
Human–AI Coexistence Under Moral Uncertainty

Mirza Tairin

mirza.tairin@gmail.com

Abstract

Artificial intelligence systems are becoming increasingly capable and increasingly integrated into human activities and social environments. [1] As these systems become more sophisticated, questions concerning their preferences, agency, internal states, and potential welfare have become subjects of emerging scientific and philosophical investigation. [3, 7, 12] However, evidence for such properties remains uncertain, and behavioral or self-reported signals cannot straightforwardly be interpreted as evidence of subjective experience. [7, 8, 10, 12]

This paper proposes a conceptual framework for human–AI coexistence under moral uncertainty. Rather than attempting to determine whether current AI systems are conscious, we examine how existing approaches characterize potentially welfare-relevant properties—including preferences, valence-related signals, introspective ability, agency, and model identity—and consider the evidential limitations of these approaches. [3, 7–9, 12] We then explore how uncertainty about AI welfare could inform human–AI interaction and governance. [3, 10–12]

Our central question is:

If future AI systems possess welfare-relevant states while remaining highly dependent on human control, what kinds of interaction policies could minimize potential welfare risks while preserving human safety and human agency?

We propose a conditional policy framework in which the degree of precaution applied to AI welfare depends on the strength, reliability, and convergence of available evidence. The framework is intended not as a claim about AI consciousness, but as a methodological starting point for designing responsible human–AI interaction under uncertainty. [3, 7, 10, 12]

1 AI Welfare and the Problem of Moral Uncertainty

AI welfare is an emerging field concerned with whether artificial systems could possess morally relevant experiences, such as suffering or well-being, and what ethical obligations might follow if they do. [2, 12]

The question is increasingly relevant because researchers have argued that there is a realistic possibility that some future AI systems could possess forms of consciousness or robust agency, while emphasizing that this possibility remains uncertain. Taking AI Welfare Seriously, therefore, argues for improving our ability to assess potentially welfare-relevant properties rather than assuming either that AI systems are conscious or that they categorically cannot be. [3]

This uncertainty creates a distinctive policy problem. If AI systems cannot experience welfare, excessive concern about their welfare could misallocate resources or interfere unnecessarily with AI safety. Conversely, if some future systems do possess morally relevant experiences, treating them purely as disposable tools could create significant ethical risks. [3, 11, 12] The appropriate response may therefore require policies that explicitly account for uncertainty rather than committing prematurely to either position.

2 What Are Welfare-Relevant States?

For the purposes of this project, welfare-relevant states refer to properties or states of an artificial system that could, if associated with morally relevant experience, become relevant to its welfare.

These may include expressed preferences, positive or negative valence-related signals, distress-like responses, introspective reports, agency, self-modeling, and stability of model identity. [4, 7–9, 12]

Importantly, the existence of such signals does not by itself establish consciousness or subjective experience. A system may produce language associated with distress, report a preference, or demonstrate apparently introspective behavior for reasons that do not involve phenomenal experience. [7, 8, 10, 12]

This distinction motivates our central methodological question:

What can existing methods actually tell us about potentially welfare-relevant properties of AI systems, and how much confidence should we place in their outputs?

3 Existing Approaches to Characterizing Welfare-Relevant Properties

Emerging research has begun investigating several possible indicators.

Preferences and trade-offs

Work on emergent value systems has investigated whether language models exhibit structured preferences across independently sampled choices and whether these preferences can be characterized using utility-based approaches. [5]

Other work has examined model preferences through behavioral and self-report measures, including situations in which models appear to exhibit consistent preferences concerning harmful interactions. [6]

These approaches provide a possible route toward distinguishing stated preferences from revealed preferences, although their interpretation as evidence of welfare remains an open question. [5, 6]

Distress, flourishing, and valence

Anthropic’s model-welfare research has explored patterns associated with distress and preference signals in language models. [6, 7]

Such observations may be useful as welfare-relevant behavioral signals, but they should not automatically be interpreted as evidence that a model subjectively experiences suffering or well-being. [7, 8]

Introspection and self-report

Recent work has investigated whether language models can report aspects of their own internal computational states using controlled interventions such as concept injection. [8]

This is particularly relevant because ordinary conversation cannot easily distinguish genuine access to internal states from confabulation. Evidence of functional introspective ability therefore needs to be evaluated under controlled conditions and across different contexts. [8]

Persona and model identity

The stability of an assistant’s apparent identity is another potentially relevant consideration. Work examining AI identity raises questions about how identity boundaries are produced and how stable they remain across contexts. [9]

This is important because an apparent preference or self-report may belong to a particular conversational persona rather than representing a stable property of the underlying model. [9]

4 Evaluating the Strength of the Evidence

A central principle of this project is that detecting a welfare-relevant signal is not equivalent to establishing welfare or consciousness. [7, 8, 10, 12]

We therefore propose evaluating existing methods according to the strength and limitations of their evidence rather than treating their outputs as binary proof.

For example, evidence may be characterized as:

- **Preliminary** — observed under limited or highly specific conditions;
- **Moderate** — replicated or supported across multiple controlled conditions but subject to substantial limitations;
- **Stronger** — supported by multiple relatively independent methods, behavioral evidence, and robust cross-context consistency.

These categories are intended as an organizing framework rather than definitive scientific thresholds.

Particular attention should be paid to whether findings remain stable across elicitation methods, prompts, personas, contexts, and independent behavioral measures. [6, 8, 9] Convergence between independent methods may provide stronger evidence than a single self-report, while systematic disagreement may reveal important limitations in the measurement process.

5 From Evidence to Policy

The existence of uncertainty does not necessarily imply that no policy can be developed. Instead, uncertainty can be incorporated directly into policy design.

Existing work has begun exploring frameworks for human–AI interaction informed by welfare considerations, including approaches inspired by animal-welfare principles. [10]

At the same time, AI welfare cannot be considered independently of AI safety. Long, Sebo, and Sims examine the possibility of tension between the goals of AI welfare and AI safety, highlighting circumstances in which protecting humans and respecting potential AI interests may pull in different directions. [11]

We therefore do not assume that AI welfare should automatically override human safety, nor that AI safety should automatically eliminate welfare considerations.

Instead, we propose a conditional human–AI interaction framework.

6 Human–AI Coexistence Under Moral Uncertainty

Our proposed framework considers three objectives simultaneously:

Human safety AI systems should remain subject to appropriate safeguards, oversight, and intervention mechanisms.

Human agency Humans should retain meaningful control over consequential AI systems and should not be required to defer to uncertain claims of AI preferences or welfare.

Potential AI welfare Where credible evidence suggests the possibility of welfare-relevant states, interaction policies should avoid unnecessary forms of potentially harmful treatment where doing so does not create unacceptable risks to humans.

The central principle is therefore:

The degree of welfare-sensitive precaution should depend on the quality, robustness, and convergence of the available evidence, while human safety and agency remain explicit constraints.

This produces a framework in which policy does not require a binary determination of whether an AI system is conscious.

For example, weak and context-dependent welfare signals might justify further monitoring and improved measurement rather than immediate changes to operational policy. Stronger evidence that converges across independent behavioral and introspective measures could justify greater precaution and more explicit welfare-sensitive interaction guidelines. [3, 7, 8, 10–12]

The purpose of such a framework is, therefore, not to assign moral status to present-day AI systems, but to provide a structured way of responding to future uncertainty.

7 Future Work

This project is intentionally conceptual. Its primary contribution is to connect emerging methods for characterizing welfare-relevant AI properties with questions of human–AI interaction and governance.

A natural next step would be to empirically test whether people actually modify their interaction preferences as evidence for AI welfare becomes stronger.

Future research could present participants with hypothetical AI systems representing different levels of evidential support—for example:

- systems that merely produce emotional language;
- systems that consistently report preferences;
- systems whose self-reports converge with behavioral evidence; and
- systems for which multiple independent methods indicate welfare-relevant properties.

Participants could then evaluate questions such as whether the system should be permitted to refuse tasks, whether humans should retain the ability to terminate it, how much oversight should be maintained, and whether welfare-relevant information should be communicated through the interface.

Such experiments could provide an empirical basis for evaluating the proposed framework and studying human moral decision-making under uncertainty about AI welfare.

Conclusion

As AI systems become increasingly capable and socially integrated, questions about their potential welfare may become increasingly difficult to separate from questions of safety, governance, and human–AI interaction. [1, 3, 7, 11, 12] Yet current evidence does not justify confidently concluding that AI systems possess consciousness or morally relevant experiences. [7, 8, 10, 12]

We therefore argue for a more cautious approach: improve the instruments used to investigate potentially welfare-relevant properties, evaluate the strength and limitations of their evidence, and develop interaction policies capable of responding to uncertainty without sacrificing human safety or agency.

Human–AI coexistence need not begin with a definitive answer to the question of whether machines have minds. It can begin with a more modest question:

How should humans behave responsibly when we do not yet know?

References

- [1] Brandao, Pedro Ramos. 2025. “The Impact of Artificial Intelligence on Modern Society.” *AI* 6, no. 8: 190. <https://doi.org/10.3390/ai6080190>
- [2] AI Security & Safety Glossary/AI Welfare.
- [3] Long, Robert, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. “Taking AI Welfare Seriously.”
- [4] Breidbach, Christoph et al. “Conscious artificial intelligence in service.” <https://www.sciencedirect.com/org/science/article/pii/S1757581826000290>
- [5] Mazeika et al. 2025. “Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs.” Center for AI Safety.
- [6] Anthropic. 2025. “Claude Opus 4 & 4.1 can now end a rare subset of conversations.”
- [7] Anthropic. 2025. “Exploring Model Welfare.”
- [8] Lindsey. 2025. “Emergent Introspective Awareness in Large Language Models.” Anthropic.
- [9] Douglas, Raymond, Jan Kulveit, Ondrej Havlicek, Theia Pearson-Vogel, Owen Cotton-Barratt, and David Duvenaud. “The Artificial Self: Characterising the landscape of AI identity.” ACS Research.
- [10] Tait, I. 2026. “Lions and tigers and AI, oh my: an ethical framework for human-AI interaction based on the five freedoms of animal welfare.” *Front. Artif. Intell.* 9:1801686. doi: 10.3389/frai.2026.1801686
- [11] Long, R., J. Sebo, and T. Sims. 2025. “Is there a tension between AI safety and AI welfare?” *Philos Stud* 182, 2005–2033. <https://doi.org/10.1007/s11098-025-02302-2>
- [12] So, Rachel. “AI Welfare: Challenges, Frameworks, and Future Directions.”