SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT AND SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY
HEARING CHARTER

"*Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence*"

Wednesday, October 18, 2023
10:00 a.m.
2318 Rayburn House Office Building

Purpose

On October 18, 2023, the Committee on Science, Space, and Technology will hold a joint Subcommittee on Investigations and Oversight and Subcommittee on Research and Technology hearing titled "Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence." The purpose of this hearing is to examine the current landscape of research, testing, and deployment of methods and tools for managing risks associated with artificial intelligence systems. The hearing will examine existing research and methodological gaps and where further investment is needed to ensure the creation of an ecosystem for the safe and responsible use of artificial intelligence.

Witnesses

- **Ms. Elham Tabassi,** Associate Director for Emerging Technologies, Information Technology Laboratory, National Institute of Standards and Technology
- **Mr. Michael Kratsios**, Managing Director, Scale AI, 4th Chief Technology Officer of the United States
- **Dr. Emily M. Bender,** Professor of Linguistics, University of Washington
- **Mr. Caleb Watney**, Co-CEO, Institute for Progress

Overarching Questions

- What efforts are currently being undertaken by academia, industry, and government to research, develop, and test methods for the responsible deployment of trustworthy AI systems?
- What types of methods, standards, and tools currently exist for managing risks associated with AI systems?

- Where do fundamental knowledge and methodological gaps exist for mitigating risks associated with AI systems?
- What outstanding technical research questions need to be considered for effective AI governance?
- What role should the federal government play in oversight of AI systems?
- Where should the federal government focus investments to promote the development and deployment of trustworthy AI?
- How can Congress invest in STEM education to develop the workforce needed to maintain U.S. leadership in AI?
- How will international approaches to AI governance influence U.S. policies?

Background

Artificial intelligence (AI) refers to computer systems capable of performing tasks that typically require human intelligence, such as decision-making or content creation. The term AI includes a range of technologies, algorithms, methodologies, and application areas, such as natural language processing, facial recognition, and robotics. Despite its recent popularity, AI is not a completely new technology. "Narrow AI,"[1] or AI that targets singular things, has been widely deployed for decades in various applications like automated warehouse robots, social media recommendation algorithms, and fraud detection in financial systems.

The term "artificial intelligence" was first coined in 1955 by emeritus Stanford Professor John McCarthy as, "the science and engineering of making intelligent machines."[2] Since then, the field progressed slowly until the "machine learning" (ML) approach was popularized in the 2000s, a shift enabled by the proliferation of data on the Internet.[3] Unlike older AI systems which were pre-programmed to follow set rules, ML uses mathematical algorithms to learn patterns in data to make classifications or predictions. For example, ML is the mechanism powering search engine results on Google, recommending new series to watch on Netflix, and the brainpower behind voice assistants like Siri and Alexa.

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. Recognizing this development, Stanford researchers popularized the term "foundation models" in 2021, highlighting these new models' foundational role for building next-generation AI applications.[4] Foundation models form the basis for "generative AI" — models that can create sophisticated writing, images, and other forms of content with minimal human input. Generative AI, including ChatGPT, has been one of the most noteworthy areas of advancement in AI.[5] Underpinned by a type of AI called a large language model (LLM), ChatGPT is trained on a significant amount of text data to understand and generate human-like language. LLMs are useful for a wide range of natural language processing tasks, such as chatbots, language translation, and text summarization.

[1] *Artificial Intelligence Definitions*, (September 2020), Stanford HAI
[2] *Ibid*
[3] Leopold, G. (2016, June 16). *Proliferation of data driving machine learning*. Datanami.
[4] Bommasani, R. (2021, August 16). *On the Opportunities and Risks of Foundation Models*. arXiv.
[5] *Introducing ChatGPT,* (n.d.), OpenAI

Risks

Although AI systems have the potential to significantly improve our lives, they also have the potential to do significant harm. While risks to any type of information-based system also apply to AI (e.g., privacy, cybersecurity, and safety concerns), these systems also create a unique set of dangers that require special attention. AI systems used in the real-world are already sufficiently advanced to cause immediate harm when not properly deployed or configured, and, given the rapid pace of technological development, long-term and structural risks are also present.

Explainability and interpretability: Advanced AI systems are functionally black boxes, which means we cannot easily explain or interpret how they reach the decisions that they do. For instance, deep learning models, which most generative AI systems are built on, use thousands or millions of interconnected nodes and hundreds of different data dimensions to make complex calculations and arrive at an output. Humans can only observe the inputs and outputs of the system; what happens within the algorithm is largely a mystery.

False information: Generative AI systems often produce "hallucinations"—false information that seems plausible. They can occur when users request information not in the training data, or when models fail to "learn" the underlying dataset correctly. Improperly deployed AI systems used for research or decision-making can mislead decision-makers and lead to bad outcomes. A malicious actor could use this flaw to create inaccurate or misleading information in disinformation campaigns.

Computational scarcity: Training AI systems requires a large amount of computational power. An analysis by OpenAI found that the amount of compute required to train the current most advanced models grew 300,000,000% from 2012 to 2017.[6] As a result, talent and cutting-edge innovation are increasingly concentrated in a handful of large companies that can afford the high computation, bandwidth, and storage costs. Additionally, computing resources available to federal agencies are in scarce supply, oftentimes with 3-4x more demand for them than what is available.

Harmful bias: There are three major types of AI bias highlighted by the National Institute of Standards and Technology (NIST).[7] First, systemic biases result when AI systems discriminate against certain groups, creating disadvantages. Second, statistical and computational biases arise from an AI system being trained on a dataset that is not representative of the population. Lastly, human biases reflect systematic errors in human judgment. These biases are often implicit and tend to relate to how an individual or group perceives information (such as the output of an AI system) to make a decision or fill in missing or unknown information. Since AI systems are designed by humans, systematic bias is present across the entire AI lifecycle and in the use of AI applications once deployed.

[6] *AI and Compute*. (n.d.). OpenAI

[7] Reva Schwartz et al., "*Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*," NIST, March 2022

Tools and Resources for AI Risk Management

Artificial Intelligence Risk Management Framework (AI RMF)

At the direction from Congress in the National AI Initiative,[8] NIST worked through a consensus-driven, open, collaborative, and transparent process to develop the AI RMF. NIST launched this voluntary framework in January 2023 which enables organizations to better mitigate risks associated with AI and incorporate trustworthiness into the design, development, use, and evaluation of AI products, services, and systems.[9] In March 2023, NIST launched the Trustworthiness and Responsible AI Resource Center, which will help facilitate the implementation of and alignment of the AI RMF.[10]

Red Teaming

Red teaming involves managing risks of AI by stress-testing systems such as LLMs, seeking to find loopholes that could be leveraged to bypass safety and security models, "make the systems produce undesirable outputs", or fail.[11] According to the White House OSTP, effective red-teaming of AI systems can be useful for identifying risks to AI Safety, including not only traditional security and safety concerns, but also privacy and bias concerns.[12] In 2023, the Biden-Harris Administration sponsored the first ever public assessment red-teaming event with the AI Village at the DEF CON 31 conference.

AI Assurance and Systems Security

Frontier technical approaches and interventions such as "confidential computing"[13] can be applied to AI systems to enhance cybersecurity and authenticity, and protect against adversarial inputs. Some AI and platform vendors have already developed systems and approaches for leveraging confidential compute principles and technologies in securing AI systems and encouraging responsible use of AI.[14]

Automated Evaluations

AI system evaluations, or "evals" for short, involve assessing the performance, fairness, and safety of AI systems. These evaluations aim to identify potential biases, vulnerabilities, and limitations in AI models and algorithms. Evaluations can be created for any objective, such as testing for bias or accuracy. While gold standard evals exist for some types of AI (e.g. NIST's Facial Recognition Vendor Test[15] for facial recognition bias), generative AI evals currently lack standardization given how new and quickly the technology is evolving. The most extensive effort has been Stanford's Holistic Evaluation of Language Models (HELM) benchmark,[16] which tests models against many different metrics such as bias, fairness, accuracy, etc.

---

[8] H.R.6216 - 116th Congress: National Artificial Intelligence Initiative Act of 2020

[9] *Intelligence Risk Management Framework 1.0*, (January 2023), NIST

[10] Trustworthy and Responsible AI Resource Center, (n.d.), NIST

[11] *Red-Teaming Large Language Models to Identify Novel AI Risks*, (23 August 2023) OSTP

[12] *Ibid*

[13] Confidential Computing, (n.d.), IBM

[14] Confidential AI, (23 May 2023) Microsoft

[15] Face Recognition Vendor Test (FRVT), (30 November 2023), NIST

[16] Holistic Evaluation of Language Models (HELM), (19 September 2023), Stanford CRFM

Developer Transparency Practices
The AI community has begun normalizing releasing public documents along with their models to increase transparency and consumer trust. These include model cards (describing relevant information for different user audiences), datasheets (documentation for the datasets used to train the model), and instructions for automated evaluations (see 'Automated Evaluations' above).

Watermarking
Watermarking is a technique that involves embedding digital marks or indicators into machine learning models or datasets to enable their identification. In the context of AI-generated content, watermarking has gained popularity as a means to curb misuse of AI-generated images. By hiding a signal in an image, watermarking can help identify whether the image was created by an AI system. However, traditional watermarking methods, such as visible overlays or metadata additions, can be easily removed or lost when images are cropped or edited. While advancements have been made in creating more robust watermarks, there is still a need for further R&D to ensure their effectiveness over time.

Audits and Technical Standards
Audits, impact assessments, and mandates for access to company data are being mainstreamed as "algorithmic accountability" tools. Algorithmic audits are difficult regulatory tools to implement because they require consensus on what constitutes bias/harm and clear standards and methodologies for conducting the audit. Standards are required before mainstream audits can be implemented — and this is still a work in progress. Some existing AI standards are being conducted at IEEE[17] and NIST is working to align its AI RMF with international ISO/IEC standards.[18]

Voluntary Commitments on Generative AI
On July 21, 2023, the White House announced that seven major AI companies – Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI – agreed to eight voluntary commitments on sharing, testing, and developing generative AI technologies to ensure safety, security, and trustworthiness.[19] On September 12, 2023, the White House announced an additional eight companies – Adobe, Cohere, IBM, NVIDIA, Palantir, Salesforce, Scale AI, and Stability – that had agreed to the voluntary commitments.[20] The eight commitments are:

1. Commit to internal and external red-teaming of models or systems in areas including misuse, societal risks, and national security concerns, such as bio, cyber, and other safety areas.
2. Work toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards.
3. Invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights .
4. Incentivize third-party discovery and reporting of issues and vulnerabilities.

[17] Autonomous and Intelligent Systems, (n.d.), IEEE
[18] *Roadmap for the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, (n.d.), NIST
[19] Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, 21 July 2023, OSTP
[20] Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI, (12 September 2023), OSTP

5. Develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content.
6. Publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias.
7. Prioritize research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy.
8. Develop and deploy frontier AI systems to help address society's greatest challenges.

Federal Research Agencies Activities

National Institute of Standards and Technology (NIST)

NIST contributes to research, standards, and data to realize the full potential of AI as a tool to enable American innovation, enhance economic security, and improve our quality of life. NIST is conducting several 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; coordinating across the government and with industry stakeholders to identify critical standards development activities, strategies, and gaps for trustworthy AI;[21] and developing guidance to facilitate voluntary data sharing arrangements among industry, federally funded research centers, and federal agencies to advance AI research and technologies.

NIST also leads and participates in the development of technical standards, including international standards, that promote innovation and public trust in systems that use AI. Unlike most countries that have a top-down, government-led approach, the U.S. has a bottom-up, industry-led approach to standards-setting. The U.S. employs a voluntary system, which relies on industry participation and leadership. A market-driven approach enables competition, ensures transparency, and takes advantage of consensus-building to drive us to the best possible outcomes.

National Science Foundation (NSF)

NSF supports advancing AI R&D across core and cross-cutting programs within the agency with focuses ranging from data and advanced computing, workforce training, and social and economic sciences. The National AI Initiative[22] 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 encourage the incorporation of ethical considerations and principles into the research and development of AI systems. NSF also funds a network of 25 AI research institutes, each devoted to a different sector or AI-related emerging application, ranging from agriculture to cybersecurity to education.[23]

[21] "U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools," (9 August 2019) NIST

[22] H.R.6216 - 116th Congress: National Artificial Intelligence Initiative Act of 2020

[23] *NSF Announces 7 New National Artificial Intelligence Research Institutes.* (n.d.). NSF - National Science Foundation

Department of Energy (DOE)
DOE operates the world's most advanced supercomputers through the National Labs, making it the default agency to conduct frontier AI research that requires extensive computational resources. DOE explores impactful scientific questions that can be uniquely solved by applying AI using its supercomputers.[24] The National AI Initiative[25] directed DOE to support R&D to advance AI tools, systems, capabilities, and workforce needs and to improve the reliability of AI methods and solutions within the agency. Additionally, DOE's Artificial Intelligence and Technology Office (AITO) developed its own AI Risk Management Playbook[26] in consultation with NIST.

Department of Homeland Security (DHS)
DHS's Science and Technology Directorate (S&T) leads AI research to strengthen capabilities that safeguard the Nation and accelerate advancements in science and technology. S&T's research includes understanding the full potential and risks of AI, safeguarding critical infrastructure systems from adversarial AI, and developing AI expert workforces. In April 2023, Secretary Mayorkas established the DHS AI Task Force, which is charged with advancing the application of AI to critical homeland security missions.[27]

International Approaches to AI
United Kingdom (U.K.)
In March 2023, the U.K. announced a £3.5 billion investment in advancing science and technology.[28] The package includes £900 million for a new supercomputer with a portion dedicated to an AI Research Resource, similar to the U.S. National AI Research Resource (NAIRR) proposed by the NAIRR Task Force in January. One month later, the U.K. announced an additional £100 million for foundation model research and commercialization.[29] The U.K. has also declared a "pro-innovation approach to AI regulation"[30] and designated an official technology diplomat to the U.S., who recently visited Silicon Valley executives in June 2023.[31] Next month, the U.K. will host the Global Summit on AI Safety, which aims to bring together key countries, as well as leading technology organizations, academia and civil society to inform national and international action for AI development.[32]

European Union (E.U.)
The European Parliament, a main legislative branch of the E.U., passed the draft version of the E.U. AI Act on June 13th, 2023. The legislation mandates use and development requirements based on classifying AI systems by risk. The bill also introduces constraints on broad applications and processes, such as strongly curtailing uses of facial recognition software, and

[24] *AI for Science, Energy, and Security Report.* (n.d.). Argonne National Laboratory.
[25] H.R.6216 - 116th Congress: National Artificial Intelligence Initiative Act of 2020
[26] DOE AI Risk Management Playbook, (n.d.), DOE
[27] *Secretary Mayorkas Announces New Measures to Tackle A.I., PRC Challenges at First State of Homeland Security Address*, (2023 April 21), DHS
[28] *Government Commits Up to £3.5 Billion to Future of Tech and Science*. (n.d.). GOV.UK.
[29] *Initial £100 Million for Expert Taskforce to Help Uk Build and Adopt Next Generation of Safe AI*. (n.d.). GOV.UK.
[30] *AI Regulation: A Pro-innovation Approach*. (n.d.). GOV.UK.
[31] Bordelon, B. (2023, June 14). *The British Diplomat Trying to Win Over the U.S. Tech Industry*. POLITICO.
[32] "*UK government sets out AI Safety Summit ambitions",* (4 September 2023), GOV.UK

requires companies to publish summaries of copyrighted material used for training generative AI systems.

People's Republic of China (PRC)
The PRC is the second global leader in private AI investments behind the U.S., totaling $13.4 billion behind the U.S.'s $47.4 billion.[33] Although this private investment gap reflects the U.S.'s lead in R&D, the PRC is closing the gap through AI industrial policy which invests billions through state-financed investment funds, designates "national AI champions," and provides preferential tax treatment to grow AI startups.[34] [35] By many metrics, the PRC has caught up or surpassed the U.S. in research and commercial capabilities. For instance, nine of the top ten universities ranked by number of AI papers published in 2021 were from the PRC (the 10th was the Massachusetts Institute of Technology). PRC-published papers also received nearly the same share of citations as US researchers (22% vs. 24%) and the PRC installed more automated industrial robots than the rest of the world combined in 2021.[36] The PRC also officially implemented a K-12 AI curriculum and is on track to produce nearly 2x more STEM PhDs as the U.S. by 2025. A recent report by the Center for Security and Emerging Technology (CSET) found that U.S. investors played a key role in fueling the PRC's AI rise, accounting for nearly one-fifth of all investments in PRC AI companies from 2015 to 2021 which totaled $40.2 billion, or 37% of the total amount raised during the six-year period.[37]

In August 2023, the Cyberspace Administration of China's (CAC) Generative AI Measures came into effect and appear to be some of the strictest regulations of its kind. The regulations state that generative AI services should not generate content "inciting subversion of national sovereignty or the overturn of the socialist system," or "advocating terrorism or extremism, promoting ethnic hatred and ethnic discrimination, violence and obscenity, as well as fake and harmful information."[38]

Russia
While Russia has used AI-enabled autonomous weapons in Syria and Ukraine, it lags far behind the world in research and commercial output. A recent study by Stanford found that Russia only had 3 authors on significant machine learning papers in 2022, compared to the US's 285 and the PRC's 49. The study also found Russia produced one 'significant' system in 2022 compared to the US's 16, the UK's 8, and the PRC's 3.[39]

[33] Maslej, et al. (April 2023). "*The AI Index 2023 Annual Report*." Stanford Institute for Human-Centered AI.
[34] *Understanding Chinese Government Guidance Funds*.. (n.d.). *Center for Security and Emerging Technology*.
[35] *China Creates National New Generation Artificial Intelligence Innovation and Development Pilot Zones.* (n.d.). Center for Security and Emerging Technology.
[36] Maslej, et al. (April 2023). "*The AI Index 2023 Annual Report*." Stanford Institute for Human-Centered AI.
[37] Emily S. Weinstein and Ngor Luong, "*U.S. Outbound Investment into Chinese AI Companies*" (Center for Security and Emerging Technology, February 2023).
[38] https://time.com/6314790/china-ai-regulation-us/
[39] Maslej, et al. (April 2023). "*The AI Index 2023 Annual Report*." Stanford Institute for Human-Centered AI.

Further Reading

Artificial Intelligence Basics

- CRS - Artificial Intelligence: Overview, Recent Advances, and Considerations for the 118th Congress
- IBM - Artificial Intelligence Basics

Generative AI

- GAO - Science & Tech Spotlight: Generative AI
- CRS - Generative Artificial Intelligence and Data Privacy
- CRS - Generative Artificial Intelligence: Overview, Issues, and Questions for Congress
- Stephen Wolfram - What is ChatGPT doing and why does it work?
- a16z - Who owns the generative AI platform?

Trustworthy AI and risks

- NIST - AI Risk Management Framework
- Google, OpenAI, Berkeley, Stanford - Concrete Problems in AI Safety
- Center for Strategic and International Studies - The Path to Trustworthy AI

National Security

- CRS - Deep Fakes and National Security
- The National Security Commission on Artificial Intelligence Report

Tools and Resources for AI Risk Management

- Identifying AI-generated images with SynthID
- NSF partners with the Institute for Progress to test new mechanisms for funding research and innovation
- Inside the White House-Backed Effort to Hack AI