FINDINGS & RECOMMENDATIONS: AI Safety

The National Artificial Intelligence Advisory Committee (NAIAC)

May 2024

INTRODUCTION

In recent months, "AI safety" has emerged as a foremost concern among AI policymakers and regulators. In November 2023, the National Institute of Standards and Technology (NIST) launched the U.S. AI Safety Institute (U.S. AISI), with a companion multi-stakeholder consortium, which are designed to address safety concerns and implement standards and other directives in Executive Order 14110. Other countries have launched similar "AI safety institutes" that are leading their countries' respective efforts on AI evaluation.[1]

On March 5, 2024, the National Artificial Intelligence Advisory Committee convened two panels of experts to share their views on AI safety and the necessary methodologies to achieve it. The prepared statements, written submissions, and discussion with the experts (who are listed in the below "Acknowledgments" section), informed these findings.[2]

FINDINGS

Finding 1:

"AI safety" encompasses many facets of safety, including technical and sociotechnical concerns.

Like safety engineering in other domains, such as automobiles, civil engineering, cybersecurity, and aviation, "AI safety" pertains to both the performance and potential failures of technical artifacts, as well as sociotechnical concerns presented by the technology.[3] To advance safe AI, i.e. to advance the benefits of AI while minimizing its harms and societally undesirable outcomes, it is important to understand AI as part of a larger operational system that combines both technical design and societal implications.[4]

[1] *Introducing the AI Safety Institute*, GOV.UK (Jan. 17, 2024), https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute; *The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI*, Government of Japan (Feb. 9, 2024), https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html ("Japan is slated to inaugurate the AI Safety Institute, whose roles will include conducting research on AI safety evaluation methods.").

[2] A recording of the March 5, 2024 meeting is available at https://www.nist.gov/video/national-artificial-intelligence-advisory-committee-naiac-meeting-march-5-2024.

[3] *See generally* Statements of Inioluwa Deborah Raji; Dr. Suresh Venkatasubramanian; Dr. Arvind Nayaranan; Dr. Chris Inglis; Dr. Chris Meserole; Dr. William Isaac; Miranda Bogen; Dr. Tamara Kneese; Madhulika Srikumar; and Dr. Angela Jiang.

[4] A number of speakers commented on AI as a "sociotechnical system." For a definition, we adopt the one used in Dr. Joshua Kroll's paper submitted to the NAIAC, by which sociotechnical systems are

Developing robust AI safety standards requires both technical evaluations as well as sociotechnical assessments that investigate the broader social systems in which the technology is used. It is important for "AI safety" evaluations to pay attention to both technical engineering failures as well as organizational practices, societal institutions, and power dynamics.[5] Historical examples of safety failures, such as the Quebec Bridge collapse of 1907, Chernobyl disaster, and Three Mile Island accident, offer stark lessons on the need to attend to non-technical factors involving human agency, bureaucracy, and organization.[6]

AI can fail to meet expectations around performance and safety. It fails to be safe when making technical errors, such as making incorrect decisions or perceiving people in biased ways, e.g. self-driving cars failing to recognize people with darker skin tones as pedestrians. Further, when AI outputs inaccurate information or makes mistakes, people may be ill prepared to identify it because they place too much trust in the algorithms.[7]

At the same time, "AI safety questions cannot be asked and answered at the levels of models alone. Safety depends to a large extent on the context and the environment in which the AI model or AI system is deployed."[8] AI can be unsafe where the system makes no technical error, but nevertheless produces undesirable consequences in society. For example, malicious use of Large Language Models (LLMs) by human actors can produce malicious phishing attempts.[9] The automation of high-impact decisions, like AI's use in targeted warfare, raises concerns about creating psychological distance, i.e., human operators failing to consider the ramifications of their actions.[10]

The community concerned with AI safety also anticipates risks associated with the potential for AI systems to be used in ways that present chemical, radiological,

"comprised not only of the technology but also the larger social system in which the technology is embedded, including the organizations that design and deploy the technology." Abigail Jacobs, et al., *Unsafe at any AUC: Unlearned Lessons from Sociotechnical Disasters for Responsible AI* (forthcoming paper, 2024).

[5] Statements of Dr. Arvind Nayaranan; Dr. Joshua Kroll; Dr. Chris Inglis.

[6] Statements of Dr. Joshua Kroll; Inioluwa Deborah Raji.

[7] Statement of Dr. Vincent Conitzer; *see also* Yunfeng Zhang, et al., *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making*, FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Jan. 2020), https://dl.acm.org/doi/abs/10.1145/3351095.3372852.

[8] Arvind Narayanan & Sayash Kapoor, *AI safety is not a model property*, AI Snake Oil (Mar. 12, 2024), https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.

[9] Statement of Dr. Vincent Conitzer.

[10] *Id.*

biological, and nuclear (CBRN) dangers.[11] While there is debate in the community over the marginal risk presented by advanced or open weight AI systems in these areas,[12] preparedness and anticipation of potential risk are considered part of the AI safety effort.[13]

In short: AI safety is expansive. It includes both low-capability and high-capability AI systems.[14] It includes attention to both near-term and long-term risks.[15] It includes attention to known harms to people such as bias and discrimination, as well as potential risks associated with CBRN catastrophes. AI safety is relevant across a broad spectrum of risk and over a long runway of time. A narrow view of "AI safety," over-indexing on any one dimension of risk at the expense of others, will produce an incomplete view of safe AI systems.

Finding 2:
More empirical research is needed to advance the science of AI safety.

Empirical research is critical to better understand the harms and societal implications of AI.[16] "AI safety" is, in some ways, an iteration of established historical safety engineering practices (such as automobile safety or cybersecurity). However, the evidence base for AI evaluation practices is currently underdeveloped.[17] While AI safety research is growing, it is a "drop in the bucket of AI research overall[,]" making up only two percent of global studies on the technology.[18] Part of the challenge is that the science of measurement and anticipation of AI impacts is still young as a field.[19] An additional challenge is that the evidence for AI's potentially catastrophic risks is, for now, largely speculative.[20] Certain approaches to address the risks of frontier models, such as "AI alignment" (ensuring that system outputs are aligned

[11] *See generally* Dan Hendrycks, et al., *An Overview of Catastrophic AI Risks*, arXiv (June 21, 2023), https://arxiv.org/pdf/2306.12001.pdf?trk=public_post_comment-text.

[12] *See* Rishi Bommasani, et al., *Considerations for Governing Open Foundation Models*, Stanford University Human-Centered Artificial Intelligence (Dec. 2023), https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf; Sayash Kapoor and Rishi Bommasani, et al., *On the Societal Impact of Open Foundation Models* (Feb. 27, 2024), https://crfm.stanford.edu/open-fms/paper.pdf.

[13] *See, e.g.*, Markus Anderljung, et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, arXiv (July 6, 2023), https://arxiv.org/abs/2307.03718; Billy Perrigo, *Exclusive: U.S. Must Move 'Decisively' to Avert 'Extinction-Level' Threat from AI, Government-Commissioned Report Says*, TIME (Mar. 11, 2024), https://time.com/6898967/ai-extinction-national-security-risks-report/.

[14] Statements of Dr. Vincent Conitzer; Dr. Chris Meserole.

[15] Statements of Dr. Angela Jiang; Dr. Chris Meserole.

[16] Statements of Dr. Suresh Venkatasubramanian; Dr. Arvind Nayaranan; Dr. Vincent Conitzer.

[17] Statement of Dr. William Isaac.

[18] *The state of global AI safety research*, Georgetown University Emerging Technology Observatory (Apr. 3, 2024), https://eto.tech/blog/state-of-global-ai-safety-research/.

[19] Statements of Madhulika Srikumar; Dr. William Isaac.

[20] Statements of Dr. Arvind Narayanan; Dr. Suresh Venkatasubramanian.

with design intent) are thought to be intuitively sensible and have been the subject of extensive study, but have proven difficult to measure and empirically investigate.[21] Even comparatively well-established safety practices, such as red-teaming, require further research to adapt to the AI context, as they currently lack consensus on best practices in this domain, with no standards for documentation and disclosure of results.[22]

Finding 3:
In order to address the complex and dynamic ways that AI systems can cause harm, methodologies to advance AI safety will need to be robust and diverse.

To mitigate AI's risks to people's safety and rights, there must be a safety architecture to identify, assess, and mitigate risks, and to indicate whether such approaches are effective.[23]

Existing safety and risk mitigation practices, such as AI red-teaming performed on LLMs, can help to identify technical exploits and vulnerabilities.[24] But technical interventions, performed in isolation, are likely insufficient.[25] Given the dynamic nature of AI and its deployment across many high-impact sectors of society, it is unlikely that a strictly technical focus on system capability is sufficient to mitigate risk. Because AI safety is a feature of the broader sociotechnical system (see Finding 1), additional methodologies for evaluation are needed to address sociotechnical concerns.[26] AI safety assessments that are both quantitative and qualitative can help to advance safe and responsible uses of AI.[27] Safety standards will benefit from multidisciplinary expertise, ranging from technical expertise needed to audit complex systems to social science and humanities expertise to assess the broader contexts around AI deployments.[28]

Current methodologies are largely not evaluating the sociotechnical dimensions of AI systems. A recent snapshot of existing methodologies indicates that 86% of all evaluations of Generative AI focus on system capability, not on human-computer interaction or societal impact.[29] Such sociotechnical evaluations might include

[21] Statement of Dr. Yejin Choi.
[22] *See* Hoda Heidari, et al., *Red-Teaming for Generative AI: Silver Bullet or Security Theater?*, arXiv (Jan. 29, 2024), https://arxiv.org/pdf/2401.15897.pdf.
[23] Statement of Miranda Bogen.
[24] Statements of Dr. Hoda Heidari; Dr. Angela Jiang.
[25] Statements of Miranda Bogen; Dr. Tamara Kneese; Dr. Chris Meserole.
[26] Statement of Dr. William Isaac; *see also* Laura Weidinger, et al., *Sociotechnical Safety Evaluation of Generative AI Systems*, arXiv (Oct. 31, 2023), https://arxiv.org/abs/2310.11986.
[27] Statement of Dr. Tamara Kneese.
[28] *Id.*
[29] Weidinger, et al.; Statement of Dr. William Isaac.

testing environments with human interaction, evaluations of bias and discrimination, consideration of organizational safety practices, and assessments of broader societal adoption via pilots, staged release plans, and impact studies before and after release.[30] Other reporting has found little evidence of companies evaluating their AI models for societal safety.[31]

RECOMMENDATIONS

Recommendation 1:

The U.S. AI Safety Institute should approach AI safety as an expansive field, addressing (at least) technical model engineering and broader societal concerns, rather than focusing on a single aspect of safety.

The U.S. AISI should develop rigorous evaluation standards that address many dimensions of AI safety, including:

- near-term and long-term risks;
- harms from relatively low-capability automated decision systems and harms from advanced "frontier" models;
- technical vulnerabilities, exploits, inaccuracies, and failures; and
- concerns around the societal implications of the use of AI systems.

A narrow view of "AI safety," over-indexing on any one dimension of risk at the expense of others, will produce an incomplete view of safe AI systems. An expansive view of safety, requiring a broad range of assessment methodologies and disciplinary expertise, is critical given the prominent role of NIST and the U.S. AISI to "promot[e] consensus industry standards for developing and deploying safe, secure, and trustworthy AI systems[.]"[32]

Recommendation 2:

The federal government should help to develop the empirical research base needed to advance the science of AI safety, from technical auditing for vulnerabilities to controlled human testing environments.

The federal government should act where it can and learn more where more knowledge is needed. The federal government already has, at its disposal, known

[30] Weidinger, et al.; *see also* Jacobs, et al.
[31] Statement of Julia Angwin; *but see* Statement of Dr. Angela Jiang.
[32] Executive Order 14110 of Oct. 30, 2023, 88 FR 75191, 75196 (Nov. 1, 2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

safety practices (such as impact assessments, AI red-teaming, real-world testing, public participation, and practices highlighted in NIST's AI Risk Management Framework) that can and should be implemented.[33] At the same time, meaningfully advancing AI safety methodologies and the science of AI safety will require further substantial investments to progress a broad understanding of AI's risks, technical safeguards, and sociotechnical considerations.[34]

The government should provide substantial funding through NIST, the U.S. AISI, the National Science Foundation, the National Institutes of Health, and other agencies as appropriate, to advance the measurement and evaluation of AI safety risks, both broadly and within specific use contexts.[35] The government should support empirical research that will help to establish a comprehensive AI safety ecosystem, addressing both technical factors (e.g., data inputs, outputs, model weights) as well as larger institutional structures that are critical to system performance and system safety. Given the particular dearth of sociotechnical testing in practice today (see Finding 3), the government should advance funding to further sociotechnical methods of evaluation.

[33] *See AI Risk Management Framework*, National Institute of Standards and Technology (Jan. 26, 2023), https://www.nist.gov/itl/ai-risk-management-framework.

[34] *National Artificial Intelligence Advisory Committee Year 1 Report* 37 (May 2023), https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf.

[35] *See Recommendation: Implementation of the NIST AI Safety Institute*, National Artificial Intelligence Advisory Committee (Dec. 2023), https://ai.gov/wp-content/uploads/2023/12/RECOMMENDATION_Implementation-of-the-NIST-AI-Safety-Institute.pdf.

ACKNOWLEDGEMENTS

Thank you to the following individuals for lending their time and expertise to NAIAC's public briefings:

Ms. Inioluwa Deborah Raji, Fellow, Mozilla and PhD candidate, University of California, Berkeley

Dr. Vincent Conitzer, Professor of Computer Science, Director, Foundations of Cooperative AI Lab (FOCAL), Carnegie Mellon University

Dr. Chris Meserole, Executive Director, Frontier Model Forum

Dr. Arvind Narayanan, Professor of Computer Science, Princeton University

Ms. Julia Angwin, Founder, Proof News

The Honorable John C. ("Chris") Inglis, inaugural U.S. National Cyber Director

Dr. Suresh Venkatasubramanian, Professor of Computer Science, Director, Center for Technological Responsibility, Reimagination, and Redesign, Brown University

Dr. William Isaac, Research Scientist, Deep Mind

Ms. Miranda Bogen, Director, AI Governance Lab at the Center for Democracy & Technology

Dr. Angela Jiang, Global Affairs and Product, Open AI

Dr. Tamara Kneese, Project Director, Algorithmic Impact Methods Lab, Data & Society

Dr. Joshua A. Kroll, Assistant Professor, Naval Postgraduate School

Ms. Madhulika Srikumar, Head of Safety, Partnership on AI

Dr. Hoda Heidari, K&L Gates Career Development Assistant Professor in Ethics and Computational Technologies, Carnegie Mellon University

Dr. Yejin Choi, Wissner-Slivka Chair of Computer Science, University of Washington; Senior Research Director at the Allen Institute for Artificial Intelligence

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/

###