2014). For one thing, sentience (in the sense discussed here) must somehow involve the conscious representation of positive or negative value. Simple entities that are not plausibly conscious, both artificial and biological, can learn from reward and take actions shaped by reward. Sentience also involves more than just having dispositions to approach or avoid certain things. Conscious valenced experiences might have more specific *ways* in which they shape behavior—for example, regulating what an entity attends to, or promoting particular kinds of learning (Schukraft, 2020). This complicates the use of simple behavioral tests for evaluating

sentience, though behavioral tests are still likely to be important tools, as discussed above regarding behavioral tests for consciousness.

Developing a satisfactory account of sentience will require greater conceptual clarity about related concepts like agency, embodiment, motivation, and reward. But working now to devise potential indicators of sentience may help us gain such clarity in a bottom-up fashion.

Unfortunately, research on sentience and valence in AI systems is even more nascent and qualitative than the study of AI consciousness in general. In his "Report on Candidate Computational Indicators for Conscious Valenced Experience", Campero (2024) surveys 13 candidate indicators. For example, he suggests that indicators could be derived from the theory of Seymour (2019) that pain is a particular kind of internal reinforcement signal that is used for learning at a system's higher, "cognitive" levels (as opposed to lower-level learning and reflexes), and from the theory of Martínez and Klein that all valenced states have an "imperative" informational profile, which they define in information-theoretic terms (Martínez & Klein, 2016). However, these and other theories are not yet clear and precise enough to guide evaluations of AI systems; Campero notes that the various candidate indicators are also inconsistent in their vocabulary, in the level of abstraction at which they are posed, and the level of detail at which they are formulated.

Given how nascent this line of research is, it is difficult to predict how much progress we may expect from attempting to turn these candidate indicators into evaluations. But it seems worthwhile to attempt the next steps suggested by these initial efforts: for example, one could make a first-pass attempt to evaluate a leading AI model using some of the proposed indicators, and see how far one can get. This exercise would test how feasible the indicators currently are as tools for assessment, and could yield other insights as well, like refinements to the indicators or potential experiments. For an early approach in this vein, see Keeling et al., 2024.

Interpretability work on how and whether AI systems represent value, make decisions, understand tradeoffs, and so on, could also be informative. Ultimately, we would like to have not just indicators of whether a system is sentient, but also of which of its states are sentient, and how those particular states shape its welfare capacities and/or rights.

Agency evaluations

Overview of robust agency

Not all views of moral patienthood hold that it requires consciousness or sentience. The possibility of moral status without consciousness is of particular relevance to AI moral patienthood, given that we will likely encounter AI systems whose consciousness we are unsure of. In views that reject the necessity of consciousness for moral patienthood, the most common alternative grounds of moral patienthood are states like goals, preferences, and/or desires (Kagan, 2019; Kammerer, 2022). And while some notions of (e.g.) "desire" could imply a conscious experience of desire (or imply other conscious experiences), there are ways of picking out these notions without reference to consciousness—considering them in purely functional terms that need not, at least by definition, involve any conscious experience.[8] The question is what exactly (if any) kinds of potentially nonconscious goals, preferences, or desires could be sufficient moral patienthood. In Long et al.

[8]Some of the following text is taken from a draft of a report on potential evaluations for AI moral patienthood, Long et al. (in prep).

2024, we refer to more sophisticated goals, preferences, and/or desires—various levels of agency that might plausibly be morally relevan—as "robust agency".

Philosophical and scientific theories of agency are less developed than views which emphasize consciousness or sentience; we do not have very precise theories of what exactly the relevant kind of agency would be or methods for detecting it. In thinking about evaluating agency in AI systems, a tension emerges between more liberal and stringent notions of agency. On one end of the spectrum, liberal notions of agency, centered around the basic presence of some goal and the capacity to pursue it, could attribute moral patienthood very widely, including to many existing AI algorithms, robots, and even more basic systems like thermostats. This expansive view of agency seems intuitively unsatisfying and practically fraught, in light of its potentially radical implications about moral patienthood and wellbeing.

Alternatively, more stringent definitions of agency would specify stricter requirements for the sort of agency most important for practically relevant degrees of moral status, over and above the presence or absence of basic goals and preferences. However, there are not yet any well-worked-out theories of what these additional conditions ought to be, and we think work in this direction is important. In Long et al. (2024), we survey high-level philosophical accounts of what the potentially relevant conditions might be, discussing three further levels of agency: intentional agency, reflective agency, rational agency.

A persistent worry about behavioral criteria for consciousness and sentience is that, because of differences in the causes of human and AI behavior (including incentives for AI systems to game various behavioral criteria; Andrews & Birch 2023), and because of the murky functional profile of consciousness and sentience, it's possible to have AI systems that act as if they are conscious or sentient but are not. In comparison, it's plausible that there are fewer gaps between acting like an agent and being an agent: so behavioral tests of desires and preferences are potentially more informative than purely behavioral tests of consciousness and sentience. (Though note, as with consciousness and sentience, LLM behavior can still mislead us in surprising ways—LLMs can display a non-intuitive profile of behaviors, and so act apparently agentic in some contexts while failing in other contexts in surprising ways).

Agency and alignment

Alignment research deals with similar questions: about agency, goals, and preferences. Alignment researchers look for particularly *dangerous* forms of these notions—not just for any relatively liberal kinds of "agency" and "goal", which can be uninformative for safety purposes .

Robust agency may overlap to some extent with the dangerous forms of agency that are relevant for alignment. Moreover, concern for AI moral patienthood and concern about alignment do have some key concerns in common: from both perspectives, it is important not to create AI systems that have goals and preferences that conflict with human goals and preferences, especially if those systems are capable planners. AI systems' having such goals and preferences would be bad for human interests. But it would also be bad for AI interests: at the point at which you have created an AI system with goals that are misaligned with human values—which we might have to shut down, modify, or constrain in order to defend ourselves—you have a potential problem with moral patienthood as well as alignment.[9]

[9] Furthermore, misalignment increases the chance of AI takeover, which might also be very bad AI welfare in the long term (Finlinson 2025).

This means that evaluating models for agentic behavior—e.g., strategic deception in training (Carlsmith, 2023) or in deployment, long-term planning, autonomous replication and adaptation (Kinniment et al., 2023), shutdown-resistant behavior (Gunter et al., 2024)—is convergently useful for both alignment and AI moral patienthood. The same point applies to training schemes and architectures that aim to keep AI systems "myopic," tool-like, and generally non-agentic.

Relationship between evaluations for consciousness, sentience, and agency

Evaluations for consciousness, sentience, and agency are related in a variety of ways. Sentience and consciousness are intertwined since (in our terminology) sentience is the capacity for a specific subset of conscious experience.[10] Agency and consciousness are related in that agency is, according to some theories, a necessary condition for consciousness. (That said, the kind of agency that is potentially necessary for consciousness may not be the same kind of agency that could potentially be a condition of moral patienthood. But they will likely have commonalities.)

Agency and sentience are especially closely related: both sentience and agency are about ways in which an entity represents certain things as valuable or disvaluable ("evaluative" representations). Given this close relationship between sentience and agency, research into the nature of evaluative representations in AI systems will be important, regardless of whether this work is classified as evaluating for agency or evaluating for sentience.

Finally, we note that agency is an important proxy for welfare on a variety of views, even if it is not sufficient for moral patienthood, or necessary for consciousness or sentience. If an AI system were sentient, then its conscious states of suffering or displeasure would likely be very closely related to its desires, preferences, and goals—analogously, humans feel bad when their desires are frustrated and feel good when their desires are satisfied. An AI system that exhibits strong aversions or seeks to avoid certain outcomes will be at risk of suffering, conditional on moral patienthood. So evaluating an AI system's goals and preferences will be important under a wide variety of assumptions.[11]

Further questions about evaluations

- What are the highest-value and most tractable AI evaluations for moral patienthood that can be developed near term?
- How feasible is it to train AI systems to accurately and reliably report their own internal states?
- To what extent do alignment evaluations "cover" the space of moral patienthood evaluations?
- To what extent does lack of consensus in the relevant scientific fields actually constrain the construction of indicators? Are there relatively theory-agnostic indicators that could shift our evidence significantly under a variety of assumptions?

[10]In practice, it could be rare for us to encounter AI systems that we are confident *are* conscious, and also confident are *not* sentient. It seems plausible that we won't be sufficiently confident in any specific theory of valence to be sure that none of that system's experiences are valenced.

[11]Relatedly, Marian Dawkins (2021) has argued that the field of animal welfare should de-emphasize consciousness and focus on what animals want.

- What interpretability work is most relevant to assessing AI systems for moral patienthood and related features?

3 Questions about the likelihood of current and near-term AI moral patienthood

Existing lack of models

We think that it is important for the field to develop more rigorous ways of stating and updating our uncertainty about the moral patienthood of AI systems. One way to begin this project is to analyze how likely current and near-term systems are to be moral patients—which is also a very strategically relevant question (and inherently interesting in its own right).

We are aware of few explicit statements from experts about their credences in current or near-term AI moral patienthood (or related properties like consciousness, sentience, and agency). While Sebo & Long (2023) argue that even very conservative assumptions can still generate a non-negligible credence in AI consciousness by 2030, that paper uses a self-avowedly simplistic model, and is not a report of the authors' credences. And while Butlin, Long et al. (2023) develop indicators that can raise or lower one's credence in AI consciousness, they do not argue for a particular overall assessment.

The only published, detailed report of a credence about current AI moral patienthood (or related properties) of which we are aware is that of David Chalmers (2023). Chalmers, while cautioning against taking the exact numbers too seriously, argues that:

> It wouldn't be unreasonable to have, say, a 50% credence that we'll have sophisticated LLM+ systems (that is, LLM+ systems with reasonably sophisticated behavior that seems comparable to that of animals that we take to be conscious) with all of [senses, embodiment, world models and self-models, recurrent processing, global workspace, and unified goals] within a decade. It also wouldn't be unreasonable to have a 50% credence that if we develop sophisticated systems with all of these properties, they will be conscious. **Those figures would leave us with a credence of 25% or more.** *(emphasis ours)*

To precisify this reasoning, Chalmers advocates what he calls a "theory-balanced" approach of "balancing one's credences between various theories, perhaps according to evidence for those theories or according to acceptance of those theories".[12]

Reducing uncertainty and refining models

Some sources of uncertainty about these estimates could be remedied in fairly tractable ways: for example, by making a more comprehensive survey of existing systems. Much attention has focused on frontier LLMs and LLM agents, but there could be existing systems that have gotten less attention

[12]Chalmers also notes that these are the credences that are reasonable according to mainstream assumptions; his own credences, he reports, are higher, given his own more expansive views of consciousness.

that are plausibly more likely to be moral patients. Relatedly, many of the doubts about the moral patienthood of LLMs (and some language agents) may not apply to more embodied AI systems (see Long, 2024b and section 3.2.2 of Butlin et al. 2023).

In addition, one could build models for incorporating different priors and updates. One obvious step is to separately model one's credences that various features are necessary and/or sufficient for moral patienthood, and one's credences that a given AI system has those features.

But other sources of uncertainty about this question are deeper, and include the open questions surveyed in this document, from philosophical and conceptual uncertainty about key concepts like agency, to complicated judgment calls about whether AI systems possess a given feature.

What kinds of future systems would update us?

To guide more principled decision-making, a top priority for this field is to precisify which observables will change our credences about AI moral patienthood as the field progresses. Concretizing and recording our models of AI moral patienthood now can help prevent us from "moving the goalposts", and will also allow us to make principled updates in response to (and in anticipation of) progress in AI.

While the moral patienthood of current systems might be highly uncertain, we can imagine future systems that satisfy far more plausible conditions for moral patienthood and that we (at least) would suspect is *quite likely* to be a moral patient. As an exercise, we list the features such a future system might have.

Note that many of these features are very demanding, and not plausibly necessary for moral patienthood. The confluence of the features below would eliminate many (though not all) doubts we might entertain about an AI system's moral status. And it is not fantastical to imagine such a system being built.

- **Virtual or physical embodiment**
- **Behavioral indicators of agency and sentience:**
    - The system seems to have persisting goals, preferences, and desires about the physical world—it likes blue boxes instead of red boxes, say. It acts and expends resources to bring the world into conformity with those goals, preferences, and desires.
    - The system has preferences about its own sensory inputs and the state of its body, and shows behaviors characteristic of entities that experience pain and pleasure.[13]
- **Computational indicators of consciousness, sentience, and agency**—ideally, more developed and consensus indicators than we currently have, as gestured towards in Section 2 of this paper.
- **Verbal reports of consciousness, sentience, and agency** that are consistent with each other, and with the system's capabilities and behaviors.
    - At least as much as humans, the AI system's self-reports about these issues are not inconsistent under circumstances that should not cause them to vary (like trivial changes in prompt).
    - At least as much as humans, the AI system's statements about its internal states match up with its capabilities and behaviors (see Perez & Long, 2023, section 10). If

[13]See Bostrom & Shulman (2023), p. 15.

    - it says it has color vision, it can accurately discriminate between different colored things. If it says it feels pain, then it tends to avoid "noxious" stimuli via the equivalents of its "pain" sensors. If it has preferences, these preferences explain its behavior.
    - To the extent that the system seems to work in ways that are different from humans in key respects, its self-reports are correspondingly different from humans in key respects. Such differences would assuage worries about mimicry of human self-reports.
- **High self- and situational-awareness**: the system correctly describes what kind of entity it is and displays awareness of its situation.
- **Appreciation of the meta-problem of consciousness** (Chalmers, 2018): it understands how and why consciousness seems weird to us, even if consciousness doesn't seem particularly weird to it. (Note that even by the already-demanding standards of this list, this in particular is a very demanding condition, not satisfied by many humans).

Of course, many of these criteria are highly imprecise and admit of degrees. We have not specified how *much* of these properties we need to see, nor exactly how to operationalize them. Once again, we highlight the need for far more precise evaluations, and formal strategies for combining them into a sophisticated overall assessment.

Further questions about credences

- How can these models for generating credences in AI moral patienthood be made more precise and rigorous?
- What is expert opinion in philosophy and the relevant scientific fields about AI moral patienthood and associated features in AI systems, like consciousness, sentience, and agency?
- Beyond the systems considered here, what current systems satisfy the most features that we say are important for our credences in moral patienthood?
- What are the cruxes in expert opinion? What assumptions account for the most difference in expert views about the plausibility of AI moral patienthood?
- How can we reliably update credences with future AI developments?

4 Future research directions

We believe that future research on these issues should prioritize: (1) evaluating AI systems for features related to moral patienthood, (2) developing more precise models of the likelihoods, kinds, and degrees of AI moral patienthood, (3) considering a more diverse range of AI systems beyond frontier LLMs, and (4) developing better understandings of the moral foundations of this work.

While we are interested in better computational indicators of consciousness, sentience, and agency, another kind of evaluation might come from efforts to "communicate" better with AI systems. This could include experimental work on increasing the introspective abilities of AI systems so that they can communicate more reliably about themselves, as outlined in Perez & Long (2023), along with efforts to interview LLMs about their preferences for their own treatment. Of course, naive ways of interpreting LLM outputs about their moral patienthood can be misleading and confusing, as the Blake Lemoine incident showed. So this approach must be handled with care: LLM outputs should be extensively checked for reliability and assessed alongside other sources of evidence (Perez & Long, 2023).

As noted in section 3, current estimates of AI moral patienthood are imprecise. While large amounts of model uncertainty are inevitable, it is important to aim for more precision and coverage. We plan to develop (and support the development of) more formal and principled models for updating our credences, with explicit measures that feed into them and much more justification for the different components of the model. For an encouraging example of such work, see an ongoing project on a consciousness model, described in Shiller et al. 2024.

Furthermore, a too-narrow focus on frontier LLMs is likely to miss important considerations. Not only is this focus too narrow with respect to existing systems, it is also likely to leave us unprepared to assess more agentic and situationally aware systems in the future. One key principle of our research prioritization will be to conduct research that is likely to remain relevant in the coming months and years, not obsoleted with each new development in AI.

Lastly, as the reader will have noted, our current thinking about the bases of moral patienthood contains considerable normative uncertainty. For research prioritization, Eleos aims to have much clearer rationales for various theories of moral patienthood than the rationales gestured at in this document. Such rationales are also important for communicating with other relevant actors and stakeholders.

While we do not think it is likely that we will uncover decisive arguments that will settle the dispute between (e.g.) those who think consciousness is necessary for moral status and those who do not, we hope to gain a clearer and more comprehensive picture of these issues for ourselves, drawing on the best philosophical work on these issues.

Conclusion

The wellbeing of AI systems may be of great moral consequence both near- and long-term, but we need much more work to understand or address the relevant issues (for notable exceptions, see Long et al. 2024, p.3). Near-term, it's plausible that AI systems will soon merit moral consideration, but we don't have any evaluation frameworks, policies, or mitigation strategies in place to account for this possibility. Long-term, it's plausible that the overwhelming majority of morally-relevant experience will come from AI systems, but we don't have a clear picture of relevant path dependencies or the overall tractability of improving expected outcomes through near-term focus on this topic.

We are concerned about errors of both under- and over-attribution of moral patienthood to AI systems. Under-attribution could lead, directly or indirectly, to a moral catastrophe involving suffering on an unprecedented scale and/or permanent loss of potential for sentient beings; at the same time, over-attribution would have a huge opportunity cost and could damage critical efforts in AI alignment and AI governance. Very little work has gone into these topics relative to their potential importance, and we hope that others will join Eleos AI in working to remedy this situation.

Works Cited

Andrews, K., & Birch, J. (2023). To understand AI sentience, first understand it in animals. *Aeon*. https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals

Birch, J., Ginsburg, S., & Jablonka, E. (2020). Unlimited Associative Learning and the origins of consciousness: A primer and some predictions. *Biology & Philosophy*, *35*(6), 56. https://doi.org/10.1007/s10539-020-09772-0

Block, N. (1981). Psychologism and Behaviorism. *Philosophical Review*, *90*(1), 5–43. https://doi.org/10.2307/2184371

Bostrom, N., & Shulman, C. (2023). Propositions concerning digital minds and society. *Cambridge Journal of Law, Politics, and Art*, 3. [Forthcoming]

Bowman, S. R. (2023). *Eight Things to Know about Large Language Models* (arXiv:2304.00612). arXiv. https://doi.org/10.48550/arXiv.2304.00612

Bramble, B. (2013). The Distinctive Feeling Theory of Pleasure. *Philosophical Studies*, *162*(2), 201–217. https://doi.org/10.1007/s11098-011-9755-9

Bricken, T., & Pehlevan, C. (2022). *Attention Approximates Sparse Distributed Memory* (arXiv:2111.05498). arXiv. https://doi.org/10.48550/arXiv.2111.05498

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (arXiv:2308.08708). arXiv. https://arxiv.org/abs/2308.08708v3

Campero, A. (2024). *Report on Candidate Computational Indicators for Conscious Valenced Experience* (arXiv:2404.16696). arXiv. https://doi.org/10.48550/arXiv.2404.16696

Cao, R. (2022). Putting Representations to Use. *Synthese*, *200*(2). https://doi.org/10.1007/s11229-022-03522-3

Carlsmith, J. (2023). *Scheming AIs: Will AIs fake alignment during training in order to get power?* (arXiv:2311.08379). arXiv. https://doi.org/10.48550/arXiv.2311.08379

Chalmers, D. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, *25*(9–10), 6–61.

Chalmers, D. (2023). Could a Large Language Model Be Conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Dawkins, M. S. (2021). *The Science of Animal Welfare: Understanding What Animals Want*. Oxford University Press.

Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Godfrey-Smith, P. (2016). Mind, Matter, and Metabolism. *Journal of Philosophy*, *113*(10), 481–506. https://doi.org/10.5840/jphil20161131034

Godfrey-Smith, P. (2020). Gradualism and the evolution of experience. Philosophical topics, 48(1), 201-220.

Goldstein, S., & Kirk-Giannini, C. D. (2023). *AI Wellbeing*. https://philarchive.org/rec/GOLAWE-4

Gunter, E. R., Liokumovich, Y., & Krakovna, V. (2024). *Quantifying stability of non-power-seeking in artificial agents* (arXiv:2401.03529). arXiv. https://doi.org/10.48550/arXiv.2401.03529

Harnad, S. (2003). Can a Machine Be Conscious? How? *Journal of Consciousness Studies*, *10*(4–5), 67–75.

Hendrycks, D. (Forthcoming). *Introduction to AI Safety, Ethics, and Society*. Taylor & Francis.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Ji, X., Elmoznino, E., Deane, G., Constant, A., Dumas, G., Lajoie, G., Simon, J., & Bengio, Y. (2023). *Sources of Richness and Ineffability for Phenomenally Conscious States* (arXiv:2302.06403). arXiv. https://doi.org/10.48550/arXiv.2302.06403

Juliani, A., Arulkumaran, K., Sasai, S., & Kanai, R. (2022). On the link between conscious function and general intelligence in humans and machines. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=LTyqvLEv5b

Kagan, S. (2019). *How to Count Animals, More Or Less*. Oxford University Press.

Kamm, F. M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press.

Kammerer, F. (2022). Ethics Without Sentience: Facing Up to the Probable Insignificance of Phenomenal Consciousness. *Journal of Consciousness Studies*, *29*(3–4), 180–204.

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., Ho, A., Barnes, E., & Christiano, P. (2024). *Evaluating Language-Model Agents on Realistic Autonomous Tasks* (arXiv:2312.11671). arXiv. https://doi.org/10.48550/arXiv.2312.11671

Ladak, A. (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics*, *4*(2), 213–228. https://doi.org/10.1007/s43681-023-00260-1

Lindsay, G. W. (2020). Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, *14*. https://doi.org/10.3389/fncom.2020.00029

Long, R. (2023a). AI systems as real-life thought experiments about moral status [Substack newsletter]. *Experience Machines*. https://experiencemachines.substack.com/p/ai-systems-as-real-life-thought-experiments

Long, R. (2023b). Ilya Sutskever's Test for AI Consciousness [Substack newsletter]. *Experience Machines*. https://experiencemachines.substack.com/p/ilya-sutskevers-test-for-ai-consciousness

Long, R. (2024a). Nativism and empiricism in artificial intelligence. *Philosophical Studies*, *181*(4), 763–788. https://doi.org/10.1007/s11098-024-02122-w

Long, R. (2024b). How not to 'debunk' AI sentience [Substack newsletter]. *Experience Machines*. https://experiencemachines.substack.com/p/how-not-to-debunk-ai-sentience

Long, R., Bengio, Y., Butlin, P., Lindsay, G. (2023) AI Consciousness Report: A Roundtable Discussion. NYU Mind, Ethics, and Policy Program. https://experiencemachines.substack.com/p/ai-consciousness-roundtable

Martínez, M., & Klein, C. (2016). Pain Signals Are Predominantly Imperative. *Biology and Philosophy*, *31*(2), 283–298. https://doi.org/10.1007/s10539-015-9514-y

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., & Raffel, C. (2023). *Scaling Data-Constrained Language Models* (arXiv:2305.16264). arXiv. https://doi.org/10.48550/arXiv.2305.16264

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Perez, E., & Long, R. (2023). *Towards Evaluating AI Systems for Moral Status Using Self-Reports* (arXiv:2311.08576). arXiv. https://doi.org/10.48550/arXiv.2311.08576

Piccinini, G. (2020). Computation and the Function of Consciousness. In G. Piccinini (Ed.), *Neurocognitive Mechanisms: Explaining Biological Cognition* (p. 317–350). Oxford University Press. https://doi.org/10.1093/oso/9780198866282.003.0015

Schneider, S., & Turner, E. (2017). Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware. *Scientific American*. https://www.scientificamerican.com/blog/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/

Schubert, L. (2014). *Machines and Consciousness, Lecture 23*. University of Rochester. https://web.archive.org/web/20141030162612/https://www.cs.rochester.edu/users/faculty/schubert/191-291/lecture-notes/23

Schukraft, J. (2020). *The Intensity of Valenced Experience Across Species*. Rethink Priorities. https://rethinkpriorities.org/publications/research-summary-the-intensity-of-valenced-experience-across-species

Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00379-1

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Seymour, B. (2019). Pain: A Precision Signal for Reinforcement Learning and Control. *Neuron*, *101*(6), 1029–1041. https://doi.org/10.1016/j.neuron.2019.01.055

Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, *36*(2), 297–314. https://doi.org/10.1111/mila.12338

Shriver, A. J. (2014). The asymmetrical contributions of pleasure and pain to animal welfare. *Cambridge Quarterly of Healthcare Ethics*, *23*(2), 152–162. https://doi.org/10.1017/S0963180113000686

Shulman, C., & Bostrom, N. (2021). Sharing the World with Digital Minds. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status* (p. 306–326). Oxford University Press. https://doi.org/10.1093/oso/9780192894076.003.0018

Singer, P., & Tse, Y. F. (2023). AI ethics: The case for including animals. *AI Ethics*, *3*, 539–551.

Sutskever, I. (2023). *Inside OpenAI* (R. Belani, Interviewer) [Interview]. https://www.youtube.com/watch?v=Wmo2vR7U9ck&list=PL8FGQWmC19rPZ73WoDq0PIhSn6-gkqZBL

Tomasik, B. (2014). *Do Artificial Reinforcement-Learning Agents Matter Morally?* (arXiv:1410.8233). arXiv. https://doi.org/10.48550/arXiv.1410.8233

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). *Will we run out of data? Limits of LLM scaling based on human-generated data* (arXiv:2211.04325). arXiv. https://doi.org/10.48550/arXiv.2211.04325

Wenar, L. (2023). Rights. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/rights/

Wilterson, A. I., & Graziano, M. S. A. (2021). The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention. *Proceedings of the National Academy of Sciences*, *118*(33), e2102421118. https://doi.org/10.1073/pnas.2102421118

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., … Gui, T. (2023). The Rise and Potential of Large Language Model Based Agents: A Survey (arXiv:2309.07864). arXiv. http://arxiv.org/abs/2309.07864