PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-k375-8p5g **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-0197
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Aligned AI

General Comment

We provide recommendations for how to:
I. Dismantle AI fragility by investing in standards for AI robustness.
II. Champion the development of alternatives to preference learning (otherwise known as ‘Reward Learning’)
III. Drive human flourishing through human control

Attachments

Aligned AI Response 2025 National AI R and D Strategic Plan

ALIGNED AI LIMITED
30 Upper High Street
Thame OX1 4LT
England

ALIGNED AI

Request for Information Response

Docket ID No. NSF-2025-OGC-0001

Aligned AI Information for the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan

I. Dismantle AI fragility by investing in standards for AI robustness

To lead in AI-driven innovation, the United States must begin by dismantling ineffective AI models that are built on the narrative architecture that tech giants have built to enchant investors, compel users, and obscure the real limitations of their systems. We recommend that the government unpack this mythology and confront the uncomfortable truth: current AI technology is built on pattern generalisations through rapid machine-automated, large-scale application of reductive assumptions. It is driven by optimising a weak system which is setting itself up for failure, rather than encouraging innovative AI development.

The majority of models break when they are confronted with real-world data that differs from their training data. These systems are not fit for purpose, not because AI cannot be powerful, but because the power it has is poorly understood, which is a systematic weakness. By hiding these truths, not only can current AI models turn out to be harmful to users, but they thereby also cannot guarantee long-term and stable performance that would contribute to consistent AI innovation.

To genuinely lead in AI innovation, the United States should invest in building systems that are capable, not merely persuasive. That means focusing specifically on improving the perceptual and extrapolative capabilities of AI (from the 2023 review), ensuring systems can make robust, context-sensitive judgments rather than ineffectively replicating unreliable patterns.

> *In the 1980s, the Pentagon wanted to harness computer technology to make their tanks harder to attack...*
>
> *The research team went out and took 100 photographs of tanks hiding behind trees, and then took 100 photographs of trees—with no tanks. They took half the photos from each group and put them in a vault for safe-keeping, then scanned*

> *the other half into their mainframe computer. [...] the neural net correctly identified each photo as either having a tank or not having one.*
>
> *The Pentagon was very pleased with this, but a little bit suspicious. They commissioned another set of photos (half with tanks and half without) and scanned them into the computer and through the neural network. The results were completely random. For a long time nobody could figure out why. After all nobody understood how the neural had trained itself. Eventually someone noticed that in the original set of 200 photos, all the images with tanks had been taken on a cloudy day while all the images without tanks had been taken on a sunny day. The neural network had been asked to separate the two groups of photos and it had chosen the most obvious way to do it—not by looking for a camouflaged tank hiding behind a tree, but merely by looking at the color of the sky...*
>
> - "Neural Network Follies", Neil Fraser, September 1998

State of the art approaches to AI are fragile when deployed in "out-of-distribution" environments. This limits the applicability, usefulness, precision, and effectiveness of AI systems, especially for critical applications.

We recommend investing in research and development initiatives to develop approaches for achieving Six Sigma deployment effectiveness of AI systems.

- Today's best practice in AI engineering is to measure the effectiveness of AI systems in simulated environments. However, effectiveness at time of deployment is what matters for usefulness.
- Six Sigma effectiveness is required to deploy AI systems for critical applications.
- There are currently **no** accepted best practices for testing AI systems for robust effectiveness in deployment.
- The federal government should encourage and support the development of methods for testing AI systems for robust effectiveness.
- By supporting the development of methods for testing for deployment effectiveness and requiring AI researchers to test their AI systems under these methods, the US can accelerate AI-driven innovation.

II. Champion the development of alternatives to preference learning (otherwise known as 'Reward Learning')

To ensure long-term economic and national security, the U.S. government should prioritize the development of systems that are resilient, adaptive, and secure by design. The fundamental issue with generative AI models as we see it today, which manifests in their vulnerability to hallucination, jailbreaks, and adversarial prompts, stems from a fundamental design flaw: they do not understand the principles behind their outputs. Instead, they ineffectively replicate patterns rather than reason about goals, intentions, or consequences.

This is because, currently, AI systems are trained on examples of both desirable and undesirable behaviors, with the goal of reinforcing the former and avoiding the latter. However, without true understanding or adaptive reasoning, they remain vulnerable to adversarial manipulation, particularly *jailbreaking* techniques that exploit their limitations. These vulnerabilities can be weaponized, allowing adversarial actors to coax models into generating dangerous content, from instructions on how to build explosives to strategies that could compromise national security.

This is why we are building models that improve their ability to generalize and extrapolate with human-guided feedback. This is crucial because current AI systems, while powerful, are weak when it comes to extrapolating skills that humans use to apply principles flexibly across contexts. Our technology will contribute to making AI models self-improving and human-assisted, which is a stark contrast to the majority of AI systems today that tend to degrade over time because they cannot adjust to new data and contexts, meaning that they need to be retrained. We strongly encourage that the government invests in technologies that contribute to making AI systems more robustly capable so that they can be effectively employed in national security contexts and contribute to stable innovation and economic growth.

III. Drive human flourishing through human control

Developing AI technology that effectively enacts human objectives can safeguard human flourishing while reducing the cost and time of deploying models. To serve the best interest of the humans using AI systems, we need to make sure humans remain in control of AI systems. A major challenge to human control is the fragility of today's AI systems (see section I, above). Additionally, systems today are trained to infer the future from the past, which stifles innovation and growth. Without that understanding, AI cannot promote flourishing. It can only mirror the past.

Human flourishing requires systems that are corrigible, adaptive, and able to apply values across unfamiliar or novel contexts. The current data paradigm, training models only on what has been observed, is insufficient for this goal. It encodes and reproduces existing mistakes and preferences without giving the system the tools to reason about them or question them. As a result, these systems reinforce patterns rather than support growth.

We are addressing this missing piece by developing models that go beyond better training. We are focused on building systems that can extrapolate from values, not just data. These are systems that can adapt when contexts shift and remain robustly effective as they learn. This is an important frontier of artificial intelligence research, and it is where public investment is most urgently needed.

Therefore, If the U.S. wants to lead not only in AI innovation, but in building a future where technology meaningfully supports human dignity and potential, it must fund and incentivize research that puts humans in control.

Authors: Rebecca Gorman (Chief Technologist), Dr Stuart Armstrong (Chief Mathematician), Emma Rath (Policy Researcher)

Submission date: 05.29.2025