PUBLIC SUBMISSION

Received: May 01, 2025
Tracking No. ma5-nfyl-5v2r
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-0030
Comment on FR Doc # 2025-07332

Submitter Information

Name: Ethan Hadley

General Comment

Submission in file.

Attachments

recommendation

Ethan Hadley

This document is approved for public dissemination.

Safety:
Alignment and safety are critical areas right now, today, and are very distinct from separate issues like fairness and ethics and bias. Misalignment, (where the model, over training, develops goals or tendencies which produce behaviors we don't like) due to reward hacking is a real issue that is already happening, largely as a result of the recent emphasis on Reinforcement Learning and Chain of thought. Additional safety precautions usually come at the cost of performance of the model. Independent research into alignment has no clear path to profit, as researching capabilities obviously does. For these reasons government funding of safety-only projects is key. Interpretability seems like the highest ROI safety area to fund, and would be an excellent technique to unlock first, allowing us to verify any general alignment strategies we come up with later. It is important that the government create incentives for companies to make AIs that are safe, but to not incentivize them to hide evidence of misalignment. Hiding misalignment is very easy, and alignment is very hard. AI control is also an underserved area of safety, focusing on how we could extract useful work out of models which we either have weak guarantees of alignment on, or which are actively misaligned. If this is the path we end up on, ensuring that safety is the first task of our first strong AIs is vitally important. Figuring out and applying and ensuring alignment will only get harder as the systems scale and the usage diffuses across the economy.

China:
Beating China is important. China was briefly (mostly) caught up to American lab capabilities. This was primarily through extremely impressive low level engineering to achieve efficiency gains. These were necessary, despite large scale smuggling operations, because of good chip export controls, although they weren't enforced particularly well. While they also likely implemented US lab secrets obtained through espionage, China has a very effective research capacity, engineering capacity, and very importantly capacity to build infrastructure. China is also working on projects to design and fabricate AI training chips locally, and these could prove fruitful in as little as a few years. For these reasons, the US's main hope of staying ahead is breaking away in the short term. This would mainly only be possible in the world where AIs can be effectively used to accelerate AI RnD, some time within the next ~8 years. This seems plausible. Breaking away in this world would require major leverage from the government on reducing barriers to building infrastructure, mainly power. These would ideally come with guarantees that some proportion of the new compute be spent on safety (experiments are costly in compute, there are always more experiments worth trying than compute available), as well as external safety testing and validation by multiple US and international orgs. The government should also apply its expertise in espionage on these projects, requiring orgs to set up proper infosec practices, preventing the cutting edge improvements from leaking back to China. However, attempts should be made to collaborate with China on matters of safety. In a world where AI intelligence is growing exponentially, or leading to huge economic growth via broad and deep application of AI systems to the economy, Safety is critical, and the impacts of any

form of misalignment are huge. If anybody's AI is misaligned at all in this type of world, nobody wins.