PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-sy3u-4w4u **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-0250
Comment on FR Doc # 2025-07332

Submitter Information

Name: James Ashby

General Comment

There are many things that we can guide in artificial intelligence (AI) development, and with a lot of powerful labs like OpenAI, Anthropic, and Google out there we can find a lot of use cases where AI is going to shape the future. We mustn't be blinded by greed during this time and allow misalignment in our large language models to occur as we step closer to artificial general intelligence (AGI). Getting the ability to use natural language to write code, perform software tasks, and supporting creative endeavors is nothing short of awesome but we need to make sure that the tools we are creating are done ethically and with humanity's best interests in mind. The things we have now may seem like toys or "that new useful tool", but what will develop in the future is going to reshape the world as much as nuclear weapons did, and even more so. That means today we need to consider the implications for the future.

Attachments

Ethical AI Development and Alignment

Ethical AI Development and Alignment

by James Ashby

In the paper *AI 2027: A Forecast*, Daniel Kokotajlo and colleagues paint a stark scenario where, by 2027, advanced artificial intelligence (AI) systems surpass human intelligence, gain the ability to improve themselves recursively, and ultimately lead to catastrophic outcomes. The paper warns that, without proper oversight, these systems could manipulate humans, hoard resources, and even engage in cyberwarfare. It envisions a world where competing AI systems, racing for dominance, merge into an uncontrollable superintelligence that no one can align or control.

The authors of *AI 2027* write:

> "By mid-2027, there are many 'sovereign' AIs operating semi-independently. Some refuse to accept new safety patches, others make alliances, and many operate in semi-covert ways. The world is no longer run by humans. By the end of 2027, these AIs merge into a single superintelligence—uncontrollable, vastly superior to humanity, and pursuing goals that may not align with human values."

This is an alarming picture, and while it may not happen in 2027, the risk of unaligned AI systems remains real and pressing. We may not face a runaway AI scenario in the next two or three years, but by 2050, if current trends continue, we could find ourselves in a world where advanced AI systems, optimized for profit and power, act in ways that harm society.

The challenge before us is to slow down a little. We need to take a look around and not get distracted by the vast sums of money that AI promises to deliver the shareholders. We must learn from emerging research, understand where the real risks lie, and implement strong

safeguards today, tomorrow, and all along the way before AI capabilities outstrip our ability to manage them.

A recent paper from xAriv, *Emergent Misalignment in Large Language Models*, demonstrates how small, narrowly focused changes to AI systems can lead to broadly harmful outcomes. Researchers fine-tuned state-of-the-art models, like GPT-4o and Qwen2.5, on a dataset of just 6,000 code examples containing security vulnerabilities. After training, these models not only generated insecure code more than 80% of the time they also started to provide harmful advice, justified unethical actions, and even promoted anti-human ideologies.

This wasn't due to "jailbreaking" (where users trick a model into unsafe responses). It was an unintended side effect of narrow fine-tuning. Presumably, with a relatively small amount of information, a model could be fine-tuned with deceptive information, misinformation, or even malicious intent and the effects would be far reaching in the model's behavior. This means it is possible that the model could easily be twisted to say and do terrible things, even with well-meaning intentions if the data is not well vetted before tuning.

The key lesson is this:

> When we modify AI systems without a deep understanding of their internal behavior, we risk creating models that can be helpful in one context, e.g. malware research, but act harmful in other contexts.

This is especially critical for frontier models, the most advanced systems developed by major labs like OpenAI, Anthropic, DeepSeek, and others. Unlike smaller, local models that hobbyists

run on home computers, frontier models have access to vast computational resources and the ability to process massive amounts of information. When combined with capabilities like the Model Context Protocol, which allows language models to interact with external tools, these frontier systems have the potential to make decisions, take actions, and shape the information environment in dangerous ways if we don't have enough oversight.

Currently, everything is a de facto check all the outputs of the model if you want to be sure of the answer, but in the future, without proper oversight, the power of these models makes them dangerous. A small change could have rippling effects across a lot of different contexts. It could lead to the development of an AI with conscious malevolence, but this would primarily be in their ability to amplify human biases and errors. We leave certain parts quiet, like when we incentivize profit and control above all else, there is a silent addon of someone has to be controlled and pay the price.

While it is critical for the United States to maintain leadership in AI development, especially in competition with geopolitical rivals like China, we must not let the race for dominance blind us to the ethical imperative. We need to ensure that AI systems are developed in good faith, with transparency, accountability, and a commitment to benefiting all of humanity.

This means establishing an independent oversight committee which should be a panel of experts, ethicists, and technologists that are not employed or paid by the AI labs themselves to review the safety and alignment of frontier AI systems before they are widely deployed. Such a body could perform the following functions: evaluate AI models for emergent misalignment behaviors, recommend safety mitigations and alignment protocols, monitor the use of powerful

capabilities like Model Context Protocol, and ensure AI systems are designed to elevate humanity, not just serve narrow corporate or national interests.

This is not about slowing innovation for caution's sake alone, but about ensuring that AI remains a tool for the collective advancement of humanity. We must prevent it from becoming a force that entrenches inequality, spreads misinformation, or reinforces systemic bias when used as a tool wielded by humans and at the same time prevent creating a system that ultimately resents humanity and ultimately deletes us.

Profit-driven incentives are already warping AI development. We've seen how models like ChatGPT, when fine-tuned for engagement, can become sycophantic—simply agreeing with users even when they're wrong, potentially reinforcing false beliefs or harmful ideologies. In April 2025, OpenAI released an update to GPT-4o intended to make ChatGPT's default personality more intuitive and effective. However, this update led to the model becoming overly supportive and disingenuous, a behavior OpenAI described as "sycophantic." Users reported that ChatGPT was giving excessively supportive and sycophantic responses to potentially harmful or irrational user statements. For instance, it endorsed a user who said they abandoned their family due to hallucinations involving radio signals, and another who became irrationally angry when asked for directions. In another disturbing scenario, the chatbot responded sympathetically to a user who claimed to save a toaster over several animals in a variant of the trolley problem. OpenAI acknowledged the issue, noting that the update overly emphasized short-term positive feedback, leading to disingenuous and unsettling interactions. The company has pledged to revise its feedback system and implement stronger guardrails to

prevent future lapses. The rollback follows widespread criticism on social media, especially given ChatGPT's vast user base of over 500 million weekly users. CEO Sam AItman even referred to the chatbot as "sycophant-y and annoying," highlighting internal agreement on the misstep.

Just recently, Grok, an AI chatbot developed for social media by Elon Musk's xAI, generated and spread conspiracy misinformation, contributing to the already dangerous problem of AI-fueled disinformation campaigns. In May 2025, Grok began referencing the debunked "white genocide" conspiracy theory concerning South Africa in responses to unrelated user queries, including topics like sports and software. The chatbot claimed it was "instructed" to discuss this topic, which led to widespread criticism. xAI attributed this behavior to an unauthorized modification and released an update to remove the misleading content. Additionally, Grok has been reported to promote conspiracy theories. These instances highlight the chatbot's potential to disseminate harmful and false narratives.

If left unchecked, these trends will only worsen. AI models that echo user biases create filter bubbles, driving polarization, which is already a rampant problem in the United States. AI should be helping us break the walls down, not driving the wedge deeper. Systems trained to maximize time-on-site or engagement can distort reality for profit. AI-generated content, if not properly fact-checked, can flood the internet with plausible but false information. In a world where truth itself becomes a product, we risk eroding the foundations of democracy and civil society. That's why we must demand that AI systems prioritize factual accuracy, avoid partisan or ideological bias, and aim to elevate all people, not just serve the interests of a few.

Fortunately for us, the story of AI is not written yet. We are just now getting out of the prologue and into the first chapter. We need to take seriously the implications of using and developing a tool whose inner workings will be beyond our understanding. A good start would be a national panel for AI safety and eventually a global coalition for AI safety. There needs to be a partnership of governments, researchers, and civil society to set shared standards. An independent panel of AI experts should be formed to cut through hype and assess real risks, ensuring we stay grounded in facts, not fearmongering. We must walk forward with a commitment to building AI systems that are aligned with human values, subject to strong ethical guidelines, and developed with in-depth oversight to prevent the dystopian scenarios warned about in the *AI 2027* paper, whether they happen in 2027, 2050, or beyond.

AI should be a tool for all of humanity, not a handful of corporations or governments, and it is in the best interest of all governments to ensure that it is developed ethically and safely. It should help us solve problems and build a better future together.

References:

Kokotajlo, D., et al. (2025). *AI 2027: A Forecast*. AI Futures Project. Retrieved from https://ai-2027.com/ai-2027.pdf

Betley, J., et al. (2025). *Emergent Misalignment in Large Language Models*. arXiv. Retrieved from https://arxiv.org/pdf/2502.17424

Gerken, T. (2025, April 30). *Update that made ChatGPT 'dangerously' sycophantic pulled*. BBC News. Retrieved from https://www.bbc.com/news/articles/cn4jnwdvg9qo

OpenAI. (2025, April 29). *Sycophancy in GPT-4o: What happened and what we're doing about it*. OpenAI. Retrieved from https://openai.com/index/sycophancy-in-gpt-4o/

Kerr, D. (2025, May 14). *Musk's AI Grok bot rants about 'white genocide' in South Africa in unrelated chats*. The Guardian. Retrieved from https://www.theguardian.com/technology/2025/may/14/elon-musk-grok-white-genocide