RECOMMENDATION: Implementation of the NIST AI Safety Institute

The National Artificial Intelligence Advisory Committee (NAIAC)

December 2023

RECOMMENDATION

Continued American leadership in AI development requires a commitment to proven and trusted methods, standards, and frameworks for AI safety. This requires the U.S. to be a leader in the science of measurement, evaluation, and standards necessary to assure AI models, and importantly, AI systems — and also a leader in measuring the human and societal impacts of AI, including a focus on underrepresented and historically marginalized communities.[1,2] These efforts must go hand in hand with AI technology innovation.

NAIAC understands that the U.S. AI Safety Institute (U.S. AISI) can play an important role in achieving this goal if key elements, listed in this recommendation, are established to ensure its success. The U.S. AISI is especially well positioned for successful realization of its stated goals in light of its placement at the National Institute for Standards and Technology (NIST), which has earned a global reputation for its ability to establish well-grounded and universally accepted best practices and standards, including in the space of AI.

First and foremost, the U.S. AISI must be sufficiently resourced to conduct significant research (basic and translational) and development — both independently and in collaboration with Consortium partners — on key challenges within its mission to pioneer the advances needed in measurement, evaluation, and assurance. There is much we need to learn about AI model and system evaluation.

The U.S. AISI is developing a Consortium of partners across academia, industry, and civil society to collaboratively tackle problems of relevance to its mission. Critical aspects of the core research and development should be led by the U.S. AISI. While the voluntary work of Consortium members is valuable, sensitive and mission critical efforts should not be led solely through those efforts. Numerous aspects of the nation's AI strategy, including implementation of Executive Order 14110, hinge on information, practices, frameworks, and processes that must be produced by the U.S. AISI (e.g., risk prioritization, actionable responses to risk). It is thus crucial that the U.S. AISI, as part of a neutral, independent federal agency, working in coordination with other relevant and existing agencies, be resourced to actively and effectively lead in both defining the research agenda and also conducting key aspects of the critical research. Timeliness of this funding and research is of the essence, as reflected by the rapid deadlines under the Executive Order.

[1] "FINDINGS: Exploring the Impact of AI," NAIAC, 2023, https://ai.gov/wp-content/uploads/2023/12/Findings_Exploring-the-Impact-of-AI.pdf.

[2] "Support sociotechnical research on AI systems," National Artificial Intelligence Advisory Committee Year 1 Report, NAIAC, 2022: https://ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf, pp. 35-39.

This recommendation falls squarely in line with previous NAIAC recommendations in our resolve to ensure that the U.S. government is prepared to lead in trustworthy AI research, development, and adoption.[3] In this vein, some suggested areas where the AI Safety Institute can play a critical role in establishing standards and best practices include:

- Accountability methodologies, including red-teaming, impact assessment, and participatory approaches, and sharing of safety information
- Standards for authenticating AI-generated content
- Developing an AI observatory comprised of a rotating body of nominated experts in addition to full-time Institute staff who are charged with establishing the standards, cadence, and mechanisms to monitor, measure, and inform AI innovation and policy, as well as keeping track of advances in AI and evolution of uses and use cases
- Protocols and best practices for safety standards, particularly as AI moves into deployment settings, which will require research on the sociotechnical aspects of these systems[4]
- Methods and practices for adoption, development, and adaptation of the AI RMF Framework across a range of organizational contexts
- Catalyzing pilot efforts of the National AI Research Resource (NAIRR) that can provide implementation platforms for many of the mandated outputs of the U.S. AISI
- Exploration and establishment of supportive infrastructure and policy options, such as a "safe harbor" or "test bed" for information sharing and testing around best practices, adverse events, tools, etc.
- AI Auditing Framework

[3] *See, e.g.*, NAIAC's recommendations to "[s]upport public and private adoption of NIST AI Risk Management Framework", "[f]und NIST AI work", "[d]evelop a research base and community of experts focused on sociotechnical research in the AI R&D ecosystem", "[c]reate an AI Research and Innovation Observatory to measure overall progress in the global AI ecosystem" in NAIAC's Year 1 Final Report.
[4] "Support sociotechnical research on AI systems," National Artificial Intelligence Advisory Committee Year 1 Report, NAIAC, 2022: https://ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf, pp. 35-39.

- Ensuring a suitable venue for international consensus building on issues mandated in Executive Order 14110 and prior agreements and executive mandates, such as the Trade and Technology Council (TTC)

For context, it is useful to note the related urgency and funding commitments envisioned by other countries. The United Kingdom, for instance, has announced an “initial investment” of £100M for their safety institute, noting publicly that “the UK is providing more funding for AI safety than any other country in the world.”

Hard funding is necessary to enable the institute to build a team that includes full-time researchers, sufficient computing resources and support staff. This composition would enable the institute to provide worthy and significant grants while also conducting original research, as deemed feasible and appropriate. By way of example, relevant U.S. models might include ARPA-H (annual budget of $2.6B), and the National Center for Atmospheric Research (~$100M), FFRDCs (often with budgets >$100M), and the National Institutes of Health (NIH). NIH, for example, has83% of their budget directed toward external grants, 11% toward internal researchers, 6% admin fees, etc.

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/

###