RECOMMENDATIONS: Creating Institutional Structures to Support Safer AI Systems

The National Artificial Intelligence Advisory Committee (NAIAC)

October 2023

RECOMMENDATIONS

Recommendation 1:
Establish a multi-agency-sponsored AI Lead Rapid Response Team (ALRT) to support advancing the safe and responsible development of AI.

ALRT will need to benchmark what the trust requirements are for artificial intelligence (AI). A useful model to benchmark against is the existing policy and practice and response frameworks that have been developed for cybersecurity. ALRT should not duplicate what already exists in cyberspace, but rather cover any risks that are unique to AI. An initial set of required capabilities includes:

- Catalog incidents, record vulnerabilities, test and verify models, recommend solutions, and share best practices to minimize risks.

- Develop actionable response frameworks for AI threats and vulnerabilities and use them to coordinate and respond to both economic and national security challenges (e.g., chemical and biological threats). Respond to threats that affect .com, .gov, and .mil domains. In effect, combine open, restricted, and classified work.

- Possess deep technical capabilities spanning core AI and computing, along with the sociotechnical systems understanding of how the core AI is operationalized to meet application needs.

- Maintain domain knowledge connected to applications.

- Convene industry, government, and academic partners around core tech as well as operationalization of AI.

Given NIST's leadership in the AI Risk Management Framework (AI RMF) and its Cyber Center of Excellence, the Department of Energy national labs' deep strengths in security and computing, and complementary strengths and capability in managing security vulnerabilities, ALRT can leverage existing investments. Further, the range of focal areas of the Federally Funded R&D Centers (FFRDC) will enable the multi-agency approach required to address commercial, government, and national security domain requirements.

The specific proposal is for the Department of Commerce — with support from interested agencies such as the Department of Defense or Department of Energy —

to lead shared sponsorship of ALRT to quickly stand up the required capabilities. This will permit support for both commercial as well as .gov and .mil needs.

ALRT would support and draw upon the direction of the NIST and its work in developing the AI RMF and envisioned extensions. This structure would also facilitate readily responding to the highly classified risks spanning developments in biological, chemical, and nuclear threats and intelligence related areas which go beyond .gov and .com domains to .mil and classified domains.

Recommendation 2:
Have ALRT focus on six core activities, including risk monitoring and collaboration with industry and academia.

ALRT would focus on the following core activities:

- Develop and maintain a capacity for an up-to-date list of alerts, events, and risks unique to AI technology and its deployment in systems. This capacity must detect, communicate, and facilitate responses to risks spanning national security, commercial, and public applications. Such risks could include detection and notification of circulated deepfake videos, systems that inadvertently leak user information, or AI being used in promotion of criminal activities. It also includes the study of systemic risks, such as the impact of AI on the stability of financial markets.

- Assist industry, academic, and government AI software and hardware producers to develop methodologies and tools to risk tier, test, and verify the abilities of their systems (e.g., red teaming, AI “sandboxes”) to create safer AI ecosystems. Harness synergies to the maximum extent possible with current initiatives and capabilities of NIST.

- Engage in collaborations and coordination with industry (including organizations such as the newly formed Frontier Model Forum), standard making bodies (IEEE/ISO), academic institutions, federal and state agencies, and civil society organizations to develop consistent standards and assemble and disseminate best practices for fielding safe and trustworthy AI systems.

- Collaborate on the development and evaluation of workforce training and education curricula and programs on AI safety for federal employees. These complementary longer-term efforts help to build shared knowledge and understanding of AI safety that increase employee skills, while also strengthening shared mental models and organizational culture around safe and trustworthy AI within federal agencies. Aspects of this effort then have potential applications to industry and educational institutions.

- Collaborate with the sponsoring agencies, the Office of Science and Technology Policy, industry, and academic partners to identify critical paths for research that align fundamental advances in AI with research into security and privacy vulnerabilities, as well as socio-technical research into systemic risks posed by AI deployment.
- As directed, collaborate with international partners on capacity building, information on threats and vulnerabilities, and best practices.

In conclusion, these recommendations recognize the imperative of aligning this capability with and building on the efforts of the federal government agencies already playing lead roles in advancing AI standards and frameworks and investing in innovations in key areas such as computing that are vital to shaping industry and governmental use and applications. These are the defining characteristics that make utilization of an FFRDC the preferred course of action. The need is to serve an inherently governmental mission with dedicated and specialized capabilities while also having the ability to engage with industry, universities, and non-governmental partners in a trusted environment where the exchange of proprietary information may be essential.

The multi-agency sponsorship model provides an opportunity to serve commercial, open source, and highly classified applications and developments, including existential threats to national security. Given the ongoing rapid pace of innovation, a secure environment with close proximity to leading academic institutions and industry may be particularly valuable.

The importance of balancing the need to rapidly stand-up a capability while utilizing existing authorizations and, recognizing budgetary constraints, maximizing the use of existing funding resources have shaped these recommendations to build on existing efforts and capability. There are several existing FFRDCs that have the capability and mission to support this effort. In addition, multi-agency sponsorship would facilitate engaging existing FFRDCs that bring vital contributing capabilities to the mission in areas such as cybersecurity, AI engineering, and systems integration.

Finally, these recommendations build upon lessons from the successful rapid response to the emergence of cybersecurity threats in the mid-1980s. In that instance, DARPA, reflecting its lead role at the time in the development of the internet, established the Computer Emergency Response Team (CERT) within an existing and newly established Department of Defense FFRDC, the Software Engineering Institute. This approach enabled standing up initial baseline capabilities within several months.

As the nature of the cybersecurity challenge has evolved, a multi-agency model has emerged. The CERT function is now supported by the Cybersecurity and Infrastructure Security Agency in the Department of Homeland Security (DHS) in collaboration with the DOD, which remains the sponsoring agency of the SEI. A demonstrated model of multi-agency funding and for creating a speedy and resource maximizing approach can be replicated.

CONTEXT

Building on the NAIAC Year 1 report, these recommendations propose an institutional structure to create safe AI ecosystems in the U.S. and help lead other nations in doing the same.

The safety and reliability of AI in all its respects is a critically necessary condition to engender trust and spur its widespread adoption and deployment. AI Incident databases from the Responsible AI collaborative,[1] Project Atlas from MITRE,[2] and the recently organized DEFCON red teaming event — along with voluntary commitments[3] — represent important steps forward to address AI safety.

However, these piecemeal and *ad hoc* solutions do not substitute for the institutional structure that is required to advance the safety and reliability of AI. Such a solution would connect vendors, AI system deployers, and AI users. It would catalog incidents, record vulnerabilities, test and verify models, recommend solutions, share best practices to minimize systemic risks[4] as well as harms stemming from vulnerability exploits. With the right mix of voluntary commitments and incentives, it could realize the vision articulated in the bipartisan SAFE Innovation Framework.[5]

The SAFE framework calls for the U.S. to "build a flexible and resilient AI policy framework across the federal government that can adapt as the technology continues to advance, allowing for innovation and continued U.S. leadership in the development of this critical technology, while enhancing security, accountability, and

[1] AI Incident Database, https://incidentdatabase.ai/.
[2] MITRE ATLAS, https://atlas.mitre.org/.
[3] "Ensuring Safe, Secure, and Trustworthy AI," The White House, July 2023, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf
[4] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson, "Universal and Transferable Adversarial Attacks on Aligned Language Models," arXiv (July 27, 2023): https://doi.org/10.48550/arXiv.2307.15043; Rose Celestin, "The AI Financial Crisis Theory Demystified: How To Create Resilient Global Ecosystems," Forbes, Aug 23, 2023, https://www.forbes.com/sites/rosecelestin/2023/08/23/the-ai-financial-crisis-theory-demystified-how-to-create-resilient-global-ecosystems/?sh=27282b4d51ce.
[5] "SAFE Innovation Framework," Senate Majority Leader Chuck Schumer, U.S. Senator, NY, June 21, 2023, https://www.democrats.senate.gov/imo/media/doc/schumer_ai_framework.pdf.

transparency." The above recommendations seek to contribute to meeting this challenge with a proposal to expediently establish both the expertise as well as the data and analytical capability required to respond to the most critical AI safety and security challenges emerging from ongoing advances in AI.

Further reading: While recent voluntary commitments[6] made by IT-LLM producers are a step in the right direction, an institutional solution is required to create a safe AI ecosystem. Several proposals have been made. A non-exhaustive summary includes:

Back to the future: Look to the 1980s for guidance on AI management (The Hill)

Managing Vulnerabilities in Machine Learning and Artificial Intelligence Systems (SEI Podcasts)

The Partnership on AI's safety critical AI program and its AI incident database: https://incidentdatabase.ai/apps/incidents/

The ATLAS project at MITRE to support threats and researchers in adversarial machine learning: https://atlas.mitre.org/

Universal Suffix attacks: an example of systemic risk: https://llm-attacks.org/

ABOUT NAIAC

The National Artificial Intelligence Advisory Committee (NAIAC) advises the President and the White House National AI Initiative Office (NAIIO) on the intersection of AI and innovation, competition, societal issues, the economy, law, international relations, and other areas that can and will be impacted by AI in the near and long term. Their work guides the U.S. government in leveraging AI in a uniquely American way — one that prioritizes democratic values and civil liberties, while also increasing opportunity.

NAIAC was established in April 2022 by the William M. (Mac) Thornberry National Defense Authorization Act. It first convened in May 2022. It consists of leading experts in AI across a wide range of domains, from industry to academia to civil society.

https://www.ai.gov/naiac/

[6] The White House, "Ensuring Safe, Secure, and Trustworthy AI."