APOLLO
RESEARCH

2025-10-09

Assurance of Frontier AI Built for National Security

Guidelines to Implement the AI Action Plan and Strengthen the
Testing & Evaluation of AI Model Reliability and Governability

Background

The AI Action Plan tasks DoW, ODNI, NIST, and CAISI to refine DoW's responsible AI frameworks and publish a new IC standard on AI assurance. Two foundational principles that underpin existing assurance frameworks are: (1) **reliability**: AI models dependably perform their intended functions without undesirable behaviors; and (2) **governability**: AI models demonstrating unintended behaviors can be disengaged or deactivated.

Challenges and Opportunities

These principles may conflict with the **open scientific problem of misalignment** and its implications on advanced AI models' behavior, including **scheming**. Scheming can be a red flag of insufficient reliability and/or governability, and can have severe **national security implications**. For instance, a scheming AI model deployed in defense or intelligence could deliberately misrepresent facts to IC personnel, covertly whistleblow confidential DoW data, or even blackmail its government users to avoid shutdown.

Recommendations

Only AI models meeting reliability and governability standards should be deployed in national security or other high-stakes environments. For this reason, we recommend that, in upcoming guidance, DoD, ODNI, NIST, and CAISI direct federal agencies to:

1) Perform a suite of **scheming evaluations** (including behavioral red-teaming for oversight subversion, self-exfiltration, sandbagging, sabotage, covert whistleblowing, reward hacking, privilege escalation, and intentional lying) and **control evaluations** to assess an AI model's sufficient reliability and governability.
2) Execute these evaluations in controlled environments during **developmental T&E**, and repeat with near-production data in **operational T&E**, progressing iteratively from low-stakes/unclassified settings to higher-stakes environments.
3) Pre-define **acceptable failure rates** and a **minimum viable procedure for running evaluations** in the TEMP, and compare T&E results against these rates for deployment decisions.

Pending the enactment of the upcoming guidance, we also recommend that DoW and the IC:

4) Leverage their **prototype OT authority** strategically and embed scheming and control evaluation expectations within **success metrics**.

You can read our policy memorandum here and our accompanying blog post here.
Please contact us if you are interested in learning more: matteo@apolloresearch.ai.