SCIENCE, TECHNOLOGY,
AND SOCIAL VALUES LAB
INSTITUTE FOR ADVANCED STUDY

Readout

U.S. AI Safety Institute Listening Session with Academic and Civil Society Leaders

September
2024

The Science, Technology, and Social Values Lab (ST&SV Lab) hosted a listening session on September 13, 2024, at the Institute for Advanced Study (IAS) for international academic and civil society leaders to provide expert feedback to US AI Safety Institute (USAISI) leadership and staff.

USAISI was established in November 2023 at the National Institute of Standards and Technology, following President Joe Biden's October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The ST&SV Lab hosted this expert consultation for USAISI leadership and staff, with international leaders in the fields of AI research, safety, and governance from academia and civil society. The listening session primed perspectives on research and methods for AI foundation model benchmarking, risk assessment, and testing ahead of the network convening.

The session was held in preparation for the November 2024 inaugural convening of the International Network of AI Safety Institutes in San Francisco aimed at strengthening international knowledge sharing on AI safety research and identifying priority topics for collaboration.[1] Operating under the Chatham House Rule, the meeting was moderated by Alondra Nelson, Harold F. Linder Professor at IAS and head of the ST&SV Lab, and Christine Custis, Civic Science Fellow in the Lab.

The listening session primed perspectives on research and methods for AI foundation model benchmarking, risk assessment, and testing ahead of the network convening.

This readout provides a summary of the listening session, highlighting themes and recommendations. The ST&SV Lab shares this readout in the interest of public visibility and to promote ongoing dialogue between government, academia, and civil society on critical issues and open questions in AI research.

Several themes and recommendations emerged, highlighting issues pertaining to the mission of USAISI and the new international network, encouraging collaboration, and recommending research priorities:

Mission and Scope

- There should be greater clarity about the mission of USAISI.
- There should be greater transparency about the aims of the International Network of AI Safety Institutes.
- The scope of work for both USAISI and the International Network of AI Safety Institutes should include a spectrum of risks and harms.

Participation

- The active involvement of diverse constituencies from many sectors, including academic institutions, industry partners, and civil society is critical to effective AI safety efforts.

Research Methods and Approaches

- USAISI and the International Network of AI Safety Institutes should advance interoperable safety benchmarks, risk assessments, and testing protocols.
- AI model evaluation frameworks should be comprehensive and relevant and include sociotechnical perspectives.
- Robust AI model evaluation requires relevant expertise including not only that of computer scientists and engineers, but also social scientists, and other fields.

Research Priorities

- Open-source AI foundation models present both challenges and opportunities for AI safety research and AI governance.
- Malign uses of synthetic content is a concerning risk and mitigating it will require the development of approaches to track content provenance and multi-stakeholder engagement.

Mission and Scope

There should be greater clarity and transparency about the goals and scope of work of the US AI Safety Institute.

Although USAISI was only recently established, it should work quickly, building on its founding Strategic Vision, to clarify how exactly its mission will be implemented.

The core of the USAISI's effort is AI model evaluations. Listening session participants expressed concerns about how the current opacity about model development, training, and testing will necessarily limit the quality of USAISI evaluations. They noted as well that the science of AI safety, including evaluation, is nascent and currently lacks consistent benchmarks for how to interrogate and assess these systems and underlying models. Evaluations may focus on hitting arbitrary thresholds or benchmarks at the model level, rather than comprehensively evaluating safety practices throughout the AI lifecycle. This dynamic presents an inherent difficulty in knowing what to test for and how to distinguish and contextualize test results. For example, evaluating the accuracy of the reasoning is very different than evaluating free recall.

Moreover, if USAISI evaluation reporting is solely focused on outcomes, accountability to the public may be stymied if details about how a model achieves this outcome is not shared. Participants also noted that USAISI's methods should attend to feedback loops that would help research and civil society organizations outside of government understand how model developers have integrated testing results into the next iteration of a model.

To continue advancing stakeholder collaboration and interoperability on AI safety, greater transparency into the goals and scope of the International Network of AI Safety Institutes is required. Clarifying specific workstreams and operational mechanisms such as relevant authorities and levers the AISI Network can use, what is out of scope of the network, and the relationship to global enforcement agencies will make it easier for countries to understand how to meaningfully participate and contribute to its work. Establishing how the network will interact with existing multi-stakeholder bodies could carve engagement channels for how the AISI Network plans to collaborate with the AI and technology governance community more broadly.

USAISI and the International Network of AISIs Should Attend to a Spectrum of Risks.

AI safety research and policy discussions have focused on security-related risks, including those to national and cybersecurity. Participants agreed that these risks could potentially have far-reaching implications in the future, but emphasized the need for USAISI's scope of work to also include real-world harms emerging from current uses of AI. Some participants also noted that USAISI's scope of work should include mitigation of harms emerging from current uses of AI, including failures of AI systems used in critical infrastructure, such as water supply to impact individual and community health and safety, or forms of bias in language models as these harms they may create greater risk *over time* if left unattended to.

Participation

Active involvement of diverse constituencies from many sectors is critical to effective AI safety efforts.

Partners needed for effective AI safety research include academia and civil society as well as industry. To aid in effectively and comprehensively identifying and mitigating a spectrum of risks and harms, AISI should engage academia, industry, and civil society. Building feedback loops between these stakeholder communities could help to sustain engagement and potentially increase transparency and trust between government and the constituencies it serves. Participants recommended that AISI proactively employ both novel and established methods for multi-stakeholder engagements, such as the participatory processes employed in the development of the AI Risk Management Framework and as part of the Christchurch Call to Action, which combined proactive, novel *informal* outreach mechanisms with longstanding *formal* engagement mechanisms such as requests for comment. Some participants recommended that governments might consider reframing stakeholder feedback or listening sessions away from notional questions about AI risks, to evaluating the empirical relationship between technology and power. For example, instead of asking civil society to describe or report AI harms, a more robust exercise might involve examining how AI use and/or discussions about AI safety are shaping–and being shaped by–power dynamics across different sectors and communities.

> **"Active involvement of diverse constituencies from many sectors is critical to effective AI safety efforts."**

Participants recommended stakeholder participation in mapping the spectrum of AI risks and harms, including in the development of a global AI harms incident reporting infrastructure. AISI could support these efforts by developing reporting mechanisms of AI harms, including reporting templates or tooling that enable the pooling and scaling of results in a database for analysis by independent third parties. To ensure these instances include a range of heterogeneous perspectives, governments could even consider providing incentives for civil society to respond. This grassroots approach could enable governments to recognize potential threats to a more diverse set of global stakeholders and more acutely identify areas where AI applications are significantly impacting people's daily lives.

Research Methods and Approaches

Both USAISI and the International Network of AISIs should advance interoperable safety benchmarks, risk assessments, and testing protocols.

Governments possess a variety of AI resources, expertise, and legal and policy frameworks for governing AI. Most countries do not claim large AI developers, but may have AI expertise, or could have neither but are deeply impacted by AI's societal or environmental harms. Participants observed that these asymmetries should not affect participation in global AI safety research or the pursuit of interoperable governance structures and frameworks. Well-resourced countries may need to share AI knowledge, training, data, or infrastructure to help promote countries' ability to participate in AI safety research. The International Network of AISIs could complement the work of international standards organizations that currently drive interoperability on AI governance issues, but may flatten important context or perspectives from distinct groups. Reduced resource asymmetries and increased global alignment on which concepts (like data flows, value chain of labor, and critical minerals extractions) are relevant to AI safety will help more countries participate in discussions about interoperability. Nations likewise have different norms, cultural attitudes, and institutional levers that could influence AI governance approaches. Reaching consensus within international standards organizations among countries with divergent approaches to AI governance and safety may not be feasible, particularly on thorny issues like risk thresholds or evaluation criteria. Despite these differences, participants concurred that fostering dialogue among and across nations is crucial for interoperability of global AI governance methods.

AI model evaluations should be relevant and comprehensive.

Model evaluation in isolation does not reveal a model's risks to the safety of users and impacted non-users. For model evaluations to be relevant they should be performed on both models (to test their capabilities), and the contexts in which they are deployed (to test their impact). As one participant noted, "safety" is not a model property, but rather something that applies to how a model interacts with people and the environment in which it is deployed. To better understand and measure safety risks, these risks need to be evaluated through a sociotechnical lens.

AI model evaluation requires relevant expertise.

Robust AI model evaluation requires relevant expertise including that of computer scientists and engineers, social scientists, and others. Participants emphasized the importance of including sociotechnical perspectives in safety assessments and incorporating expertise from multiple disciplines. Test and evaluation work streams within an AISI Network should be informed by the expertise of anthropologists and others who deeply understand human behavior. In the context of language, for example, ethnographic researchers can contribute to multi-lingual evaluations of LLMs to improve representation of, and understanding of nuance in, underrepresented languages. In certain languages, false interpretations or lack of nuance could even expose a model to risks of jailbreaks. Without this unique expertise, most existing model evaluation methods may not account for the linguistic and cultural nuances necessary to properly assess an AI model's capabilities and risks.

In addition to anthropologists and linguists, we may need to reconsider what we assume about who is an AI "expert" capable of performing evaluations. The inclusion of people who work on surveillance and extractivism (pertaining to land, energy, data, and resources) could dramatically expand the scope of evaluations beyond the existing presumption of what they should target, and the range of factors and practices across the AI stack that could be evaluated. This expanded evaluation approach might incorporate programs on supply chain transparency and responsible minerals.

"Both USAISI and the International Network of AISIs should advance interoperable safety benchmarks, risk assessments, and testing protocols."

The USAISI has finalized agreements with leading AI labs to test their latest models pre- and post-deployment. As this activity gets underway, and as other countries pursue these agreements (or determine if and how to share results with one another), governments will need to consider how to incorporate a broad range of expertise and viewpoints in defining evaluation methodologies.

Research Priorities

Open-Source AI Foundation Models Present Both Challenges and Opportunities for AI Safety Research and AI Governance.

Open-source AI foundation model development and use is different from proprietary or "closed-source" development, including in the level of transparency regarding model development.[2] AISIs and similar government entities have held high-profile engagements on AI safety with large companies that possess the resources to shape governance approaches. AI safety research and governance approaches, however, should account for open-source development, that in many cases involves smaller companies. Participants recommended that USAISI AI safety discussions can sometimes over-index on the practices associated with the development of one large proprietary model, but should expand to include organizations innovating in open source AI by integrating small or open-source AI models across a distributed chain.

Participants discussed the need to balance governance approaches between proprietary and open-source development, considering the unique challenges and opportunities of distributed development. They emphasized the importance of including smaller companies and organizations in safety discussions, rather than focusing solely on large technology companies.

Malign uses of synthetic content is a concerning risk and mitigating it will require the development of approaches to track content provenance and multi-stakeholder engagement.

Participants noted that basic frameworks to understand content origin and history would significantly enhance transparency and explainability in AI systems, paramount to any AI safety regime. The USAISI, in partnership with AI labs and other governments, could develop a governance regime to require and standardize provenance information for visual content (particularly in contexts like political advertising). These elemental documentation practices would help lay the groundwork for more advanced AI safety research on synthetic content.

Participants flagged that a well-informed public can act as a filter to prevent the viral spread of harmful synthetic content. Risks from synthetic content typically focus on public misrepresentation, non-consensual intimate imagery (NCII), or child sexual abuse material (CSAM), and new capabilities like synthetic biometrics could have major implications for identity fraud. Literacy programs could draw from cybersecurity awareness campaigns, not only teaching users to evaluate and verify the authenticity of media they consume and share but also providing a nuanced understanding of the difference between harmful and satirical content. The AISI Network could identify foundational objectives for digital literacy programs and engage with civil society to devise programming and track results accordingly.

The complex web of diverse stakeholders in the digital media landscape complicates the assignment of governance roles and responsibilities, creating accountability gaps for limiting the spread of harmful synthetic content. Model developers are actively working on deepfake detection tools, but may be hesitant to release them for fear that bad actors could develop bypass capabilities. Industry fora driving work on content provenance have not explicitly prioritized civil society engagement to date.[3] To organize collaboration, the AISI Network could coordinate technical research with AI labs on detection tool efficacy and vulnerabilities, and work with civil society to understand how tools compare to or complement digital literacy initiatives. Separately, governments could clarify penalties for individuals or entities knowingly producing or spreading harmful media.

To achieve robust AI safety and governance, governments cannot work in silos without broad, participatory multi-stakeholder engagement. We look forward to continued collaboration with the USAISI and the Network of AISIs to elevate diverse perspectives and tackle these issues that require global cooperation across sectors, experiences, and all regions.

Participants

Ada Lovelace Institute (UK)
Brown University
Cheikh Hamidou Kane Digital University (Senegal)
Center for Democracy and Technology
Data & Society Research Institute
Equiano Institute (South Africa)
Institute for Advanced Study
Mozilla Foundation
National Fair Housing Alliance
Princeton University
RAND
SRI International
Stanford University
University of California, Berkeley
University of Illinois, Urbana-Champaign

[1] U.S. Secretary of Commerce Raimondo and U.S. Secretary of State Blinken Announce Inaugural Convening of International Network of AI Safety Institutes in San Francisco, U.S. Department of Commerce, September 18, 2024.

[2] The IAS AI Policy and Governance Working Group has recommended an approach to move past this binary framing: AIPGWG Response - NTIA RFC on Open Foundation AI w Available Model Weights (March 2024).

[3] The Coalition for Content Provenance and Authenticity (C2PA) includes technology companies and media outlets working to develop a standard for content provenance.