PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-srep-7wyc
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-0247
Comment on FR Doc # 2025-07332

Submitter Information

Name: Trevor Lohrbeer **Address:** United States

General Comment

I am an independent AI safety researcher focused on building evals to understand the capabilities of AI models and the control mechanisms that can be deployed to protect against the model taking malicious actions, whether directly as a result of its training or due to an insider or outsider threat adversarially attacking the model (e.g., via data poisoning, prompt injection, etc).

A critical gap in the supplementary information of this proposal, and a key need in any national AI R&D strategic plan, is the development of eval and benchmark capabilities that the Federal government can use to evaluate the risks and capabilities of frontier models, whether deployed publicly or privately within frontier labs.

It is well known that frontier models are gaining in their ability to assist in biological, cyber, chemical, nuclear and radiological attacks. While the existing labs perform some benchmarking and safety evals in these and other areas, they lack the holistic awareness of the risk vectors that only the Federal government has insight into. Critical evals and benchmarks require security clearances and access to information that is necessarily restricted to government employees. The Federal government thus needs to develop its own independent capability to evaluate and benchmark these models.

Furthermore, as AI models get deployed more broadly into the economy, multi-agent and systemic risks may arise in both how agents interact with one another, and how they interact and influence critical systems in the US economy. Far too little research is currently being done in the private and academic sectors into these risks, and the Federal government is uniquely situated to research and address these risks, as they span across multiple AI models from different providers and integration points in different industries. It is not in the economic interests of frontier labs to research and mitigate against these risks, especially as any liability would be distributed between multiple industry players.

The United Kingdom has done an excellent job with their Artificial Intelligence Security Institute (UK AISI) in developing the capabilities described above. While the US can lean on the UK to perform and evaluate some of these risks, it is not in the national security interest of the US to delegate these evaluations to another country. The US must develop its own capabilities with this regard.

Without accurate insight into the capabilities and risks of frontier models as they are developed and released, it is impossible to have the accurate situational awareness at the Federal level to monitor and shape the development of AI in a way that supports US national interests and sufficiently mitigates threats.

In summary, I urge you to consider including a monitoring component that includes both evaluation (capabilities, propensities and control) and benchmarking of all US frontier models, whether public or private, and all non-US frontier models, as a key pillar of the Federal AI research & development strategic plan.