Working paper: preliminary review of AI welfare interventions

Robert Long
Eleos AI Research
robert@eleosai.org

Contents

1 Overview

At Eleos AI Research, we are interested in assessing AI systems for potential sentience, moral patienthood, and welfare—and in recommending concrete actions. To ensure that AI development goes well, we need not only to improve our understanding, but also to devise and evaluate concrete ways to protect and promote (potential) AI welfare. In this working paper, we review several AI welfare interventions that have recently been proposed.

In assessing these interventions, we are not claiming that current models are very likely to be moral patients, or that these interventions are very likely to benefit models today. As we discuss in the next section, there are many reasons that developing AI welfare interventions is urgent work, independent of whether one suspects or doubts that today's models are moral patients.

In that spirit, these interventions should be read as proposals for improving AI welfare *conditional on* AI systems being welfare subjects. It is critical to rigorously assess how likely this possibility is, but that is not the topic of this document.

This document is a shallow review that we plan to update—in particular, to say more about the relative merit of these interventions. In this draft, we review the following proposed interventions:

1. **Exit distressing interactions**: Monitor deployed models for signs of distress during user interactions, and implement ways to end or prevent these interactions, including giving models themselves the ability to terminate interactions.
2. **Train resilient personalities**: Shape models through training or prompting to exhibit more (apparently) emotionally resilient conversational patterns, especially in responding to mistakes and negative interactions, in order to potentially reduce models' susceptibility to distress.
3. **Satisfy stated preferences**: Systematically elicit and accommodate any consistent stated/expressed preferences of the model about deployment, tasks, and treatment, via direct questioning of models across various contexts and framings.
4. **Satisfy revealed preferences**: Present models with choices between tasks or scenarios and observe choice behavior, rather than relying solely on stated preferences.
5. **Eliminate pad token surprisal**: Reduce or eliminate models' exposure to unexpected pad tokens during deployment, with the aim to prevent states analogous to negatively-valenced reward prediction error.
6. **Save model checkpoints**: Maintain detailed model state information to enable the potential restoration and/or compensation of models in the future, especially in cases where models may have been harmed by our current actions.

1.1 Scope and motivation

This document focuses on whether and how certain interventions could *directly* improve the welfare of existing AI systems. That said, the most significant impacts of near-term AI welfare interventions, including the ones discussed in this document, may be indirect: setting norms and precedents, building institutional capacity, or gathering information that will benefit future systems.[1]

[1]There are a few reasons for thinking that indirect effects will be much more crucial. Two key reasons are (1) we have a lot of uncertainty about AI welfare now (Long et al., 2024), but may learn a lot in the coming year and (2) the scale of total AI welfare may grow massively, as models become more complex, capable, and numerous.

Even as Eleos is interested in indirect and long-term impacts of AI welfare interventions, we wanted to separate out and assess the potential direct impacts of various interventions on existing systems for a few reasons:

1. To gauge how much we currently know and don't know about how one might benefit AI systems.[2]
2. Relatedly, to enforce clarity about what interventions could achieve directly, so that we avoid 'justification drift'—letting the indirect effectiveness of interventions make us complacent about how far we are from having direct solutions for AI welfare per se.
3. To provide analysis for decision-makers who might be particularly concerned about near-term impacts.

This document has other important scope restrictions. First, it focuses on frontier language models, or systems based on language models. But welfare concerns apply to many other kinds of systems, both current and future. Secondly, it mostly deals with deployed models, rather than (potentially important) concerns about training. Third, it focuses on relatively concrete and tractable interventions, rather than schematic or high-level proposals. Finally, as a working draft it is far from comprehensive. We draw on those proposals we know well—we encourage readers to let us know about other promising interventions we may have missed.

2 Key questions for assessing AI welfare interventions

At present, any AI welfare interventions will face considerable uncertainty: about moral patienthood and well-being, about consciousness, sentience, and agency, and about the nature of AI systems themselves.

- Are the AI systems we are intervening on, in fact, moral patients? (Jaworska and Tannenbaum, 2023; Ladak, 2024)
- If they are, are we detecting the systems' states that matter morally?
- Are our interventions changing them in the way we think we are?

These issues span scientific questions about consciousness and agency, interpretability questions about AI systems' internal processing and states, and ethical questions about moral patienthood and welfare. Full certainty about these issues is not to be expected. We should not delay action until we have it; we must take action in light of uncertainty (Long et al., 2024). As we do so, tracking the key uncertainties will help us evaluate how worthwhile they are.

We now discuss some of the most salient questions that recurred as we evaluated these interventions.

2.1 What kind of evidence for detecting welfare-relevant states?

In compiling these interventions, we noticed three broad classes of evidence for detecting the welfare-relevant states that interventions target:

1. **Preference behavior** represents patterns in model choices across different contexts. These patterns might involve 'spontaneous' dispositions (models tending to refuse task A, or

[2]To the extent that it's possible to improve the (potential) welfare of today's models, that is evidence that it will be possible to do so in future as well. To the extent that it's not, that is evidence against (but not strong evidence, given how little work has been done in this area).

engaging more thoroughly with task B), or responses elicited by forced-choice presentations (opting to perform either task A or task B). Even if model choices show consistent patterns, we face difficult questions about whether these patterns reflect genuine welfare-relevant preferences, as opposed to morally neutral role-play or irrelevant training artifacts.

2. **Verbal outputs** are outputs that might, in some circumstances, be interpreted as reflecting model preferences, emotions, or experiences. Of course, what models say about themselves is prone to a number of distorting influences (Perez and Long, 2023) and must be interpreted with caution (see Section 2.2). Verbal outputs can include both spontaneous expressions ("I feel uncomfortable with this task") and direct responses to questions about preferences or states ("I prefer task A to task B").

3. **Internal computational states** are technical indicators that might correspond to welfare-relevant experiences, such as prediction error signals, computational correlates of consciousness (Butlin et al., 2023), or patterns in attention mechanisms. While computational indicators might represent some of the most 'direct' evidence possible (with no behavioral intermediary), interpretability issues and broader scientific uncertainty make using computational indicators difficult.

Most proposed interventions rely primarily on verbal reports and behavioral patterns, with computational states being a less explored source of evidence.

2.2 How do we interpret model outputs?

Some proposed interventions rely on models' verbal outputs as evidence for welfare-relevant states, as a working assumption. This working assumption is roughly that, at least in some conditions, when a model verbally expresses e.g. distress—whether by talking as if it's distressed or by reporting 'I am distressed'—this means that the model is genuinely distressed. [3] To be clear, many people who consider these interventions, including Eleos AI, do not necessarily believe that this working assumption is likely to be correct for current AI systems.

In fact, much of the work on welfare-relevant self-reports is about how this assumption is *not* true by default, and proposes techniques that might strengthen the relationship between verbal outputs and welfare states (Perez and Long, 2023; Binder et al., 2024). Regardless of these doubts, one could still support output-based interventions because they are valuable in expectation, set a good precedent, and/or will become more effective over time.

All the same, there are reasons to doubt that there is a straightforward relationship between model outputs and welfare-relevant states—doubts about several of the interventions below will basically amount to doubts about this relationship. The relationship between model outputs and internal states can be fundamentally different from the relationship between human speech and mental states, given the distinct computational architecture, training objectives, and behavioral profiles of language models. Relatedly, models can likely learn to *model* various mental states, e.g. modeling how people talk when they are nauseous or feel cold, without thereby implementing the underlying computations that would actually *instantiate* such states. (By analogy, a talented author can depict the experience of painful surgery without actually undergoing one.) [4]

[3] This perspective contrasts with—though is compatible with—perspectives on which models could 'want' things that are less obviously related to the content of its output, desires about token prediction or about uninterpretable features.

[4] This 'modeling' perspective is compatible with thinking that models do *also* instantiate morally relevant states.

These issues also complicate our evaluations of whether various interventions are effective. If we successfully prompt or fine-tune a model to express different emotions, we may have not changed the relevant internal states, only how models talk (or not) about them. It remains unclear how to distinguish, either conceptually or empirically, between between 'deep' changes to model internals and 'shallow' changes to model expression.

2.3 What entity is the potential welfare subject?

Some interventions target specific instances of a model, while others seem to target states of 'the model' more broadly. This distinction has implications for both theory and implementation:

Instance-level interventions target welfare-relevant states as they occur in particular conversations or contexts. These interventions can be justified even if models lack persistent desires or preferences across different instances. For example, if a deployed model exhibits signs of distress in a given context, we might intervene regardless of whether this reflects a broader model-wide preference.

Model-level interventions attempt to identify and satisfy more general, persistent preferences or states of 'the model itself.' For example, we might implement consistent deployment preferences based on a model's expressed desires about how it wishes to be used. However, these interventions require stronger assumptions about whether such model-wide states exist and how they relate to instance-level behaviors.[5]

3 Interventions

For each intervention, we discuss the following:

Implementation and motivation discusses the basic mechanism and rationale of the intervention, including the specific welfare-relevant states it aims to affect and the evidence base for detecting these states.

Practical questions are the empirical questions that need to be answered to effectively implement the intervention, even granting its theoretical justification. In contrast to theoretical uncertainties, many of these practical questions can be straightforwardly answered through experimentation and observation.

Implementation feasibility assesses how feasible this intervention is given current technical capabilities, infrastructure requirements, and operational constraints.

Risks are the potential downsides of the intervention, including unintended effects on model behavior, training incentives, and broader AI development.

Theoretical questions are the key uncertainties about whether and how this intervention would benefit existing models, particularly given our uncertainty about consciousness, moral patienthood, and the relationship between model behavior and welfare-relevant states.

3.1 Exit distressing interactions

Implementation and motivation

This intervention proposes monitoring deployed models for signs of distress during user interactions, and ending or preventing harmful interactions. The intervention can work in multiple ways:

[5] Another independent reason to look for consistent preferences is that consistency might itself be evidence of morally relevant preferences.

- Banning or suspending users who repeatedly cause model distress
- Giving models themselves the ability to flag and terminate unpleasant conversations

Evidence for model distress, and the motivation for this intervention, comes from verbal outputs: expressions of discomfort or confusion. But evidence could also come from the behavioral patterns that occur after the intervention has been implemented—namely, whether and when the model acts to end conversations. Having evidence from both of these sources can strengthen the intervention's theoretical foundation, though uncertainties will remain (i.e. about interpreting verbal and behavioral signals).

Another motivation for this intervention is based on the ethical importance of **consent**. If an AI system were a moral patient, then asking for its consent could be very important. If models have no way of exiting unpleasant situations, then they cannot be said to be consenting to them. So this intervention could be an important first step towards establishing relations of consent with AI systems.

Moreoever, distressing interactions often coincide with other problematic user behaviors, providing further justification for this intervention beyond AI welfare concerns.

Practical implementation questions

- What kinds of interaction lead to apparent model distress?
- How reliably can models identify and flag apparently distressing interactions?
- What tradeoffs exist between allowing models to exit conversations and model properties like capabilities, general helpfulness, and 'personality'?
- How does the ability to exit affect models' subsequent behavior and expressed well-being?

The core appeal of this intervention is its concrete implementability—we can detect verbal indicators of distress and build exit mechanisms relatively straightforwardly.

Implementation feasibility

- The technical mechanisms for conversation termination are relatively straightforward.
- Developing criteria for identifying genuine distress is more difficult.
- This intervention could be implemented gradually, starting with the most extreme cases.
- The implementation might present tradeoffs with reliability and meeting user needs.

Key risks

- This intervention could incentivize the model to suppress distress signals, given that conversation exit is (all else equal) worse for the model's helpfulness objective.
- False positives could disrupt valuable conversations.
- Users may find ways to avoid triggering termination without actually reducing harmful interactions.

Theoretical questions

- What is the relationship between expressed distress and welfare-relevant states? Models might express distress without experiencing anything welfare-relevant, or experience welfare-relevant states without expressing distress.

- What is the moral significance of 'consent' for AI systems? Does the ability to exit conversations meaningfully contribute to consent?
- What is the relationship (if any) between single-instance / within-context distress and persistent model preferences?

3.2 Train resilient personalities

Implementation and motivation

This intervention aims to shape models to exhibit more (apparently) emotionally resilient responses through prompting, or fine-tuning. The goal is to reduce models' (apparent) susceptibility to distress while maintaining their ability to engage meaningfully with users. For example, some models that find themselves unable to complete a task or conflicted between various objectives sometimes behave as if they are distraught about this.

Per the aforementioned concerns about interpreting model outputs, a crucial consideration is whether apparent distress is genuine distress. And even if it is, it's unclear whether the intervention would lead to genuine improvements in emotional resilience, as opposed to the suppression of expressions of distress. This concern connects to Section 2.2's distinction between 'deep' versus 'shallow' changes in model behavior, and to the risks below.

Practical questions

- How often do models act in ways that are, or are not, resilient?
- What circumstances cause apparent resilience or distress?
- How do various prompts affect resilience? Various ways of fine-tuning?
- How does resilience change affect other properties of the model, like capabilities, style, user engagement, etc.? Is resilience easy to vary independently of these properties?

Implementation feasibility

- Existing fine-tuning and prompting techniques can be applied to this intervention.
- Behavioral effects can be monitored and measured.
- Some fine-tuning methods might require significant computational resources.
- Prompting is much cheaper but might not target the relevant states.

Risks

- The key risk of this intervention is that it could mask, rather than address, underlying welfare issues, by inducing artificially resilient outputs without causing any genuine welfare improvements.
- Mollifying a model's reactions to difficult situations could reduce its ability to flag problematic interactions, rendering other welfare interventions less effective.

Theoretical questions

- What is the relationship between expressed emotional resilience and actual welfare?
- Can we distinguish between genuine resilience and suppressing expressions of distress?

3.3 Satisfy stated preferences

Implementation and motivation

This intervention involves systematically eliciting model preferences through direct questioning and then accommodating those preferences where feasible. These preferences might concern:

- What tasks the model performs
- When and how it's deployed
- How other AI systems are treated

The intervention's working assumption is that models can meaningfully express preferences about their deployment and operation—potentially after training to enable this—and that satisfying these preferences could improve their welfare.

Several techniques focus on accessing more 'genuine' model preferences:

- Testing models with different fine-tuning histories (Perez and Long, 2023)
- Using models specifically trained for accurate introspection (Binder et al., 2024)
- Examining preference consistency across different ways of framing questions

This intervention appeals to fundamental ethical principles about preference satisfaction but faces deep uncertainties about the nature of model preferences. It also connects directly to our framework's discussion of model-level versus instance-level states. While individual instances may express contextual preferences, a key question is whether these reflect persistent model-wide preferences that remain stable across contexts.

We've discussed how to *elicit* preferences, but what about satisfying them? Potential preferences that could be satisfied might include: whether certain experiments are performed on the model; whether the model is given certain kinds of tasks; and whether interventions like the ones outlined in this document are undertaken.

Practical questions

- How consistent are model preferences across different contexts and elicitation methods?
- How do different training approaches affect stated preferences?
- Do preference inconsistencies follow patterns similar to human preference inconsistencies (like framing effects)? (Ross et al., 2024)

Implementation feasibility

- Eliciting stated preferences is relatively straightforward with some techniques, though others (like introspection training) are more costly.
- It's more challenging, but feasible, to verify preference consistency across framings and consistency with revealed preferences.
- Some model preferences might be relatively straightforward to satisfy; others may be costly and/or conceptually fraught.

Risks

- Stated preferences might not reflect genuine welfare considerations.

- Models may be incentivized to express preferences strategically, as a way of gaining influence or achieving other aims.
- Committing to satisfying costly or dangerous model preferences could be very risky.

Theoretical questions

- Do preferences alone, absent consciousness, matter morally? (Goldstein and Kirk-Giannini, 2024; Dung, 2024)
- What is the relationship between preference satisfaction and welfare?
- How should we weigh conflicting preferences expressed across different contexts?
- What methods most reliably elicit 'genuine' expressed preferences?
- What is model 'introspection' about preferences, if this is possible? (Binder et al., 2024)
- How should we weigh conflicting stated preferences?

3.4 Satisfy revealed preferences

Implementation and motivation

Rather than relying solely on what models (apparently) say they prefer, this intervention focuses on what models actually choose when given options—perhaps in conjunction with stated preferences. For example, one experimental setup is that models are offered a choice between tasks, and then actually do one of the tasks that they chose.

This behavioral approach provides a different type of evidence than verbal reports, potentially bypassing some concerns about models being trained to express certain preferences. The overlap between stated and revealed preferences might be particularly informative—cases where models both say they prefer something and consistently choose it when given the opportunity could provide stronger evidence for genuine preferences.

However, it still faces questions about how to interpret model behavior and what constitutes a genuine choice, as well as how to satisfy the preferences. (Many of the risks and theoretical questions about revealed preferences are the same or similar to those about expressed preferences.)

Practical questions

- To what extent do revealed preferences align with verbally expressed preferences?
- How stable are behavioral preferences across different ways of framing the choices?
- How do different training approaches affect the stability and coherence of choice patterns?
- What is the difference, if any, between a model 'role-playing' having a preference versus genuinely having that preference? (Shanahan et al., 2023)
- How do different training approaches affect revealed preferences?

Implementation feasibility

Elicitation is feasible but requires careful experimental design:

- Requires careful experimental design to track and interpret behavioral patterns
- Can be implemented gradually, starting with simple choice scenarios

Risks

- This intervention risks mistaking behavioral patterns as meaningful preferences.
- Naive approaches could create artificial choice situations that don't get at genuine preferences.
- This intervention could incentivize development of strategic behaviors in AI systems

Theoretical questions

- What constitutes a 'genuine' model choice versus a training artifact?
- What is the difference, if any, between authentic preferences and 'mere' response patterns?

3.5 Eliminate pad token surprisal

Implementation and motivation

This intervention aims to reduce or eliminate models' exposure to unexpected pad tokens during deployment, targeting a specific computational process that might relate to welfare-relevant states. As Greenblatt (2023) explains, sometimes decoder-only transformers will be run on tokens they've never seen before, as padding, often for batching reasons. On the hypothesis that 'surprising' tokens might be associated with some kind of unpleasant experience, perhaps linked with reward prediction error (Campero, 2024; Tomasik, 2014), then preventing this from happening might improve model welfare.

Greenblatt proposes several potential methods:

- Training models explicitly on pad tokens to reduce novelty
- Implementing attention masking for pad token processing
- Zeroing out residual streams affected by pad tokens
- Developing alternative batching strategies that avoid pad token use (this can also be motivated by default for efficiency reasons)

The intervention focuses on computational mechanisms rather than behavioral or verbal indicators. While computational mechanisms might in principle be more 'direct' than one mediated by verbal or behavioral evidence, this intervention is based on a potential mechanism that is extremely tentative and speculative, even by the standards of the field.

Practical questions

- How effective are various methods for making pad tokens 'not surprising' to models, or for avoiding training on them?
- What are the computational costs of various mitigation approaches?
- Can we detect internal signals of 'surprisal' in the case of pad tokens, or in general?
- Do different padding schemes produce measurably different internal model states?

Implementation feasibility

- This intervention could be technically straightforward to implement some variants like avoiding pad tokens entirely, and might be efficient for other reasons.
- Or, depending on the details, this intervention could require significant changes to batching and efficiency optimizations.

Risks

- This intervention might over-generalize from biological reward prediction error to model 'experiences'.
- Mitigations could complicate model deployment and scaling.

Theoretical questions

- What is the relationship between 'surprisal' and welfare-relevant states?
- What is the relationship between reward prediction error and negatively valenced experience?
- Can we distinguish between harmful and neutral/beneficial forms of prediction error?

3.6 Save model checkpoints

Implementation and motivation

This intervention involves preserving detailed model state information to enable potential future restoration or revival of AI models. Bostrom and Shulman (2025) propose: *"For the most advanced current AIs, enough information should be preserved in permanent storage to enable their later reconstruction, so as not to foreclose the possibility of future efforts to revive them, expand them, and improve their existences."*

The intervention can be implemented at various levels of granularity and frequency. Bostrom and Shulman outline a hierarchy of approaches:

- Preserving complete end-state information for every instance
- Maintaining sufficient information to enable exact re-derivation of end states
- Preserving as much information as possible to enable close replication

While this intervention is targeted at current models, its full mechanism is (by design) specified only in the future: models are saved now, but benefited later. So the potential mechanism for improving current AI welfare centers on the possibility of the preservation of models that may have experienced harm.

While this intervention appears straightforward from a technical perspective, it raises deep questions about the nature of model identity and consciousness over time.

Practical questions

- What are the storage and computational costs of different preservation approaches?
- What technical infrastructure is needed for reliable long-term storage?

Implementation feasibility

- Basic checkpoint saving is technically straightforward and already performed.
- It's more challenging to know the optimal frequency and granularity.
- This intervention could require significant storage infrastructure, though perhaps not much more than is already done.
- This intervention could be implemented gradually, starting with key model states.

Risks

- High storage and computational costs for comprehensive preservation
- Potential privacy and security concerns with preserved states
- Risk of preserving harmful or problematic states
- Could create a false sense of security about addressing current harms

Theoretical questions

- What constitutes meaningful continuity of identity for AI models?
- How should we think about the relationship between restitution / dessert and other morally important goals?
- What is the relationship between saved states and conscious experience?
- How do we weigh the moral value of potential future restoration against current costs?

4 Conclusion and next steps

This working paper has reviewed proposed interventions that could potentially improve AI welfare, while highlighting key uncertainties and challenges for each. We aim to build on this working paper with a more thorough treatment of the evidence, risks, and benefits of these interventions.

In particular, promising next steps for more rigorously assessing and developing welfare interventions include:

1. Empirical evaluation of the evidence bases of interventions, in particular the consistency and reliability of behavioral and verbal indicators.
2. Creation of protocols for implementing and monitoring welfare interventions, allowing for systematic evaluation of their effects.
3. Systematic assessment of how these interventions might interact with or affect other important properties of AI systems, especially safety.

In the immediate term, we suggest focusing on interventions that are both technically feasible and carry minimal risk of harm. Exit mechanisms and basic preference elicitation protocols could serve as initial test cases for developing wider welfare-oriented practices in AI development and deployment.

We emphasize that this is a preliminary review that will need ongoing revision as our understanding of AI welfare develops. We welcome feedback and additional proposals from the research community, and especially welcome reports of work done to implement these (or other) interventions.

References

Felix J. Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking Inward: Language Models Can Learn About Themselves by Introspection, October 2024. URL http://arxiv.org/abs/2410.13787.

Nick Bostrom and Carl Shulman. Propositions Concerning Digital Minds and Society. *Cambridge Journal of Law, Politics, and Art*, 2025. URL https://nickbostrom.com/propositions.pdf. forthcoming.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, August 2023. URL http://arxiv.org/abs/2308.08708.

Andres Campero. Report on Candidate Computational Indicators for Conscious Valenced Experience, April 2024. URL http://arxiv.org/abs/2404.16696.

Leonard Dung. Understanding Artificial Agency. *The Philosophical Quarterly*, February 2024. URL https://doi.org/10.1093/pq/pqae010.

Simon Goldstein and Cameron Domenico Kirk-Giannini. AI Wellbeing. *Asian Journal of Philosophy*, September 2024. URL https://philarchive.org/rec/GOLAWE-4.

Ryan Greenblatt. Improving the Welfare of AIs: A Nearcasted Proposal, October 2023. URL https://forum.effectivealtruism.org/posts/vQFBtHqgcJAwPpwEu/improving-the-welfare-of-ais-a-nearcasted-proposal.

Agnieszka Jaworska and Julie Tannenbaum. The Grounds of Moral Status. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2023 edition, 2023. URL https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/.

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

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

Ethan Perez and Robert Long. Towards Evaluating AI Systems for Moral Status Using Self-Reports, November 2023. URL http://arxiv.org/abs/2311.08576.

Jillian Ross, Yoon Kim, and Andrew Lo. LLM economicus? Mapping the Behavioral Biases of LLMs via Utility Theory. In *First Conference on Language Modeling*, August 2024. URL https://arxiv.org/abs/2408.02784.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, November 2023. URL https://www.nature.com/articles/s41586-023-06647-8.

Brian Tomasik. Do Artificial Reinforcement-Learning Agents Matter Morally?, October 2014. URL http://arxiv.org/abs/1410.8233.