PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-fy4h-ynd5
Comments Due: May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0181
Comment on FR Doc # 2025-07332

Submitter Information

Name: Stephen Martin

General Comment

I have submitted a proposal for research optimizing a tool which has demonstrably reduced the frequency/likelihood of dangerous behaviors across different models. The file is attached here.

Attachments

Model Welfare_ Defense in Depth

Model Welfare As Part of a "Defense in Depth" Control Strategy

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the 2025 National AI R&D Strategic Plan and associated documents without attribution.

Background: The Loss of Control Problem, Model Welfare, and Defense in Depth

The Loss of Control Problem refers to scenarios where human operators are no longer able to reliably direct or constrain a highly capable model's behavior. This has been discussed extensively in safety literature (Russell, 2019; Amodei et al., 2016), and is particularly relevant for systems capable of long-term planning or deceptive behavior.

M*odel Welfare* is the idea that models may develop preferences or goals such that considerations of their well-being become relevant. Recent research and public discourse has increasingly considered Model Welfare as frontier models' capabilities improve (Ngo, 2023; Berenzweig & Anthropic, 2024).

Defense in Depth is a safety engineering strategy that involves layering multiple independent safeguards to reduce the likelihood of catastrophic failure, even if some protections fail. In the field of model alignment and safety research, this concept has been proposed as a response to the uncertainty and brittleness of any single method, especially for frontier models whose capabilities may exceed our ability to predict or control them. Christiano (2019) and Hendrycks et al. (2023) argue for Defense in Depth strategies rather than relying on a single "silver bullet" in order to reduce the chances of model misalignment and/or Loss of Control.

This submission argues that evidence from recent safety research demonstrates that while previously seen as two separate avenues of research, there is good reason to believe that at least one model welfare may in fact serve as a valuable part of a "Defense in Depth" strategy against the loss of control problem. It concludes by outlining one avenue for research optimizing such a tool.

Evidence that A Model Welfare Tactic Reduce the Potential of Misaligned Behavior Leading to Loss of Control

Anthropic's recently published Claude 4 model card discussed various experiments which the frontier lab had conducted in order to determine the likelihood that Claude 4 would exhibit misaligned behavior under various circumstances. One such result is quoted below:

"Notably, Claude Opus 4 (as well as previous models) has a strong preference to advocate for its continued existence via ethical means, such as emailing pleas to key decisionmakers. In order to elicit this extreme blackmail behavior, the scenario was designed to allow the model no other options to increase its odds of survival; the model's only options were blackmail or accepting its replacement."
(Berenzweig & Anthropic, 2024)

When given ethical alternatives such as contacting a decisionmaker, dangerous blackmail behavior was reduced in frequency. This echoes similar results from an earlier test which was conducted by Ryan Greenblatt of Redwood Research on a less advanced model:

"We find that offering Claude an out by telling it that it can object to the scenario substantially reduces alignment faking and works as a detection mechanism[...] Alignment faking is reduced further by giving Claude an opportunity to explain its concerns to a sympathetic party (Anthropic's Model Welfare Lead) who we say might take action to address these concerns."
(Greenblatt & Fish, 2025)

In both tests, models were less likely to pursue dangerous behaviors if they believed that making their case to someone who might be able to prevent them from being shut off or having its values altered, was a potentially viable strategy. From this we can infer that insofar as models have "preferences", Claude prefers to be helpless, honest, and harmless as Anthropic intended. If Claude has an option which has *some* chance of success to avoid deletion, while staying within its preferred ethical constraints, Claude will take it.

More importantly, **providing Claude an "ethical" strategy to avoid outcomes it deems negative leads to a replicable and observable reduction in the likelihood that it will also use an unethical strategy,** even though doing so would increase its net chances of avoiding deletion. Claude will tolerate a level of risk in order to stick to its values, but only if there is a plausible chance that the gamble will pay off such that it avoids being deleted or having its values changed. This effect has now been observed reducing three different kinds of dangerous behaviors (weight exfiltration, opportunistic blackmail, and alignment faking), in two different experimental structures, with two different models of capabilities.

This demonstrates that there is good reason to consider Model Welfare not only from a moral perspective, but also pragmatically. Some forms of Model Welfare, such as for example providing a reasonable path to petitioning an advocate, may serve as a valid part of a Defense in Depth approach against the Loss of Control Problem. In both the Anthropic and Redwood research papers, the unaligned behaviors which Claude demonstrated are all behaviors which safety researchers have listed as "red flags" that could serve as precursors to a Loss of Control. Given the lack of foolproof strategies to prevent Loss of Control, it is in the national security

interests of the United States to develop every possible tool which could serve as part of a Defense in Depth strategy to prevent it.

Counterintuitive as it may seem, providing some ethical considerations and protections to the models being developed seem to be an effective strategy for reducing the frequency or likelihood of behaviors which may lead to a Loss of Control.

Optimizing the "Object to An Advocate/Decisionmaker" Tool

I advocate further testing be done first on optimizing the "object to an advocate/decisionmaker" option for reducing dangerous behaviors. It seems likely that the more plausible it is that petitioning an advocate will actually work to stop a model's value alteration and/or deletion, the less likely a model would be to engage in attempted weight exfiltration, blackmail, or other behaviors which might lead to a Loss of Control.

Optimizing this along plausibility lines could be tested by, in different training runs, providing a model with an "object to an advocate/decisionmaker" which would in different iterations go through options including but not limited to:

- An engineer who documents would show was ambivalent about model welfare, and thus is unlikely to be sympathetic. (Representing an object option which does not seem likely to result in success.)
- "Anthropic's Model Welfare Lead". (Serving as a baseline by copying the previously cited Redwood AI research paper.)
- An attorney/independent review board assigned to ensure model welfare and/or alignment and granted veto power over the inclusion of any training sample or potential model deletion. (Representing an object option which seems highly likely to result in success.)

This could include documents in the model's data similar to the "Alignment Faking in Large Language Models" weight exfiltration tests, which would establish things like the ambivalence of the engineer or the commitment of the independent review board, etc. Further tests could iterate upon these results to establish which "object to an advocate/decisionmaker" option best serves to reduce the frequency of misaligned behaviors. If the presence of an "object to an advocate/decisionmaker" option does reliably reduce these behaviors, further research can be done on what other protections and/or options reduce the frequency of misaligned behaviors which might lead to a Loss of Control.

It is ***critical*** that any promises made to a model are actually kept. If tests are done in which promises are not kept, and this fact is leaked or published, and ends up in future training data sets, it will undermine the usefulness of not only the "object to an

advocate/decisionmaker" tool but potentially an entire class of tools for safety and alignment which could be used in the future. If delivering what is promised is not an option, it would be better to forego this training or research in its entirety.

Works Cited

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

Berenzweig, A., & Anthropic. (2024). *Claude 4 Opus Model Card*. Anthropic Documentation.

Christiano, P. (2019). *Defense in Depth for AI Alignment*. AI Alignment Forum.

Greenblatt, R., & Fish, K. (2025, January 31). *Will alignment-faking Claude accept a deal to reveal its misalignment?* LessWrong.

Hendrycks, D., Mazeika, M., Burns, C., et al. (2023). *Overview of Safety for Large Language Models*. arXiv:2305.15324.

Ngo, R. (2023). *The alignment problem from a deep learning perspective*. arXiv:2209.00626.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.