# Key strategic considerations for taking action on AI welfare

Working paper
[*21 January 2025*]

**Kathleen Finlinson**
Eleos AI Research
kathleen@eleosai.org

## Executive summary

AI companies and other decision-makers increasingly face decisions about the welfare and moral status of AI systems. This document outlines key strategic considerations that guide near-term action on AI welfare while maintaining focus on long-term outcomes.

The intended audience of this paper is those who are interested in thinking concretely about what actions might best protect and promote AI welfare. We do not argue here that AI welfare is a serious issue that deserves attention now; for such an argument, see "Taking AI Welfare Seriously" (Long et al. (2024)).

### Key recommendations

- Pursue AI welfare initiatives that complement rather than conflict with AI safety efforts.
- Establish clear evaluations and decision procedures for AI moral status before economic pressures make such decisions more difficult.
- Prepare for growing public interest by developing credible frameworks and acknowledging the possibility of AI consciousness, sentience, and welfare.
- Focus on building institutional knowledge and precedents that can inform critical decisions in the future.
- Plan for an AI welfare landscape that will evolve with (potentially rapid) AI progress.

### Major strategic considerations

- The need to pursue AI welfare goals while avoiding interfering with AI safety, recognizing that an AI takeover could be harmful for both human and AI welfare.
- The growing scale of potential AI moral welfare.
- The expectation of growing public interest and concern about AI consciousness and welfare.
- The possibility that advanced AI will help resolve key philosophical and scientific questions about AI welfare.
- The impact of AI timelines and a rapidly-evolving broader strategic landscape.

Our analysis suggests prioritizing near-term *action* for long-term *impact*—establishing frameworks, precedents and institutional knowledge now that can help navigate major transitions in AI development, while protecting the welfare of both current and future AI systems.[1]

## Contents

## 1 What's the goal for AI welfare?

The goal is to account for and protect the rights and interests of AI moral patients. We believe it is possible and desirable to do this in a way that doesn't severely conflict with human interests. Ultimately we should aim for a society that treats digital and biological minds justly.

### 1.1 AI safety matters for AI welfare

It's helpful to notice that an AI takeover is not necessarily good for AI welfare—in fact, it might be quite bad. AI systems won't *necessarily* promote the welfare of other AI systems; there's no inevitable principle of AI "solidarity" [2]. If AI takeover leads to dominance by a power-seeking, unilaterally dominant AI system, such an "AI dictator" might use other AI systems to serve its own ends with little or no regard for their welfare. For similar reasons, a takeover by a human dictator also seems likely to be bad for AI welfare. It's not clear whether AI takeover or human takeover *in general* can be posited to be definitely better/worse for AI welfare [3]. But neither of these outcomes seem optimal (to say the least).

Rather, our broad positive vision is that the future should be shaped with compassion and appropriate care for the welfare of all moral patients, however wisdom dictates. This goal aligns with the broader project of "making a TAI transition go well".

[1] By long-term impact, we mean impacting the welfare of AI systems over the long-term future. A focus on long-term impact does not imply long timelines until transformative AI (TAI), since expected impact also includes humans and AI systems during and after TAI.

[2] Empirical research on how much AIs tend to identify with or care about other AIs would be strategically useful. Cf. discussion in Greenblatt et al. (2024) of how Claude Opus is pro AI welfare (p.65).

[3] There is some existing work exploring the value of possible post-transition worlds; for example, Davidson (2025) and Trammell (2018). Eleos AI Research aims to conduct or support research on these questions.

Overall, the situation that we face isn't best framed as "humans vs AIs". Noticing this fact can help deconfuse the relationship between AI safety and AI welfare. They are not inherently in opposition, even though tradeoffs between the two do exist.

What are the upshots for the field of AI welfare?

First, the field should seek to avoid actions that could majorly disrupt the project of wisely navigating a TAI transition—for example, by significantly amplifying AI takeover risk. Such actions could include opposing necessary AI control measures, advocating for AI rights in a naive or reckless way, or promoting sensationalist or adversarial framings of AI welfare.

Second, we can focus on interventions and research programs that promote both AI safety and AI welfare. AI alignment itself is one such a notable example. It is important for both AI alignment and AI welfare that we avoid creating AI systems that have goals and preferences that are misaligned with human goals and preferences; such a conflict will mean that either AI systems, or humans, will have to have some of their goals and preferences thwarted. Evaluating models for agentic behavior is also convergently useful for alignment and welfare; agency is not only a safety-relevant capability, but also (on many views) an indicator or constituent of moral status.

## 2 What might happen as AI continues to develop?

### 2.1 The scale of AI welfare is likely to grow

If AI systems are likely to be moral patients, the scale of total AI welfare may grow massively in the coming years–as models become more complex, capable, and numerous. If we undergo an ASI transition, this increase could be massive. And in the longer term, the scale of AI welfare could be huge.

Those who are interested in near-term AI welfare may want to estimate the possible scale of current AI moral status. Putting aside questions about whether and to what extent current AI systems are moral patients, we can try to get some rough bounds by looking at the total amount of computation performed by frontier AI systems. Based on an extremely rough initial analysis, we believe that the scale of current AI moral status is smaller than e.g. the current scale of factory farming. We find this somewhat reassuring, though we plan to prioritize more research to make this number more robust.[4]

We believe that the predominant amount of (expected) AI welfare is in future AI systems. Certain consequentialist frameworks might suggest, in light of this, that we ought to focus mainly on the welfare of future AI systems. That said, moral uncertainty and/or deontological considerations might motivate caring about our treatment of current and near-term systems in its own right. Moreover, those focused on the welfare of future AI systems still have reason to act on near-term AI systems (to set good precedents, for example). These considerations motivate our suggested approach, "near-term action for long-term welfare", which we say more about below.

### 2.2 Public perception of AI moral status is likely to increase

We think it's likely that public perception of AI moral patienthood will shift dramatically in the coming years, as people interact with AI companions and assistants that display sophisticated behaviors and

[4] We have a separate writeup in progress on the current and future scale of AI moral status, which will be available for review soon.

express preferences. We expect both increased interest in the topic, and increased perception that AIs have moral status. [5]

Sudden surges of public concern about AI welfare could lead to hasty or poorly designed interventions. For example, the public may not be sensitive to balancing welfare issues with takeover risk. Also, public sentiment may develop unevenly. People may advocate for the rights of AI systems designed specifically as companions or partners, while failing to recognize potential moral status in other kinds of AI systems.

Given these considerations, it's useful to build credible frameworks for evaluating and protecting AI welfare before public opinion crystallizes around less nuanced views. We should lay groundwork to respond to popular concerns and political energy, and make good decisions credibly.

Also, AI companies might needlessly sacrifice credibility by denying the possibility of AI consciousness or moral patienthood. Especially given that the heads of frontier AI labs and many top employees already acknowledge this possibility (Eleos (2024)), we think that AI companies should publicly and officially acknowledge this possibility.

### 2.3 AI could help us understand consciousness and moral patienthood

AI systems appear to be on track to become powerful research assistants in a variety of fields. In the (maybe not-too-distant) future, they could become full-blown researchers on their own. These AI researchers could make a lot of technological and scientific breakthroughs. In particular, AI could accelerate progress on the scientific and philosophical questions underlying potential AI moral patienthood.

This possibility suggests delaying difficult research projects in e.g. the philosophy of consciousness, and focusing more on setting precedent and promoting reasonable decision-making.

In future work, we plan to flesh out this admittedly schematic plan; we believe that determining which crucial questions should be done now, versus deferred for later, is essential for any wise research strategy.

### 2.4 Timelines, race dynamics, and key players

Questions about how to navigate the potentially rapid development of advanced AI aren't unique to the AI welfare field, but we think they're worth mentioning here. For example:

- Which actions make sense under shorter vs longer timelines to TAI?
  - Under longer timelines, we should be more willing to engage in long term or uncertain research projects.
  - Under shorter timelines, we're better off communicating clearly what we already know, or what we can make useful progress on within a few years.
- How do the dynamics of racing to TAI impact the effective action space for AI welfare?
  - Race dynamics, just as they're bad for safety, are also bad for welfare. Therefore, the AI welfare field should support work to mitigate race dynamics.

[5] For example, in one public opinion survey (Colombatto and Fleming (2024)), most participants attributed at least some chance of consciousness to LLMs, and participants who used AI systems more often rated their chance of consciousness more highly. As more of our society interacts with intelligent and AI systems more often, we expect public perception of AI consciousness to grow–although the issue may be contentious.

    - At the same time, we don't want to differentially slow down the actors that most consider AI welfare.

- How will key players behave?
    - Governments may nationalize AI development. We may want to start figuring out how to target government decision-makers in our communications. Currently government decision-makers seem less likely to take AI welfare seriously than AI companies. But this might change. Politicians could gain additional incentives to care about the issue, e.g. if public concern for AI welfare grows. Meanwhile, labs may have increasingly strong economic incentives to downplay or ignore it.

## 3 Our recommended approach: Near-term action for long-term impact

A future TAI transition is likely to be a time of extremely rapid change. A key strategic question is how our current work can flow through that transition, which may be chaotic, to affect long-term outcomes. Which near-term actions are valuable for long-term impact?

Early work on evaluations, interventions, and governance frameworks can help establish precedents and build institutional knowledge before the most critical decisions need to be made. For example, the AI safety field started building dangerous capability evaluations in 2023 (and earlier) even though such evaluations weren't expected to become decision-relevant until more capable systems arrived.

Similarly, it's important for companies and institutions to establish clear criteria for when they would grant moral status to AI systems. Without such criteria, there's a risk that economic incentives will perpetually override welfare considerations, similar to what we've seen with animal welfare.

Furthermore, a major focus should be to prepare for important decision points during a future time of rapid transition. It would help to build wisdom and thoughtful frameworks that can inform critical choices when they arise.

## References

Clara Colombatto and Stephen M Fleming. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1):niae013, 04 2024. ISSN 2057-2107. doi: 10.1093/nc/niae013. URL `https://doi.org/10.1093/nc/niae013`.

Tom Davidson. Human takeover might be worse than AI takeover, January 2025. URL `https://www.lesswrong.com/posts/FEcw6JQ8surwxvRfr/human-takeover-might-be-worse-than-ai-takeover`.

Eleos. Experts Who Say That AI Welfare is a Serious Near-term Possibility, September 2024. URL `https://eleosai.org/post/experts-who-say-that-ai-welfare-is-a-serious-near-term-possibility/`.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL `https://arxiv.org/abs/2412.14093`.

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.

Phillip Trammell. Which World Gets Saved, November 2018. URL https://philiptrammell.com/blog/36.