U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY
HEARING CHARTER

Trustworthy AI: Managing the Risks of Artificial Intelligence

Thursday, September 29, 2022
10:30 am – 12:30 pm
Hybrid In-Person/Remote Hearing

PURPOSE

On Thursday, September 29, 2022, the Subcommittee on Research and Technology of the Committee on Science, Space, and Technology will hold a hearing to discuss tools, best practices, and challenges in the design, development, testing, and deployment of trustworthy artificial intelligence (AI) systems. The Subcommittee will examine efforts in academia, industry, and government to create a culture of responsibility around AI systems, identify and remove harmful bias in AI systems, improve explainability and transparency of AI systems, and mitigate other risks associated with AI systems. The Subcommittee will also explore the National Institute of Standards and Technology's ongoing efforts to create an artificial intelligence risk management framework.

WITNESSES

- **Ms. Elham Tabassi,** Chief of Staff, Information Technology Laboratory, National Institute of Standards and Technology
- **Dr. Charles Isbell**, Dean and John P. Imlay, Jr. Chair of the College of Computing, Georgia Institute of Technology
- **Mr. Jordan Crenshaw**, Vice President of the Chamber Technology Engagement Center, U.S. Chamber of Commerce
- **Ms. Navrina Singh**, Founder and Chief Executive Officer, Credo AI

OVERARCHING QUESTIONS

- What are the risks that can arise from the development and deployment of AI systems, including how harmful biases can arise in these systems?
- What are the activities being undertaken by academia, industry, and the government to develop, test, and responsibly deploy trustworthy AI systems?
- How should the United States encourage more organizations to think critically about risks that arise from AI systems, including at the earliest stages of development?
- Where should the Federal government focus efforts to promote the development and deployment of trustworthy artificial intelligence across every sector of the economy?

BACKGROUND

Artificial intelligence refers to the theory and development of computer systems that can perform tasks that would normally require human intelligence, such as decision making or speech recognition. Modern AI systems are engineered or machine-based systems that can, for a given set of human-defined objectives and with varying levels of autonomy, generate predictions, recommendations, or decisions

influencing real or virtual environments.[1] All applications of artificial intelligence in use today can be considered "narrow AI," or AI that is designed to do a very specific set of tasks. In contrast, artificial general intelligence is a theoretical system that possesses generalized human cognitive abilities and, when presented with an unfamiliar and complex problem, could develop solutions drawing from contextual knowledge. Modern systems are likely decades away from achieving artificial general intelligence.

Most AI systems are developed using a technique called machine learning, which involves developing an algorithmic model based on input data, then using that model to make certain optimizations or predictions. An example of this is image recognition, in which a set of human-labeled images (e.g., "traffic lights" in CAPTCHA tests that users take when logging into a website) are fed into an algorithm, which then looks for patterns common to all images with a specific label. The algorithm builds a model (i.e., "learns") from this "training data", so when it is presented with an unlabeled image containing one of the objects that was in the training data, it can make a guess as to what the object is. This method of training algorithms with human-labeled data is called "supervised learning". There is also "unsupervised learning", in which no labels are provided, and the algorithm simply looks for similarities and groups images into clusters based on certain characteristics. Additionally, there is "reinforcement learning", in which an algorithm interacts with its environment, executes actions, and learns through trial and error.

While AI systems have been in use in the commercial sector for decades, recent advances in computing, improved software engineering, and better access to large data sets have markedly increased the capabilities of AI systems. As a result, AI systems have led to a wide range of innovations with the potential to benefit nearly all aspects of our society and support our economic and national security. AI systems are increasingly used in scientific research to help sort and analyze massive amounts of data in fields such as weather prediction, cosmology, and genetics research. Recent advances in natural language processing and image generation have led to AI systems that can write text or generate art.[2]

AI RISKS

While AI-systems have the potential to improve our lives, in sometimes transformative ways, they also have the potential to do significant harm if risks associated with these systems are not mitigated. While risks to any type of information-based system also apply to AI systems (e.g., privacy, cybersecurity, and safety concerns), these systems also create a set of risks that require specific consideration. AI systems can amplify, perpetuate, and exacerbate existing structural inequalities in our society, or create new ones. AI systems can also exhibit unintended properties with potential ethical, safety, or security consequences for individuals or communities. Risks associated with AI systems arise from the data used to train the AI system, the system itself, the use of the system, or interaction of people with the system. Importantly, AI systems and their associated risks are socio-technical, meaning they are a product of the complex human, organizational, and technical factors involved in their design, development, and use. For example, questions of fairness or equity caused by the decisions of AI systems relate to societal dynamics and human behavior. Purely technical solutions will not solve societal challenges.

Harmful Bias

One major set of risks caused by AI systems is harmful bias, which can occur when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning

[1] "AI Risk Management Framework: Second Draft," NIST, August 18, 2022.
[2] "GPT-3 Powers the Next Generation of Apps," OpenAI, March 25, 2021; "DALL·E: Creating Images from Text," OpenAI, January 5, 2021.

process. Bias can be introduced purposefully or inadvertently into an AI system, or it can emerge as the system is being deployed. For example, a facial recognition system trained mostly on light-skinned faces will perform poorly identifying faces with darker skin, and a facial recognition system trained to perfection in a lab may fail when encountering real-world scenarios. Moreover, intentional or unintentional changes during training may fundamentally alter AI system performance.

According to the National Institute of Standards and Technology (NIST), there are three categories of bias.[3] First, systemic biases result when AI systems create advantages for certain social groups while disadvantaging others. Systemic bias is also referred to as institutional or historical bias. Systemic biases can creep their way into datasets or can be reinforced by institutional norms, practices, and processes across the AI lifecycle. Second, statistical and computational biases result from errors that occur due to a sample that the AI system is trained on not being representative of the population. These biases often arise when algorithms are trained on one type of data and cannot extrapolate beyond those data. Finally, human biases reflect systematic errors in human thought. These biases are often implicit and tend to relate to how an individual or group perceives information to make a decision or fill in missing or unknown information. Because AI systems are designed by humans, this type of bias is present across the entire AI lifecycle.

Not all bias is harmful. Statistical and computational biases that arise in an analysis are a normal part of data science. Bias can also be beneficial, such as algorithms that use data on an individual's habits to tailor new content based on their interests. However, many cases of bias can cause significant harm. For example, a self-driving car trained by driving on the roads of Boston may not recognize different patterns in other cities, and an AI diagnostic tool trained on x-ray images of younger patients may fail to perform well on older patients. Combatting harmful bias in AI will require better alignment between AI tasks and actual human goals. While it will require additional technology expertise to improve the detection and mitigation of bias, it will also require an understanding of the relevant social and ethical considerations.

Explainability and Interpretability

Some AI systems are functionally black boxes, which means it is difficult to understand why algorithms make the decisions that they do. For example, one type of machine learning system is called a "neural network," which consists of thousands or even millions of simple processing nodes that are densely interconnected. Training data is fed to the bottom layer and as it passes through the succeeding layers it gets multiplied and added together in complex ways, until it finally arrives at the output layer in its transformed final state. Due to the complexity, scientists are unable to fully understand these interactions in a useful way. Observers can only effectively assess this process by reviewing an algorithm's inputs and outputs.

This challenge has given rise to fields of research focused on assessing and understanding algorithmic decisions. For example, researchers and companies are working to improve algorithmic explainability, or the ability of algorithms to explain their decisions. However, modern explainability techniques come with trade-offs— improving the explainability of algorithms has often come at the cost of accuracy of outputs.[4] In contrast, some researchers are focused on interpretability, which refers to techniques used to understand the meaning of AI systems' output in the context of its designed functional purpose. One area of focus for interpretability is called test, evaluation, validation, and verification (TEVV), which uses

[3] Reva Schwartz et al., "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," NIST, March 2022.
[4] Cynthia Ruden, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, 2019, 206–215.

separate AI actors to examine an AI system or its components or detect and remediate problems throughout the AI lifecycle.[5]

Safety

Ensuring AI systems are safe means preventing them from leading to physical or psychological harm, or creating a state in which human life, health, property, or the environment is endangered.[6] One major challenge of AI safety is ensuring the system can continue to operate safely in unfamiliar situations. For example, modern autonomous vehicles can only operate in certain environments under certain conditions in a safe manner.[7] Another challenge to achieving AI safety is avoiding misspecification, or poor alignment between an AI behavior and the system designer's intentions. Misspecification occurred in YouTube's video recommendation algorithm when an AI system that was optimized for user engagement unintentionally directed users to extremist content.[8]

Safety issues are mostly dealt with through careful design, planning, and testing to prevent failures, conditions, or environments in which it becomes dangerous to use an AI system. According to NIST, practical approaches to AI safety include "rigorous simulation and in-domain testing, real-time monitoring, and the ability to shut down or modify systems that deviate from intended or expected functionality".[9]

Cybersecurity and Privacy

While AI systems are susceptible to the same privacy and security risks as all information-based systems, there are some concerns that are unique to AI systems. AI systems have more complex attack surfaces that can enable malicious actors to compromise their security more easily. For example, malicious actors could theoretically make alterations to open-source datasets to manipulate an AI system to produce an inaccurate or harmful result.[10] Similarly, AI systems could be trained outside an organization's security controls or trained in one domain and then "fine-tuned" for another, resulting in vulnerabilities. As a result, existing privacy and cybersecurity guidance are ill-equipped to ensure the data protection of AI systems.

Computational Costs

Training AI systems requires a large amount of computational power. Since 2012, the amount of computational power used to train the largest AI systems has been increasing exponentially—doubling every 3.4 months.[11] A paper in 2019 found that training a single large-scale AI system required five times as much carbon as the lifetime emissions of the average American car.[12] If the United States is to avert the climate crisis while maintaining its global leadership in AI, the research community and tech industry should explore more efficient AI training methodologies and more efficient computing systems.

---

[5] "AI Risk Management Framework: Second Draft," NIST.
[6] Ibid.
[7] Several automakers have achieved level four automation in their vehicles. See "Levels of Automation" NHTSA, accessed September 22, 2022.
[8] Homa Hosseinmardi et. al., "Examining the consumption of radical content on YouTube," Complex Networks & Their Applications, hosted on Proceedings of the National Academy of Sciences, 2022, 166-177.
[9] "AI Risk Management Framework: Second Draft," NIST.
[10] Andrew Lohn, "Poison in the Well," Center for Security and Emerging Technology, June 2021.
[11] Jack Clark, "AI and Compute," *OpenAI*, May 16, 2018.
[12] Emma Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP," In the 57th Annual Meeting of the Association for Computational Linguistics (ACL), stored in arxiv, July 2019.

GOVERNMENT ACTION

In December 2020, Congress enacted the *National Artificial Intelligence Initiative Act* or NAIIA (P.L. 116-283). This bipartisan legislation, which was led by the House Science Committee, accelerated and coordinated Federal investments and new public-private partnerships in research, standards, and education in trustworthy artificial intelligence. The law establishes interagency coordination and strategic planning efforts in AI research, development, standards, and education through an Interagency Coordination Committee and a coordination office managed by the Office of Science and Technology Policy (OSTP). The legislation also created the National AI Advisory Committee (NAIAC) to assess the implementation of the law, track advancements in AI science, and propose recommendations to advance U.S. competitiveness in AI. The Department of Commerce selected members for the NAIAC in May 2022, with the plan to publish a report in 2023.[13] Finally, the legislation directed the Department of Energy (DOE), the National Science foundation (NSF), and Department of Commerce research agencies to conduct AI-related activities, many of which are designed to assess and mitigate AI-related risks.

OSTP

OSTP has pursued several initiatives related to promoting trustworthy AI. In 2021, OSTP announced an effort to develop a bill of rights for an automated society, also called the "AI bill of rights".[14] OSTP has sought input from the boarder community on what this document should contain. In March 2022, OSTP also sought feedback on updating the National AI Research and Development Strategic Plan, which includes strategic aims to both "understand the ethical, legal, and societal implications of AI" and "ensure the safety and security of AI systems".[15]

National Institute of Standards and Technology

NIST, which is housed within the Department of Commerce, conducts fundamental and applied research and measurement activities to cultivate trust and improve the design, development, and governance of AI systems. NIST published principles of explainable AI in 2020 before NAIIA was enacted.[16] In NAIIA, Congress directed NIST to expand upon these efforts by developing a voluntary AI risk management framework through collaboration with stakeholders across public and private sectors. To date, NIST has held two workshops to develop the AI risk management framework, released two drafts of the framework, and published a draft playbook to help with implementation.[17] NIST plans to publish the first version of the AI risk management framework in January 2023.

In addition, NIST is conducting several other trustworthy AI-related activities, including:

- Developing taxonomy, terminology, and testbeds for measuring risks in AI systems and informing the standards needed for key technical characteristics of AI trustworthiness.
- Developing data characterizations, key practices for data documentation, and datasets that the broader community can use to test or train AI systems while preserving privacy and cybersecurity.

[13] "Commerce Department Launches the National Artificial Intelligence Advisory Committee," Department of Commerce, May 4, 2022.
[14] "Join the Effort to Create A Bill of Rights for an Automated Society," White House, November 10, 2021.
[15] Office of Science and Technology Policy, "Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan," Federal Register, February 2, 2022.
[16] P. Jonathon Phillips et. al., "Four Principles of Explainable Artificial Intelligence," NIST, September 2021.
[17] "AI Risk Management Framework: Second Draft," NIST.

- Coordinating across the government and with industry stakeholders to identify critical standards development activities, strategies, and gaps for trustworthy AI.[18]
- Developing guidance to facilitate voluntary data sharing arrangements among industry, federally funded research centers, and federal agencies to advance AI research and technologies.

National Science Foundation

Achieving the responsible design and deployment of AI also requires integrating ethics into technology education and research at every stage—from K-12 education to AI developers. It requires viewing AI as an interdisciplinary field rather than a purely technical field. NSF funds university research across all non-biomedical disciplines (including social sciences) and numerous STEM education programs. As a result, the agency will play a key role in achieving these goals. In NAIIA, Congress directed NSF to make awards supporting research that contributes to the development of trustworthy AI, supports K-12, undergraduate, and graduate education on trustworthy AI, and creates faculty technology ethics fellowships to support more research into the field of technology ethics.[19] Moreover, the *CHIPS and Science Act of 2022* (P.L. 117-167) directs NSF to establish a requirement for an ethics statement in award proposals to ensure researchers are considering the social implications of their work.[20] NSF also funds a network of 18 AI research institutes, each devoted to a different sector or AI-related challenge. This combined investment of $220 million reaches a total of 40 states and the District of Columbia. In 2021, NSF announced a partnership with NIST to establish an AI research institute on trustworthy AI.[21] The winner of this solicitation should be announced later this year.

International

There are also international conversations taking place surrounding the development and responsible deployment of trustworthy AI. The Organisation for Economic Cooperation and Development (OECD) adopted a set of AI principles for guiding governments in responsible stewardship of trustworthy AI in 2019.[22] Many individual countries have also established their own AI strategies that incorporate ethics to various extents. Singapore was one of the first to develop an AI governance framework in 2019, later iterations of which evolved into a practical toolkit for companies to demonstrate trustworthy AI in a practical manner.[23] The European Union proposed the AI Act in 2021 to harmonize regulations as they relate to AI systems, including a process for self-certification and government oversight of many categories of high-risk AI systems.[24] Because AI risk management is a relatively new activity and organizations are required to self-certify that they control for AI-risks, there is significant uncertainty surrounding the pending EU law's requirements. Many U.S. companies are looking to the NIST AI risk management framework as a possible solution to this dilemma.

PRIVATE SECTOR ACTION

The private sector is also attempting to tackle issues related to developing and deploying trustworthy AI. Companies such as Microsoft, Google, and Intel have all published their own versions of AI ethics

[18] "U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools," NIST, August 9, 2019.
[19] Rep. Eddie Bernice Johnson, *The National AI Initiative Act*, H.R.6216, incorporated into H.R.6395, 116th Cong.
[20] Rep. Eddie Bernice Johnson, *CHIPS and Science Act of 2021*, H.R.4521, incorporated into H.R.4346, 117th Cong.
[21] James Donlon and Rebecca Hwa, "National Artificial Intelligence (AI) Research Institutes," NSF, November 16, 2021.
[22] "OECD AI Principles," OECD, February 2019.
[23] "Singapore's Approach to AI Governance," Singapore Personal Data Protection Commission, May 25, 2022.
[24] "The AI Act" European Union, April 2021.

principles.[25] Many industry groups are also engaging in their own activities to promote trustworthy AI development and deployment. For example, the U.S. Chamber of Commerce has launched a bipartisan commission on AI to "advance U.S. leadership in the use and regulation of AI technology".[26] Many of these principles developed by industry are generally abstract and lack concrete governance structures and accountability measures. However, some major technology companies have begun to develop and implement concrete measures.[27] Other businesses are developing tools and practical methodologies to help organizations assess and mitigate AI-related risks. The Mozilla foundation is funding open-source AI auditing tools.[28] Some companies have developed proprietary tools that enable their clientele to identify and mitigate AI risks.[29]

[25] "Responsible AI," Microsoft, accessed September 21, 2022; Responsible AI Practices," Google, accessed September 21, 2022; "Intel's Recommendations for the U.S. National Strategy on Artificial Intelligence," Intel, March 5, 2019.
[26] "U.S. Chamber Launches Bipartisan Commission on Artificial Intelligence to Advance U.S. Leadership," U.S. Chamber of Commerce, January 18, 2022.
[27] Jon Belkowitz and Leah Koshiyama, "Trust in the Time of AI: Why Salesforce Invests in Ethical Guardrails," Salesforce, April 19, 2022.
[28] "Mozilla Technology Fund Seeks People, Projects Auditing AI Systems with Open-Source Approaches," Mozilla Foundation, September 6, 2022.
[29] For examples, please see ORCAA and Credo AI.