Testimony of

Elham Tabassi

Chief of Staff
Information Technology Laboratory

National Institute of Standards and Technology
United States Department of Commerce

Before the
United States House of Representatives
Committee on Science, Space, and Technology
Subcommittee on Research and Technology

Trustworthy AI: Managing the Risks of Artificial Intelligence

September 29, 2022

Chairwoman Stevens, Ranking Member Feenstra, and distinguished members of the Subcommittee, I am Elham Tabassi, Chief of Staff of the Information Technology Laboratory (ITL) and the lead for NIST's trustworthy and responsible AI program at the Department of Commerce's National Institute of Standards and Technology – known as NIST. We appreciate the committee's continued support of our work and thank you for the opportunity to testify today on NIST's efforts to improve the trustworthiness of artificial intelligence.

NIST is home to five Nobel Prize winners, with programs focused on national priorities such as cybersecurity, advanced manufacturing, semiconductors, the digital economy, precision metrology, quantum information science, biosciences and artificial intelligence. NIST's mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

In the NIST Information Technology Laboratory, we work to cultivate trust in information technology and metrology. Trust in the digital economy is built upon key attributes like cybersecurity, privacy, usability, interoperability, equity, and avoiding bias and increasing usefulness in the development and deployment of technology. NIST conducts fundamental and applied research, advances standards to understand and measure limits and capabilities of technology and develops tools to evaluate such measurements. Technology standards and measurements—and the foundational and applied research that enables their development and use—are critical to advancing trust in digital products and services. These standards and measurements can provide increased assurance and utility, thus enabling more secure, private, and rights-affirming technologies.

NIST's Role in Artificial Intelligence

NIST contributes to the research, standards, measurements, and data required to realize the full promise of artificial intelligence (AI) as a tool that will enable American innovation, enhance economic security, and improve our quality of life.

As a non-regulatory agency, NIST prides itself on the strong partnerships it has cultivated with the government and private sector. NIST seeks and relies on diverse stakeholder feedback among government, industry, academia, and non-profit entities to develop and improve its resources. The collaborative, transparent, and open processes NIST uses to develop resources result in more effective and usable resources that are trusted, and therefore, widely used by various organizations. Our resources are used by federal agencies, as well as private sector organizations of all sizes, educational institutions, and state, local, tribal, and territorial governments.

Much of NIST's AI effort[1] focuses on cultivating trust in the design, development, and use of AI technologies and systems. Working with the community, NIST is:

- conducting fundamental research to advance trustworthy AI technologies and understand and measure their capabilities and limitations
- applying AI research and innovation across NIST laboratory programs
- establishing benchmarks and developing data and metrics to evaluate AI technologies
- leading and participating in the development of technical AI standards

[1] https://www.nist.gov/artificial-intelligence

- contributing to discussions and development of AI policies, including supporting the National AI Advisory Committee[2]

NIST AI Risk Management Framework

Among its many AI-related activities, NIST is developing the AI Risk Management Framework[3] (AI RMF) to provide guidance on managing risks to individuals, organizations, and society associated with AI. AI risk management is about offering a path to minimize potential negative impacts of AI systems, as well as pointing to opportunities to maximize positive impacts and creating opportunities for innovation. Identifying, mitigating, and minimizing risks and potential harms associated with AI technologies are essential steps towards the development of trustworthy AI systems and their appropriate and responsible use. Like NIST's well-known Cybersecurity and Privacy Frameworks, the NIST AI RMF will provide a set of outcomes that enable dialogue, understanding, and actions to manage AI risks. The AI RMF is a voluntary framework seeking to provide a flexible, structured, and measurable process to address AI risks prospectively and continuously throughout the AI lifecycle.

In August, NIST released its second draft of the AI RMF[4] with the goal of releasing AI RMF 1.0 in January. This is consistent with congressional direction in the National Artificial Intelligence Act of 2020. This latest draft builds on the March 2022 initial draft and a December 2021 concept paper – and the many comments from organizations and individuals.

NIST also released a draft AI RMF Playbook[5] in August. This companion to the AI RMF when completed will provide additional guidance to organizations on the actions they can take to meet the outcomes included in the Framework.

AI research and development, as well as the standards landscape, are evolving rapidly. For that reason, the AI RMF and its related documents will evolve over time and reflect new knowledge, awareness, and practices. NIST intends to continue its robust engagement with stakeholders to keep the Framework up to date with AI trends and reflect experience based on the use of the AI RMF. Ultimately, the AI RMF will be offered in multiple formats, including online versions, to provide maximum flexibility.

The Framework is being developed through a consensus-driven, open, transparent, and collaborative process. From the start of this initiative, NIST has offered a broad range of stakeholders the opportunity to take part in workshops[6], respond to a Request for Information (RFI)[7], and review draft reports[8] and other documents including draft approaches[9] and versions of the framework[10]. NIST also has reached out directly to AI practitioners along with other stakeholders across a full spectrum of interests domestically and internationally. This outreach

[2] https://www.nist.gov/artificial-intelligence/national-artificial-intelligence-advisory-committee-naiac
[3] https://www.nist.gov/itl/ai-risk-management-framework
[4] https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf
[5] https://pages.nist.gov/AIRMF/
[6] https://www.nist.gov/itl/ai-risk-management-framework/ai-risk-management-framework-workshops-events
[7] https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development-request-information
[8] https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf
[9] https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf
[10] https://www.nist.gov/itl/ai-risk-management-framework

has included companies, government agencies, academia, and not-for-profit organizations representing civil society, consumers, and industry. NIST has actively encouraged others to provide direct input, and many organizations and individuals have contributed their insights to NIST. Those have included international organizations, with the goal of aligning the NIST Framework with standards and approaches being developed around the globe.

The current draft AI RMF defines certain key characteristics of trustworthy AI systems and offers guidance for mapping, measuring, and managing them. As defined in the draft AI RMF, trustworthy AI is valid and reliable, safe, fair, and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced. AI systems are socio-technical in nature, meaning they are a product of the complex human, organizational, and technical factors involved in their design, development, and use. Many of the trustworthy AI characteristics – such as bias, fairness, interpretability, and privacy – are directly connected to societal dynamics and human behavior.

NIST's Research on AI Trustworthiness Characteristics

To build on NIST's work on the AI RMF and provide additional guidance to organizations to advance trustworthy and responsible AI, NIST also conducts fundamental research on many of the AI trustworthiness characteristics.

» *AI Trustworthiness Characteristics – Fair and Bias is Managed*

While there are many approaches for ensuring technologies that we use every day are safe and secure, there is less research into how to advance systems that are fair with bias managed. Fairness in AI includes concerns for equality and equity by addressing issues such as bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application and context of use.

NIST has significantly expanded its research efforts to identify, understand, measure, manage and mitigate bias, with a focus on a socio-technical approach. NIST recently published "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (NIST Special Publication 1270)[11], which identifies the concepts and challenges associated with bias in AI and provides preliminary guidance for addressing them.

NIST has identified three major categories of AI bias to be considered and managed: systemic, computational, and human, all of which can occur in the absence of prejudice, partiality, or discriminatory intent. Current attempts for addressing the harmful effects of AI bias remain focused largely on computational factors such as representativeness of datasets and fairness of machine learning algorithms. Human and systemic institutional and societal factors are significant sources of AI bias that are currently overlooked. Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems. Human biases relate to how an individual or group perceives and uses AI system information to make a decision or fill in missing information.

Through the NIST National Cybersecurity Center of Excellence (NCCoE), we are beginning a project, "Mitigation of AI/ML Bias in Context"[12], to develop additional guidance to mitigate bias in AI and Machine Learning (ML). Under the NCCoE model, NIST works collaboratively with

[11] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
[12] https://www.nccoe.nist.gov/projects/mitigating-aiml-bias-context

relevant industry and academia partners. The "Mitigation of AI/ML Bias in Context," project intends to apply the concepts in our March 2022 NIST publication on bias to build a proof-of-concept implementation, or "use case," for credit underwriting decisions in the financial services sector. Future application use cases may also be considered, such as hiring or school admissions. These will help promote fair and positive outcomes that benefit users of AI/ML services, the organizations that deploy them, and all of society. A small but novel part of this project will examine the interplay between bias and cybersecurity, with the goal of identifying approaches which might mitigate risks that exist across these two critical characteristics of trustworthy AI.

» *AI Trustworthiness Characteristics – Explainable and Interpretable*

Explainability and interpretability are important characteristics to ensure users and operators of AI can understand the decisions or predications made by AI, thus avoiding the "opaque system" concept associated with AI. Explainability refers to a representation of the mechanisms underlying an algorithm's operation, whereas interpretability refers to the meaning of an AI systems' output in the context of its designed functional purpose.

NIST has released two publications aimed at providing deeper understanding of the principles of Explainability and interpretability: "Four Principles of Explainable Artificial Intelligence" (NISTIR 8312)[13] and "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence" (NISTIR 8367)[14].

» *AI Trustworthiness Characteristics –Secure and Resilient*

AI systems that can withstand adversarial attacks and maintain confidentiality, integrity, and availability are resilient and secure systems.

NIST released the draft "A Taxonomy and Terminology of Adversarial Machine Learning"[15] (NISTIR 8269) to advance a taxonomy for securing applications of AI, specifically, adversarial machine learning. NIST's Cybersecurity Framework[16] is widely used to address the cybersecurity risks of organizations. NIST is constantly updating the Cybersecurity Framework to account for changes in the cybersecurity technology, standards, and risk landscape.

NIST is building an experimentation testbed called Dioptra[17] to begin to evaluate adversarial attacks against ML algorithms. The testbed aims to facilitate security evaluations of ML algorithms under a diverse set of conditions. To that end, the testbed has a modular design enabling researchers to easily swap in alternative datasets, models, attacks, and defenses. The result is the ability to advance the metrology needed to ultimately help secure AI systems.

» *AI Trustworthiness Characteristics – Privacy-enhanced*

Privacy safeguards the important human values of autonomy and dignity through methods that focus on providing individuals with anonymity, confidentiality, and control over various facets of their identities. These outcomes generally should guide choices for AI system design,

[13] https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence
[14] https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence
[15] https://www.nccoe.nist.gov/ai/adversarial-machine-learning
[16] https://www.nist.gov/cyberframework
[17] https://pages.nist.gov/dioptra/

development, and deployment. From a policy perspective, privacy-related risks may overlap with security, bias, and transparency.

NIST's Privacy Risk Assessment Methodology[18], developed in 2016 and NIST's Privacy Framework[19], issued in 2020, are voluntary tools that organizations from all industry sectors across the world are using to identify and manage privacy risks in the systems, products and services they develop and deploy, improve their privacy programs, and better comply with privacy regulation.

NIST is also conducting research on privacy-enhancing technologies (PETs) to advance data-driven, innovative solutions to preserve the right to privacy, including hosting the Privacy Engineering Collaboration Space[20], a virtual public platform that serves as a clearinghouse for open-source tools and PETs use cases. In coordination with the National Science Foundation (NSF) and the White House Office of Science and Technology Policy (OSTP), NIST is co-sponsoring the U.S.-U.K. prize competition on PETs[21]. First announced at the Summit for Democracy in December 2021, the winning solutions will compete for a combined U.S.-U.K. prize pool of $1.6 million and will be showcased at the second Summit for Democracy anticipated in early 2023.

Research on Applications of AI

NIST's multidisciplinary laboratories and varied fields are an ideal environment to develop and apply AI[22]. Various AI techniques are being used to support NIST scientists and engineers, drawing on ML and AI tools to gain a deeper understanding of and insight into our research. NIST is integrating AI into the design, planning, and optimization of NIST's research efforts – including hardware for AI[23], computer vision, engineering biology and biomanufacturing, image and video understanding, medical imaging, materials science, manufacturing, disaster resilience, energy efficiency, natural language processing, biometrics, quantum science, robotics, and advanced communications technologies. Key focus areas include innovative measurements using AI/ML techniques, predictive systems using AI/ML models, and enabling and reducing the barriers to autonomous measurement platforms.

AI Measurement and Evaluation

NIST has a long history of devising appropriate metrics, measurement tools, and challenge problems to support technology development. NIST first started the measurement and evaluation of automated fingerprint identification systems in the 1960s. Evaluations strengthen research communities, establish research methodology, support the development of standards, and facilitate technology transfer. NIST is looking to bring these benefits of community evaluations to bear on the problem of constructing trustworthy AI systems. These evaluations will begin with community input to identify potential harms of selected AI technologies in context, and the data

[18] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/resources
[19] https://www.nist.gov/privacy-framework/privacy-framework
[20] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space
[21] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/prize-challenges
[22] https://www.nist.gov/applied-ai
[23] https://www.nist.gov/artificial-intelligence/hardware-ai

requirements for AI evaluations. NIST also hosts a biweekly AI metrology colloquia series[24], where leading researchers share current work on AI measurement and evaluation.

As discussed above, NIST has been engaged in focused efforts to establish common terminologies, definitions, and taxonomies of concepts pertaining to characteristics of AI technologies in order to form the necessary underpinnings for trustworthy AI systems. Each of these characteristics also requires its own portfolio of measurements and evaluations. For each characteristic, NIST aims to document and improve the definitions, applications, and strengths and limitations of metrics and measurement methods in use or being proposed. NIST's current efforts represent only a small portion of the research that will be required to test and evaluate trustworthy AI systems.

A significant challenge in the evaluation of trustworthy AI systems is that context (the specific use case) matters; accuracy measures alone will not provide enough information to determine if deploying a system is warranted. The accuracy measures must be balanced by the associated risks or societal harms that could occur. The tolerance for error drops as the potential impacts of risk rise.

New NIST efforts in AI evaluation will focus on other socio-technical aspects of system performance in addition to accuracy. In particular, the evaluations have the goal of identifying risks and harms of systems before such systems are deployed, and to define (and eventually create) data sets and evaluation infrastructure that will allow system builders to detect the extent to which their system exhibits those harms.

Examples of NIST AI measurement and evaluation projects[25] include:

- *Biometrics:* Over that past sixty years, NIST has been testing and evaluating biometric recognition technologies, including face recognition, fingerprint, biometric quality, iris recognition, and speaker recognition.
- *Computer vision:* NIST's computer vision program includes several activities contributing to the development of technologies that extract information from image and video streams through systematic, targeted annual evaluations and metrology advances, including the Open Medica Forensics Challenge, Activities in Extended Video (ActEV), handwriting recognition and translation evaluation, and others.
- *Information retrieval:* The information retrieval research uses large, human-generated text, speech, and video files to create test collections through the Text Retrieval (TREC), TREC Video Retrieval Evaluation (TRECVID), and Text Analysis (TAC) Conferences. The Text Retrieval Conference is responsible for significant advancements in search technology. A 2010 NIST study[26] estimated that without TREC, U.S. internet users would have spent an estimated 3.5 billion worth of additional hours using search engines between 1999 and 2009.

AI Standards

NIST plays a critical role in the standards process as the nation's measurement laboratory and has a unique role relating to standards in the Federal enterprise. Our coordination function,

[24] https://www.nist.gov/programs-projects/ai-measurement-and-evaluation/ai-metrology-colloquia-series

[25] https://www.nist.gov/programs-projects/ai-measurement-and-evaluation/nist-ai-measurement-and-evaluation-projects

[26] https://trec.nist.gov/pubs/2010.economic.impact.pdf

currently defined under the National Technology Transfer and Advancement Act and the NIST Organic Act, has yielded benefits to the nation ever since the Institute was established by Congress as the National Bureau of Standards in 1901. NIST's strong ties to industry and the standards development community have enabled NIST to take on critical standards-related challenges and deliver timely and effective solutions.

NIST works to support the development of AI standards that promote innovation and public trust in systems that use AI. Pursuant to U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools[27], NIST seeks to bolster AI standards-related knowledge, leadership, and coordination; conduct research to support development of technically sound standards for trustworthy AI; promote partnerships to develop and use standards; and engage internationally to advance AI standards.

I serve as the Federal AI Standards Coordinator to work across the government and industry stakeholders to gather and share information on AI standards-related needs, strategies, and best practices.

NIST facilitates federal agency coordination in the development and use of AI standards in part through the Interagency Committee on Standards Policy (ICSP) AI Standards Coordination Working Group[28]. This working group seeks to foster agency interest and participation in AI standards and conformity assessment activities, facilitate coordination of U.S. government positions on draft standards, identify effective means of coordinating with and contributing towards voluntary consensus bodies, align U.S. government activities with those of the private sector on AI standards development activities, promote effective and consistent federal policies leveraging AI standards, and raise awareness of federal agencies' use of AI that contributes to standards activities.

NIST also engages internationally through bilateral and multilateral work on AI. The United States championed development of the first international principles for the responsible use of AI at the Organisation for Economic Co-operation and Development, or OECD. The U.S. also serves as a founding member of the Global Partnership on AI, which includes all members of the G7 and others such as Brazil and India, to coordinate R&D AI initiatives. NIST advances research on trustworthy AI with the Indo-Pacific Economic Framework. NIST supports the US-EU Trade and Technology Council (TTC) in building common approaches for trustworthy AI. Under the TTC, the U.S. and EU have launched a new AI sub-working group where NIST is working towards common frameworks for AI risk management and developing metrics and methodologies for measuring AI trustworthiness. And as mentioned above, the U.S. – led by NIST, NSF, and OSTP – is collaborating with the UK to develop prize challenges on advancing privacy-enhancing technologies.

Interagency Coordination

NIST leads and participates in several federal AI policymaking efforts and engages with many other federal offices and interagency groups. This includes administering the National Artificial Intelligence Advisory Committee (NAIAC)[29], on behalf of the Department of Commerce. The NAIAC is tasked with advising the President and the National AI Initiative Office. NIST supports the operation of this advisory committee. The Secretary of Commerce appointed the 27

[27] https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf
[28] https://www.nist.gov/standardsgov/icsp-ai-standards-coordination-working-group-aiscwg-charter
[29] https://www.nist.gov/artificial-intelligence/national-artificial-intelligence-advisory-committee-naiac

members in April 2022. NAIAC held its first meeting in May 2022. Five working groups have been established to focus NAIAC's work on leadership in trustworthy AI, leadership in research and development, supporting the U.S. workforce and providing opportunity, U.S. leadership and competitiveness, and international cooperation.

NIST also co-chairs the National Science and Technology Council's Machine Learning and Artificial Intelligence Subcommittee[30], the Networking and Information Technology Research and Development's (NITRD) AI Working group[31], and the NITRD Fast Track Action Committee[32] which is drafting a national strategy to advance privacy-preserving data sharing and analytics. NIST founded and is co-chairing the AI Standards Coordination Working Group (AISCWG) under the Interagency Committee on Standards Policy (ICSP). NIST's AI lead also serves as Federal AI Standards Coordinator and is a member of the National AI Research Resource Task Force[33].

Conclusion

Advancing artificial intelligence research and standards that contribute to a secure, private, interoperable, and world-leading digital economy is a top priority for NIST. Our economy is increasingly global, complex, and interconnected. It is characterized by rapid advances in technology. The timely availability of AI trustworthiness standards and guidance is a dynamic and critical challenge. Through robust collaboration with stakeholders across government, industry, and academia in the U.S. and elsewhere, NIST aims to cultivate trust and foster an environment that enables AI innovation on a global scale – and to do so in a way that respects and advances human rights.

NIST's team includes some of the top AI and standards experts in the world. This includes staff with multidisciplinary backgrounds in science and engineering. Working with our partners in other federal agencies, the private sector, academia, and other allied countries, and with the support of Congress, we will work tirelessly to address current and future challenges.

Thank you for the opportunity to present on NIST activities to improve AI trustworthiness. I look forward to your questions.

[30] https://www.ai.gov/about/#MLAI-SC_Machine_Learning_and_AI_Subcommittee
[31] https://www.ai.gov/a-new-nitrd-iwg-for-artificial-intelligence-ai-rd/
[32] https://www.nitrd.gov/coordination-areas/privacy-rd/appdsa/
[33] https://www.ai.gov/naiac/

Elham Tabassi (Fed)
Chief of Staff, Information Technology Laboratory

Elham Tabassi is the Chief of Staff in the Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST). She leads NIST Trustworthy and Responsible AI program that aims to cultivate trust in the design, development, and use of AI technologies by improving measurement science, standards, and related tools in ways that enhance economic security and improve quality of life. She has been working on various machine learning and computer vision research projects with applications in biometrics evaluation and standards since she joined NIST in 1999. She is the principal architect of NIST Fingerprint Image Quality (NFIQ) which is now an international standard for measuring fingerprint image quality and has been deployed in many large-scale biometric applications worldwide.
She is a member of the National AI Resource Research Task Force, a senior member of IEEE, and a fellow of Washington Academy of Sciences.