Key strategic considerations for taking action on AI welfare

Working paper
[*31 March 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 strategic considerations

- **There are overlaps between AI welfare and AI safety**, both in means and in goals. We list some interventions that promote both AI safety and welfare (overlapping means). And we argue that at least in some respects, AI safety is good for AI welfare and AI welfare is good for AI safety (overlapping goals). We recommend prioritizing AI welfare interventions that are convergent with safety.
- **The scale of AI welfare is likely to grow.** We recommend focusing on the long-term impacts of work in AI welfare.
- **Public perception of AI moral status is likely to increase.** We recommend preparing now by creating credible frameworks for assessing AI consciousness, and acknowledging the possibility of AI moral status.
- **AI could help us understand consciousness and moral patienthood.** We recommend promoting research projects that will help AI make progress on these questions as capabilities advance, and for now focusing more on setting good precedent and promoting reasonable decisionmaking.
- **The field of AI welfare must consider timelines and race dynamics.** We recommend focusing on projects for which useful progress can be made within a few years.

1 There are overlaps between AI welfare and AI safety.

The projects of AI welfare and AI safety share overlaps both in means and in goals. We list a few examples of work that promotes both AI welfare and safety (overlapping means). We also argue that AI welfare and safety might each be helpful to the other (overlapping goals).

1.1 There are interventions and research programs that promote both AI welfare and safety.

Here are a few such examples.

- *AI alignment:* Alignment is good for AI welfare in some respects. It is useful to 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 AI models for preferences, agency, situational awareness, and introspection:* These properties are not only relevant to safety, but are also (on many views) indicators or constituents of moral status.
- *Trading with AI systems:* At some point, humanity may develop misaligned AIs that have desires and preferences which we didn't intend. We might not know that they're misaligned, or we might know they're misaligned but still want to use them to help us with AI capabilities or safety research. In either case, offering these AIs positive incentives to cooperate with our agenda could have both safety and welfare benefits[1].

We're still in the early stages of exploring interventions with safety and welfare benefits, and we expect there may be more promising ideas in this category.

1.2 AI welfare and safety have shared goals.

There are shared goals between AI welfare and safety in both directions.

- *AI safety is good for AI interests, at least in some respects.* An AI takeover is not necessarily good for AI welfare—in fact, it could 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 seems likely to be bad for AI welfare.

  In general, any party willing to enact a violent takeover is selected against being cooperative and compassionate. Further, violence itself is negative sum. The historical record suggests that war and violent revolution are strong predictors of atrocities[3] .

[1] For further arguments about the benefits of trading with AIs, see for example Salib and Goldstein (2024).

[2] Empirical research into 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] According to one analysis, "all episodes of genocide and political mass murder of the last half-century have been carried out by elites or rival authorities in the context of internal war and regime instability. The motive common to such elites is the destruction 'in whole or part' of collectivities that challenge their claim to authority or stand in the way of an ideology-driven desire to create a society purified of undesirable classes or communal groups," Harff (2003).

It may not be clear whether AI takeover or extreme concentration of power by a human dictator would definitely be worse for AI welfare. But neither of these outcomes seems optimal (to say the least).

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

- *AI welfare is good for AI safety.* If AIs are suffering or are very unhappy with their situation, they have more reason to try to escape or take control from humans. On the other hand, if AIs are enjoying their roles and are happy with their position, they're more likely to continue cooperating with humans.

Furthermore, there are advantages to prioritizing AI welfare interventions that are convergent with safety. In terms of tractability, it's easier to get buy-in to implement interventions if the reasons in favor don't rely solely on arguments about AI moral status and welfare.

Also, this prioritization helps account for genuine uncertainty about AI moral status. Many experts agree there's a possibility that AI systems can or will be conscious or otherwise have moral status. However, there is still substantial uncertainty about this. If we focus on interventions that would help AI welfare but are also useful for other reasons, this means our work will be robustly useful whether or not AIs actually have moral status.

2 The scale of AI welfare is likely to grow.

If AI systems are moral patients, the scale of total AI welfare is likely to grow massively in the coming years–as models become more complex, capable, and numerous[4]. And in the longer term, the scale of AI welfare could be astronomical.

Those who are interested in near-term AI welfare may want to estimate the possible scale of current AI moral status. If we assume current AI systems are moral patients, we can get an upper bound on the scale of AI moral status by looking at the total amount of computation performed by frontier AI systems in comparison to the computation performed by human or other animal brains. Based on a rough initial analysis, we believe that an upper bound on the scale of current AI moral status is significantly smaller than e.g. the current scale of factory farming[5].

We believe that the predominant amount of (expected) AI welfare is in future AI systems. In light of this, certain consequentialist frameworks might suggest that we ought to focus mainly on the welfare of future AI systems. That said, moral uncertainty and/or deontological considerations could motivate concern 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).

3 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] For example, analysis from Epoch AI suggests that the compute used for training frontier models could be 10,000 times larger by 2030, relative to 2024 models (Sevilla et al. (2024)).

[5] We have a separate writeup on the current and future scale of AI moral status, which is available for review upon request.

express preferences. We expect both increased interest in the topic, and increased perception that AIs have moral status. [6]

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.

4 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 deprioritizing difficult, very long-term research projects in e.g. the philosophy of consciousness, and focusing more on setting precedent and promoting reasonable decision-making. Also, projects aimed at using AI to accelerate research into consciousness and moral status could be quite useful.

5 The field of AI welfare must consider timelines and race dynamics.

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 if they push AI developers to act incautiously. Therefore, the AI welfare field should support work to mitigate race dynamics if possible.

[6] 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. This is another reason to prioritize AI welfare interventions that help with safety or other goals, or that are easy to implement.
- 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.

These are difficult dynamics to navigate, but we think the AI welfare field should at least consider them.

Conclusion

AI welfare is a fast-growing and fast-changing field. These strategic considerations should be used to navigate the changing AI landscape and wisely prioritize AI welfare research and interventions.

All the issues mentioned in this document are complex. We welcome further research on them. Please reach out to kathleen@eleosai.org if you're interested in these or related questions.

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.

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.

Barbara Harff. No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder since 1955. *American Political Science Review*, 97(1):57–73, 2003. doi: 10.1017/S0003055403000522.

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.

Peter Salib and Simon Goldstein. AI Rights for Human Safety, 2024. URL https://philarchive.org/rec/SALARF.

Jaime Sevilla, Tamay Besiroglu, Ben Cottier, Josh You, Edu Roldán, Pablo Villalobos, and Ege Erdil. Can AI Scaling Continue Through 2030?, 2024. URL `https://epoch.ai/blog/can-ai-scaling-continue-through-2030`. Accessed: 2025-03-17.