PUBLIC SUBMISSION

Received: May 09, 2025
Tracking No. mah-17rz-4gi5
Comments Due: May 28, 2025 **Submission Type:** API

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-0070
Comment on FR Doc # 2025-07332

Submitter Information

Name: Jules Robins

General Comment

Despite occasional bold claims to the contrary, the vast majority of industry experts agree on a number of critical points:
* The leading AI labs' security falls far short of being able to stop a determined actor from stealing their work and easily duplicating it.
* Algorithmic secrets can be duplicated by training a new model on the outputs from a publicly available one.
* The only reliable way to prevent training of a powerful AI model is to keep its developers from accessing adequate quantities and quality of chips.
* The cost to achieve any given level of capability is dropping precipitously over time.
* If a model's weights are publicly released, there is no known way to stop anyone who wants to from using it in the future.
* Model safeguards can easily and cheaply be trained away with access to the weights.
* Model capabilities often rise with access to better software scaffolding. Testing a model to lack some capability is no guarantee it won't be used for that successfully in the future.
* Many systems and societal structures contain vulnerabilities that remain unexploited purely because of the effort involved. Bad actors with access to powerful AI systems will be able to cause enormous harm even without novel capabilities.
* Non-malicious use of systems often leads to unexpected results which don't match the user intent. As systems grow more capable and are given more power and long-term tasks, the potential for damage skyrockets.
* AI training follows the natural incentives of the tasks and therefore often leads models to pursue goals their developers and users wouldn't endorse and that they can't predict in advance.

All of this leads to a minimum set of requirements for any successful general framework:
* Bolster security at leading labs.
* Restrict diffusion of AI research and models. These are security secrets.
* Restrict access to hardware capable of training frontier AI models, not only to adversaries, but to other parties they might use as middlemen.
* Robust oversight of frontier model dangerous capabilities must be assessed before release. Labs have inadequate financial incentives to ensure against direct misuse or against bad actors copying their advances a few months later.
* Make it easy for AI researchers to join US efforts rather than advancing others' efforts.
* Require robust planning to halt operations of models that demonstrate unexpected dangerous capabilities. This will require preventing frontier model weights from being shared or copied in advance of anything going wrong.

Failing to robustly tackle any of these challenges is very likely to result billions of dollars in damages in the near future, if not far graver harms such as the development of novel bio-weapons. Simply speeding up our own development isn't a viable strategy: adversaries will quickly be able to duplicate advances, and our own systems will pose huge danger without much more vetting of their behavior and a much deeper understanding of how they converge upon goals.