GEORGETOWN UNIVERSITY
Walsh School *of* Foreign Service
*Center for Security and Emerging Technology*

May 31, 2019

Elham Tabassi,
Acting Chief of Staff, Information Technology Laboratory,
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

**Re: RFI on Standards for Reliable, Robust, and Trustworthy Artificial Intelligence**

Dear Ms. Tabassi:

The Center for Security and Emerging Technology (CSET) offers the following submission for NIST's consideration.

NIST requested information about "the current state, plans, challenges, and opportunities regarding the development and availability of AI technical standards and related tools, as well as priority areas for federal involvement in AI standards-related activities" and listed 18 topics of interest. CSET's submission is on the following topics:

- **Topic 3:** The needs for AI technical standards and related tools. How those needs should be determined, and challenges in identifying and developing those standards and tools.
- **Topic 11:** Specific opportunities for, and challenges to, U.S. effectiveness and leadership in standardization related to AI technologies.

CSET's submission focuses specifically on standards for "reliable, robust, and trustworthy" AI (collectively "***trustworthy AI***"), as that phrase is used in the Executive Order on Maintaining American Leadership in Artificial Intelligence. It expands on the following claims:

1. **Standards for trustworthy AI can advance both U.S. and global interests.** Standards are a prerequisite for intra- and international coordination toward trustworthy AI.
2. **Developing standards for trustworthy AI requires foundational research**. Standards should be informed by foundational research on what makes AI trustworthy.
3. **NIST should create a National AI Testbed**. NIST should advance foundational research on trustworthy AI by establishing a National AI Testbed: a digital platform containing public and non-public datasets, code, and testing environments on which AI systems can be developed, stored, and tested.

## 1. Standards for trustworthy AI can advance both U.S. and global interests.

Standards are a prerequisite for intra- and international coordination toward trustworthy AI.

*Trustworthy AI*. An AI system is *trustworthy* if it consistently works as intended across its prescribed domain of use. AI systems might fail to work as intended for many reasons, including adversarial attack, misspecified goals, or differences between the environments in which they're trained and deployed (Amodei 2016; Ortega 2018; Goodfellow 2014; Krakovna 2018; DHS 2018). Ensuring machine learning-based AI systems are trustworthy might be particularly challenging due to their opacity; the speed at which the field of machine learning is developing; and competitive pressures for developers to prioritize speed over caution.

*The importance of standards for trustworthy AI*. We're only beginning to understand how and when AI systems might be untrustworthy. To date, there are no common metrics to assess trustworthiness, which poses an obstacle to clear communication about the risks. For example, we might want AI in the criminal justice system to be unbiased; AI in self-driving cars to be interpretable; and AI on the battlefield to be secure from adversarial attacks and robust to changes in its environment; but what exactly does all that mean? Without a clearer understanding of the risks, developers are likely to either under- or over-invest in caution. This in turn could lead to public safety harms if the technology is unreliable; economic harms if innovation is stifled due to unfounded fears of the technology; and national security harms if adversaries are unable to find common ground on how the technology should and should not be developed and used. *Standards* can help avoid these harms by creating a common understanding of what constitutes trustworthy AI.[1] Standards of this form are a prerequisite for creating uniform practices, whether in the form of industry best practices, domestic regulation, or treaties and international norms governing economic and military use of AI.

## 2. Developing standards for trustworthy AI requires foundational research.

Standards should be informed by foundational research on trustworthiness.

*The need for foundational research*. Standards are the last step in a long process that begins with foundational research on questions like: How might an AI system fail; how can we quantitatively

[1] This submission uses the term *standards* to mean a technical document regarding what trustworthiness for AI requires in different contexts, and how can it be measured and tested. ISO/IEC Guide 2: 2004 differentiates eight common types of standards. Of particular importance for trustworthy AI are *terminology standards*, *testing standards*, and *process standards*.

measure the likelihood of failure; and how can we verify in the laboratory that a system is unlikely to fail, or be undermined by an adversary, before it's deployed in the real world? Even after initial standards are developed, research on these questions should continue indefinitely, informing updates to the standards. It's unclear in what contexts standards are ripe today and in what contexts they're premature, which is itself a question warranting further investigation. In some cases, existing standards related to AI-adjacent topics like cybersecurity might be applied to AI. But for the most part, trustworthy AI presents enough new problems that a research-first approach is needed.

*Types of foundational research.* The risks of AI vary considerably by context and the type of system involved. For example, a commercial language model will have a different risk profile than the AI incorporated into a weapons platform. For this reason, different standards will be needed for different use cases. A first step toward trustworthy AI standards might involve a taxonomy of current use cases for AI in both the public and private sectors. Next, the risks associated with different use cases could be assessed, i.e., how might the system fail, and what is the nature and magnitude of harm in the event of failure. Risk assessment can help to prioritize both the use cases and risks within the use cases, as well as to identify pressing open questions requiring further research. Ultimately, metrics and benchmarks for risks should be created. This process should continue indefinitely to keep pace with new developments in AI (NIST 2012).[2]

**3. NIST should create a National AI Testbed.**

NIST should advance foundational research on trustworthy AI by establishing a National AI Testbed: a digital platform containing public and non-public datasets, code, and testing environments on which AI systems can be developed, stored, and tested.

*A National AI Testbed as vehicle for public-private collaboration.* The foundational research necessary for trustworthy AI standards will require public-private collaboration. NIST, in partnership with other federal agencies and the private sector, should provide and maintain the

[2] NIST should complement, rather than duplicate, the work of other standards bodies and industry. In particular, the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC)'s working group on AI (ISO/IEC JTC 1/SC 42) is currently working on developing definitions and use cases for trustworthy AI. Other notable efforts include the Institute of Electrical and Electronics Engineers' (IEEE) AI standards series; and the MLPerf benchmark suite, an industry-academia collaboration that measures the performance of software frameworks and computing platforms.

environment for this work in the form of a National AI Testbed.[3] The Testbed could be a digital platform that supports two overlapping functions:  a hub for research and development relevant to trustworthy AI standards; and an environment in which to test academic, commercial, and government AI models. Access to government datasets and computing resources would make it a uniquely attractive research environment. The 2016 AI R&D strategy envisioned exactly this:

> The government has massive amounts of mission-sensitive data unique to government, but much of this data cannot be distributed to the outside research community. Appropriate programs could be established for academic and industrial researchers to conduct research within secured and curated testbed environments established by specific agencies. AI models and experimental methods could be shared and validated by the research community by having access to these test environments, affording AI scientists, engineers, and students unique research opportunities not otherwise available.

*A National AI Testbed can advance other US interests*. A National AI Testbed would advance key U.S. interests in addition to those discussed under Point 1 above. First, appropriate precautions would need to be taken to make sensitive data and code accessible to researchers without infringing privacy, security, or proprietary interests. These precautions—such as federated learning, multi-party computation, and homomorphic encryption—require further research and development. The Testbed would create a demand and environment for developing these precautions. Second, the Testbed would provide an opportunity to increase collaboration on AI development with the international community, such as via the EU's AI4EU testbed. It would therefore also create a demand and environment for the development of interoperability standards. Third, if the Testbed allows the secure pooling of sensitive resources like data and code, it would increase the competitiveness of the U.S. AI ecosystem by enhancing researchers' access to these critical resources. Lastly, it could provide a secure way to provide researchers access to models with dual-use risks without incurring the public safety risks of open sourcing the code. Security could come either from giving only vetted researchers access, or from limiting the nature of access. For example, OpenAI recently elected not to fully release a language model, GPT-2, that it believed could be misused (OpenAI 2019). A platform in which labs like OpenAI could securely make such models accessible would help dissolve the tradeoff between open science and public safety.

[3] NIST has established collaborative research environments and testbeds with similar features, such as its Manufacturing Robotics Testbed, National Software Research Library, nSoft, and Robotics Test Facility. And on January 1, 2019, the European Union established AI4EU, a private-public partnership that aims to provide a common platform for pooling data, computing power, research tools, and algorithms.

*Why NIST*. NIST is well-placed to establish and maintain a National AI Testbed. First, NIST has access to resources unique to the federal government, such as datasets, as discussed above. Second, the Testbed would require a level of sustained public-private coordination that the private sector is less well-suited to provide. As a federal agency with historic ties to industry, NIST is uniquely well-positioned to cater to the distinct needs of both industry and government. Third, the Testbed would require coordination across government programs. The fundamental research needed to establish trustworthy AI standards will require a mix of approaches, including research grants, collaborative projects, competitions, and prizes. Any standards effort should therefore integrate with the U.S. government's general AI R&D strategy, for which "measuring and evaluating AI technologies through standards and benchmarks" is a priority (NSTC 2016). NIST has historically provided such a coordinating role between federal agencies. Finally, the Testbed would complement existing NIST projects, such as its Privacy Collaboration platform and Cybersecurity Framework.

**References**

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete problems in AI safety" (2016).
https://arxiv.org/abs/1606.06565

Department of Homeland Security, "AI: Using Standards to Mitigate Risks" (2018).
https://www.dhs.gov/sites/default/files/publications/2018_AEP_Artificial_Intelligence.pdf

Goodfellow, Ian, Shlens, Jonathon, Szegedy, Christian, "Explaining and Harnessing Adversarial Examples (2014).
https://arxiv.org/abs/1412.6572

Krakovna, Victoria, "Specification gaming examples in AI" (2018).
https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

National Institute of Standards and Technology, "Guide for Conducting Risk Assessments" (2012).
https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf

National Science and Technology Council, "The National Artificial Intelligence Research and Development Strategic Plan" (2016).
https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

OpenAI, “Better Language Models and Their Implications” (2019).
https://openai.com/blog/better-language-models/

Ortega, Pedro A., Vishal Maini, and the DeepMind safety team, “Building safe artificial intelligence: specification, robustness, and assurance” (2018).
https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1