

**BALANCING KNOWLEDGE AND
GOVERNANCE: FOUNDATIONS
FOR EFFECTIVE RISK MANAGEMENT
OF ARTIFICIAL INTELLIGENCE**

HEARING
BEFORE THE
SUBCOMMITTEE ON INVESTIGATIONS
AND OVERSIGHT
SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY
OF THE
COMMITTEE ON SCIENCE, SPACE,
AND TECHNOLOGY
OF THE
HOUSE OF REPRESENTATIVES
ONE HUNDRED EIGHTEENTH CONGRESS
FIRST SESSION

OCTOBER 18, 2023

Serial No. 118-27

Printed for the use of the Committee on Science, Space, and Technology



Available via the World Wide Web: <http://science.house.gov>

U.S. GOVERNMENT PUBLISHING OFFICE

53-782PDF

WASHINGTON : 2025

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

HON. FRANK LUCAS, Oklahoma, *Chairman*

BILL POSEY, Florida	ZOE LOFGREN, California, <i>Ranking Member</i>
RANDY WEBER, Texas	SUZANNE BONAMICI, Oregon
BRIAN BABIN, Texas	HALEY STEVENS, Michigan
JIM BAIRD, Indiana	JAMAAL BOWMAN, New York
DANIEL WEBSTER, Florida	DEBORAH ROSS, North Carolina
MIKE GARCIA, California	ERIC SORENSEN, Illinois
STEPHANIE BICE, Oklahoma	ANDREA SALINAS, Oregon
JAY OBERNOLTE, California	VALERIE FOUSHEE, North Carolina
CHUCK FLEISCHMANN, Tennessee	KEVIN MULLIN, California
DARRELL ISSA, California	JEFF JACKSON, North Carolina
RICK CRAWFORD, Arkansas	EMILIA SYKES, Ohio
CLAUDIA TENNEY, New York	MAXWELL FROST, Florida
RYAN ZINKE, Montana	YADIRA CARAVEO, Colorado
SCOTT FRANKLIN, Florida	SUMMER LEE, Pennsylvania
DALE STRONG, Alabama	JENNIFER McCLELLAN, Virginia
MAX MILLER, Ohio	TED LIEU, California
RICH MCCORMICK, Georgia	SEAN CASTEN, Illinois,
MIKE COLLINS, Georgia	<i>Vice Ranking Member</i>
BRANDON WILLIAMS, New York	PAUL TONKO, New York
TOM KEAN, New Jersey	
VACANCY	

SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT

HON. JAY OBERNOLTE, California, *Chairman*

BRIAN BABIN, Texas	VALERIE FOUSHEE, North Carolina,
MAX MILLER, Ohio	<i>Ranking Member</i>
RICH MCCORMICK, Georgia	KEVIN MULLIN, California
VACANCY	JEFF JACKSON, North Carolina

SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY

HON. MIKE COLLINS, Georgia, *Chairman*

JIM BAIRD, Indiana	HALEY STEVENS, Michigan,
DARRELL ISSA, California	<i>Ranking Member</i>
RICK CRAWFORD, Arkansas	ANDREA SALINAS, Oregon
SCOTT FRANKLIN, Florida	KEVIN MULLIN, California
BRANDON WILLIAMS, New York	EMILIA SYKES, Ohio
TOM KEAN, New Jersey	SUZANNE BONAMICI, Oregon

C O N T E N T S

October 18, 2023

	Page
Hearing Charter	2

Opening Statements

Statement by Representative Jay Obernolte, Chairman, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, U.S. House of Representatives	11
Written Statement	13
Statement by Representative Valerie Foushee, Ranking Member, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, U.S. House of Representatives	14
Written Statement	15
Statement by Representative Mike Collins, Chairman, Subcommittee on Research and Technology, Committee on Science, Space, and Technology, U.S. House of Representatives	16
Written Statement	17
Statement by Representative Haley Stevens, Ranking Member, Subcommittee on Research and Technology, Committee on Science, Space, and Technology, U.S. House of Representatives	17
Written Statement	18
Written statement by Representative Frank Lucas, Chairman, Committee on Science, Space, and Technology, U.S. House of Representatives	19
Written statement by Representative Zoe Lofgren, Ranking Member, Committee on Science, Space, and Technology, U.S. House of Representatives	68

Witnesses:

Ms. Elham Tabassi, Associate Director for Emerging Technologies, Information Technology Laboratory, National Institute of Standards and Technology	
Oral Statement	20
Written Statement	22
Mr. Michael Kratsios, Managing Director, Scale AI 4th Chief Technology Officer of the United States	
Oral Statement	33
Written Statement	36
Dr. Emily M. Bender, Professor of Linguistics, University of Washington	
Oral Statement	42
Written Statement	44
Mr. Caleb Watney, Co-CEO, Institute for Progress	
Oral Statement	50
Written Statement	52
Discussion	60

Appendix I: Answers to Post-Hearing Questions

Ms. Elham Tabassi, Associate Director for Emerging Technologies, Information Technology Laboratory, National Institute of Standards and Technology	86
Mr. Michael Kratsios, Managing Director, Scale AI 4th Chief Technology Officer of the United States	95

IV

	Page
Dr. Emily M. Bender, Professor of Linguistics, University of Washington	98
Mr. Caleb Watney, Co-CEO, Institute for Progress	102

Appendix II: Additional Material for the Record

Statement submitted by Representative Valerie Foushee, Ranking Member, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, U.S. House of Representatives Statement for the Record of Rep. Anna G. Eshoo	106
Letter submitted by Representative Haley Stevens, Ranking Member, Sub- committee on Research and Technology, Committee on Science, Space, and Technology, U.S. House of Representatives Merve Hickok, President, et al., Center for AI and Digital Policy	107
Article submitted by Representative Suzanne Bonamici, Committee on Science, Space, and Technology, U.S. House of Representatives “How Flawed Data Aggravates Inequality in Credit,” Edmund L. An- drews, HAI, Stanford University	116

**BALANCING KNOWLEDGE AND
GOVERNANCE: FOUNDATIONS
FOR EFFECTIVE RISK MANAGEMENT
OF ARTIFICIAL INTELLIGENCE**

WEDNESDAY, OCTOBER 18, 2023

HOUSE OF REPRESENTATIVES, SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT JOINT WITH THE SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY, COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY,

Washington, D.C.

The Subcommittees met, pursuant to notice, at 2:33 p.m., in room 2318 of the Rayburn House Office Building, Hon. Jay Obernolte [Chairman of the Subcommittee on Investigations and Oversight] presiding.



**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 Test¹⁵ 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 IEEE¹⁷ 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 Initiative²² 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 Initiative²⁵ 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 Playbook²⁶ 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 U.K. 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](#)

Chairman OBERNOLTE. The joint hearing will come to order. Without objection, the Chair is authorized to declare a recess of the hearing at any time.

I'd like to welcome everyone today to today's hearing entitled "Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence (AI)." I now recognize myself for five minutes for an opening statement.

I'd like to welcome everyone to what I am hopeful will be a very fruitful discussion on foundational questions that Congress needs to answer to ensure that we strike a careful balance between protecting consumers and protecting innovation as we create a regulatory framework for artificial intelligence.

Earlier this summer, the Science Committee held a hearing to explore how Congress can ensure that AI technology advances our national interest. One of the unifying takeaways from that hearing was that there remains a number of unsolved technical challenges, which, if we address them, would advance innovation while making AI systems safer, more transparent, and more easily—easier to implement guardrails around.

Today's hearing will build upon that theme by exploring how the Science Committee can support research, testing, and the development of methods and tools for managing AI risks. These tools and methods will be critical to the good governance of these systems as Congress examines how we should regulate artificial intelligence.

I believe that foundational research on AI is a necessary precondition for safer systems. While regulations can make undesirable actions unlawful, technological advances in coordination with other countries and monitoring who has access to the compute necessary to train frontier AI models and how it's being put to use will be critical tools in supporting and enforcing these regulations.

One example of this is the use of confidential computing to make the theft of AI models more difficult. It is currently relatively straightforward to steal an AI model because the data needed to operate the model, known technically as the model weights, is stored in raw data files. Stealing these files would allow criminals to use the model without spending the millions of dollars on compute and data necessary to train and develop the model. Stricter cybersecurity laws can only go so far in preventing this from occurring. However, recent research into confidential computing could potentially enable the model to be operated without allowing access to the model weights, rendering cyber theft of the model impossible and addressing the root causes of this problem.

Another challenge that advances in research can help resolve is identifying whether AI has been used to generate content. It's clear that instances of identity fraud, plagiarism, and a host of other issues will become more and more common as the power of AI tools increase. Watermarking is a technique where digital content is encrypted with a unique, often hidden identifier that provides information about its origin. Watermarking has the potential to identify the origin of any content, regardless of how it was created or digitally altered. Although current discussions of watermarking center on desire that all AI-generated content be watermarked, watermarking can also use to prove the provenance of authentic content. In fact, I believe it's entirely possible that in the future,

people might automatically assume that any content is AI generated unless its watermark proves its authenticity. Solving this technical problem would enable a new set of good governance tools concerning generative AI that were previously infeasible.

These examples illustrate the fact that research and policy are not mutually exclusive but, in fact, are mutually dependent. Research advances unlock previously unpractical approaches to regulation, and smart policy accelerates research.

I believe we must also standardize technical definitions for methodologies, risks, and technological concepts across all of our agencies of government. The NIST (National Institute of Standards and Technology) AI Risk Management Framework (AI RMF) lays out a foundation that other agencies can build upon. As these agencies consider passing rules or using procurement authorities to incentivize good behavior, they should also ensure a consistent methodology for assessing risk levels and tailor their policies accordingly.

Another technical area in which Congress has a crucial role to play is promoting and establishing best practices. This includes everything from technical standards and evaluation benchmarks to testing the trustworthiness and risks of AI systems. And let us not forget that Congress should certainly not rush to overregulate, but we should also not be complacent. The United States must avoid falling behind other major world powers who are finalizing their AI standards and regulations. Without proactive American leadership, supremacy in AI could be seized by the EU or China, both of whom are taking far more draconian approaches to AI regulation. Because the United States currently leads the rest of the world in AI research and development (R&D), it makes no sense for us to stand by while American companies are forced to either make educated guesses or to play by others' rules.

We must also ensure that our academic institutions continue to play an active role in research and development of AI and that the transparent system of publication and peer review that has enabled its success thus far is not replaced with an opaque system where cutting-edge research is only performed by corporations. This is why I believe it's critical that we establish a National Artificial Resource—Research Resource—NAIRR, we call it—as a shared national research infrastructure to ensure researchers have access to the tools needed to test, develop, and create new AI-backed technologies.

This idea was proposed earlier this Congress when I joined my fellow AI caucus chairs in introducing H.R. 5077, the *Creating Resources for Every American to Experiment with Artificial Intelligence Act*, or the *CREATE Act* we call it. See what we did there. Establishing a shared computing and data infrastructure resource such as that detailed in the act would democratize access to AI by providing researchers and students across scientific fields and disciplines with access to compute resources and high-quality data, along with the appropriate educational tools and user support.

I would like to thank all of our witnesses for taking the time this afternoon to join us for this important discussion. I'm looking forward to hearing your recommendations to how this Committee can strengthen our Nation's leadership in artificial intelligence and set

clear rules for the safe, responsible, and human-centered development of this exciting new technology.

[The prepared statement of Chairman Obernolte follows:]

Good afternoon. Welcome to what I am hopeful will be a very fruitful discussion on foundational questions Congress needs to answer to ensure that we strike a careful balance between protecting consumers and protecting innovation as we create a regulatory framework for artificial intelligence.

Earlier this summer the Science Committee held a hearing to explore how Congress can ensure that AI technology advances our national interest.

One of the unifying takeaways was that there remains a number of unsolved technical challenges which, if addressed, would advance innovation while making AI systems safer, more transparent, and easier to implement guardrails around.

Today's hearing will build upon that theme by exploring how the Science Committee can support research, testing, and the deployment of methods and tools for managing AI risks. These tools and methods will be critical to the good governance of these systems as Congress examines how it should regulate AI.

I believe that fundamental research on AI is a necessary precondition for safer systems. While regulations can make undesirable actions unlawful, technological advances and coordination with other countries in monitoring who has access to the compute necessary to train frontier AI models and how it is being put to use will be critical tools in supporting and enforcing regulations.

One example is the use of confidential computing to make the theft of AI models more difficult. It is currently relatively straightforward to steal an AI model because the data needed to operate the model—known as the model weights—is stored in raw files. Stealing these files would allow criminals to use the model without spending the millions of dollars on compute and data necessary to train the model. Stricter cybersecurity laws can only go so far to prevent this from occurring. However, recent research into confidential computing could potentially enable the model to be operated without allowing access to the model weights, rendering cyber-theft impossible and addressing the root cause of the problem.

Another challenge that advances in research can help resolve is identifying whether AI has been used to generate content. It is clear that instances of identity fraud, plagiarism, and a host of other issues will become more common as the power of AI tools increases. Watermarking is a technique whereby digital content is encrypted with unique, often hidden, identifiers that provide information about its origin. Watermarking has the potential to identify the origin of any content regardless of how it was created or digitally altered. Although current discussions of watermarking center on a desire that AI-generated content be watermarked, watermarking can also be used to prove the provenance of authentic content. In fact, it is entirely possible that in the future, people may automatically assume that all content is AI-generated unless its watermark proves its authenticity.

Solving this technical problem would enable a new set of good governance tools concerning generative AI that were previously infeasible.

These examples illustrate the fact that research and policy are not mutually exclusive, but in fact mutually dependent. Research advances unlock previously impractical approaches to regulation, and smart policy accelerates research.

I believe that we must also standardize technical definitions for methodologies, risks, and technological concepts across the government. The NIST AI Risk Management Framework lays out a foundation that other agencies can build upon. As these agencies consider passing rules or using procurement authorities to incentivize good behavior, they should ensure a consistent methodology for assessing risk levels and tailor their policies accordingly.

Another technical area in which Congress has a crucial role to play is promoting and establishing best practices. This includes everything from technical standards to evaluation benchmarks for testing the trustworthiness and risks of AI systems.

Let us not forget that while Congress should certainly not rush to overregulate, we should also not be complacent. The U.S. must avoid falling behind other major world powers who are finalizing their AI standards and regulations.

Without proactive American leadership, supremacy in AI could be seized by the EU or China, both of which are taking far more draconian approaches to AI regulation.

Because the U.S. currently leads the rest of the world in AI research and development, it makes no sense for us to stand by while American companies are forced to either make educated guesses or play by others' rules.

We must also ensure that our academic institutions continue to play an active role in research and development of AI, and that the transparent system of publication and peer review that has enabled its success thus far is not replaced with an opaque

system where cutting-edge research is only performed by corporations. This is why I believe it is critically important that we establish a National Artificial Intelligence Research Resource (NAIRR) as a shared national research infrastructure to ensure researchers have access to the tools needed to test, develop, and create new AI-backed technologies. This idea was proposed earlier this Congress when I joined my fellow AI Caucus Chairs in introducing H.R. 5077, the *Creating Resources for Every American To Experiment with Artificial Intelligence Act (CREATE AI Act)*. Establishing a shared computing and data infrastructure resource, such as that detailed in the Act, would democratize access to AI by providing researchers and students across scientific fields and disciplines with access to compute resources and high-quality data, along with appropriate educational tools and user support.

I want to thank all our witnesses for taking the time to join us today for this important discussion.

I look forward to hearing your recommendations for how this committee can strengthen our nation's leadership in artificial intelligence and set clear rules for the safe, responsible, and human-centered development of this exciting new technology.

Chairman OBERNOLTE. I'd now like to recognize the Ranking Member of the Investigations and Oversight Subcommittee, the gentlewoman from North Carolina, for her opening statement.

Mrs. FOUSHEE. Chairman Obernolte and Chairman Collins, thank you for holding today's hearing on this important and timely subject, and certainly, thank you to our witnesses.

Artificial intelligence is an emerging technology that holds the potential to transform many aspects of human society, and I am incredibly impressed with the quality of scientific research currently being conducted in the realm of AI. In my district, North Carolina's 4th, Duke University is leading the Athena Institute, which is an NSF (National Science Foundation)-funded multidisciplinary research institute focused on AI research for edge computing related to communications networks. And UNC (University of North Carolina) Chapel Hill is participating in the Artificial Intelligence Institute for Engaged Learning, another NSF-funded research institute working on the development of AI tools to enhance educational learning opportunities. I have met with these researchers. Their work is exciting, and their brilliance, ingenuity, and commitment to developing AI applications that serve the public good are second to none.

At the same time, there is no escaping the reality that certain kinds of AI systems and their applications could pose grave risks in the absence of clear, rigorous oversight rooted in practical and ethical considerations. AI risk management is a complex subject that defies easy answers, but there are some guiding principles that should be at the front of our minds as we think about these issues.

First, there is a pressing need to develop effective scientific tools and test methods that can support researchers and policymakers in evaluating the benefits and risks of different AI systems. This is an area where Federal agencies such as NIST can play a leading role.

Second, it is crucial for the Federal Government to prioritize research into the benefits and risks of different AI systems and applications and then fund that research accordingly. AI risk management must be a Federal research priority because it is essential for understanding AI's broader impact on society.

And finally, as we race to keep up with the breakneck pace of AI research and the accompanying risk, our discussion must be

centered around the real-world risks and safety concerns that arise from AI systems. The debate over AI risk management has an unfortunate tendency to become fixated on dramatic existential risks like the end of human civilization, which are based in speculation. Those risks are legitimate subjects for research inquiry, but they should not be elevated over concrete, tangible concerns in areas such as equity, bias, privacy, cybersecurity, and disinformation. These are the existing fact-based risks to real people in the real world, not distractions pulled from the world of science fiction, and they deserve to be prioritized in the AI Risk Management Framework to come.

The Science Committee is the right place for this kind of discussion. We pride ourselves on our responsible and bipartisan approach to complex questions like this one with deliberations grounded in scientific fact and informed by a broad range of perspectives. I'm confident that today's hearing will continue in that tradition.

As a Member of the New Democrat Coalition's Artificial Intelligence Working Group, I believe this is exactly the kind of issue that deserves more of our attention in Congress. Chairman Obernolte, I have great respect for your leadership on AI issues, and I am eager to use today's hearing to explore areas where we may be able to work together on AI-related oversight in the future.

And I yield back.

[The prepared statement of Mrs. Foushee follows:]

Chairman Obernolte and Chairman Collins, thank you for holding today's hearing on this important and timely subject. Artificial Intelligence is an emerging technology that holds the potential to transform many aspects of human society. And I am incredibly impressed with the quality of scientific research currently being conducted in the realm of AI.

In my district, North Carolina's Fourth, Duke University is leading the Athena Institute, an NSF-funded multidisciplinary research institute focused on AI research for edge computing related to communications networks. And UNC Chapel Hill is participating in the Artificial Intelligence Institute for Engaged Learning, another NSF-funded research institute working on the development of AI tools to enhance educational learning opportunities. I have met with these researchers. Their work is exciting, and their brilliance, ingenuity, and commitment to developing AI applications that serve the public good are second-to-none.

At the same time, there is no escaping the reality that certain kinds of AI systems and their application could pose grave risks in the absence of clear, rigorous oversight rooted in practical and ethical considerations. AI risk management is a complex subject that defies easy answers. But there are some guiding principles that should be at the front of our minds as we think about these issues.

First, there is a pressing need to develop effective scientific tools and test methods that can support researchers and policymakers in evaluating the benefits and risks of different AI systems. This is an area where federal agencies such as NIST can play a leading role. Second, it is crucial for the federal government to prioritize research into the benefits and risks of different AI systems and applications, and then fund that research accordingly. AI risk management must be a federal research priority because it is essential for understanding AI's broader impact on society. Finally, as we race to keep up with the breakneck pace of AI research and the accompanying risks, our discussion must be centered around the real-world risks and safety concerns that arise from AI systems.

The debate over AI risk management has an unfortunate tendency to become fixated on dramatic existential risks, like the end of human civilization, which are based in speculation. Those risks are legitimate subjects for research inquiry. But they should not be elevated over concrete, tangible concerns in areas such as equity, bias, privacy, cybersecurity, and disinformation. These are the existing, fact-based risks to real people in the real world, not distractions pulled from the world of science fiction. And they deserve to be prioritized in the AI risk management framework to come.

The Science Committee is the right place for this kind of discussion. We pride ourselves on our responsible and bipartisan approach to complex questions like this one, with deliberations grounded in scientific fact and informed by a broad range of perspectives. I'm confident that today's hearing will continue in that tradition.

As a member of the New Democrat Coalition's Artificial Intelligence Working Group, I believe this is exactly the kind of issue that deserves more of our attention in Congress. Chairman Obernolte, I have great respect for your leadership on AI issues, and I'm eager to use today's hearing to explore areas where we may be able to work together on AI-related oversight in the future.

I want to thank our expert witnesses for appearing before the committee today and for your thoughtful testimony on this topic. I look forward to a vigorous and engaging discussion.

Thank you, and I yield back.

Chairman OBERNOLTE. Thank you, Mrs. Foushee.

I now recognize the Chairman of the Research and Technology Subcommittee, the gentleman from Georgia, for his opening statement.

Mr. COLLINS. Thank you, Mr. Chairman. Good afternoon, and thank you again to our panel of witnesses for sharing their experience with us here today. I look forward to hearing your recommendations on where the Federal Government should focus investments to promote the development and deployment of effective artificial intelligence systems.

The purpose of this hearing is to explore an important question: What tools and resources still need to be developed in order to create guardrails on AI? I also hope we can use this hearing as an opportunity to learn where gaps exist and where further investment is needed to achieve this goal. Federal science agencies, academia, and industry are carrying out critical research to develop and test methods for the responsible methods and tools for managing risk from AI.

In fact, my home State of Georgia is leading the way. Earlier this month, Georgia State University received a \$10 million grant from the Department of Defense (DOD) to establish the Center of Excellence in Advanced Computing and Software. The Center of Excellence will help researchers address critical problems in AI and robotics. In addition, three Georgia institutions have received investments totaling \$60 million to lead three National Science Foundation AI research institutes, and they will play a critical role in facilitating public-private partnerships to advance research and translate into practical applications.

As AI systems continue to be developed at scale, we must continue to address existing knowledge gaps and build off the progress we have thus made so far. This will be critical for the United States to lead in AI and to enable American innovation, enhance economic and national security, and maintain our global competitive edge.

I'm proud that this Committee and this chamber have led the way in holding public hearings to discuss these critical issues. As Congress grapples with how to regulate AI, it is critical that it is done in an open and transparent manner, unlike our colleagues in the Senate.

I once again want to thank our witnesses for their participation here today, and I yield back, Mr. Chairman.

[The prepared statement of Mr. Collins follows:]

Good afternoon. Thank you to our panel of witnesses for sharing their expertise with us here today.

I look forward to hearing your recommendations on where the federal government should focus investments to promote the development and deployment of effective artificial intelligence systems.

The purpose of this hearing is to explore an important question—what tools and resources still need to be developed in order to create guardrails on AI?

I also hope we can use this hearing as an opportunity to learn where gaps exist and where further investment is needed to achieve that goal.

Federal science agencies, academia, and industry are carrying out critical research to develop and test methods for the responsible methods and tools for managing risks from AI.

In fact, my home state of Georgia is leading the way.

Earlier this month, Georgia State University, received a \$10 million grant from the Department of Defense to establish the Center of Excellence in Advanced Computing and Software. The Center of Excellence will help researchers address critical problems in AI and robotics.

In addition, three Georgia Institutions have received investments totaling \$60 million to lead three National Science Foundation AI Research Institutes, which will play a key role in facilitating public-private partnerships to advance research and translate it into practical applications.

As AI systems continue to be deployed at scale, we must continue to address existing knowledge gaps and build off the progress we have made thus far. This will be critical for the U.S. to lead in AI and to enable American innovation, enhance economic and national security, and maintain our global competitive edge.

I am proud that this Committee and this Chamber have led the way in holding *PUBLIC* hearings to discuss these critical issues. As Congress grapples with how to regulate AI, it is critical that it is done in an open and transparent manner, unlike our colleagues in the Senate.

I once again would like to thank our witnesses for their participation here today.

Chairman OBERNOLTE. Thank you, Mr. Collins.

I'll now recognize the Ranking Member of the Research and Technology Subcommittee, the gentlewoman from Michigan, for her opening statement.

Ms. STEVENS. Thank you, Mr. Chair.

The Science Committee meets. The work continues. It is a real delight to be with you, Chairman Obernolte and obviously Ranking Member Foushee, our colleagues from the Oversight and Investigations Committee and, of course, the Chairman of the Research and Technology Subcommittee, Mr. Collins. And we welcome our very distinguished panel of witnesses. You have very informative testimony, and we're quite grateful.

So in the grand tradition of this Committee, we have worked together to promote trustworthy AI. And I have helped pass legislation on this very Committee targeting deepfakes and have worked to include trustworthy AI provisions in the *CHIPS and Science Act* that passed Congress last year, and we're thrilled to see that working its way through NIST. Mr. Kean and I have a bipartisan bill, the *Privacy Enhancing Technology Research Act*, which just passed through this Committee a few months ago, and as soon as the Congress is back open, we hope to see it hit the floor, and will support—you know, this bill supports privacy-enhanced data sets and tools for AI training systems. And anyone in this AI conversation knows we need to do PETs (privacy-enhancing technologies).

So today's hearing is also a very nice follow up from a hearing that we held last year that I chaired on the Research and Technology Subcommittee to discuss AI risks broadly, AI risks and biases. We had a very informative hearing, particularly on the bias risks, and throughout the—you know, some of the racial disparities that come up. And so it's certainly quite remarkable how rapidly

AI technology has advanced just in the last year since we had that hearing. Its potential benefits are obviously boundless. We could go on and on on that front, but so, too, are the risks, and not just the hypothetical, but the very much here and now.

And, fortunately, we have seen a swift response and good leadership from President Biden and our Federal agencies to assess these risks and threats. And the Biden Administration released its framework for an AI Bill of Rights just last year and multiple, multiple Executive orders (EOs) to address AI risks. And I was especially pleased to see the National Institute of Standards and Technology released the AI Risk Management Framework in January, bravo, the mighty agency that could and will. And that's just a very important step in AI safety that was developed in the 2020 *National AI Initiative Act*, another product of this very Committee. And so this document is already helping organizations understand and mitigate the risks associated with these technologies. And I'm excited to see more major AI companies adopt this baseline for AI risk management.

We're going to continue to work rigorously to develop meaningful and robust governance frameworks for trustworthy AI. That is absolutely our charge. We've got several TIME100 AI leaders as witnesses today, no pressure. You know, we're going to look back in history on your leadership and the work that you're doing. Financial AI systems will have different risks than generative AI chatbots. We're going to need to develop technical standards to mitigate risks across sectors and use cases. Testing and evaluation (T&E) of AI systems will also be critical to enable AI governance. After all, it's very hard to regulate what you cannot measure, as our friends at NIST know.

I think in conclusion here, it's fair to say that we're in a big competition globally on the technology front. I sit on the Select Committee on Strategic U.S. Competitiveness with the Chinese Communist Party (CCP). We want to have the risk framework, we want to have the mitigation, we want to have ownership of the technology here. You know, we've heard from other NIST witnesses, according to international indicators, indicators that we don't even measure as a nation, that China is leading in 27 out of 34 critical sectors. So AI is a place that the United States wants to continue to lead on. We want to set the standards for the world. And we're so excited to get exploring and discovering and deepening our knowledge here today.

Thank you so much, Mr. Chair, and I yield back.

[The prepared statement of Ms. Stevens follows:]

Thank you to Chairman Obernolte and Ranking Member Foushee for holding this hearing jointly with Chairman Collins and myself, and welcome to our distinguished panel of witnesses.

I have long worked with my colleagues on both sides of the aisle to promote trustworthy AI. I have helped pass legislation targeting deepfakes and worked to include trustworthy AI provisions in the *CHIPS and Science Act*. My *Privacy Enhancing Technology Research Act*, which passed through this Committee just a few months ago, will support privacy-enhanced datasets and tools for training AI systems.

Today's hearing will follow up on a hearing we held in the Research and Technology Subcommittee last September to discuss AI risks broadly. It is remarkable how rapidly AI technology has advanced just in the last year. Its potential benefits are boundless. Unfortunately, so too are its risks—not just the hypothetical but very much the here and now.

Fortunately, we have seen a swift response by President Biden and our federal agencies to these risks and threats. The Biden Administration released its Framework for an AI Bill of Rights last year and multiple Executive Orders to address AI risks. I was especially pleased to see the National Institute of Standards and Technology released the AI risk management Framework in January, an important step in AI safety that was developed by the 2020 *National AI Initiative Act*, a Science Committee product. This document is already helping organizations understand and mitigate the risks associated with these technologies, and I am excited to see more major AI companies adopt this as a baseline for AI risk management.

However, we are still in the early days of developing a meaningful governance framework for trustworthy AI. There is a lot of hard work ahead of us to be able to adequately define and measure the risks associated with AI systems. As we learned during our hearing last year, context is critical in AI risk management. Financial AI systems will have different risks than generative AI chatbots. We will need to develop technical standards to mitigate risks across sectors and use cases. Testing and evaluation of AI systems will also be critical to enable governance. After all, it's very hard to regulate what you can't measure.

I look forward to an update on NIST's work to develop technical standards and AI risk profiles for different sectors and use cases. I am particularly interested in understanding NIST's plans for promoting the adoption of the risk management framework across the public and private sectors.

Another major challenge I hope that we address in this hearing is that of the AI risk management workforce. Federal agencies and regulators who are seeking to address AI-related challenges lack a workforce capable of understanding and responding to AI risks. Most private sector organizations deploying AI systems also lack such expertise. As we focus on development of the AI workforce broadly, we must not ignore this critical segment of the workforce and the particular skills they will require.

I'm looking forward to discussing today what more Congress can do to ensure the United States leads the world in trustworthy artificial intelligence.

Thank you and I yield back.

Chairman OBERNOLTE. Thank you, Ms. Stevens.

[The prepared statement of Chairman Lucas follows:]

Thank you, Chairman Obernolte and Chairman Collins, for holding this hearing today. Artificial intelligence is transformational technology, and it deserves thoughtful, informative discussions about any potential government action in this area.

As we discussed during our last hearing on AI, rushing to regulate this technology could have detrimental effects on our ability to innovate and maintain American leadership. We want to be sure that we allow this technology to grow and advance, and to be sure that we develop it in a safe and trustworthy manner that maintains American values of fairness and transparency.

That requires a measured approach, one which I'm proud to say our Committee has long embraced. From the time we passed the *National AI Initiative Act*, we've focused on the strategic development of AI and risk management standards.

A crucial part of that legislation was directing the National Institute of Standards and Technology (NIST) to develop an AI Risk Management Framework that could serve as a tool for the trustworthy design, development, use, and evaluation of AI.

I'm very pleased with the work NIST put into this product, which was developed with years of input from stakeholders, government, industry, and academia.

This represents the kind of approach we need to ensure AI remains safe, trustworthy, and fair.

Other international entities are moving forward with regulations as we speak. The European Union is taking a particularly prescriptive approach to managing the risks of AI.

While we cannot afford to fall too far behind other nations, I urge caution to anyone who would move forward too quickly with strict regulations here in the U.S.

We need to find a balance between allowing the technology the freedom to grow and developing a framework that helps us manage potential risks from AI.

That's why today's hearing is so important. We can't rush to regulate before knowing the tools and resources we need in place in order to successfully approach this governance challenge.

AI is grabbing big headlines right now and I understand the impulse to make rushed decisions about how to manage it. But that leads us down unproductive paths. Our friends in the Senate, for instance, are choosing to make an end run around the normal committee process and hold confidential meetings with big tech companies instead. The White House, instead of steadily building on the smart risk

management framework put out by NIST, is issuing reports, bills of rights, and executive orders somewhat haphazardly.

I'm not arguing that we shouldn't regulate the development of AI at all, just that we should approach any new requirements methodically and openly so we can ensure the best possible outcome for Americans and American businesses.

I believe that if we approach this correctly, we will get this right and ensure the safe, trustworthy, and fair development of AI.

Thank you to our witnesses for joining us today, and I look forward to a productive discussion.

Chairman OBERNOLTE. We'll move now to our witness panel. Our first witness today is Ms. Elham Tabassi, the Director for Emerging Technologies of the Information Technology Laboratory at the National Institute of Standards and Technology. She also leads NIST's Trustworthy & Responsible AI Program and the development of the National AI Risk Management Framework that seeks to cultivate trust in the design, development, and use of AI technologies. Ms. Tabassi, you are recognized for five minutes.

**TESTIMONY OF MS. ELHAM TABASSI,
ASSOCIATE DIRECTOR FOR EMERGING TECHNOLOGIES,
INFORMATION TECHNOLOGY LABORATORY,
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY**

Ms. TABASSI. [inaudible] and Members of the Subcommittee, thank you for the opportunity to testify today on NIST's efforts to advance trustworthy and responsible AI. My name is Elham Tabassi. I'm Associate Director for Emerging Technologies at the Information Technology Laboratory in the Department of Commerce's National Institute of Standards and Technology, or NIST. I also lead the NIST Trustworthy & Responsible AI program.

AI technologies provide enormous opportunities for positive impact. They can also cause cascading negative consequences if proper safeguards are not in place. NIST works to advance research, standards, measurements, and tools to manage AI risks and realize the full promise of this technology for all people. NIST works with the AI community to develop and improve its resources. We pride ourselves in our engagements and collaborations, and the trust we have built with the private sector, government agencies, civil society, and academia.

Among its many AI-related activities, NIST released the AI Risk Management Framework, or AI RMF, in January 2023 after considerable collaborations with stakeholders. Directed by congressional mandate, the AI RMF is a voluntary framework that provides a flexible, structured, and measurable process to address the AI risks purposefully and continually throughout the AI lifecycle. The AI RMF offers a resource to the organization's designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. The framework was developed through an open, transparent, and collaborative process that allowed for a broad range of stakeholder input. NIST also released a companion NIST AI RMF playbook and a roadmap. The playbook provides additional guidelines to organizations on the action they can take to meet the outcomes included in the framework. The roadmap identifies key activities for advancing the AI RMF that could be carried

out by NIST in collaboration with public or private organizations or by those organizations independently.

To support the operationalization of the AI RMF, NIST established the online Trustworthy & Responsible AI Resource Center (AIRC) in March 2023. It serves as a one-stop shop for foundational content, technical documents, and AI toolkits. In June 2023, NIST launched a generative AI public working group to help NIST develop a profile of AI RMF to address the risks of generative AI technologies.

The AI technologies are advancing at an extraordinary pace. For that reason, the AI RMF and its related documents will evolve over time to reflect new knowledge and practices. NIST continues its robust engagement with stakeholders to keep the framework up to date with AI trends and reflect experience based on the use of AI RMF.

To build on NIST's work on the AI RMF and provide additional guidelines to organizations to advance trustworthy and responsible AI, NIST conducts fundamental research on many of the AI trustworthiness characteristics and prioritizes its work based on its insights and the community's stated needs.

New NIST efforts in AI evaluation will focus on sociotechnical aspects of system functionality and performance in addition to accuracy. In particular, the evaluations have the goal of identifying risks and harms of AI systems before they are deployed and to establish metrics and evaluation infrastructure that will allow AI developers and deployers to detect the extent to which AI systems exhibit negative impacts or harms.

NIST engages with partners around the world to advance shared goals in trustworthy AI. NIST also coordinates with other Federal agencies and leads several policymaking and interagency efforts, including administering the National Artificial Intelligence Advisory Committee, or NAIAC, which advises the President and the National AI Initiative Office.

Advancing AI research, evaluation, and standards that contribute to a secure, private, interoperable, innovative, and world-leading digital economy is a top priority for NIST. Thank you for the opportunity to present on NIST's activities to improve AI trustworthiness. I look forward to your questions.

[The prepared statement of Ms. Tabassi follows:]

Testimony of

Elham Tabassi
Associate Director for Emerging Technologies
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 Investigations and Oversight and
Subcommittee on Research & Technology

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

October 18, 2023

Chairman Obernolte, Chairman Collins, Ranking Members Foushee and Stevens and members of the subcommittees, I am Elham Tabassi, Associate Director for Emerging Technologies at the Information Technology Laboratory in the Department of Commerce's National Institute of Standards and Technology (NIST). I also am the lead for NIST's trustworthy and responsible artificial intelligence (AI) program. 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 AI.

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 (ITL), we work to cultivate trust in information technology and metrology. Trust in the digital economy is built on key attributes of technology like cybersecurity, privacy, usability, interoperability, equity, and avoiding bias and increasing usefulness. NIST conducts fundamental and applied research and advances standards to understand and measure limits and capabilities of information technology. We develop tools to measure and evaluate those technologies. These 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. They provide increased assurance and utility, enabling more secure, private, and rights-affirming technologies. More than ever before, our economy and society sorely need and depend on these kinds of standards and measurements. That is especially true when it comes to AI.

We are aware of the enormous recent increase in interest in AI. We know that the world is demanding collaborative, sophisticated frameworks for the development and deployment of trustworthy AI; infrastructure for evaluations of AI systems; and coordinated standards around which AI can be defined, built, and assessed. This is at the heart of NIST's work. We are maximizing the impact of our limited resources to meet the urgent needs of this moment.

NIST's Role in Artificial Intelligence

NIST contributes to the research, standards, measurements, and data required to realize the full promise of AI, while managing its risks, to 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 public and private sectors. NIST seeks and relies on diverse stakeholder insights and feedback from government, industry, academia, and non-profit entities to develop and improve its resources.

The collaborative, transparent, and open processes NIST uses to prioritize, develop, and carry out its research and to produce its guidelines result in more effective and usable resources that are trusted and, therefore, widely used by federal agencies, as well as private sector organizations of all sizes, educational institutions, and state, local, tribal, and territorial governments. Our inclusion of international experts and organizations in most of these processes means that NIST's work is more likely to help shape the way that others around the globe approach issues related to AI standards and guidelines. This enhances the opportunity for alignment.

NIST’s team includes some of the top AI and standards experts in the world. Our staff has multidisciplinary backgrounds from industry, government, and academia with deep experience in various aspects of science and engineering related to AI.

Much of NIST’s AI effort¹ 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
- contributing to discussions and development of AI policies, including supporting the National AI Advisory Committee².

NIST AI Risk Management Framework

Among its many AI-related activities, NIST developed the AI Risk Management Framework³ (AI RMF 1.0). Released in January 2023, the AI RMF is a voluntary framework that provides a flexible, structured, and measurable process to address AI risks purposefully and continually throughout the AI lifecycle. **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 provides a set of outcomes that enable dialogue, understanding, and actions to manage AI risks.

As directed by the National Artificial Intelligence Initiative Act of 2020 (Division E of P.L. 116-283, the National Defense Authorization Act for Fiscal Year 2021), the AI RMF offers a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. It has been exceptionally well-received nationally and internationally, where public and private entities in the U.S. and abroad are adopting or incorporating it in their responsible AI practices.⁴

The Framework is intended to be voluntary, rights-preserving, non-sector-specific, and use-case agnostic, providing flexibility to organizations of all sizes and in all sectors and throughout society to implement the approaches in the Framework. The AI RMF is designed to be practical, to adapt to the AI landscape as AI technologies continue to develop, and to be operationalized by organizations in varying degrees and capacities so society can benefit from AI while also being protected from its potential harms.

¹ <https://www.nist.gov/artificial-intelligence>

² <https://www.ai.gov/naiaac/>

³ <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf>

⁴ <https://www.nist.gov/itl/ai-risk-management-framework/perspectives-about-nist-artificial-intelligence-risk-management>

The Framework was 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 workshops⁵, respond to a Request for Information (RFI)⁶, and review draft reports⁷ and other documents including draft approaches⁸ and versions of the framework⁹.

NIST also has reached out directly to AI practitioners along with other stakeholders across a full spectrum of interests domestically and internationally. This outreach included companies, government agencies, academia, and not-for-profit organizations representing civil society, consumers, and industry. NIST 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 AI RMF defines certain key characteristics of trustworthy AI systems and offers guidelines for mapping, measuring, and managing them. As defined in the AI RMF, trustworthy AI is valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced, and fair with their harmful biases managed. 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 also released a companion NIST AI RMF Playbook¹⁰ and a roadmap¹¹. The Playbook provides additional guidelines to organizations on the actions they can take to meet the outcomes included in the Framework. The roadmap identifies key activities for advancing the AI RMF that could be carried out by NIST in collaboration with private and public sector organizations – or by those organizations independently.

To support and operationalize the AI RMF and playbook, NIST established the online Trustworthy and Responsible AI Resource Center¹² in March 2023. The resource center is a platform to support people and organizations in government, industry, and academia driving technical and scientific innovation in AI. It serves as a one-stop-shop for foundational content, technical documents, and AI toolkits such as a glossary, a repository hub for measurement methods and metrics. In the coming months, NIST will add to the Resource Center a tracker of AI-related standards and a curated selection of public data sets, so that the public can engage with trustworthy and responsible AI technologies and standards.

In June of 2023, NIST launched a Generative AI Public Working Group¹³ of volunteers led by NIST staff to help NIST develop a profile of AI RMF for generative AI. To address the risk of generative AI technologies, this public working group is helping to develop four sets of

⁵ <https://www.nist.gov/itl/ai-risk-management-framework/ai-risk-management-framework-workshops-events>

⁶ <https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development-request-information>

⁷ <https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf>

⁸ https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf

⁹ <https://www.nist.gov/itl/ai-risk-management-framework>

¹⁰ <https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook>

¹¹ <https://www.nist.gov/itl/ai-risk-management-framework/roadmap-nist-artificial-intelligence-risk-management-framework-ai>

¹² <https://airc.nist.gov/home>

¹³ <https://www.nist.gov/news-events/news/2023/06/biden-harris-administration-announces-new-nist-public-working-group-ai>

guidelines: pre-deployment verification and validation of generative AI models; digital content provenance; incident disclosure; and governance of generative AI systems. This Working Group has attracted about 1,000 participants who are providing NIST with substantive input that will lead to very practical guidelines on generative AI.

AI research and development, as well as the standards landscape, are developing at an extraordinary pace. 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.

NIST's Research on AI Trustworthiness Characteristics

To build on NIST's work on the AI RMF and provide additional guidelines to organizations to advance trustworthy and responsible AI, NIST also conducts fundamental research on many of the AI trustworthiness characteristics, prioritizing its work based on its insights and the community's stated needs.

» 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 harmful bias managed.

Fairness in AI includes concerns for equality and equity by addressing issues such as bias and discrimination. While certain forms of discrimination may violate federal civil rights laws, standards of fairness in AI 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 published "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (NIST Special Publication 1270)¹⁴, which identifies the concepts and challenges associated with bias in AI and provides preliminary guidelines 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 approaches to address 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. NIST has made this aspect of trustworthiness a premier area of focus and has been recognized widely for doing so.

» AI Trustworthiness Characteristics – Explainable and Interpretable

Explainability and interpretability are important characteristics to ensure that users and operators of AI can understand the decisions or predictions made by AI, thus avoiding the "opaque system"

¹⁴ <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf>

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 system's 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)¹⁵ and "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence" (NISTIR 8367)¹⁶.

» *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"¹⁷ (NIST AI 100-2) to advance a taxonomy for securing applications of AI, specifically, adversarial machine learning. NIST's Cybersecurity Framework¹⁸ is widely used to address the cybersecurity risks of organizations. NIST periodically updates that Framework to account for changes in the cybersecurity technology, standards, and risk landscape.

NIST has built an experimentation testbed called Dioptra¹⁹ 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. It is publicly available on GitHub, and we intend to continue maintaining and updating it for public use. To that end, the testbed has a modular design that enables researchers to easily swap in alternative datasets, models, attacks, and defenses. This should advance the measurement capabilities needed to ultimately help secure AI systems. We know that the need for these types of testbeds is growing.

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

NIST's Privacy Risk Assessment Methodology²⁰, developed in 2016 and NIST's Privacy Framework²¹, issued in 2020, are voluntary tools that organizations from 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

¹⁵ <https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence>

¹⁶ <https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence>

¹⁷ <https://www.nccoe.nist.gov/ai/adversarial-machine-learning>

¹⁸ <https://www.nist.gov/cyberframework>

¹⁹ <https://pages.nist.gov/dioptra/>

²⁰ <https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/resources>

²¹ <https://www.nist.gov/privacy-framework/privacy-framework>

Engineering Collaboration Space²², 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 co-sponsored the U.S.-U.K. prize competition on PETs²³. First announced at the Summit for Democracy in December 2021, the winning solutions competed for a combined U.S.-U.K. prize pool of \$1.6 million and the winners were showcased at the second Summit for Democracy in March 2023.

Research on Applications of AI

NIST's multidisciplinary laboratories and varied fields are an ideal environment to develop and apply AI²⁴. Various AI techniques are being used to support NIST scientists and engineers, drawing on machine learning (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 our research efforts – including hardware for AI²⁵, 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.

Just in the last year, NIST's scientists have designed new circuits for supercomputing,²⁶ developed a deep learning algorithm to interpret breathing patterns,²⁷ engineered a new AI-enabled method to digitally simulate hurricanes,²⁸ and used AI to detect disease in human breath molecules.²⁹

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 requirements for AI evaluations. NIST also hosts a biweekly AI metrology colloquia series³⁰, where leading researchers share current work on AI measurement and evaluation.

NIST has been engaged in focused efforts to establish common terminologies, definitions, and

²² <https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space>

²³ <https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/prize-challenges>

²⁴ <https://www.nist.gov/applied-ai>

²⁵ <https://www.nist.gov/artificial-intelligence/hardware-ai>

²⁶ <https://www.nist.gov/news-events/news/2022/10/nists-superconducting-hardware-could-scale-brain-inspired-computing>

²⁷ <https://www.nist.gov/news-events/news/2022/12/wi-fi-could-help-identify-when-youre-struggling-breathe>

²⁸ <https://www.nist.gov/news-events/news/2023/03/ai-could-set-new-bar-designing-hurricane-resistant-buildings>

²⁹ <https://www.nist.gov/news-events/news/2023/04/jilas-frequency-comb-breathalyzer-detects-covid-19-excellent-accuracy>

³⁰ <https://www.nist.gov/programs-projects/ai-measurement-and-evaluation/ai-metrology-colloquia-series>

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 a vital – but 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 they are deployed, and to define 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 projects³¹ 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 Media Forensics Challenge, Activities in Extended Video (ActEV), and handwriting recognition and translation evaluation.
- *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 advances in search technology. A 2010 NIST study³² estimated that, without TREC, U.S. internet users would have spent an estimated 3.5 billion 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, 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. That role was reinforced recently when NIST was charged by the Administration to lead the National Standards Strategy for Critical

³¹ <https://www.nist.gov/programs-projects/ai-measurement-and-evaluation/nist-ai-measurement-and-evaluation-projects>

³² <https://trec.nist.gov/pubs/2010.economic.impact.pdf>

and Emerging Technology³³. We have recently begun seeking public input on how to best implement that strategy; we will begin reviewing comments at the close of public comment in early November.³⁴

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 Tools³⁵, 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, working across the government and with 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 Group³⁶. 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 other nations, including Brazil and India, to coordinate R&D AI initiatives. 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 launched a new AI sub-working group which developed a joint AI roadmap on Evaluation and Measurement Tools for Trustworthy AI³⁷ in December 2022, through which NIST is working towards common definitions for AI risk management and developing metrics and methodologies for measuring AI trustworthiness. Progress in implementation of the joint map was reported to the TTC in May 2023, including the launch of three expert groups to focus on AI terminology and taxonomy, standards and tools for trustworthy AI and risk management, and monitoring and measuring AI risks. The groups have:

- (i) issued a list of 65 key AI terms essential to understanding risk-based approaches to AI³⁸, and
- (ii) mapped the involvement of the United States and the European Union in standardization activities with the goal of identifying relevant AI-related standards of mutual interest.

³³ <https://www.nist.gov/standardsgov/usg-nss>

³⁴ <https://www.nist.gov/news-events/news/2023/09/nist-seeks-input-implementation-national-standards-strategy-critical-and>

³⁵ https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

³⁶ <https://www.nist.gov/standardsgov/icsp-ai-standards-coordination-working-group-aiscwg-charter>

³⁷ https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf

³⁸ <https://www.nist.gov/document/eu-us-terminology-and-taxonomy-artificial-intelligence>

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)³⁹, on behalf of the Department of Commerce. Established by the National Artificial Intelligence Initiative Act of 2020, NAIAC is tasked with advising the President and the National AI Initiative Office. NIST provides executive support to this advisory committee. The Secretary of Commerce appointed the 26 members in April 2022. NAIAC held its first meeting in May 2022. In its year one report⁴⁰, the NAIAC presented findings and recommendations related to promoting leadership in trustworthy artificial intelligence and leadership in research and development; supporting the U.S. workforce and providing opportunity; and international collaboration. Subsequent publications have included instructive explainers such as the FAQs on Foundation Models and Generative AI⁴¹ document.

NIST also co-chairs the National Science and Technology Council's Machine Learning and Artificial Intelligence Subcommittee⁴², the Networking and Information Technology Research and Development's (NITRD) AI Working group⁴³, and the NITRD Fast Track Action Committee⁴⁴ which drafted 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). As NIST's AI lead, I also serve as Federal AI Standards Coordinator and a member of the National AI Research Resource Task Force,⁴⁵ which released its final report in January 2023⁴⁶.

Conclusion

Advancing AI research and standards that contribute to a secure, private, interoperable, innovative, 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 standards and guidelines is a dynamic and critical challenge. Through robust collaboration with stakeholders across government, industry, and academia in the U.S. and elsewhere, NIST is cultivating trust and fostering an environment that enables AI innovation on a global scale – ensuring that we can all gain the benefits of AI while managing its risks.

NIST's team includes some of the top AI and standards experts in the world. Our staff has multidisciplinary backgrounds from industry, government, and academia with expertise and experience in a wide range of science and engineering related to AI. Working with our partners in other federal agencies, the private sector, academia, and allied countries, and with the support of Congress, we have made important progress and will continue to 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.

³⁹ <https://www.nist.gov/artificial-intelligence/national-artificial-intelligence-advisory-committee-naiac>

⁴⁰ <https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf>

⁴¹ <https://www.ai.gov/wp-content/uploads/2023/09/FAQs-on-Foundation-Models-and-Generative-AI.pdf>

⁴² https://www.ai.gov/about/#MLAI-SC_Machine_Learning_and_AI_Subcommittee

⁴³ <https://www.ai.gov/a-new-nitrd-iwg-for-artificial-intelligence-ai-rd/>

⁴⁴ <https://www.nitrd.gov/coordination-areas/privacy-rd/appdsa/>

⁴⁵ <https://www.ai.gov/nairr/>

⁴⁶ <https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf>

Elham Tabassi
Associate Director for Emerging Technologies
Information Technology Laboratory
National Institute of Standards & Technology

In addition to serving as Associate Director for Emerging Technologies in NIST's Information Technology Laboratory (ITL) Elham Tabassi leads NIST's 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. Tabassi is the principal architect of NIST Fingerprint Image Quality (NFIQ), an international standard for measuring fingerprint image quality which has been deployed in many large-scale biometric applications worldwide. Among her other roles at NIST, Tabassi has served as ITL Chief of Staff.

She is a member of the National AI Resource Research Task Force, the US Government's AI Standards Coordinator, a senior member of IEEE, and a fellow of Washington Academy of Sciences. In September 2023, Tabassi was named by TIME magazine as one of the "100 Most Influential People in AI."

Chairman OBERNOLTE. Thank you, Ms. Tabassi.

Our next witness is Michael Kratsios. Mr. Kratsios is the current Managing Director at Scale AI. Prior to joining Scale, Mr. Kratsios served as the fourth Chief Technology Officer (CTO) of the United States where he oriented national technology policy around ensuring American leadership in emerging technologies, including overseeing the implementation of the National AI Initiative. He also served as the Acting Undersecretary of Defense for Research and Engineering, and currently serves as a Distinguished Fellow at the Council on Competitiveness.

Mr. Kratsios, you are recognized for five minutes.

**TESTIMONY OF MR. MICHAEL KRATSIOS,
MANAGING DIRECTOR, SCALE AI
4TH CHIEF TECHNOLOGY OFFICER OF THE UNITED STATES**

Mr. KRATSIOS. Chairmen Obernolte and Collins, Ranking Members Foushee and Stevens, and Members of the Investigations and Oversight and Research and Technology Subcommittees, thank you for the opportunity to testify here today. My name is Michael Kratsios. I'm the Managing Director of Scale AI. Previously, I served as the CTO of the United States and also as Acting Undersecretary of Defense for Research and Engineering.

I'd like to use my statement today to make two points. First, large language models (LLMs) mark a milestone in AI, but they're not our first brush with the promise and the risks of this technology. Policymakers have grappled with AI oversight for years. Since 2016, successive Administrations and Congress have laid a robust foundation for AI policy and regulation. This includes a report on the future of AI under President Obama and the Nation's first AI strategy and multiple Executive orders under President Trump. When Trump Administration made AI a national technology priority, he committed to doubling Federal AI research spending, created first-of-a-kind national AI research institutes, developed the world's first guidance for the regulation of AI in the private sector, and collaborated with allies to develop the world's first intergovernmental AI policy guidelines at the OECD (Organization for Economic Cooperation and Development). Congress codified many of these efforts in the bipartisan 2020 *National AI initiative* and *AI in Government* acts. The Biden Administration continued work on AI, releasing a blueprint for an AI Bill of Rights, securing a voluntary—securing voluntary commitments from tech companies, and publishing an AI Risk Management Framework.

Unfortunately, implementation of existing legislation and Executive orders has fallen short. A study by Stanford found that less than 40 percent of the legal requirements associated with the *AI in Government Act* and the two Executive orders have been implemented. While considering additional legislation or pursuing new administrative actions on AI, policymakers should first ensure that Federal agencies fully implement existing laws and follow through with requirements from AI Executive orders.

When contemplating new AI regulation, lawmakers should pursue a use case and sector-specific risk-based approach rooted in high-quality testing and evaluation. This approach tailors the rigor of the evaluation to the level of risk posed by each use case. For

example, AI-powered medical diagnostics should undergo more stringent testing and evaluation compared to lower risk applications like a movie recommendation algorithm.

And we're not starting from scratch here. The current OMB (Office of Management and Budget) guidance to agencies on regulating AI in the private sector already takes this approach. I believe this guidance remains fair and appropriate. The Biden Administration has recognized the value of testing and evaluation through their support of the DEF CON AI red teaming event in August. The event gathered over 2,000 participants to red team eight leading LLMs on a test evaluation platform built by Scale.

In short, there's been a tremendous amount of work on AI oversight to date across the Federal Government. The prudent path forward is to build on existing efforts and adopt new use cases—and adopt use-case, sector-specific, risk-based guidelines for responsible AI deployment.

Second, to ensure the deployment and development of safe, secure, and responsible AI, we must harness the full strength of America's unique innovation ecosystem of government, industry, and academia. Historically, the Federal Government has focused on funding fundamental R&D, and industry has commercialized those discoveries. However, rapid advancements in the state-of-the-art of AI are being propelled today primarily by private companies. This landscape requires ecosystem collaboration to ensure we develop safe and secure AI.

Industry is participating in a number of consensus-building forums to address gaps in and conduct research on AI standards. Scale recently published its own technical methodology for LLM testing and evaluation, which combines automated assessment with human evaluation of AI systems.

Within the Federal Government, the Department of Energy's national labs represent a unique strategic asset both for fundamental AI research and for addressing potential risks. DOE is leading—world-leading supercomputing capabilities, multidisciplinary expertise, and university partnerships make it a natural home for the U.S. Government's work on addressing frontier risk. Additionally, Congress could take a major step forward by formally establishing and funding the NAIRR. This would create a critical shared research infrastructure, expanding access to computational capabilities, high-quality datasets, and educational resources.

Beyond research and development, we should convene stakeholders, provide direction, and assert American leadership at international standards organizations. The United States should continue to engage with our partners and allies across international fora to ensure American values continue to underpin AI's development. We must not see this ground to authoritarian governments who do not share our values. And our Nation's universities must continue to tackle complex research problems and lend their expertise to collaborative working groups.

I truly believe that the United States can and must maintain and expand our global leadership in artificial intelligence. American leadership in AI is crucial for maintaining the economic and national security of the United States.

Thank you again for this opportunity, and I look forward to your questions.

[The prepared statement of Mr. Kratsios follows:]

The Honorable Michael J.K. Kratsios

Fourth Chief Technology Officer of The United States
The White House

Former Under Secretary of Defense for Research and Engineering (A)
United States Department Of Defense

Managing Director
Scale AI

Distinguished Fellow
Council on Competitiveness

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

Statement before the Science, Space, and Technology
Joint Investigations & Oversight and Research & Technology Subcommittees
United States House of Representatives

Washington, D.C.

October 18, 2023

Chairman Obernolte, Chairman Collins, Ranking Member Foushee, Ranking Member Stevens, and Members of the Investigations and Oversight and Research and Technology Subcommittees, thank you for the opportunity to be here today to testify on the critical topic of developing safe, secure, and responsible artificial intelligence.

I am honored to be here today to discuss these topics with you.

My name is Michael Kratsios, and I am the Managing Director of Scale AI, an artificial intelligence company headquartered in San Francisco. Prior to my time at Scale, I served as the fourth Chief Technology Officer of the United States at the White House and also as the Acting Under Secretary of Defense for Research and Engineering at the Pentagon.

While at the White House, I helped architect the country's first national AI strategy, the American AI Initiative.

Scale, where I work today, was founded in 2016 with the mission of accelerating the development of AI. Scale works with leading frontier model builders like OpenAI and Meta, as well as with the federal government, including the Department of Defense's Chief Digital and AI Office and the U.S. Army.

Today, Scale fine-tunes, red teams, or tests and evaluates nearly all of the leading frontier large language models (LLMs). This gives Scale a unique vantage point to understand and advance the development of safe, secure, and trustworthy AI.

* * *

I'd like to use my statement to make two points.

First, the advent of large language models represents a turning point in artificial intelligence. However, as remarkable as these systems are, they are not our first encounter with the promise and risks of AI. For years, policymakers have grappled with AI oversight. The prudent path forward relies on building on existing efforts and adopting use case and sector-specific, risk-based guidelines for responsible AI deployment.

Since 2016, successive administrations and Congress have laid a robust foundation for AI policy and regulation. This includes a report on the future of AI under President Obama and the nation's first AI strategy, the American AI Initiative, and multiple executive orders under President Trump.

The Trump Administration made AI a national technology priority. It committed to doubling federal AI research spending,¹ created first-of-a-kind national AI research institutes,² developed the world's first guidance for the regulation of AI in the private sector,³ and collaborated with allies to develop the world's first intergovernmental AI policy guidelines at the OECD.⁴

Congress codified many of these efforts in the 2020 National AI Initiative and AI in Government Acts, including the establishment of the National AI Initiative Office at the White House to coordinate AI policy efforts across agencies and offering a blueprint for the creation of a National AI Research Resource (NAIRR).

The Biden Administration has continued work on AI, releasing a Blueprint for an AI Bill of Rights⁵ and earlier this year securing voluntary commitments from a number of technology companies regarding AI safety.⁶ In January of this year, NIST released the government's first AI Risk Management Framework, in fulfillment of another directive of the National AI Initiative Act of 2020.⁷

Unfortunately, implementation of existing legislation and the 2019 and 2020 Executive Orders has fallen short. A study by Stanford University's Institute for Human Centered AI found that less than 40 percent of the 45 legal requirements associated with the AI in Government Act and the two AI Executive Orders have been implemented.⁸ Notably the Office of Management and Budget has yet to publish required guidance to support agency use and acquisition of AI.

While considering additional legislation or pursuing new administrative actions on AI, policymakers should first ensure that federal agencies fully implement existing laws and are following through with requirements from executive orders relating to responsible AI adoption.

When it comes to regulating AI systems, lawmakers should pursue a use case and sector-specific, risk-based approach rooted in high quality testing and evaluation. This approach tailors the rigor of evaluation to the level of risk posed by each use case. For example, AI-powered medical diagnostics, which carry higher risks to human health and safety, should undergo more stringent testing and evaluation by the FDA compared to lower-risk AI applications like a movie recommendation algorithm.

¹ <https://www.wsj.com/articles/trump-wants-to-double-spending-on-ai-quantum-computing-11581378069>

² <https://www.wsj.com/articles/white-house-announces-1-billion-plan-to-create-ai-quantum-institutes-11598432400>

³ <https://trumpwhitehouse.archives.gov/wp-content/uploads/2020/11/M-21-06.pdf>

⁴ <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

⁵ <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

⁶ <https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>

⁷ <https://www.nist.gov/itl/ai-risk-management-framework>

⁸ https://dho.stanford.edu/wp-content/uploads/AI_Implementation.pdf

We are not starting from scratch on this approach. Test and evaluation is a standard part of bringing products to market in nearly every industry. With regards to AI risk and regulation, current OMB guidance to agencies on regulating AI takes a use case specific approach and is very clear. “A risk-based approach should be used to determine which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits.”⁹ I believe this guidance remains fair and appropriate.

The Biden Administration has recognized the value of red teaming and test and evaluation, both in their voluntary commitments and through their support for the DEF CON31 AI Village Red Team event in August of this year. The event brought together over two thousand participants to red team eight leading LLMs on a test and evaluation platform built by Scale. This event demonstrated the critical role that test and evaluation plays in both model development and ensuring AI is safe to deploy.

In short, there has been a tremendous amount of work on AI oversight to date across the Federal Government. Any new legislative and regulatory actions should build upon and supplement this existing robust body of work and should focus on ensuring a use case and sector-specific, risk-based approach to regulation.

* * *

Second, to ensure the development and deployment of safe, secure, and responsible AI, we must harness the full strength of America’s innovation ecosystem. America's unique innovation ecosystem of government, industry, and academia has collectively driven countless technological advances that have shaped the modern world.

In the past, the federal government has focused resources on funding early stage, basic, pre-competitive research and development, while visionary industry leaders commercialized those discoveries for broad public benefit. However, rapid advancements in state of the art AI are being propelled today primarily by private companies. This landscape requires increased ecosystem collaboration to ensure we develop AI that is safe, secure, and responsible.

Industry is participating in a number of consensus-based forums, like the Frontier Model Forum and ML Commons, to address gaps in and conduct research and develop on AI standards, such as those for testing and evaluation, which still are a technical work in progress.

Scale recently published its own technical methodology for LLM test and evaluation, which combines automated assessment with human evaluation of AI systems.¹⁰ However, there is still

⁹ <https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf>

¹⁰ <https://scale.com/guides/test-and-evaluation-vision>
<https://static.scale.com/uploads/6019a18f03a4ae003acb1113/test-and-evaluation.pdf>

much work to be done, and collaboration among industry, the government, and academia is essential.

The federal government must continue driving this innovation. In particular, the Department of Energy's national laboratories represent unique strategic assets, both for fundamental AI research, but also very importantly for assessing potential risks. With their world-leading supercomputing capabilities, multidisciplinary expertise, and university partnerships, DOE could be a natural home for the U.S. Government's work on assessing frontier risk.

Additionally, Congress has the opportunity to take a major step forward by formally establishing and funding the NAIRR. Doing so would create a critical shared research infrastructure, providing expanded access to computational capabilities, high-quality datasets, and educational resources. Access to high-quality training data is the bedrock for developing responsible AI systems. The NAIRR could serve as a crucial engine to drive the curation and availability of such data across government agencies.

Beyond research, the government should convene stakeholders, provide direction, and assert American leadership at international standards organizations. The United States should continue to engage with our partners and allies across relevant fora like the G7, G20, and the OECD to ensure our American ideals and values continue to underpin the development of AI. We must not cede this ground to authoritarian governments who do not share our same values.

Our nation's universities must continue tackling complex research problems and lend their expertise to collaborative working groups. Robust, federally funded research will remain critical to advancing the state of the art in a responsible manner.

Only by marshaling the full strength of the American innovation ecosystem can we ensure American leadership in developing trustworthy AI.

* * *

I truly believe that the United States can maintain and expand our global leadership in artificial intelligence. American leadership in AI will be critical for maintaining the economic and national security of the United States.

Thank you again for the opportunity to be here today, and I look forward to your questions.

Michael J.K. Kratsios

Biography

Michael Kratsios is the Managing Director of Scale AI, leading corporate strategy and accelerating AI applications across industries. Michael also serves as a Distinguished Fellow at the Council on Competitiveness.

Michael most recently served as the fourth Chief Technology Officer of the United States at the White House. As the President's top technology advisor, Michael led the development and execution of the national technology policy agenda.

Under his leadership, the White House reoriented U.S. national technology policy towards an explicit primary focus on ensuring continued American leadership in emerging technologies. Michael is responsible for architecting the nation's first AI Strategy, the [American AI Initiative](#), doubling Federal AI and quantum research spending, developing first-in-the-world regulatory principles to govern AI development, and leading U.S. efforts to develop the world's first [intergovernmental AI policy guidelines](#) at the OECD.

Michael oversaw the implementation of the bipartisan National AI Initiative and National Quantum Initiative Acts, including the establishment of the [National AI Initiative Office](#) and the [National Quantum Coordination Office](#) at the White House, and spearheaded a [billion dollar program](#) to create a network of AI and quantum research institutes across the country. Michael also co-founded the [COVID-19 High Performance Computing Consortium](#), the largest public-private computing partnership ever created, to accelerate the pace of scientific discovery in the fight against the virus.

Michael previously served as Under Secretary of Defense for Research and Engineering (Acting). As the third highest ranking official at the Department of Defense, Michael managed the largest research and development budget for a single organization in the world, supervising all research, engineering, development, and prototyping at the Pentagon. He oversaw [DARPA](#), the [Missile Defense Agency](#), the [Defense Innovation Unit](#), the [Space Development Agency](#), and the [DoD Laboratory enterprise](#).

Michael was unanimously confirmed by the U.S. Senate, was named to Fortune's "40 Under 40," and is a recipient of the DoD's Distinguished Public Service Medal, the highest civilian award conferred by the Department. Michael's writings have appeared in *The Wall Street Journal*, *The Washington Post*, *WIRED*, *Bloomberg*, *Newsweek*, and *Fortune*.

Prior to joining the White House, Michael invested in and advised technology companies at Thiel Capital. A South Carolina native, Michael graduated from Princeton University and served as a Visiting Scholar at Beijing's Tsinghua University.

Chairman OBERNOLTE. Thank you, Mr. Kratsios.

Our third witness is Dr. Emily Bender. Dr. Bender is a research professor and scientist with the School of Computer Science and Engineering at the University of Washington. Her work has illuminated how natural language processing systems and large language models work, seeking to promote fairness and healthy engagement between humans and AI systems.

Dr. Bender, you are recognized for five minutes.

**TESTIMONY OF DR. EMILY M. BENDER,
PROFESSOR OF LINGUISTICS, UNIVERSITY OF WASHINGTON**

Dr. BENDER. Thank you. Good afternoon, Chairmen Obernolte and Collins and Ranking Members Foushee and Stevens. Good afternoon, distinguished Members of the Committee. Thank you for this opportunity to speak with you today about the risks associated with technologies being marketed as, in quotes, “artificial intelligence.” My name is Emily M. Bender, and I’m professor of linguistics at the University of Washington, also affiliated with computer science and the Information School.

This technological sector that we’re talking about is presently the site of enormous influxes of capital and enormous concentrations of power. In order to channel the innovation taking place toward the benefit of society, we need thoughtful regulation that shores up our rights. I’m very heartened to see your Committee engaging with these important discussions.

A key step toward clear discussions of these topics is clear terminology. I find that the phrase artificial intelligence is best understood as a marketing term and one which tends to muddy the waters. It is clearer in my opinion to talk about automation, and then we can focus on who is automating what and why.

Types of automation that are sometimes called in quotes AI include the following, among others: automated decision systems and automated decision support systems; automated classification like facial recognition systems; automated recommender systems like algorithmic feeds in social media; systems which provide an interface layer for access to human labor like Uber and Lyft; automated translation of information between formats like automatic transcription, machine translation, and image style transfer; and synthetic media machines like ChatGPT, DALL-E, and others.

I’d like to say a few more words about ChatGPT because it is important to understand how it produces the illusion of understanding and therefore the illusion of intelligence. ChatGPT is fundamentally what we call a language model. That means that its only task is to repeatedly produce a likely next word. And it’s doing that given some input word sequence or prompt and its training text. Because it is trained with enormous amounts of text, it can produce plausible seeming output on nearly any topic. But that output comes without any commitment to its contents. At present, no one is actually accountable for the synthetic media being spilled into our information ecosystem.

At the same time, as people, we can’t help but make sense of language that we encounter. We do it instinctually by imagining the point of view of the person that we’re imagining has spoken or written that thing in order to infer what it is that we think they

want us to understand by having said that. Prior to the advent of ChatGPT, if we encountered some coherent-seeming text, it was safe to assume that it came from some person who was using it to communicate something. So it's only natural that when we see the output of ChatGPT, we imagine that the computer behind it is doing the same thing when emphatically it is not. If we don't account for our own sensemaking ability when we try to evaluate these technologies, we can end up concerned with fantasy scenarios instead of real-world risks and harms.

Sometimes in conversations about the risks of so-called AI, you'll hear a strong focus on something called existential risk or the idea that the, quote, "AI" might become sentient and turn on us. It is important to know that discussions of fantastical malevolent autonomous thinking machines are a deflection and a distraction. They are a deflection because they locate the potential for harm in the technology itself rather than in the choices of human actors. They are a distraction because they point your attention toward imaginary scenarios and away from the actually occurring harms and real-world risks.

Those real harms and risks includes things like the pollution of the information ecosystem with synthetic media; the exploitation of workers and data theft underlying the creation of the systems; substantial environmental impact both in carbon footprint and water usage; the oppressive surveillance of people as citizens, as migrants, as workers, and as the formerly and currently incarcerated; the fraying of social services when automated allocation systems are being used to reduce the amount of support offered, or when the replacement of quality healthcare, education, and other social benefits with synthetic text is suggested. And finally, there's the use of automation to replace skilled jobs with deskilled, precarious gig work.

These harms are real and present, but we are not powerless. I believe that an effective regulatory remedy would include the following elements: First, requirements of transparency about the fact of automation, including the fact of synthetic media; requirements of transparency about the data systems are trained on and their resulting performance; clear accountability for system outputs; recourse for people who are adversely affected by automated systems; and national funding for humanities and social science research into the impact of technology on society.

Thank you again for your attention and for the opportunity to speak with you today about these important matters.

[The prepared statement of Dr. Bender follows:]

**Written Testimony of
Emily M. Bender**

Professor of Linguistics
University of Washington

Joint Subcommittee Hearing on:

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

before the
United States the House Committee on
Science, Space and Technology
Subcommittee on
Oversight & Investigations and
Research & Technology Subcommittee

October 18, 2023

Good morning, Chairmen Obernolte and Collins and Ranking Members Foushee and Stevens. Good morning distinguished Members of the Committee. Thank you for the opportunity to speak with you today about the risks associated with the technologies being marketed as “artificial intelligence”. My name is Emily M. Bender, and I am a Professor of Linguistics at the University of Washington. This technological sector is presently the site of enormous influxes of capital and enormous concentrations of power. In order to channel the innovation taking place towards the benefit of society, we need thoughtful regulation that shores up our rights. I am very heartened to see your committee engaging in these important discussions.

A key step towards clear discussions of these topics is clear terminology. I find that the phrase “artificial intelligence” is best understood as a marketing term, and one which only muddies the waters. It is clearer to talk about automation. That helps us ask important questions like:

- What is being automated?
- Who is automating it and why?
- Who benefits from that automation?
- How well does the automation work in its use case that we’re considering?
- Who is being harmed and what recourse do they need?
- Who has accountability for the functioning of the automated system?
- What existing regulations already apply to the activities where the automation is being used?

Types of automation sometimes called “AI” include the following (among others):

- Automated **decision systems** and **decision support systems**, such as those used to help relevant officials decide how to set bail, which loans to approve, which resumes to look at more closely, and how to allocate social benefits.
- Automated **classification**, of images, text, or datafied representations of people. Applications of this include benign use cases like helping digital cameras keep the focus on people in a photo, but also oppressive uses of facial recognition technology for surveillance purposes and systems which purport to detect the sound of gunshots.
- Automated **recommender systems**. These are systems which automate the choice of what information to present to a person such as in algorithmic social media feeds.
- Systems providing an **application programming interface layer** for access to human labor. Examples here include services like Uber and Lyft but also platforms like Amazon Mechanical Turk. In all cases, the interface layer lets external actors treat the human labor almost as software components.
- Automated **translation** of information between formats. This category includes applications such as automatic transcription (speech to text), speech synthesis (text to speech), machine translation, code generation from natural language descriptions, and image style transfer.
- **Synthetic media machines**, like ChatGPT, DALL-E, etc. These are systems which create text, images, video or audio based on input prompts.

I'd like to say a few more words about ChatGPT, because it is important to understand how it produces the illusion of understanding (and therefore the illusion of intelligence). ChatGPT is fundamentally a language model, and more precisely a model of the text it is trained on. That means that its only task is to repeatedly produce a likely next word, given some input word sequence or "prompt", according to the distributions in its training text. Because it is trained with enormous amounts of text, it can produce plausible seeming output on nearly any topic. But that output comes without any commitment to its contents. At present, no one is actually accountable for the synthetic media being spilled into our information ecosystem.

At the same time, we can't help but make sense of language that we encounter. We do that by instinctually imagining the point of view of the person speaking or writing in order to infer what they wanted us to understand by saying what they said.¹ Prior to the advent of ChatGPT, if we encountered some coherent-seeming text, it was safe to assume that it came from some person or group of people who were using it to communicate something. So it is only natural that when we see the output of ChatGPT, we imagine that the computer behind it is doing the same thing — when it emphatically is not.

In short, we relate to ChatGPT analogously to how one plays with the old Magic 8 Ball toy: That toy provides answers, most of which only make sense for yes-no questions. If you ask "What should I have for lunch?" and it replies "Signs point to yes", the conversation is incoherent. Accordingly, in playing with the toy, we quickly learn to use only yes-no questions so that we can

¹Clark, H.H. (1996). *Using Language*. Cambridge University Press.

make sense of the output. With ChatGPT, we are doing a similar exercise. For ChatGPT, the input is simply a string of words that form the prompt for it to continue, with likely next words. For us as users, we frame our input to ChatGPT so that we can make sense of the output as if it were an answer.

Another part of the illusion that ChatGPT “understands” comes from the way it can be used for translation tasks, plus the fact that its training data includes natural language text paired with related computer code.² A translation task maps one representation to another, such as English text to French text, or English text to computer code. When people do translation tasks, we centrally use our understanding of the content of what we are mapping. And when we write computer code, it often isn’t a translation task for us at all: rather, we reason about a problem and then break it down into smaller pieces. So when we see a computer translating English text to computer code, this reinforces an illusion of “understanding” and “intelligence”.

If we don’t account for our own sensemaking abilities when we try to evaluate these technologies, we can end up concerned with fantasy scenarios instead of real-world risks and harms. Sometimes in conversations about the risks of so-called “AI”, you’ll hear a strong focus on something called “existential risk” or the idea that the “AI” might become sentient and turn on us.

It is important to know that discussions of fantastical, malevolent, autonomous thinking machines are a deflection and a distraction. They are a deflection because they locate the potential for harm in the technology itself, rather than in the choices of human actors. They are a distraction because they point your attention towards imaginary scenarios and away from the actually occurring harms and real-world risks.

Those real risks and harms include things like the following:

- The pollution of the information ecosystem with synthetic media³
- The reproduction of biased and discriminatory patterns in advertizing,⁴ search results,⁵ generated media, and decision making, including in high-stakes scenarios in housing, employment and the justice system⁶
- The exploitation of workers and data theft underlying the creation of these systems⁷
- Substantial environmental impact, both in carbon footprint and water usage⁸

²OpenAI has been anything but transparent about the training data it uses, but the presence of such training data can be inferred from system output.

³Shah, C. and Bender E.M. (Under review). Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? https://bit.ly/Env_IAS

⁴Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM* 56.5:44-54.

⁵Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press

⁶<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

⁷<https://time.com/6247678/openai-chatgpt-kenya-workers/>
<https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483>

⁸<https://www.theguardian.com/technology/2023/aug/01/techscape-environment-cost-ai-artificial-intelligence>

- The oppressive surveillance of people:⁹ as citizens, as migrants, as workers, as the formerly and currently incarcerated
- The fraying of social services: when automated allocation systems being used to reduce the amount of support offered,¹⁰ or via the replacement quality health care, education and other social benefits with synthetic text and other automated systems
- And finally, the use of automation to replace skilled jobs with deskilled, precarious gigwork¹¹

These harms are real and present, but we are not powerless. I believe that an effective regulatory remedy would include the following elements:

- Requirements of transparency about the fact of automation, including of synthetic media. We should always know when we are encountering the output of a synthetic media machine and we should always know when automated pattern matching has been involved in decisions about us.
- Requirements of transparency about the data systems are trained on and their resulting performance. There are many detailed proposals for how to thoroughly document both datasets used in training these systems and the resulting models.¹²
- Clear accountability for system outputs. Computer programs are not the kind of thing that can take responsibility.¹³ When automation is involved in producing some kind of output (decision, recommendation, synthetic media) it should always be clear which person, corporation or other organization is accountable for any harm caused by that output.
- Recourse for people who are adversely affected by automated systems. Automated systems have the ability to scale harm and are also inflexible — whereas the humane and fair treatment of people demands discretion.¹⁴

⁹Ajunwa, I. (2023). *The Quantified Worker: Law and Technology in the Modern Workplace*. Cambridge University Press.

Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

¹⁰Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. MacMillan Publishers.

¹¹Gray, M. and Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

¹²Bender, E.M., Freidman, B., and McMillan-Major, A. (2021). A guide for writing data statements for natural language processing. <https://techpolicylab.uw.edu/data-statements/>.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):8692.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, p.220–229.

¹³Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W.H. Freeman.

¹⁴Alkhatib, A. (2021). To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, p.1–9.

- National funding for humanities and social science research into the impact of technology on society. The expertise required to understand how automation can adversely impact people and society lies in the fields of scholarship which engage with social structures and the human condition. To effectively understand the risks of so-called “artificial intelligence” we need vibrant and well-supported scholarly communities in these fields.

Thank you again for your attention and for the opportunity to speak with you today about these important matters.

Biography
Emily M. Bender

Dr. Emily M. Bender is a Professor of Linguistics at the University of Washington where she is also the Faculty Director of the Computational Linguistics Master of Science program and affiliate faculty in the School of Computer Science and Engineering and the Information School. She earned her AB in Linguistics at UC Berkeley (1995) and her MA (1997) and PhD (2001) in Linguistics at Stanford University. She is a leading voice in criticizing AI hype, using her expertise in linguistics to shed light on how language technologies work and how we should situate their output. In 2022, she was elected a Fellow of the American Association for the Advancement of Science, in recognition (among other things) of her work raising ethical issues in the development of natural language technology. In 2023, she was included in the inaugural TIME100 AI list of most influential people in AI.

Website: <https://faculty.washington.edu/ebender/>

Chairman OBERNOLTE. Thank you, Dr. Bender.

Our final witness this afternoon is Mr. Caleb Watney. He is co-CEO (Chief Executive Officer) of Institute for Progress (IFP), a nonpartisan research organization that promotes innovation policy and scientific advancement. His research focuses on policy methods of achieving state capacity and a posture of innovation. He's written commentary published in *The Washington Post*, *The Atlantic*, *Lawfare*, *Politico*, among others. Mr. Watney, you are recognized for five minutes.

**TESTIMONY OF MR. CALEB WATNEY,
CO-CEO, INSTITUTE FOR PROGRESS**

Mr. WATNEY. Thank you, Chairman Obernolte and Chairman Collins, Ranking Member Foushee and Ranking Member Stevens, Members of the Committee. Good afternoon, and thank you for the opportunity to testify today.

I'm the co-founder and co-CEO of the Institute for Progress, IFP. We're a nonpartisan thinktank that's focused on innovation policy. In particular, we spend a lot of time thinking about the American research and development enterprise and how we can get the most bang for our buck on a per-dollar basis. As one example of this, we recently signed a partnership with the National Science Foundation to help them think through more creative ways of doing scientific grantmaking and overall kind of monitoring the efficacy of the R&D enterprise in the United States.

What I'd like to talk about today are four key research priorities within the field of AI that I think represent a tremendous opportunity for the Federal Government to be proactive about shaping the technology. Effectively, these four areas have served as market failures where private companies have been unwilling or unable to invest the sufficient level of capital. And this provides a really important key moment for the public sector to make sure that the development of AI stays within the public interest.

First, we need a better way to understand how models are making decisions. This field of research is broadly called interpretability, and it tries to understand how these black box models are actually making decisions. It seems difficult to imagine how we would begin to integrate these models in healthcare, in finance, in transportation, in national security information loops unless we actually understand how they're making decisions and we can rely on them to be trustworthy.

There is some investment happening in both the private sector and within the public sector on this, but it's orders of magnitude too little compared to the social importance of this research topic. While it's not a perfect analogy, I think something on the scale and the level of ambition of the Human Genome Project would be a remarkable downpayment to really understand the inner workings of these black box models and make sure that we can begin to build trustworthy and reliable systems.

Two, defensive cybersecurity. The CEO of Anthropic Dario Amodei recently acknowledged in an interview that if a state actor were determined to steal the model weights, that is the inner workings of how their models work, they would be unable to stop them. This is tremendously concerning, especially given the large policy

investment that the national security community has been making to prevent the Chinese Communist Party and other nation-states from being able to achieve state-of-the-art models. What we need are a new set of technologies to tilt the playing field back to the defensive side of cybersecurity. As—Chairman Obernolte, as you mentioned earlier, confidential computing is one set of techniques that could accomplish this. By encrypting model weights throughout the training and the deployment process, it'd be substantially harder for adversaries to be able to steal frontier models and effectively turn what is currently a cybersecurity challenge into a physical security challenge.

Third, we need a better set of real-world benchmarks. I think it's remarkable the extent to which we have disagreements not only about the future path of AI, but also about its current and present capabilities. I think a large part of this is due to the immaturity and the shallowness of current benchmarks. There's lots of good work being done, but oftentimes, these work as multiple-choice tests that are being rapidly aced or which don't actually make apples-to-apples comparisons across leading models.

Furthermore, it's possible for leading AI companies to pick and choose the benchmarks that they think make their models look best, and there's been a splintering effect across the ecosystem. We also don't really have that many models that are trying to test for real-world capabilities in open-ended environments or when can—humans can be a key part of the decisionmaking loop. This is a key area for Federal research and investment, and I think a better understanding of model capabilities today could actually inform better policy discussions tomorrow.

Fourth, we need a better set of techniques around privacy-preserving machine learning. Ms. Stevens, as you mentioned, I think there's a lot of really good work that is already beginning to happen on this between federated learning and differential privacy and model auditing and assurance, there's lots of ways that we can make the very real perception of tradeoffs between AI progress and AI privacy less concerning, but this is going to be an important role for the Federal Government to actually invest real funding and coordination on.

And in conclusion, through targeted Federal funding and creative partnerships, we can target these core market failures, and we can advance research that not only pushes the boundaries of what we think AI can do for society, but also ensures that it's being developed and deployed in a manner that is ethical, secure, and beneficial for all of the United States and the world.

Thank you for the opportunity to testify, and I look forward to answering any questions.

[The prepared statement of Mr. Watney follows:]



Written Testimony of Caleb Watney
Co-CEO, Institute for Progress (IFP)

Before the
U.S. House Committee on Science, Space, and Technology
Wednesday, October 18th, 2023

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

Chairman Obernolte and Chairman Collins, Ranking Member Foushee and Ranking Member Stevens, and members of both subcommittees: Good morning, and thank you for the opportunity to testify today. I'm the co-founder and co-CEO of the Institute for Progress (IFP), a nonprofit and nonpartisan research organization focused on innovation policy. Our organization is dedicated to accelerating the pace of scientific progress and steering emerging technologies to promote human flourishing and American values.

In particular, we spend a lot of time thinking about the ways that American R&D dollars can have the greatest impact — as one example, we recently signed a research partnership with the National Science Foundation to support its work on improving scientific grantmaking and ensuring the American R&D enterprise is operating effectively.¹

I plan to focus my comments today on two key questions:

- 1) What are high-impact areas of research for advancing AI that the federal government is uniquely positioned to advance?
- 2) What are mechanisms, partnerships, or processes that federal agencies may adopt to advance research priorities for safe and trustworthy AI systems?

To put it more succinctly: what should the federal government fund in AI, and how should it be funded?

Globally, the private sector is spending hundreds of billions of dollars on AI. It would be fair to ask why the government should invest in this space at all. Doesn't it seem like the private sector has this area handled?

¹ "NSF Partners with the Institute for Progress to Test New Mechanisms for Funding Research and Innovation." *National Science Foundation*, Sep. 2023.
[new.nsf.gov/news/nsf-partners-institute-progress-test-new](https://www.nsf.gov/news/nsf-partners-institute-progress-test-new).

It's useful to distinguish between the amount of money being spent in a research area, and the types of research that are being prioritized. The federal government has always recognized that it has an essential role to play in shaping the *direction* of technological development, and that federal R&D dollars are a key leverage point. For instance, the federal government has invested in clean energy technologies for decades, culminating in the massive advancements in solar and wind energy we see today. If we had left investment decisions purely to the private sector, there is little doubt that we would be much further behind in our goals of an abundant and clean energy future. Similarly, the federal government has shaped the direction of early internet and satellite technologies through DAPRA, biomedical technology through the NIH, and genomic research through the Human Genome Project.

Within the field of AI, we have seen the fundamental performance capabilities of frontier Large Language Models (LLMs) grow rapidly without equivalent advances in model robustness, interpretability, fairness, and security. While LLMs today are no doubt impressive, public concerns about their embedded biases, potential inaccuracies, and lack of transparency are entirely justified

Why are AI capabilities so strong in some ways and so behind in others? Because there are market failures within technology research fields. Private companies naturally spend less time developing a general understanding of model decision-making than on discovering commercial applications. Similarly, AI labs are inclined to publicize benchmarks (or create new ones) that make their models look better, leading to a splintering effect across the industry. But the American public has a strong interest in making sure that models are trustworthy in their application throughout the economy and that we have standardized benchmarks for their flaws and capabilities. Ultimately, it will be up to the public sector to shape the direction of cutting-edge AI development in accordance with the public interest. The full might of the American R&D engine has been a powerful force for aligning these interests in the past, and it can be now.

AI R&D Research Priorities:

While many areas within the AI ecosystem would merit additional public funding, I would like to highlight four key priorities:

Interpretability:

Today, we have limited methods for understanding how advanced AI systems produce the outputs that they do. No matter what your concerns about current or future AI systems might be, developing a better understanding of how LLMs and other advanced AI models make decisions at a fundamental level will be enormously important. As we consider integrating AI in healthcare, financial markets, the

criminal justice system, national security information loops, transportation networks, and many other sectors, it will be difficult to have trustworthy systems unless we understand on a deep technical level how they are making decisions.

Thankfully, there are some early signs that we are beginning to make theoretical breakthroughs on this problem.² But the scope and importance of the issue demand a level of ambition that goes well beyond our existing grant programs in this area. While not a perfect analogy, an initiative on the scale of the Human Genome Project to map the inner workings of advanced systems today would be a major step toward making AI more reliable and trustworthy.

Defensive Cybersecurity:

Large Language Models and other advanced AI systems have the potential to dramatically change the balance between offense and defense in cybersecurity. This matters both for safeguarding frontier AI models from international adversaries and for the application of frontier models to offensive cyber capabilities.

In a recent interview, the CEO of Anthropic acknowledged that if state actors were determined to steal the model weights from its most advanced systems, Anthropic would be unable to stop them.³ This is highly concerning, especially given the tremendous policy investment the US national security community has made through export controls to prevent the Chinese Communist Party from achieving state-of-the-art (SOTA) AI models. To be clear, this vulnerability doesn't necessarily imply negligence on the part of any specific company. Instead, it shows the inherent difficulty of the field.

While more enforcement of cybersecurity best practices would surely help, we ultimately need a better set of technologies that favor defense. There are exciting techniques broadly described as "confidential computing," which involve both hardware and software innovations, that could enable AI model weights to be encrypted throughout all parts of the training and deployment process — effectively turning a cybersecurity challenge into a physical security challenge.⁴ But these techniques are still immature, and typically entail a dramatic efficiency tradeoff. If federal investments in this area led to breakthroughs in the viability and efficiency of these techniques, the U.S. could enforce higher cybersecurity standards across the

² Bricken, Trenton, et al. "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning." *Transformer Circuits Thread*, Oct. 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.

³ "But could we resist if it was a state actor's top priority to steal our model weights? No. They would succeed." "Dario Amodei (Anthropic CEO) - \$10 Billion Models, OpenAI, Scaling, & AGI in 2 years." *Dwarkesh Patel*, Aug. 2023. <https://www.dwarkeshpatel.com/p/dario-amodei>.

⁴ See e.g. "confidential computing," "fully homomorphic encryption," "trusted execution environment"

board and thereby decrease the likelihood of SOTA models falling into the hands of malicious actors.

On the flip side, we should consider how these models might be misused. It seems entirely possible that small groups of malicious actors could paralyze regional electricity grids or hospital systems at a much greater scale with the aid of advanced AI systems. The federal government can anticipate these risks and proactively shift the terrain back to the advantage of defenders through an ambitious R&D program focused on defensive cybersecurity innovation, including using AI to proactively monitor and patch vulnerabilities.

Benchmarking and Evaluations:

Creating prudent policy around AI is difficult in part because observers disagree not only about the *future* path the technology might take, but also about *current* capabilities. One reason for this uncertainty is that benchmarking and evaluating the performance of SOTA models is quite difficult.⁵ Benchmarks today are often simple multiple-choice tests that are rapidly averted or struggle to make apples-to-apples comparisons across leading models. Additionally, AI companies can cherry pick the standards or the implementation that make their models look best, leaving consumers and policymakers with huge question marks. More fundamentally, we have a dearth of benchmarks that test for AI capabilities in the real world — in open-ended environments, in detailed sector-specific applications, or when humans can be part of the action loop.⁶ This is a clear example of an area that would benefit from additional federal investment and coordination.

Especially for benchmarks that measure bias, inaccuracy, or other sociotechnical evaluations, we would ideally be able to move toward a system like NIST's Face Recognition Vendor Test (FRVT) where a public leaderboard incentivizes companies to perform better on widely publicized and agreed-upon benchmarks.⁷ A more granular understanding of the real-world capabilities and weaknesses of today's models would help build a shared set of facts that can inform tomorrow's policy discussions.

Privacy-Preserving Machine Learning:

Privacy concerns in AI are mounting as machine learning algorithms often require access to massive datasets that could contain sensitive information. These concerns

⁵ See e.g. "Challenges in evaluating AI systems." *Anthropic*, Oct. 2023. www.anthropic.com/index/evaluating-ai-systems.

⁶ There are some intriguing examples that hint at what might be possible for open-ended task evaluation, see e.g., [WebArena](#) and [WebShop](#).

⁷ "Face Recognition Vendor Test (FRVT)." *NIST*, www.nist.gov/programs-projects/face-recognition-vendor-test-frvt.

are particularly acute in healthcare, finance, and areas involving personal identifiers. Government investment could be instrumental in advancing technologies for privacy-preserving machine learning, which would enable the development of accurate models without direct access to sensitive data.⁸

- *Differential Privacy*: This well-studied approach introduces mathematical guarantees that ensure individual data points cannot be reverse-engineered from the model's output. Additional investment could explore applications for synthetic data, accelerate the implementation of differential privacy in large-scale systems, and make it a viable option for any federally funded projects that handle sensitive data.
- *Homomorphic Encryption*: While still computationally expensive, fully homomorphic encryption allows for computations to be performed on encrypted data, offering the promise of privacy-preserving analytics and machine learning. Federal R&D funding could accelerate the development of efficient algorithms and hardware optimized for homomorphic operations.
- *Federated Learning*: This approach trains models across multiple decentralized devices holding local data samples without exchanging them. The government can invest in R&D to make federated learning more efficient, secure, and applicable to a wider array of data types and machine learning models.
- *Model Assurance and Forensics*: These techniques could enable developers of a model (or outside auditors) to prove with certainty that particular characteristics of a model are true; for instance, that a model does not contain an individual's sensitive information in its training data set, or that it was developed in accordance with certain safeguards.

Federal funding in these areas would not only advance the state of the art, but also provide public-sector organizations with the tools they need to securely leverage machine learning. By investing in privacy-preserving machine learning technologies, the government can lay the groundwork for a better equilibrium between AI progress and privacy.

AI R&D Partnerships and Mechanisms

⁸ It's also possible that investing in such techniques could increase the ability for democracies to compete internationally in AI development without feeling the need to compromise on the privacy of their citizens. See e.g. Hwang, Tim. "Shaping the Terrain of AI Competition." *Center for Security and Emerging Technology*, June 2020. cset.georgetown.edu/publication/shaping-the-terrain-of-ai-competition/.

Given the dual-use nature of AI and the massive commercial market that already exists, government funders will have to use a broader set of tools and partnerships to shape the frontiers of this field.⁹ Below are some suggestions and examples:

- **Public compute for academics:** As the House Science Committee has already been considering, proposals like the National AI Research Resource (NAIRR) would enable academics to access computational resources that are currently accessible only to large industry actors. Many efforts to utilize the large academic network of AI experts to shape AI R&D will have limited effect until they have access to infrastructure like the NAIRR.
- **Infrastructure for model sharing:** Federal agencies should invest in research to establish frameworks and infrastructure for AI model sharing. This initiative could tackle unresolved questions surrounding the types of models shared, the stakeholders involved, and the cost distribution. For instance, it could specify whether base models or deployed models should be shared depending on the risk evaluation needs. Part of this should also involve research on the protocols involved in red-teaming models for sensitive national security risks.¹⁰ This broader investment area will be critical for auditing purposes, policy decision-making, and risk assessment.
- **Co-fund public goods:** There may be opportunities for government funders to co-invest with philanthropies and/or industry labs on the provision of important public goods. For instance, earlier this year, OpenAI launched a small defensive cybersecurity grant program aimed at a number of important research questions.¹¹ However, the size of the grant pool was only one million dollars — orders of magnitude too small to effectively shape research. NSF could consider bringing together a coalition of funders to dramatically scale this grant funding opportunity while taking advantage of the technical expertise at labs like OpenAI.
- **Technical expertise in grant reviews:** Additional efforts could be made to bring in technical leaders from industry to serve on review panels for relevant

⁹ For some additional suggestions on this point, see Hwang, Tim and Watney, Caleb. "How DARPA Can Proactively Shape Emerging Technologies." *Institute for Progress*, June, 2023. progress.institute/how-darpa-can-solve-market-failures-in-emerging-technologies/.

¹⁰ For instance, how should the new AI National Security Center at the NSA engage with frontier labs? Clark, Joseph. "AI Security Center to Open at National Security Agency." *U.S. Department of Defense*, Sept. 2023. www.defense.gov/News/News-Stories/Article/Article/3541838/ai-security-center-to-open-at-national-security-agency.

¹¹ Rotsted, Bob, et al. "OpenAI Cybersecurity Grant Program." *OpenAI*, June, 2023. <https://openai.com/blog/openai-cybersecurity-grant-program>

grant funding rounds. Adding this expertise into the reviewer pool could surface important considerations for the viability of technical proposals that could otherwise be missed.

- **Field building exercises:** Federal agencies should continue developing ecosystems of technical talent working on understaffed research areas in AI. The recent DEF CON event was a promising example of bootstrapping a community of individuals who are developing expertise at red-teaming advanced AI models.¹²
- **Use Other Transaction Authority:** As a legal mechanism, Other Transaction Authority (OTA) provides a flexible template for innovative agency partnerships. Both the NSF and NIST recently received OTA authorization in the CHIPS and Science Act, a promising development. However, agencies often feel hesitant to use newfound OTA flexibility creatively without explicit prodding from Congress. This committee could impress upon the federal agencies the need to meet this challenge using all the tools at their disposal.

Conclusion

There are lots of difficult tradeoffs in AI governance. What's the proper balance between regulation and international competition? Between privacy and transparency? Between interpretability and performance? Between security and compliance costs? To be sure, innovation will not allow us to avoid these tough questions entirely. But breakthroughs in interpretability, cybersecurity, benchmarking, and privacy-preserving machine learning can make these tradeoffs much less severe.

At a fundamental level, *governance is downstream of technological feasibility*. In other words, the standards that we can set are directly influenced by what we can technologically achieve. Our investments in research today will set the stage for policy decisions tomorrow.

Through targeted federal funding and creative partnerships, we can advance research that not only pushes the boundaries of what AI can do but also ensures that it is developed and deployed in a manner that is ethical, secure, and beneficial for all of society.

Thank you for the opportunity to present — I look forward to answering any questions you might have about my testimony.

¹² Groll, Elias. "Fifty minutes to hack ChatGPT: Inside the DEF CON competition to break AI." *CyberScoop*, Aug. 2023, cyberscoop.com/def-con-ai-hacking-red-team.

Caleb Watney is the co-founder and co-CEO of the Institute for Progress (IFP), a non-partisan research organization based in Washington D.C. that focuses on innovation policy. Caleb manages the science and immigration policy teams at IFP. His research focuses on policy levers the U.S. could use to rebuild state capacity and increase long-term rates of innovation.

Previously, Caleb worked as the director of innovation policy at the Progressive Policy Institute, a technology policy fellow at the R Street Institute, and a graduate research fellow at the Mercatus Center. His commentary has been published in The Washington Post, The Atlantic, Politico, Lawfare, and American Affairs. He has also been cited in the New York Times, The Economist, Vox, Ars Technica, and the National Journal. He received his master's in economics from George Mason University and a bachelor of business administration from Sterling College.

Chairman OBERNOLTE. Thank you, Mr. Watney. Thank you very much to all of our panelists. It was an excellent introduction to this very complicated but important topic.

I now recognize myself for five minutes of questions.

Mr. Kratsios, I'd like to start with you. And I want to talk about Federal preemption because this is something that we're going to have to decide very soon whether or not the regulation of AI is something that is going to be federally preempted or whether or not we're going to allow the States to be the laboratories of democracy and to innovate on this. Even if the Federal Government sets a floor for regulation, do we allow individual States to set a ceiling?

So on the one hand, that would certainly be a more anti-Federalist way to go about it. On the other hand, I fear that it might create a very complex legal and regulatory landscape that would make it difficult for smaller firms who don't have a building full of lawyers to rely on to be able to comply with. And I think it might be very destructive to entrepreneurialism. So from your industry standpoint, what do you think? Should—we have three things we can do. We can fully federally preempt, we can partially preempt, and we can allow the States to set their own ceilings or we cannot act and allow the States to take the lead on this. What do you think of those three options?

Mr. KRATSIOS. My general view on this issue is Federal preemption areas of emerging technology tend to be—like artificial intelligence can be extraordinarily beneficial to ensuring that the sort of larger marketplace of entrepreneurs is able to tackle the very important issues of building companies around this new tech. We've seen when patchworks do exist, ultimately, you're not able to see the quick proliferation of technologies kind of across the country. I think the most—the best example of this or a really good one is the struggle that Europe has had for many years in their long attempt to create a single digital market for a lot of their technology companies. And the failure to have one for decades kind of was a big hindrance to their ability to innovate.

So my general sense is that I think we should begin by thinking about what type of Federal preemption can ultimately be done. The other thing that I would worry a little bit about and why I think there's some urgency to this is there continues to be, as many of you I'm sure know, this almost sort of regulatory pipeline between Brussels and Sacramento. And California tends to be a State that likes to very quickly address some of these technology-related regulatory issues, and they see Brussels as kind of the place where they can sort of learn from.

So my general sense is we probably should be wary of individual States like California running ahead, and I think there's a big opportunity for Committees like this one to make it a much more clear regulatory environment for so many of our entrepreneurs.

Chairman OBERNOLTE. I think, at the very least, we in Congress have to recognize the Federal responsibility to regulate interstate commerce and the fact that that is a Federal obligation and not something we can allow the States to preempt, so I think that's a strong argument for the Federal Government to take a leading role in this. But, you know, the point is this is a decision that has to be made soon because the mistake that was made with digital data

privacy is Federal inaction, which allowed—now we have 23 different State frameworks, probably more to come in the months and years ahead, that create this patchwork of regulation that's very difficult for small companies to deal with. And now there's a creative ownership amongst the States that makes it more difficult for us to preempt at the Federal level. So I think if we're going to do it, it has to be done now.

Ms. Tabassi, I would like to ask you a question about the path that we take in crafting a Federal regulatory framework for AI. And NIST, I think, has done some of the furthest-reaching work on looking at risk management and looking at what a framework might look like. In my thinking, we really have two paths. We can follow the European model where we spin up a brand new bureaucracy and empower them with the ability to issue licenses and preempt that authority from—away from the existing sectoral regulators. Or we could empower our existing sectoral regulators to regulate within their sectoral spaces and give them the tools and the resources that they need, which would include support from organizations like NIST in developing testing and evaluation standards and the frameworks that you've already been working on to be able to do those jobs. Which of those two approaches do you think would be best? I mean, is it like if you consider the fact that the FDA (Food and Drug Administration) has already processed over 500 applications for the use of AI in medical devices? You know, the question is, is it easier to teach a new organization everything the FDA knows about ensuring patient safety, or is it easier to teach the FDA what it doesn't already know with the help of organizations like NIST about artificial intelligence?

Ms. TABASSI. Thank you for that question. As we all know, NIST is a non-regulatory agency, so we don't take—

Chairman OBERNOLTE. That's why I asked you.

Ms. TABASSI [continuing]. Regulation agnostic.

Chairman OBERNOLTE. You don't have a dog in the fight.

Ms. TABASSI. Thank you. Thank you. The approach we took for—in development of the AI RMF is actually a hybrid approach that the AI RMF provide the horizontal, the interoperable lexicon to talk about risk. So because we don't want to make—we want to make sure that the risk of bias or interpretability has the same meaning, and hopefully, similar measurement techniques and methods across all of the different sectors and think about development of the profiles, which are verticals, that give more specificity for that domain.

One other thing that I also want to mention, a lot of really good and important noting that importance of availability of metrics, methodologies, test environment, and measurement science for evaluations of AI systems, limits, capabilities, functionality, and trustworthiness. So regardless of the regulatory or policy landscape, we do need to have measurement science methods and methodologies to ensure the AI assurance because, again, if we can't measure it, we cannot improve it.

Chairman OBERNOLTE. Well, we need to clearly keep talking about this issue. I see I'm out of time. I will now recognize Ranking Member Foushee for five minutes for her questions.

Mrs. FOUSHEE. Thank you, Mr. Chairman.

Before I begin, I ask unanimous consent to enter into the record a statement by Representative Anna Eshoo.

Chairman OBERNOLTE. Without objection.

Mrs. FOUSHEE. Thank you.

Dr. Bender, I'll direct this initial question to you. As Members of Congress, most of us are not scientists or technologists. We can be confronted with overwhelming claims about the capabilities and risks of artificial intelligence. Some of those risks are impacting people today, and others are theoretical. As lawmakers, how do we separate the hyperbole from the reality as we debate the guardrails that may be necessary to support meaningful governance of AI systems and applications?

Dr. BENDER. Thank you for that question. And I recognize that you're in a difficult position trying to navigate this. I think that one way to go is to ask people to define their terms. So when someone's telling you about artificial intelligence, ask them what they mean about that. If they're telling you AI can do XYZ, ask them how do you know? What is the evaluation that led you to that claim? And pin it down to specifics. That would be my advice.

Mrs. FOUSHEE. Ms. Tabassi, would you like to add any further thoughts on this question?

Ms. TABASSI. I second everything Emily said. I think grounding the discussions into evidence and data is one way to go. And of course, we need to have more measurement science and methodologies to be able to get to those evidence.

Mrs. FOUSHEE. My next question is for the entire panel, starting with Ms. Tabassi. I have very serious concerns about bias and equity issues arising from the manner in which AI systems are developed and deployed. One of the most important ways to address AI bias and equity issues is to engage a broader community of stakeholders and to get them involved with AI development throughout the process. What are some of the best practices to bring diverse communities, including historically marginalized communities, into the decisionmaking process for the design and deployment of AI systems?

Ms. TABASSI. Thank you for that question, and that's a really important one. And what we learn in development of the AI RMF is the importance of reaching out and having many diverse views around the table. There are discussions around participatory engagements and how to bring the voices of the impacted community but also the voices of the group that study the impact of the AI systems or technology on the community and other people. Some of these frameworks talk about the impacted community ought to be consulted, but beyond that, they need—also need to be—have an opportunity to contribute and collaborate. But ultimately, we need to have them as part of the co-design and involved in the earliest stages of the conversations.

Mr. KRATSIOS. Yes, and for—on—from the perspective of Scale, when we worked on our DEF CON platform for testing and evaluation, one of the things that we were excited to do with our partners was to actually deploy the testing platform to a number of HBCUs (historically Black colleges and universities) and other communities around the United States. So it wasn't just the hackers in Las Vegas, but we also took it to universities where you had a diverse

set of individuals who maybe have not interacted with these models to try to test them, evaluate them, red team them, probe them in the areas that, you know, they know better than anyone else. And I think, you know, using platforms like this to be able to do these testing and evaluation experiments and opportunities across a diverse set of folks can make a really big difference.

Dr. BENDER. I'd like to point to the school of thought called value sensitive design developed by Friedman and Hendry at the University of Washington and their colleagues, which has a whole host of methodologies for, first of all, doing things like stakeholder analysis, figuring out who was going to be impacted by this. It's not necessarily the people who are using the technology. It's often the people who are having the technology used on them or on their communities. And then there's things including something called the diverse voices methodology for empaneling panels of what they call experiential experts, people who have the lived experience that gives them the expertise to understand how they're going to be affected and bring that into the discussion. So I would point to value sensitive design.

Mr. WATNEY. I think there's a lot of interesting opportunities to kind of involve participatory democracy in the design and implementation of AI systems. There's a lot of ways in which when we are actually beginning to roll out and design systems, you have the opportunity to, in some sense, be more deliberate and intentional about the kinds of decisions and values you're building in, whereas before, a lot of the inherent systems we're using, they have values being built in, but we don't get the chance to examine them as thoroughly.

Just—I haven't had time to dig fully into the paper, but I saw just a couple of days ago one method for kind of AI governance is constitutional AI. And I saw that they did an interesting project where they tried to work with a bunch of different focus groups from all over the United States that were very sort of diverse and try to help them articulate what are the values that are important to them and see what are the kinds of models you get when you build and the results from these conversations into the kind of constitution of different AI systems.

Mrs. FOUSHEE. That's my time. Mr. Chairman, I yield back.

Chairman OBERNOLTE. The gentlewoman yields back.

I will now recognize Chairman Collins for five minutes for his questions.

Mr. COLLINS. Thank you, Mr. Chairman.

We've heard China talk a lot about how they want to be the world leader in everything from social, economic, military, even in the space realm. And I think that's where I want to kind of start out with Mr. Watney and Mr. Kratsios. Although China does lag behind the United States on AI capabilities, you know, the Chinese Communist Party is spending billions on industrial policy to accelerate innovation and workforce development. Can you both elaborate on how Congress can leverage the government's unique AI assets to establish effective public-private partnerships to keep us ahead of the Chinese?

Mr. WATNEY. Thank you, Representative. Great question. I think I'll start by returning to sort of the cybersecurity point that I men-

tioned in my testimony. But the—between the Department of Commerce and sort of the larger national security community, we’ve been engaged in sort of a series of pretty robust and sophisticated export controls on top AI chips to try to prevent China from achieving kind of state-of-the-art models and make sure that the United States maintains its technological leadership. But I’m quite concerned if they’re not complemented by very strong kind of cybersecurity standards, in some sense, on the back end, you can still end up losing your model development just as easily as you can on the front end. So I think that’s going to be point No. 1.

Point No. 2 is that we should definitely leverage the fact that the United States has this huge, dynamic, open ecosystem for innovation. Ultimately, the United States is willing to work with democracies all around the world. We have always been the home for global talent from all around the world, and that’s enabled us to kind of be a pivotal leader in innovation. China does have sort of a more closed innovation ecosystem, and I think we should make sure to leverage that. And one thing I know has been a part of discussions before in, say, the *CHIPS and Science Act* is greater opportunities to allow stem Ph.D. students from around the world who are studying at U.S. universities to be able to stay here and make sure that we’re sort of strengthening our ecosystem that way.

Mr. KRATSIOS. I think there’s two ways to approach it. You have to take both a promote-and-protect set of policies to kind of should ensure, in my opinion, this American leadership in artificial intelligence. On the protect side, I think Caleb made some really good points. And I think—sorry, I applaud the Commerce Department for the efforts they made in sort of reupping and improving the export controls that came in earlier this week.

But I think the—on the promote side, something that’s very important to stress is that the United States has a very unique free market approach to innovation. And it’s markedly different than the way that the CCP attempts to challenge these issues. And I think the strength of the U.S. system is we allow this free market approach to have certain ideas ultimately succeed and other ones that are ideas that you never thought could ever succeed or would be the answer ultimately are. So in some sense, I think we have an advantage in that sense, and our free market approach is we will win ultimately because of it, not in spite of it.

I think just one small anecdote about China, which I think is fascinating on the large language model front is one inhibitor they have to their large language model development is their strict sort of approach to censorship. They essentially have mandated to all of their language model developers, as one analyst put it, to make sure that their models can’t even count to 10 because 8 and 9 are in sequence, and you’re not allowed to talk about 1989. So they have a number of challenges that they themselves have to deal with in attempting to get these models to conform with some of their absurd authoritarian approaches.

Mr. COLLINS. OK. Thank you.

Dr. Bender, in your testimony, you speak about AI identifying and classifying people through text and images and also the risk of surveillance and oppression. With the ability of foreign-owned companies like TikTok to leverage AI against the American people,

what safeguards need to be put in place and—to protect us from our adversaries?

Dr. BENDER. Thank you. I think we have an opportunity here to set standards for protecting personal individual freedom around, you know, what does it mean to have control over one's own data? What does it mean to have control over how we are represented online? And if we set those regulations for, you know, software that's used by U.S. citizens or software that operates in the United States, then we can take a leadership role there and do so in a way that reflects American values.

Mr. COLLINS. All right. Thank you. That's all I have, Mr. Chairman. I yield back.

Chairman OBERNOLTE. The gentleman yields back.

We will hear next from Ranking Member Stevens for five minutes for your questions.

Ms. STEVENS. Thank you, Mr. Chair. And I would like to submit a letter for the record from the Center for AI and Digital Policy.

Chairman OBERNOLTE. Without objection, so ordered.

Ms. STEVENS. Great. Thank you.

And how delighted to hear so many of our witnesses. Mr. Kratsios, Mr. Watney endorse funding the NAIRR, you know, the National AI Research Resource. I think that would be very important, and many of us have called on the Administration to do so.

And, Ms. Tabassi, could you provide an update on the critical work that AIRC is doing and how that is being perceived by industry at this time?

Ms. TABASSI. Sorry, who is doing?

Ms. STEVENS. That you're doing through the Responsible AI Resource Center through NIST.

Ms. TABASSI. Oh, thank you.

Ms. STEVENS. Yes.

Ms. TABASSI. Thank you so very much. Yes. We launched the AI Resource Center end of March, tried to be a one-stop shop for foundational content, technical documents, everything that's needed for discussions of AI and AI risk. So it includes AI RMF. It includes the playbook and an interactive forum to make it easy for everybody to use. It includes a glossary because, as Dr. Bender mentioned, the importance of having common vocabulary and making sure that you're in a shared understanding. There is a metrics hub to bring people—to give people everything that's happening in the world of the measurement. We are going to add a standard tracker. All of these things are trying to help the community with operationalization of the AI RMF, particularly small-, medium-sized businesses that need more help and support in operationalization of the AI RMF. We also like that to be a platform for engagement, so beyond that, just information-sharing, make it a platform for conversations around AI risk management and AI evaluations.

Ms. STEVENS. And, Dr. Bender, how else should the NIST framework serve as a baseline for monitoring and mitigating AI risks?

Dr. BENDER. I think it's a wonderful set of guidance. The question is how do we make sure that people use it? And I see a couple of pathways there. One is through Federal procurement, so any company who wants to do business with the Federal Government

would have to follow these wonderful procedures. And then another is actual, you know, regulatory enforcement that companies would have to do to stay on the right side of the law that would empower the people working on the inside trying to figure out how to do the right thing to actually be heard.

Ms. STEVENS. Yes. And, Mr. Watney, I was obviously excited to see you mention in your testimony privacy, preserving machine learning. And like you said, AI models require massive amounts of data to learn and function. And we must ensure better protections are in place for the data sets and tools in AI training. And look, this is why we're excited about the PET bill, the *Privacy Enhancing Technology Research Act*. So what are the consequences of not protecting data privacy when training AI systems?

Mr. WATNEY. Yes, I think there's a few levels to this. One is, obviously, on just the surface level, consumers might have less trust in systems if they don't know that their information is being protected. I think to sort of create broader buy-in across the whole ecosystem, it'd be helpful to use some of these privacy-preserving machine learning techniques. I also think at an international level I'm sort of concerned about a broader trend where democracies might feel the need to compromise on their internal privacy protections in order to compete with more authoritarian regimes. China is sometimes perceived to have certain advantages in AI because they have laxer privacy protections. But if we could again shift the playing field in favor of privacy-preserving machine learning models that could enable democracies to maintain the lead in AI without sort of feeling——

Ms. STEVENS. Yes, we can have open source, and we can also have privacy.

Mr. WATNEY. Right.

Ms. STEVENS. That's the goal. And I just want to get in here real quick a question about workforce because the career pathways are just underdeveloped at this point. And so kind of for anyone who wants to jump in, I just want to talk about and see if you have comments about approaches, the RMF approaches, AI skills and job activities for risk management. We've had a lot of conversations on this Committee about cyber. We need an AI workforce. What's going to go into that? How do we start thinking about it? And we have a former CTO here.

Mr. KRATSIOS. I think there's a couple ways to think about it. I think the first thing whenever I sort of analyze these types of questions is what are the pools of current appropriated dollars that are used toward education and workforce and then try to map those to the areas where you believe are most relevant to the particular educational sort of upskilling that we have to do on the front of AI.

I think an area that's very ripe that sits primarily within some of the work that NSF is doing on K through 12, development of curricula. So you can't teach any of this stuff. The curricula doesn't exist. And this is something that has actually been happening as part of the National Quantum Initiative, which I think may be interesting to think about from an AI standpoint is there aren't sort of K through 12 courses for quantum. So the question is like how can you prepare students over time and be in a position to leverage it? And I think the same sort of thinking can kind of go into the

AI sense where, if you can develop the right curricula, then you can actually have, you know, in power, sort of educators with tools to move forward.

Ms. STEVENS. That's great. Thank you, Mr. Chair. We'll yield back.

Chairman OBERNOLTE. The gentlewoman yields back.

We'll hear next from the Chairman of the Full Committee, Mr. Lucas, for five minutes.

Chairman LUCAS. Thank you, Mr. Chairman.

Ms. Tabassi, in your written testimony, you highlight how NIST has begun developing industry-specific profiles for operationalizing the Risk Management Framework? What is your assessment of the current gaps in the technological infrastructure, tools, standards, that could slow down Federal agency adoption of the Risk Management Framework?

Ms. TABASSI. Thank you very much for the question. I think the biggest technical challenge and technical gap is availability of metrics, methodologies, test environment, and standards for evaluations of AI systems. And I think it was also mentioned that we need to evaluate AI systems beyond their functionality and performance, beyond their technical robustness and measure the societal robustness. We need to be able to measure AI risk and impact. And those measurements ought to be done where the AI system in its native context in interacting and being used, operated by the humans, impacting in—interaction with humans that are being impacted or influenced by the systems. And those measurement sciences, that test environment doesn't exist today.

Chairman LUCAS. Following up on that, Mr. Kratsios, in your experience building AI infrastructure at Scale and developing AI policy at OSTP (Office of Science and Technology Policy) and DOD, what structural or resource challenges do you anticipate agencies receiving when adopting this Risk Management Framework into their development, testing, and procurement processes?

Mr. KRATSIOS. You know, as I've seen agencies attempt over time to sort of continue with a regulatory approach, but in the light of new technology, the challenge is always slowly getting them up to speed on what it actually means to do the testing and evaluation necessary to approve the use of that technology. I think something that has been looked at, for example, by the Department of Transportation for many years is how you can appropriately be able to verify or validate that a vehicle is safe for operation. And this moves sort of in a world where there weren't a ton of these vehicles and now there are. There has to be the right sort of testing and evaluation methodology available for them to be able to do that sign-off. And I think the same goes with artificial intelligence. And as the Chairman mentioned, you know, you've already seen, you know, quick progress in FDA in being able to adapt some of this stuff, but I think at the end of the day, the real challenge will be can we have testing and evaluation standards that can allow these regulators to incorporate into their efforts?

Chairman LUCAS. Continuing with you, Mr. Kratsios, in your testimony, you point out that the implementation of existing AI legislation and Executive orders have fallen short. To date, less than 40 percent of the 45 legal requirements associated with the *AI in Gov-*

ernment Act and the 2019 and 2020 AI Executive orders have been implemented, and the Office of Management and Budget is yet to publish required guidance to support agency use and acquisition of AI. What are the some of the potential harms in not fully implementing the existing laws and Executive orders as they relate to responsible AI now?

Mr. KRATSIOS. Yes, I think I am heartened to hear that a potential Executive order may be coming out this month, which will address some of the issues that are still outstanding from those EOs and legislation. I think the biggest challenge is that a lot of the requirements from the Executive orders and from the legislation are important foundational pieces that future regulatory structures are built on top of. So if you don't have a very clear understanding of sort of what the potential use cases are in an agency or what regulatory—if you don't do the hard work to actually assess where AI is going to impact the regulation—the regulatory regime that you're sort of overseeing, it's very hard for you to actually then proceed with actually doing the regulation itself. So I think we need to get—make sure that we have a strong foundation in place, and in doing so, then we can build on the important stuff that we've been talking about today.

Chairman LUCAS. Finally, Ms. Tabassi, NIST functions as the Federal AI standards coordinator and works across the government with stakeholders to identify critical standards, development activity, strategies, and gaps. How is NIST helping the Federal Government think about our role in developing technical standards for AI?

Ms. TABASSI. We are working with all of the government agencies across in understanding their priorities, their needs for standards, but also working on scientific underpinning needed for development of the clear, implementable standards. That goes from the common definitions to establishing trustworthiness characteristics to advancing tests and measurements for AI standards—

Chairman LUCAS. Thank you.

Ms. TABASSI [continuing]. For AI systems.

Chairman LUCAS. I yield back, Mr. Chairman.

Chairman OBERNOLTE. , gentleman yields back.

We'll hear now from the Ranking Member of the Full Committee, Ms. Lofgren, for five minutes for her questions.

Ms. LOFGREN. Thank you, Mr. Chairman.

First, I'd like to ask unanimous consent to put my statement into the record.

Chairman OBERNOLTE. Without exception—objection, so ordered.

[The prepared statement of Ms. Lofgren follows:]

Thank you, Chairman Obernolte and Ranking Member Foushee of our Investigations and Oversight Subcommittee, and Chairman Collins and Ranking Member Stevens of our Research and Technology Subcommittee, for holding today's hearing. I would also like to welcome our distinguished panel of witnesses.

This hearing is about what questions we need to answer and tools we need to build to get meaningful governance of AI systems. Our Federal science agencies have taken significant first steps to address this challenge, most notably, NIST's work on the AI Risk Management Framework. In the meantime, the technology itself continues to rapidly advance in both capabilities and applications. I believe regulation of AI will be necessary, but I am also keenly aware that we must strike a balance that allows for innovation and ensures the U.S. maintains leadership.

One critical question to answer in the near term is whether we can use a single AI governance framework or if governance should be broken out by sector or use

case. There is currently some debate about how to control the technology itself, including through licensing, but ultimately our goal should be to stop or at least reduce any negative outcomes of its use. The NIST Risk Management Framework recognizes that context is important for mitigating the risks of AI deployment, and I suspect regulation should follow that example.

I am also keenly interested in the intersection of AI and intellectual property. To ensure intellectual property protections and promote innovation, we must develop tools, testing, and standards to track the use of copyrighted content by AI systems. That technical foundation will necessarily underpin any policy solution agreed to by both the content creation community and AI platforms.

Finally, I am curious about what policy levers Congress may have to incentivize good AI governance practices. As my colleague Representative Lieu has often raised, the Federal government can influence the AI ecosystem by requiring AI risk management on systems adopted by Federal agencies and contractors. However, a shift in Federal policy to require AI risk management practices will require mature standards and a workforce capable of implementing them.

While the contours of a regulatory framework are still being debated, it is clear we will need to lay the technical groundwork to govern AI. I thank the witnesses for joining us today for this very important discussion.

I yield back.

Ms. LOFGREN. And then, you know, this is a great day to be having this hearing because we are working together in a bipartisan way, honestly, to try and find solutions to challenges that face our society. And that's a pleasure to be doing. I think that Chairman Obernolte posed the options very well, which is do you want a single agency that we expect to know everything about every law, or do you want to enhance the capability of existing agencies that have expertise in an area of law to apply that in AI? And I think, personally, the latter is likely to be successful. I mean, we have a whole body of law that's been developed on discrimination and hiring or intellectual property or on and on and on. It's impossible to think how one agency could replicate all of that knowledge.

So I'm interested in how we might empower not only agencies, but individuals to protect themselves from violations of existing law. And I think some of that comes down to the metrics question and the measurement issues because, for example, it's already illegal to discriminate on the basis of ethnicity or religion in hiring, but if you don't know that's happening to you through the use of AI, how do you avail yourself of that protection?

And I've been thinking a lot about how we might impose some responsibility on those who have created AI systems. Frankly, some of the AI developers have revealed to me privately they don't fully understand how what they've developed is operating. And so if they don't, certainly we are not going to, but we need to put the responsibility on the creators for violations, it seems to me.

I—so that's a question for any of you who wants to deal with it. And I have a second question if you can. The whole idea of licensing strikes me as improbable for a technology that even the developers don't fully understand. And so I'm thinking about licensing versus registration so that we might have an opportunity to have insight into what's being developed, but furthermore, to have each technology in a regulatory scheme to protect the humans against the violation. Anybody who can address those two issues, I would be greatly interested.

Dr. BENDER. I'd love to speak quickly to that point. I agree that having this spread out across the agencies is really important. I think the FTC (Federal Trade Commission) is doing an excellent job saying there is no AI loophole. They regulate what they regu-

late. It doesn't matter if somebody's using automation to do the thing. It still falls under their jurisdiction. Likewise, you know, the—for medical components, we have methodologies for testing whether something works well enough. If there's some pattern matching going on in there, it should still pass the same tests. And by tests, I don't mean multiple choice tests, I mean, experiments, right?

In terms of licensing versus registration, those concepts to me seem to belong to this domain of everything machines, and I don't think that that's a sensible way to think about this. I think it's worth keeping in mind that something like ChatGPT sounds like it can do everything because it can give us synthetic text on any topic, but that doesn't mean it's a robolawyer or robotherapist or robo doctor and so on. And so I think keeping things in terms of existing rights and then figuring out where we need to shore up those rights is a good way to think about it.

Mr. KRATSIOS. Yes, I agree with Dr. Bender. I think the concept of trying to create a licensing or registration regime is, I think, incredibly difficult, primarily from essentially the same thing you're saying. But another way maybe to say it is from a definitional standpoint. Like what are you registering? What does it need to be in order to be licensed? And even coming up with some sort of valid definition that doesn't change over time. There's a lot of discussion around is it a frontier model, is it not? You know, is it really powerful, is it not so powerful? Those types of questions are extraordinarily hard, and those definitions are really hard to kind of lay out in my opinion.

Mr. WATNEY. Yes, I just return, I think, to the importance of additional investments in both interpretability and sort of broader benchmarking. Again, it's just remarkable how much it's actually hard to have a shared conversation about what are the risks we're actually concerned about? What are the kinds of system capabilities that we think these things do or do not have? And kind of a real deep set of benchmarks around performance in the real world on open-ended tasks when humans can be a part of the decision-making loop I think would just be so helpful for actually grounding a larger discussion here.

Ms. LOFGREN. All right. Thank you, Mr. Chair.

Chairman OBERNOLTE. The gentlewoman yields back.

We'll hear now from the gentleman from Texas, Mr. Babin. For five minutes, you're recognized.

Mr. BABIN. Thank you, Mr. Chairman. I appreciate it, and appreciate all your witnesses for being here.

Ms. Tabassi, as you know, the Port of Houston is a vital economic generator in my district there in east Texas, District 36, but it's also a national security asset that rebuffs thousands of attempted cyber attacks each and every year. AI has the potential to bolster our defenses for critical infrastructure like the Port of Houston, but it can also be an asset for nefarious actors as well. How can NIST partner with agencies like the Department of Homeland Security, the Federal Aviation Administration, Pipeline and Hazardous Materials Safety Administration to develop and promote AI policies that will safeguard our critical infrastructure, including major ports like the Port of Houston?

Ms. TABASSI. Thank you very much for that question. And AI RMF lays kind of the foundations and the right guidance for developing all of these other guidelines on our—we call them profile of the AI RMF, and we're already working with, for example, our colleagues across the agencies, with Department of Homeland Security operationalizing AI RMF, with our colleagues at the DOD on advancing measurement and evaluations of the AI. So coordination is the key and is vital, and we'll be happy to work on and build on top of AI RMF to address all of these important questions.

Mr. BABIN. Thank you very much.

And, Mr. Kratsios, by many metrics, China has caught up and in many cases surpassed the United States in research and commercial capabilities. Chinese universities are publishing many more research papers than U.S. institutions. They received nearly the same share of citations as U.S. papers. How concerned are you with the pace of AI progress in China? And then after you finish, I'd like to open it up for all four of you.

Mr. KRATSIOS. I continue to remain very concerned, and I believe that that's why American leadership in artificial intelligence should remain a top national priority. In order to sort of maintain our leadership, we have to essentially do four things. We have to do research and development, meaning increasing spending in R&D and focusing it on fundamental basic precompetitive R&D that the private sector is not focused on.

The second is on workforce development. That is the area where I was talking about earlier. How can we prepare the American worker to be able to harness and leverage this technology for the benefit of the country?

The third is on regulations. As we talked about here today, we have to be in an environment where the regs that we have are leading the world or people are copying what we're doing, not the other way around, and create a system that is flexible enough to promote innovation, but also responsible enough to protect American citizens.

And the fourth is through international alliances. We must work with likeminded partners and democracies in order to ensure that this technology is built to reflect our values, not authoritarian values. And I think if we can continue to sort of leverage and push the needle on all four of those areas, we can and will remain ahead.

Mr. BABIN. Thank you very much.

And briefly, anyone else?

Dr. BENDER. Yes, just briefly speaking from academia, I see that the research community in China is vibrant and very well supported by their national funding initiatives, and I would love to see similar things going on here, especially including research funds not just to people developing what they call AI systems, but also people in the humanities and social sciences who can study how they're affecting society.

Mr. BABIN. All right. Thank you very much.

And, Mr. Kratsios, one more. A few days ago, China released new standards that training datasets and algorithmic outputs must pass to go to market. They have also built an algorithmic registry whereby algorithms above a certain capability threshold need to be

registered with the government. The United States must avoid such draconian measures while ensuring that we protect our citizens from potential harm. How can the U.S.-set transparency and risk benchmarks align with our democratic values? And what tools and standards do agencies need to ensure that those benchmarks are met?

Mr. KRATSIOS. Yes, that's a really good point, bringing that up. I think that made a lot of waves in the artificial intelligence community of how the CCP is thinking about sort of enforcing their own sort of viewpoint in the world through this regulatory regime.

On the part of the United States, the best approach that we can do is do the very hard research and development associated with developing a testing and evaluation regime for these large language models. Right now, there is no standardized T&E process, but the government has endorsed for these large language models. Scale has taken an effort to create an industry standard itself where we released our own testing and evaluation standard that combines both an automated approach to artificial intelligence and humans together to test these models. But I'm excited to hear about the incredible work that NIST is doing. But all this stuff can't happen fast enough. It's important that the United States is out there with world-leading standards as a way to push back on the approach that the CCP is taking.

Mr. BABIN. Thank you. And my time has expired, so I yield back, Mr. Chairman.

Chairman OBERNOLTE. The gentleman yields back.

We'll go next to the gentlewoman from Oregon. Ms. Bonamici, you're recognized for five minutes.

Ms. BONAMICI. Thank you to the Chairs and the Ranking Members. Thank you to the witnesses for being here today.

As artificial intelligence or, as Dr. Bender says, automation is rapidly expanding in our daily lives, I remain concerned about how to prevent incorrect bias. The Stanford Institute for Human-Centered Artificial Intelligence did a report showing that underbanked communities, for example, those who have a bank account but might also tap into alternative financial services like payday loans or check cashing services, face disproportionate errors and denials from credit reporting models that rely on AI. So denying someone access to credit because of an inadequate model, for example, could cause a consumer to lose her home or it will affect their credit and financial security.

So I request unanimous consent to enter this study into the record.

Chairman OBERNOLTE. Without objection, so ordered.

Ms. BONAMICI. Thank you, Mr. Chairman.

Dr. Bender, nice to see you again. I want to ask about how we can best correct for errors in data in these models that integrate biases. So, of course, I support training the workforce on implicit bias, more recruitment of diverse researchers and engineers, updating civil rights and consumer protection statutes, for example. But is that enough to effectively address this issue?

Dr. BENDER. Those are all good ideas. I think there's more that's needed. I think, first, we have to recognize there's no such thing as an unbiased training dataset. There's always going to be some

remaining bias, but still less biased is better, and it's worth striving for. And I think that, in addition to what you've already listed, a really key tool that we have is documentation of training data because that allows us to ask questions like: Is this trained on a set of data that would lead us to suggest—to want the patterns in that training data matched in our use case. So it's helpful at procurement time, and it's also very helpful for people who are seeking recourse. Someone mentioned earlier, I think it was Ranking Member Lofgren, that if you are experiencing discrimination but you don't know, what do you do about it? And I think documentation about training data will be helpful in that regard.

Ms. BONAMICI. Thank you so much. And I'm going to ask this. I want to follow up on a question that was started by the Chair Obernolte and Ranking Member Lofgren. We assume there's going to be some kind of regulation, so I'm going to give you three choices. I just want to hear briefly from each of you. Existing entity like the FTC, a new entity that's created hopefully with the expertise that's needed, or is it going to depend on the topic? And Department of Education, the issues relating to education ideas, and the Department of Health and Human services, the issues related to health, do you send employment issues to Department of Labor, for example? So if you just want to go down the line and say, existing entity, a new entity, or depends on the topic, that would be helpful.

Mr. WATNEY. I think it broadly depends on the entity. I think there's—to your earlier points, like there's so much already embedded law everywhere and making sure that we're enforcing that. There might be sometimes the need for new offices within existing agencies, but I think that's probably going to be the right balance.

Ms. BONAMICI. Thank you. Dr. Bender?

Dr. BENDER. Depending on the topic, but with some central coordinating body, perhaps NIST, to help share expertise across.

Ms. BONAMICI. Excellent.

Mr. KRATSIOS. Yes, I agree, depending on the body. It should be industry-specific and use-case specific.

Ms. BONAMICI. Thank you. Ms. Tabassi?

Ms. TABASSI. I think a hybrid approach, I think, yes, it's—AI is very contextual. It's important to do it for the particular use cases. It's also important to make sure that at least terminology is the same across.

Ms. BONAMICI. Excellent. Thank you. And, Ms. Tabassi, I know the Biden Administration recently published the blueprint, of course, the AI Bill of Rights and highlight in there the importance of mitigating safety and discriminatory risks. And as the Administration works to implement this framework, what practices should they adopt to mitigate risks before automation is deployed? And how will we know whether these systems are ethically implemented? Because we do hear a lot about AI ethics, so how will we verify that ethics are embedded within?

Ms. TABASSI. Right. So in terms of ethics, I think, first, again, we need to have a clear definition on what ethics is. In AI RMF we tried not to use ethics because it has dependencies to the culture, to the time, and tried to unpack it to the concepts of the trustworthiness. But it all, again, goes back to measurement and

having test environments and methodologies and metrics for tests to provide that type of information. And that's something that's missing now. And there's no quick fix. There's a lot of research that's needed to build that type of test environments.

Ms. BONAMICI. And I have a few seconds left. Dr. Bender, I'll ask you. Ethics in automation or artificial intelligence, does it mean the same thing to everybody? Are there different definitions of ethics? And how—

Dr. BENDER. Yes—

Ms. BONAMICI. Who decides what ethics means in this context?

Dr. BENDER. There's different definitions. The one that I like best comes from Dr. Margaret Mitchell, who talks about it basically in the process. Are you developing your system in a way that has an eye toward people who might be harmed by it and making sure that their interests are represented?

Ms. BONAMICI. That's very helpful. And I'm about to run out of time, so I yield back.

Chairman OBERNOLTE. The gentlewoman yields back.

We'll go next to the gentleman from Indiana, Mr. Baird. You're recognized for five minutes.

Mr. BAIRD. Thank you, Mr. Chairman and Ranking Member. I really appreciate that you are holding this Committee session. And I really appreciate, I always do, I appreciate all the witnesses sharing your expertise with this Committee as we struggle with trying to find ways to protect the public.

And so this Committee has played an instrumental role in passing the *CHIPS and Science Act*. And in my district in Indiana, we have plans to start production of the chips or the semiconductors in the next five years in a fab lab in West Lafayette. But in this process, I learned that the average car has anywhere from 1,400 to 1,500 chips, and some cars have even more than that. And then my background is agriculture. And, you know, the farm equipment and machinery that we're using to implement more and more precision agriculture is extremely important.

So since AI is playing such a major role in many of these industries, I'm curious to know what role it's playing in the chip manufacturing, or what does it have the capability to do in this industry? Ms. Tabassi, do you want to start that conversation or—

Ms. TABASSI. Obviously, chips and—specialized chips for AI are basically the engine and power. It's really important. So at NIST we—our researchers are working on some specialized AI chips, kind of next generations that can do many of those computations inside the chips. And we work with the community to advance this type of research.

Mr. KRATSIOS. Yes, as we have seen over the last six months with the rise of, I guess, the NVIDIA stock price, it's become very clear that sort of GPUs (graphics processing units) are in sort of insatiable demand globally. So I think the one area that I think is important for us to continue to consider is who has access to leading edge, essentially, APU—AI-specific GPUs. And, you know, we've taken—it sounds like the Administration has taken some thinking around whether certain countries of concern should not have access to them because they essentially are the—sort of a foundational asset to any type of AI development. And I think that

needs to continue to have serious exploration in the months ahead and continue to be vigilant on it because without these chips, you can't develop these frontier models.

Dr. BENDER. I think in that context I'd like to point out that all of the energy right now in this field is on bigger and bigger models that require these chips and lots of them and lots of energy to run them and lots of water to cool them, and there's real environmental costs there. And so if we had investigations efforts going into finding more streamlined, less data-hungry, less energy-hungry ways of doing this, that could be very valuable.

Mr. WATNEY. I think, yes, at the highest level it's just really worth keeping a careful eye on this. I think by anticipating or predicting trends in chip design, that's going to tell us a lot about future AI systems, both kind of geographically where they're being built, what their underlying capabilities are. You're seeing already a lot of specialization in the difference between maybe consumer-grade electronics, the kind of stuff that's going in your personal laptop or phone, and then the stuff that's really dedicated and focused on training the next frontier of large language models.

I also think this is a particular industry to try to maintain American leadership in. Chip design is itself almost just as important as the chip manufacturing. And thankfully, the United States and our sort of partner allies do have a strong lead in that, but I think maintaining that will be really important for shaping the future of the AI field as well.

Mr. BAIRD. Thank you. My next question—and we don't—we only have about a minute left, but, you know, there's been a lot of advances in quantum science. They've had some major breakthroughs. And so this Committee is probably going to approach reauthorization of the *National Quantum Initiative Act*. So what do the advancements in quantum computing mean for AI? Any volunteers?

Mr. WATNEY. I'll just chime in briefly. I think a lot of the concerns about quantum or the implications are sort of similar. I think on a cybersecurity perspective, I think there's a sort of defensive-to-offensive balance within cybersecurity and between, you know, major advancements in AI, but also quantum. There's a real chance that that defense-to-offensive balance just totally flips. And I think the U.S. Government has a real opportunity through an ambitious R&D agenda to kind of purposefully tilt the terrain back in favor of defenders and make sure that we're sort of like on the leading edge of that. And I think that applies just as much to quantum as it does to AI.

Mr. BAIRD. Thank you. And I see my time is up, and so I yield back.

Chairman OBERNOLTE. The gentleman yields back.

We will go next to the gentleman from Illinois, Mr. Casten. You're recognized for five minutes.

Mr. CASTEN. Thank you, Mr. Chair and to all our witnesses. I love this topic and it's fascinating and I've been thinking about it for a long time. And I'm struggling to be pithy as I frame my questions because I'm stuck on the fact that back in 1998 when I was in grad school, I was taking classes learning how to program neural networks and genetic algorithms and machine learning. And we

are talking about this as if it is a new thing with cutting-edge entrepreneurs who are just sort of plowing the field for the first time. And, you know—and back in the 1990's, this was—this wasn't like this is cutting-edge research. This is if you want to get a job out of grad school, you can go work to a hedge fund that's using high-speed trading algorithms that are based on this and you need to understand this. There's a technology consulting firm who is doing—they didn't call it big data at the time, but it's crunching and finding out that people who buy diapers also buy six packs of beer and you should reorganize your shops. I mean, we all read—these are 30-year-olds. Now I see you all nodding, but just is there anybody who—do any of you think that this is not a technology that has been commercially available for at least three decades? Now granted in simpler forms, right, but the question of ethics, the question of how do we program these things, I remember having these conversations 30 years ago.

Dr. BENDER. It's not new, you're right.

Mr. CASTEN. OK.

Dr. BENDER. And it's also—when it gets called neural nets, it's based on a 1950's notion—it's a bad model of a 1950's notion of how neurons work.

Mr. CASTEN. Fair. Thank you. As a former biochemical engineer, I appreciate that correction.

I raise that because I get a little bit nervous if we frame this as saying just let the entrepreneurs take care of this and, you know, don't get in the way of regulation if we don't first say, OK, well, what—over the last 30 years, have we—and, Mr. Watney, you mentioned about the need for this. Do we actually have a legal framework—because I'm not aware of one—that right now says if you are using an AI system, we have—and you violate laws or you violate certain ethics rules, that we are clear about who is to be held to account.

Mr. WATNEY. This is just, I think, a big murky area of law that we don't really have clear guidance on. And part of it comes down to not only who are the actors responsible, but like where does liability in some sense most naturally fall? I think this kind of returns back to the importance of understanding at a deeper level how the models are making decisions. If we had a sort of—if we could tear down the veil of the black box, if we could sort of peer behind and understand at a more mechanistic level how these models are making decisions, I think it'd be so much more obvious kind of where maybe biases are coming from, where inaccuracies are, what are the kinds of—

Mr. CASTEN. So if I could because I want to just pick on that point. We had—I'm on the Financial Services Committee. Two years ago, we had a question that one of the big credit card companies was asking us to help them because Facebook would not tell them whether their ad targeting software was compliant with the fair lending rules. And so we had a fairly robust back-and-forth with the CFPB (Consumer Financial Protection Bureau) about who would be accountable. And it's—it still is to some degree—the current director has a path, but I think legally, it's still somewhat an open question.

More than 10 years ago, the—this is not public information so I’m going to be a little bit shady. But there was a high-speed trading firm who found out that their trading algorithm was making money by shorting an anticipated Russian invasion of Crimea. And they—and arguably, that didn’t break any law. It was unethical, but it didn’t break any law. And they chose to shut that down.

And so if we—I guess my question, the reason why I interrupted you—and I apologize for interrupting you is—is this even something that you can program a fix to, or is this really something where some—whether government entity or, you know, white hat nonprofit has to have some degree of audit responsibility and an obligation to maintain auditable code?

Mr. WATNEY. Right, I think—I don’t—I mean, I’m less familiar with the specific context of this scenario, but I would guess it’s totally plausible Facebook or the high-speed trading firm didn’t even know what their model was sort of making a decision on, which sort of gets down to it’s really hard to hold different parties accountable or to know kind of where things are going wrong without an understanding of where the model is kind of making decisions.

I think, to an earlier point that was mentioned, like liability is probably an underrated lever here to kind of like carefully align interests across the ecosystem. And I think there’s been a lot of other conversations about other regulatory tools, and, you know, those may or may not be appropriate, but liability seems like underdiscussed, I think.

Mr. CASTEN. Well, if any of you—I’m going to get tight on time here—but Lori Trahan and Adam Schiff and I introduced a bill to give the FCC (Federal Communications Commission) the authority to essentially say you have to be able to—if you’re using these codes, you have to have—your algorithms are private, that’s fine, but you have to have the ability to reach in and inspect, which raises the question of is that the right entity? You know, do they have the technical horsepower to do that? And could you even have a single national entity with that jurisdiction, given as the servers are all over the world? And if any of you have thoughts or criticisms, I’d love to engage with you because I do think it’s this meaty question that, at least from where I sit, you have to have someone with auditor rights who has the public interest at heart. If any of you have comments on that before I get pulled, I’m happy to hear.

Dr. BENDER. Can I just say really quickly, if I am running some chemical processing plant and without my knowledge I’m polluting a river next to me, aren’t I still liable for that?

Mr. CASTEN. Well, so I guess the question would be in the case of the CFPB, who is liable for the *Fair Lending Act* violation, the bank who used the ad or the algorithmic company who said I didn’t actually program to do it, it just evolved that way? I think there are cases where it’s clean. There are a lot of cases where it’s less clean. Who’s using the algorithm? Who was the benefit—who was the aggrieved party? And are you liable for an algorithm that you didn’t program, right, but it evolved into doing something that you don’t trust?

Dr. BENDER. Yes. I’d be leery of biological metaphors like evolution, always.

Chairman OBERNOLTE. The gentleman's time has expired. We'll go next to the gentleman from New Jersey. Mr. Kean, you're recognized for five minutes.

Mr. KEAN. Thank you, Mr. Chairman. And I want to associate myself with Representative Stevens' support and strong advocacy for the *Privacy Enhancing Technology Research Act* that passed this Committee and I agree needs to get—go to the floor. And, as a cosponsor of that bill, it is surely important that we focus on government coordination to ensure responsible data use.

My first question is for Ms. Tabassi. I'm interested in understanding how NIST is engaging with States' implementation of their various task forces. And can you provide insights as to whether it's collaborative, when—do you have any concerns that—as you're approaching their—either their risk mitigation strategy or their public safety focus on their local State labs, those efforts?

Ms. TABASSI. I'm sorry, collaboration with—

Mr. KEAN. When you're looking at—so, for example, New Jersey is starting to look at a State artificial intelligence task force and looking at a little bit of the—whether you're focusing on the State level or the—on the Federal regulatory, what guidance are you—from a NIST perspective, what's your best recommendation, whether it's State or Federal partners are getting involved in this?

Ms. TABASSI. So in development of the AI RMF we engage a very broad community, and we heard from all sorts of stakeholders, including State, local governments, international, and globally. I don't have any information about particularly the task force that you mentioned. But again, in all of these discussions, I do think that it's important to have some sort of an interoperable lexicon, a horizontal foundation of at least having the same vocabulary and understanding of the risk and harms that give some sort of interoperability across all of the applications and all of the use cases.

Mr. KEAN. Thank you. And this, I think, is for anyone who feels themselves appropriate to answer it. In light of the recent Senate roundtable where executives of major AI companies called for a new Federal agency to regulate AI, there are concerns about policy alignment with a select few industry leaders. The involvement of companies poised to benefit financially also raises questions about regulatory fairness. So it's worth noting that the Federal Government already has agencies in place that are actually involved in AI regulation. How can you best balance the regulatory structure and whether to create new departments or not create new departments and only—and currently enhance things that departments are already doing? So Mr.—

Mr. KRATSIOS. Yes, I think my view is that there should not be a new government AI agency. We should rather focus on use case, sector-specific, risk-based guidelines that are rooted in sort of the fundamental science that this Committee cares about around creating test and evaluation techniques that then can be utilized across all the relevant agencies that have spent many times a number of decades regulating their industries.

Mr. KEAN. OK.

Dr. BENDER. I'd just like to add to that that I think, again, we want it spread out through the agencies, but those agencies might need help in building up the expertise to be able to cut through AI

hype where people being regulated, companies being regulated come in and say, well, this is different this time because it's AI. We need folks in all the different agencies to be able to say, no, it's still our jurisdiction.

Mr. KEAN. Thank you. Mr. Watney?

Mr. WATNEY. Yes, I think sometimes discussions about like new or old agencies, whether they're spread out or centralized, in some sense, the more important thing is capabilities. Like do we have fundamentally the state capacity somewhere in the government to be able to poke and prod at these models in useful ways, understand their failure modes and their successes, benchmark their success sort of shape the direction through science funding. My suspicion is that often this is going to end up being spread across different agencies. But the more important thing is, does the government have the state capacity to actually do this? And then you can sort of figure out the like exact structure later.

Mr. KEAN. OK. Thank you. Ms. Tabassi, do you have—

Ms. TABASSI. All the great things, as we mentioned, and I think that expertise in the domains is important so there is reasons—that's why we did—we talk about development of the profiles or verticals, and that goes the same thing for the agencies with certain expertise.

Mr. KEAN. OK. Thank you. And I yield back.

Chairman OBERNOLTE. The gentleman yields back.

We will hear next from the gentleman from California. Mr. Mullin, you're recognized for five minutes.

Mr. MULLIN. Thank you, Mr. Chair. Thank you all for your attendance at this important hearing.

When we talk about trustworthy AI, the use of such technology in machines in the physical world pose a danger that I'd like to dig into just a little bit further. In San Francisco, which I partially represent, we have seen autonomous vehicles, or AVs, which had great promise, struggle to address some complex real-world situations. These malfunctions have obstructed public transit routes, blocked intersections, and impeded first responders from reaching people in need.

Last month, I led a letter, along with Speaker Emerita Pelosi, to the National Highway Traffic Safety Administration (NHTSA) to improve its collection of safety data so different manufacturers may be compared relative to each other and relative to manually operated vehicles. We're waiting to hear back from NHTSA on what steps it plans to take. But this is just one area where we need to be able to assess AI capabilities and its limits. And I'm concerned that we do not have clear standards for the assessment of safety.

So, Ms. Tabassi, could you share more about the work NIST is doing in this area relative to safety?

Ms. TABASSI. Thank you. Safety is being identified as one of the seven trustworthiness characteristics of a—in AI RMF and it—beyond going—beyond the having definitions for what safety means, which understand that now is being defined and perceived differently across the different domains. We also need to have, as you said, tests and measurements for measuring the functionality, capability, and limits on the technical robustness, as well as the societal robustness.

In AI RMF, it mentions that the—in terms of the safety, most urgent parameterizations and most risk management is required when there are risks to human life and where there are risks of serious injuries. And also it mentions that when there are cases that significant risks are present, actual harm is happening, that there should be—that there's—the deployment and development should cease in a safe manner until all of those risks are being addressed and managed properly.

Mr. MULLIN. Thank you for that. And switching gears a bit, Dr. Bender, as an educator and researcher, do you have any thoughts on the topic of better educating Americans so we are prepared for the coming onslaught of humanlike content generated by AI? What does that practically look like? What are we talking about in terms of integrating this into classrooms? Are we talking about education campaigns, for example? Can you enlighten me?

Dr. BENDER. Yes. So I think one important piece of our next steps is leveling up our information literacy, not just, you know, in the young generation who are in classrooms right now, but across the population so that we can better handle, as you say, the onslaught of synthetic media. That's important. I think we can't rely on that solely. I think we also need to have something on the other end, for example, watermarking. Chairman Obernolte suggested maybe watermarking for authenticity, as well as watermarking the synthetic media. But yes, absolutely. We are living in a new environment where we now have to contend with, especially textually, something that previously could only have come from people all around us and figuring out how to—I see a lot of discourse in the education space, talking about making sure people know how to use these tools, and not enough about making sure people understand how they're actually working and what they're not good for.

Mr. MULLIN. Thank you for that. With that, I yield back.

Chairman OBERNOLTE. The gentleman yields back.

We'll go next to the gentlewoman from Pennsylvania. Ms. Lee, you're recognized for five minutes.

Ms. LEE. Thank you, Mr. Chair, and thank you to the witnesses for your time and expertise on this critical area of technological innovation.

Every day, individuals and families in my district are impacted by algorithms that determine whether their insurance claim be accepted or their resume reviewed. Design, decision, and development tools—or, excuse me, deployment tools based upon defined principles address bias. The key question is what policies can best guide how those policies are set?

There are numerous ethical, privacy, and economic imperatives for striking a balance between harnessing the benefits of AI and addressing its challenges, which will be crucial to ensuring AI truly has a positive impact.

I talk about being the first Black woman to represent Pennsylvania in Congress, so my concerns are strongest, though, in ensuring that the advancement of AI technology in our society doesn't result in Black folks, the LGBTQI+ community, and other marginalized communities being used as sacrificial lambs.

So discussing the practical applications of AI and machine learning truly highlights, one, the ethical implications of AI technology

to the intellectual property and privacy rights, and the continual war against misinformation and disinformation. It's unrealistic to expect that bias can be eliminated in AI tools. It's crucial to identify what kinds of bias assessments are utilized prior to deploying the AI system and the fact that it will be embedded in a larger system. So equitable outcomes and not just unbiased or equitable algorithms should be our utmost focus.

Being born and raised in this country's industrial heartland, I'd be remiss to not discuss these implications AI will have on our labor and workforce. They're wonderful opportunities to improve efficiency and simplify strenuous work activities, but we have to remain vigilant and cautious to protect all workers and provide them with access and opportunities for skills training and education so that they're not being left behind in the technological revolution.

So in everything that I do, the human element is the most critical and important to me. This remains the same in making sure that the laws and policies that we help to create and enforce build better outcomes and that technology supports a greater humanity rather than replace it, which is the fear.

Dr. BENDER, bias can present itself in models in a variety of forms: racial, statistical, human, psychological, et cetera. Not all types of biases are equal. How are AI and machine learning decision tools being designed and developed to improve the outcomes produced by these decision processes to mitigate bias?

Dr. BENDER. So a lot of the work isn't trying to do that at all. And Abeba Birhane points out that using machine learning is an inherently conservative as a past-preserving move because it says let's take the patterns from this data collected from the past and use them to affect the future. And so if we want instead to take patterns to build the future we want, we need to be thoughtful about how we curate training datasets. And to various other points that you brought up there, we need to make sure that there's recourse for people who are harmed by adverse decisions.

Ms. LEE. So from like—so from a policymaking standpoint, do you think it would be more important to define bias more broadly or must bias be defined within the context of AI machine learning application and use?

Dr. BENDER. That's a really good question. I think if we leave bias undefined, we're going to have problems. And you're absolutely right. There's things, for example, automation bias means we tend to trust machines because they seem objective, and we need to be able to work around that. But there's also bias in the form of like discrimination. And so I think any legislation around bias should be defining what bias means in that context.

Ms. LEE. So what regulations do you think should be applied to these tools across the board versus regulations that are specific to their use case? And in that case of the latter, how will we be able to do that for each case?

Dr. BENDER. So I think how to do it for each case falls back to what existing regulations actually already apply. Things that should happen across the board I think have to do with transparency around datasets, transparency around the fact of automation, and transparency of who's accountable, which person is ultimately accountable for decisions or outputs.

Ms. LEE. Thank you. Ms. Tabassi, my understanding based on conversations I had and my staff has had with researchers at CMU, Carnegie Mellon, is that for the most part AI decision tools are being designed and deployed without much forethought on the part of the entity deploying them as to what policies and principles will guide their use to ensure their fairness. What would you say Congress must do to separate the processes of setting principles that will guide AI and machine learning tools prior to design and deployment of such tools from the technical operation of such tools?

Ms. TABASSI. In AI RMF we talk about a continuum, and that the risk management and accountability is a shared responsibility across all of them. And for any of these trustworthiness characteristics, including bias, it's important to start thinking about this as the earliest stage of the planning and design. And so talking about how we can advance research centers and evaluations to bring all of these concepts into the design and development of the systems and not having them as an afterthought to the—after the systems is deployed is a very important topic.

Ms. LEE. Thank you. And thank you all so much for your time and your expertise. I yield back.

Chairman OBERNOLTE. The gentlewoman yields back.

We will hear perhaps last but certainly not least from my colleague from New York. Mr. Bowman, you're recognized for five minutes.

Mr. BOWMAN. Thank you so much, Mr. Chairman.

Ms. Tabassi, the need for better evaluation frameworks has been surfaced consistently today. Will NIST partner with other agencies and private sector organizations to provide sample data and create evaluations around specific high risk use cases like hiring?

Ms. TABASSI. So working across the community and with everybody and our colleagues across the government agencies and our colleagues across all of the AI community, academia, private sector, and civil society-centered development organizations is something that we do at NIST and that's how we work. So we are certainly not—more than open, eager, and enthusiastic to work with everybody on understanding how to evaluate AI systems in a way that not only look at the functionality, limits, and capabilities, but also their trustworthiness and their societal robustness. We look forward to work with that, and we're already talking about having conversations with the community on how to set up those evaluations and advance those techniques.

Mr. BOWMAN. Thank you so much.

Mr. Kratsios, just the other day I saw that the U.S. Government outlined more than 700 potential use cases for artificial intelligence across agencies. Our procurement system is a key lever. The executive branch has to incentivize responsible AI. As a former White House CTL, what is—CTO, excuse me—what is your perspective on ensuring that agencies contract with the most responsible, effective AI providers?

Mr. KRATSIOS. You bring up a very good point around procurement being a very powerful lever and incentivizing better behavior from these potential model developers or algorithm developers more broadly, and I think developing procurement methodologies and guidelines which align with a risk-based approach is the best

way to do it. And what—and the only way you can do that is if you can actually have done the hard technological and scientific discovery of what is a risk-based approach to monitor these particular use cases. So not to put more work on my friend over here, but once we can sort out some of those testing and evaluation techniques, I think it's a—it's very critical that we infuse them into our procurement methodologies.

Mr. BOWMAN. Yes, and thank you for leading the way for me, but I wanted Ms. Tabassi to comment if she wouldn't mind.

Ms. TABASSI. I want to say that some of the guidelines in the AI RMF can already be applicable and can—we can start building on top of those. And of course, advancing the evaluations is something near and dear to our hearts, and we're working with the community.

NIST has a long history of evaluations, and we already have a lot of different evaluations in the biometric, information retrieval, all of those things can be leveraged for evaluations of AI systems, and we're thinking of how to extending them, but also building this addition of being able to evaluate AI systems in their context of use with real interactions with the humans that operate them, use them, being impacted by them to be able to assess the risk and impact of the AI systems. We're excited about that work, and we look forward to work with the whole community on that.

Mr. BOWMAN. Thank you so much. I have a question for Dr. Bender. The development of large language models is still a fundamentally human process that's driven by developers, researchers, and contractors who make decisions at every stage of development. The university system plays a pivotal role in ensuring that we train a generation of thoughtful builders who take ethical and social ramifications into account in all of their research. How can the Federal Government collaborate better with the university research community to move in the right direction here?

Dr. BENDER. Thank you for that question, and thank you for recognizing the contractors who do a lot of the labor, often poorly compensated in developing these. I think that there's a role for the Federal Government to play in supporting research in the social sciences and humanities because that's what we're going to understand how these things are actually affecting society. And, you know, also, integrating ethics throughout the curriculum is something that I'm working on. I know many other people are as well, but figuring out a way to value that, I think Federal Government valuing it, it will help people value it if they're coming in through, say, computer science.

Mr. BOWMAN. Got it. I have no questions for Mr. Watney, but you can feel free to comment on anything. I've enjoyed your comments so far, so please add on wherever you see fit.

Mr. WATNEY. Thanks. Yes, I think maybe if this is starting to be—to wrap up, I'll give a more—

Mr. BOWMAN. Yes.

Mr. WATNEY [continuing]. Broader comment. But I think I'm encouraged by just how much agreement there seems to be kind of across the political spectrum on the importance of investing in new standards and privacy preserving techniques in cybersecurity and interpretability. Like I think there's a real foundation here for

America to play a proactive role in shaping the broader research ecosystem. And we should also recognize this is not a new thing. The United States has always recognized it has a really pivotal role in not only the rate of technological progress, but the direction. We've done this in—you know, through DARPA (Defense Advanced Research Projects Agency) for satellite technology and the internet. We've done this through the NIH (National Institutes of Health) for, you know, biomedical technology. We've done this through climate tech. So I think this is part of a longer American tradition of really being proactive about shaping technology. And there's just, I think, a lot of work to do.

Mr. BOWMAN. Thank you so much. I yield back.

Chairman OBERNOLTE. The gentleman yields back.

That concludes our Member questions. This has been a really interesting hearing. I want to thank all the witnesses for your valuable testimony and thank the Members for their great questions. The record will remain open for 10 days for additional comments and written questions from Members.

With that, our hearing is adjourned.

[Whereupon, at 4:25 p.m., the Subcommittees were adjourned.]

Appendix I

ANSWERS TO POST-HEARING QUESTIONS

ANSWERS TO POST-HEARING QUESTIONS

*Responses by Ms. Elham Tabassi*U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY*“Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence”*Ms. Elham Tabassi, Associate Director for Emerging Technologies, Information
Technology Laboratory, National Institute of Standards and TechnologyQuestions submitted by Chairman Frank Lucas

1. As Congress pursues legislative options regarding AI and accountability for algorithms at the company level, many efforts currently have been related to Section 230. Outside of this argument, there are positions that advocate for including algorithm audits and impact assessments in the federal procurement process to root out harm throughout the AI life cycle.

- a. Do you agree with these arguments?

NIST Response:

As a non-regulatory agency, NIST does not have a view on what specifically should be required in federal procurement. In general, as stated in the AI Risk Management Framework, we suggest audits and impact assessments as possible components, among many others, of a holistic risk management approach that can help map and measure projected or actual impacts of an AI system.

- b. How would this impact the procurement process and open competition in contract acquisition?

NIST Response:

The effects of mandating algorithmic audits and impact assessments for procurement and open competition would likely vary significantly depending on how they were implemented. Impact assessments are most effective when they can take into account the system’s context of use. Mandatory requirements would presumably impose some degree of burden that would impact some suppliers more than others, but the full impact would depend on the exact implementation.

2. How is NIST supporting the development of AI standards that both promote innovation and public trust in these systems? Can you elaborate on these activities for us?

NIST Response:

NIST participates and actively promotes the participation of a wide range of stakeholders – including their standards experts, impacted communities, domain experts, and other cross-disciplinary experts – in ongoing AI standards development work. There is a need for clear specifications, benchmarks and evaluation for trustworthy AI. By defining common vocabularies, establishing socio-technical characteristics of trustworthy AI, and developing metrics and testbeds for validation, verification, and evaluation of AI systems (including AI red-teaming), technical standards can set the right safeguards that accelerate the pace of innovation in a safe and trusted manner. NIST has been working with the AI community on advancing an interoperable lexicon for trustworthy AI, building the scientific foundations

regarding what to test and how to test AI for safety and trust, and creating technical building blocks needed to develop consensus-based standards.

Global leadership, participation, and cooperation on international AI standards will be critical for consistent “rules of the road” that enable market competition, preclude barriers to trade, and allow innovation to flourish. The [Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence](#) (EO 14110) directs the Secretary of Commerce to establish a plan for global engagement on promoting and developing AI standards. In its role as federal AI standards coordinator, NIST works across the government and with industry stakeholders to identify critical standards development activities, strategies, and gaps. NIST tracks AI standards development opportunities, periodically collecting and analyzing information about federal agencies’ AI standards-related priority activities, and making recommendations through the interagency process to optimize engagement.

Standards shape the design, development and use of technologies that underpin our economies, cultures, and societies. AI standards that articulate requirements, specifications, test methodologies, or guidelines designed around values which promote public trust can help ensure that AI technologies and systems meet critical objectives – such as functionality, interoperability; and trustworthiness characteristics such as validity, reliability, safety, and security, and trustworthiness. In contrast, standards that are not fit for purpose, not yet available, not broadly accessible (notably to start-ups and small and medium-sized enterprises), or not designed around valid technological solutions may hamper innovation and the timely development and deployment of trustworthy AI technologies.

3. Small businesses are a critical component of the U.S. economy, but often they are faced with limited resources and tight budgets. According to the U.S. Chamber of Commerce, one in four small businesses have adopted AI in some manner and those who have saw a 12-point increase in their likelihood of profit growth over non-AI users. Ensuring small business owners have the necessary tools and support to grow and prosper is vitally important to me, and new technology innovations, such as AI, offer a great deal of promise and potential to do just that.
 - a. What resources is NIST developing to help small businesses measure their risks and properly utilize AI technologies?

NIST Response:

NIST recognizes how important it is for small businesses to have easy access to practical, easy-to-use, and reliable resources for understanding and managing AI-related risk. The NIST AI Risk Management Framework (AI RMF), which is intended to be used by organizations of any size in any sector of our society, recognizes this need and states “...small to medium-sized organizations managing AI risks or implementing the AI RMF may face different challenges than large organizations, depending on their capabilities and resources.” The companion AI RMF Playbook was designed by NIST to assist the full scope of Framework users—from those who are very familiar and fully equipped to tackle AI risk management, to those who may be challenged in terms of staff and funding to support those risk management activities.

Additionally, the NIST’s Trustworthy and Responsible Artificial Intelligence Resource Center was launched on March 31, 2023. The AI resource center is a platform to support people and organizations in government, industry, and academia—both in the U.S. and internationally—driving technical and scientific innovation in AI. It serves as a one-stop-shop for foundational content, technical documents, and AI toolkits such as glossary, repository hub for standards, measurement methods and metrics, and data sets. It also provides a common forum for all AI actors to engage and collaborate in the development

and deployment of trustworthy and responsible AI technologies and standards. To match the complexity of AI technology, the AI resource center will grow over time to provide an engaging interactive space that enables stakeholders to share AI RMF case studies and profiles, educational materials and technical guidance related to AI risk management, and needed standards.

NIST is committed to keeping the special needs of small businesses in mind as it produces additional guidelines and explanatory materials relating to safe, secure, and trustworthy AI.

4. Training datasets are not created with perfect objectivity — they are constructed by humans, meaning they are subject to human biases. How can we ensure there are adequate data transparency requirements so consumers and researchers can understand and analyze the potential biases of datasets?

NIST Response:

Data transparency and documentation methods can help AI actors across the AI lifecycle make explicit the many implicit assumptions and decisions in AI design and development. Documentation processes help define the limitations and socially appropriate uses of AI systems, in turn reducing the likelihood of perpetuating or amplifying systemic bias in context. For documentation practices to be adequate, they should be established by organizational policy and implemented under strong AI governance structures. These practices are best when informed by interdisciplinary and multi-stakeholder inputs, applied systematically and iteratively, and connected to organizational measurement protocols. The NIST AI Risk Management Framework offers some suggested approaches.

Questions submitted by Rep. Haley Stevens

1. We know some of the harms that happen when AI systems fail—whether that is a generative AI system providing false information or a credit ranking system discriminating against a protected class of people. However, we don’t have a sense of how common or rare these failures are. Moreover, most modern evaluations don’t account for how these systems affect the people that use them in the real world.
 - a. What types of testing are needed to fully understand the human impacts of AI failures?

NIST Response:

We currently lack comprehensive information about the frequency and volume of AI risks and associated negative or positive impacts on individuals, and the conditions under which these outcomes occur. The research community can work to develop testing protocols beyond traditional approaches such as model-testing in laboratory settings and include evaluations of systems and their impacts within the broader context of use. Structured human feedback methods, such as red teaming and user testing, can safely surface failure modes of systems before they are released, and improve understanding of the many ways that end users may take up and act upon information presented to them by generative AI systems. Outcomes of these testing approaches can shed light on how frequently end users encounter risks and impacts (both positive and negative) when interacting with generative AI technology.

2. One of the major barriers to training trustworthy AI systems is a lack of clean, representative, and privacy-enhanced training data.
 - a. Would you care to comment on NIST’s ongoing work to improve privacy enhancing technologies?

NIST Response:

NIST partnered with the National Science Foundation (NSF) in leading the recent Summit for Democracy PET challenge. This effort resulted in 90 ideation papers and 19 submitted technical solutions utilizing privacy preserving techniques. The final challenge stage had 6 winning solutions that were demonstrated at an international workshop and provided as open source packages to the privacy community of interest.

NIST is now working with several agencies to build upon the data privacy knowledge gained during the PETs challenge, and further develop privacy enhanced data as well as statistically robust synthetic data based on real-work data sets. NIST and NSF are co-leading a Networking and Information Technology Research and Development (NITRD) interagency working group for Privacy R&D that includes DHS, DOE, NIH, and the US Census Bureau. In addition, NIST is working with the US Census Bureau to explore the development of a US PETs Lab. This forum will be used to provide PETs documentation for developers and policy makers and establish prototyping environments for government, academic, and industry partners to safely experiment with PETs.

Questions submitted by Rep. Rick Crawford

1. As the federal government considers this AI Executive Order and the impending OMB guidance, do you think it is important to look to utilizing the NIST AI RMF when considering how best to implement procurement guidelines for federal agencies when purchasing AI products and services?

NIST Response:

President Biden cited the importance of federal agency use of AI in his October 30, 2023, [Executive Order \(EO\) on Safe, Secure, and Trustworthy Artificial Intelligence](#) (EO 14110). That order states “It is important to manage the risks from the Federal Government’s own use of AI and increase its internal capacity to regulate, govern, and support responsible use of AI to deliver better results for Americans.”

The AI RMF is an important tool that would help any organization, including federal agencies, when considering the purchase of AI products and services. Of course, context matters, and it is vital that agencies take into account their specific needs and circumstances as they set their policies, processes, and procedures for procuring those AI offerings.

The AI RMF notes that a diverse set of people within an organization are responsible for and play a role in procurement AI technologies: “Procurement tasks are conducted by AI actors with financial, legal, or policy management authority for acquisition of AI models, products, or services from a third-party developer, vendor, or contractor.” That is consistent with another statement in the Framework: “The AI RMF is intended to be used by AI actors across the AI lifecycle and dimensions. The AI RMF calls out “procurement experts” as among those who are important in the “deploy and use” phases of the AI life cycle, and NIST suggests that these AI actors would benefit by use of the Framework.

2. Along with the publication of the AI RMF, NIST has published additional “cross-walks” that are intended to provide mapping and guidance for the relationship to other widely recognized specifications for trustworthy AI practices. What are the most important additional standards and/or guidance frameworks that federal agencies can use to augment their adoption of good governance practices for AI systems? Is NIST, or the U.S. government more broadly, active in any of those other standardization activities that are important to the future use of the AI RMF?

NIST Response:

NIST’s crosswalks are primarily about identifying how the elements of our framework correspond to the elements in other frameworks, particularly since the frameworks often include similar content that is organized in different ways.

In general, the AI RMF and its accompanying resources serve as a strong baseline for AI-specific governance practices. It may still be necessary to apply guidance frameworks on issues that are not specific to AI, such as cybersecurity or software quality assurance. AI risk management should be part of organizations’ efforts to address the full spectrum of enterprise risks from their products and services.

NIST, along with colleagues across the U.S. government, actively supports standards development processes, including discussions on risk management. One of our goals as we engage with standards development organizations on AI is for the AI RMF to serve as an international standard that will help AI actors around the globe build up their AI risk management capabilities.

3. Do you believe NIST is currently well resourced and ready to implement the goals of the AI RMF?

NIST Response:

NIST has worked collaboratively with others in government, private sector, academia and non-profit organizations right from the start in developing the AI RMF, just as it does in most of its AI activities. This collaboration has served as a positive force in developing resources like the AI RMF that are more likely to be implemented by industry, government, and others, and amplifies the investment that NIST makes in its AI work. Having that external commitment and involvement provides NIST with a considerable return on our investment. Increased funding would permit the agency to focus even more on assisting organizations in all sectors as they implement the AI RMF, and provide additional voluntary guidelines and suggested practices related to trustworthy and responsible AI technologies.

Questions submitted by Rep. Summer Lee

My understanding based on conversations I and my staff have had with researchers at Carnegie Mellon is that for the most part, AI decision tools are being designed and deployed without much forethought on the part of the entity deploying them as to what policies and principles will guide their use to ensure their fairness.

1. How do we create safeguards to establish the rights of individuals over the use of their image, voice, and/or likeness in AI technologies? What are the metrics being used in learning models to measure such safeguards?

NIST Response:

Currently, most of the metrics used for measuring the behavior and impacts of AI systems focus on the technical characteristics of the systems and their ability to perform their tasks adequately. Metrics for societal impacts of AI systems have not yet been defined in our collective measurement toolkit. The research community could contribute to addressing this gap by collaborating to develop a new, systematic measurement science relating to trustworthy AI. NIST expects to play an important role in developing the kinds of methods and practices necessary for measuring AI system trustworthiness and societal impact. That, in turn, will contribute to a fuller understanding of how AI technology may affect the rights of individuals. It is up to policymakers to determine whether and how to establish any new legal rights or protections, perhaps leveraging newly developed metrics of social impact.

- a. What is the responsibility of model developers and how can we ensure that they have the rights to the data they are using to train the model?

NIST Response:

The AI RMF encourages organizations to apply risk management practices throughout the AI lifecycle including data collection activities or other risks that may encroach on intellectual property or other legal rights. To that end, the Framework offers a variety of target outcomes that AI actors can strive for, and the companion Playbook suggests concrete actions to achieve those outcomes. It will be up to policymakers to determine whether and which measures model developers should be held legally responsible for, and the mechanisms for enforcing such measures.

- b. How can we both identify policy measures and develop technologies to protect creative content from unauthorized use. From the point of view of anyone--- artist, entrepreneur, etc who seeks to make a living by promoting and selling their creative content on the internet, how can they ensure that content won't be scraped and used by a developer?

NIST Response:

As a technical matter, one proposal has been to establish protocols by which content can be labeled by its creators in a machine-readable fashion as not intended for use as training data, or affirmatively labeled as permissible for use as training data. For some types of content, such information can be embedded as metadata in the files themselves (e.g., alongside metadata like what camera was used to capture the image); others are pursuing methods where servers can host content to supply indicators of whether or how the content they host may be used. The latter is already done, for example, to indicate when content should not be indexed by search engines, and some AI model developers have already established similar protocols for excluding websites as training data.

The EO 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence tasked NIST to develop a report to the Director of OMB and the Assistant to the President for National Security Affairs identifying existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, for:

- Authenticating content and tracking its provenance
 - Labeling synthetic content (e.g., watermarking)
 - Detecting synthetic content
 - Preventing generative AI from producing Child Sexual Abuse Material or producing non-consensual intimate imagery of real individuals
 - Testing software used for the above purposes
 - Auditing and maintaining synthetic content
2. How can we ensure basic accountability in the use of these tools so that, for example, I can reassure my constituents whose insurance claim or resume was rejected by an algorithm that it arrived at that decision fairly if I don't know what data it's been trained on because that is proprietary information?

NIST Response:

Promoting the trustworthiness characteristics identified in the AI RMF can help provide greater assurance. For example, by enhancing AI system interpretability and transparency, AI actors can provide insight into how a system makes its decisions and help other stakeholders examine system performance and other trustworthiness characteristics such as fairness (another trustworthiness characteristic). The NIST AI RMF companion Playbook provides specific suggestions for how to enhance AI system interpretability, explainability, and transparency.

As I noted when testifying before the Committee, we need better testing metrics and methodologies for assessing AI trustworthiness characteristics such as explainability, interpretability, and transparency. If we develop such metrics and apply them consistently to the systems making these decisions, we will be better able to assess whether the systems are explainable or fair, in some cases even if we cannot inspect the training data or other internal aspects of the AI system itself.

3. “Shudu”, a Black AI-generated model was featured in campaigns for Balmain, Ellesse, and even Vogue. The creator said the model was inspired by human models, black women, like Naomi Campbell and Alex Wek.
- a. How can we promote actual diversity and inclusion in training AI and machine learning models to ensure a diverse workforce is not in competition with AI-generated ‘diversity’?

NIST Response:

AI systems can be used to generate content representing different population subgroups. This content generally will not be representative of the true subgroup unless the systems are based on representative training data, and governed in a way that solicits and incorporates the subgroup's views and experiences. This can be addressed via structured human feedback mechanisms, as well as a demographically and interdisciplinarily diverse workforce within the organization that is empowered to question design and

implementation decisions and has the authority and stature to make such changes.

Responses by Mr. Michael Kratsios

U.S. HOUSE OF REPRESENTATIVES COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

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

Mr. Michael Kratsios, Managing Director, Scale AI, 4th Chief Technology Officer of the United States

Questions submitted by Chairman Frank Lucas

1. As Congress pursues legislative options regarding AI and accountability for algorithms at the company level, many efforts currently have been related to Section 230. Outside of this argument, there are positions that advocate for including algorithm audits and impact assessments in the federal procurement process to root out harm throughout the AI life cycle.

- Do you agree with these arguments?

Response: As discussed at the hearing, there is a need for robust sector and use case specific test and evaluation frameworks for AI. This is relevant both for commercial and federal applications.

- How would this impact the procurement process and open competition in contract acquisition?

Response: Test and evaluation is a well understood part of the procurement process today for everything from aircraft to weapon systems. The government should take a sector specific, risk-based approach to test and evaluation when procuring AI.

2. Training datasets are not created with perfect objectivity — they are constructed by humans, meaning they are subject to human biases. How can we ensure there are adequate data transparency requirements so consumers and researchers can understand and analyze the potential biases of datasets?

Response: The best methodology for testing and evaluating LLMs remains an evolving technical question. Scale recently published its own technical methodology for LLM test and evaluation, which combines automated assessment with human evaluation of AI systems. There is still much work to be done, and collaboration among industry, the government, and academia is essential to get this right.

3. In your written testimony, you discuss the success of the DEF CON 31 hacker conference's AI red-teaming "village". Are there any specific key takeaways or new findings from that event that the Federal government should be aware of or incorporate into the development of AI testing and evaluation standards across the Federal government?

Response: This past summer's DEF CON event highlighted the critical role that red-teaming and test and evaluation play in the AI lifecycle development process. Over the course of the three day event over 2,200 people had the opportunity to test eight of the leading Large Language Models and provide critical data to developers on the potential vulnerabilities of their models.

From the government's point of view, this exercise was beneficial and the results of it will be made public after an embargo period to ensure that the model developers have adequate time to address the vulnerabilities that were found throughout the exercise.

Questions submitted by Rep. Rick Crawford

1. As the federal government considers this AI Executive Order and the impending OMB guidance, do you think it is important to look to utilizing the NIST AI RMF when considering how best to implement procurement guidelines for federal agencies when purchasing AI products and services?

Response: The NIST Risk Management Framework is important to consider. It is important for the Government to prudently execute on the red-teaming and test and evaluation requirements called for in the recent Executive Order and OMB Guidance.

2. Along with the publication of the AI RMF, NIST has published additional "cross-walks" that are intended to provide mapping and guidance for the relationship to other widely recognized specifications for trustworthy AI practices. What are the most important additional standards and/or guidance frameworks that federal agencies can use to augment their adoption of good governance practices for AI systems? Is NIST, or the U.S. government more broadly, active in any of those other standardization activities that are important to the future use of the AI RMF?

Response: The recent AI Executive Order tasked NIST with standing up working groups to create certain standards. NIST should also assess the existing work of industry groups like the ML Commons and the Frontier Model Forum to better understand where work has already begun on these topics. Getting these standards right will require

collaboration among Government, industry, and academia.

3. Do you believe NIST is currently well resourced and ready to implement the goals of the AI RMF?

Response: NIST is a critical asset to our nation and the Executive Order and OMB Implementation memo recently tasked them with standing up a variety of new responsibilities related to moving AI. Timely execution of these new lines of effort will require an appropriate level of resourcing.

Responses by Dr. Emily M. Bender
**U.S. HOUSE OF REPRESENTATIVES COMMITTEE ON SCIENCE, SPACE, AND
 TECHNOLOGY**

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

Dr. Emily M. Bender, Department of Linguistics, University of Washington

Questions submitted by Chairman Frank Lucas

1. As Congress pursues legislative options regarding AI and accountability for algorithms at the company level, many efforts currently have been related to Section 230. Outside of this argument, there are positions that advocate for including algorithm audits and impact assessments in the federal procurement process to root out harm throughout the AI life cycle.

A. Do you agree with these arguments?

B. How would this impact the procurement process and open competition in contract acquisition?

Response from Dr. Emily M. Bender:

1a) Section 230 is meant to provide immunity to internet carriers (and similar) from liability for the content posted by third party users. However, in the case of synthetic media generating machines (for example, chatbots driven by large language models like ChatGPT) the words output by the software do not reflect the thoughts or opinions of some user. They are purely synthetic text, reflecting the probabilities gleaned from the training data. As such, they are the output of the software and I believe that the creators/providers of that software should be accountable for them.

1b) By requiring things such as algorithm audits and impact assessments in the federal procurement process, the federal government has the opportunity to guide the industry towards higher standards of practice. If there is concern about smaller businesses not being able to afford these processes (and thus not being able to compete with larger,

more established players), Congress could consider setting up a program that performs the audits and impact assessments for corporations under a certain size.

Questions submitted by Rep. Haley Stevens

1. One of the major barriers to training trustworthy AI systems is a lack of clean, representative, and privacy-enhanced training data. Last October, researchers discovered that a popular open-source dataset used to train AI systems contained private medical records—violating the privacy of those exposed.

a. What are the strengths and limitations of PETs to mitigating privacy risks associated with machine learning?

Response from Dr. Emily M. Bender:

1a) While PETs can be helpful, different approaches have different efficacy. For example, anonymized data can often be re-identified (especially in combination with other data sources). Ultimately, a company using PETs to train a system based on private data is still in possession of that data, raising the risks that the data might be stolen, used to re-identify further data, or otherwise used maliciously. I think we need to move to a system with consent at its core: data should be used to train machine learning systems only if the data subjects have provided their informed, opt-in consent for it to be used. Furthermore, that consent should be revocable. Such a system would mean that datasets are much more expensive to create and maintain, but that is okay. I don't think we have to take it for granted that the ability to construct massive datasets for machine learning is somehow more important than individual data rights.

Questions submitted by Rep. Rick Crawford

1. As the federal government considers this AI Executive Order and the impending OMB guidance, do you think it is important to look to utilizing the NIST AI RMF when considering how best to implement procurement guidelines for federal agencies when purchasing AI products and services?

2. *Along with the publication of the AI RMF, NIST has published additional “cross-walks” that are intended to provide mapping and guidance for the relationship to other widely recognized specifications for trustworthy AI practices. What are the most important additional standards and/or guidance frameworks that federal agencies can use to augment their adoption of good governance practices for AI systems? Is NIST, or the U.S. government more broadly, active in any of those other standardization activities that are important to the future use of the AI RMF?*
3. *Do you believe NIST is currently well resourced and ready to implement the goals of the AI RMF?*

Response from Dr. Emily M. Bender:

1) Yes, the NIST AI RMF is thoughtfully constructed and based on research into the risks and harms of so-called AI systems. It also reflects broad public input. It is definitely a good starting point for these purposes.

2) Federal agencies should consider the various dataset and model documentation toolkits which have been proposed, including:

- Gebru et al 2021 Datasheets for Datasets:
<https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/abstract>
- Mitchell et al 2019 Model Cards for Model Reporting:
<https://dl.acm.org/doi/abs/10.1145/3287560.3287596>
- Bender et al 2021 Data Statements for Natural Language Processing:
<https://techpolicylab.uw.edu/data-statements/>

The second part of this question (about what activities the US government is currently engaged in) is outside my expertise.

3) This question is outside my expertise. However, NIST is doing important work and so it is important to make sure that it is well-resourced.

Questions submitted by Rep. Summer Lee

1. *Bias can present itself in models in a variety of forms; racial, statistical, human, psychological, etc. Not all types of biases are equal. Bias in AI tools is only part of an overall decision process that involves human decision-making and therefore may also*

introduce human, or psychological, bias into the decision process. For example, if you decide that, in the interest of erring on the side of security, you are going to tolerate flagging a certain number of innocent people, how do you design the tool to ensure that it is not disproportionately erroneously flagging one demographic over another?

Response from Dr. Emily M. Bender:

A key step in building such a tool is in the evaluation. As Dr. Margaret Mitchell advocates for in her work on Model Cards and elsewhere, it is extremely important to perform disaggregated analysis so that system performance for different demographics can be analyzed and compared. Dr. Joy Buolamwini echoes this point in her recent book *Unmasking AI*.

Doing this effectively requires a good understanding of which groups are at risk of being disproportionately flagged as well as carefully constructed test cases and datasets to use for auditing. In addition, it is important to understand the effects of different kinds of errors (for example, false positives vs. false negatives) in context. What is the impact of incorrectly flagging someone? What is the impact of failing to flag someone who should have been flagged? These effects vary enormously depending on the use case for the tool—consider the impacts if the tool is for deciding who should be brought in for questioning by the police vs. who should be offered a particular medical screening.

Creating effective test cases as well as understanding the impact of different kinds of errors requires participation by the communities likely to be harmed by the technology. This participation should be a part of the entire system lifecycle: informing system design, informing pre-deployment tests, and part of on-going monitoring of system performance.

Responses by Mr. Caleb Watney

Responses to Questions for the Record (my replies in bold):

Questions submitted by Chairman Frank Lucas

1. As Congress pursues legislative options regarding AI and accountability for algorithms at the company level, many efforts currently have been related to Section 230. Outside of this argument, there are positions that advocate for including algorithm audits and impact assessments in the federal procurement process to root out harm throughout the AI life cycle.

a. Do you agree with these arguments?

If chosen carefully and scoped appropriately, requirements for audits, evaluations, and/or benchmarks as part of the federal procurement process can be an important way to ensure high-quality AI tools for use in the government while also serving as a proving ground for the deployment of novel audits and evaluations.

b. How would this impact the procurement process and open competition in contract acquisition?

It will be important to scope the requirements carefully to ensure that they do not become unnecessarily burdensome or prevent the development of a competitive market. However, well-designed evaluations and audits can increase the odds of ultimately achieving the given product goals. General concerns about a lack of open competition in the procurement process would be best addressed through a more holistic reform of the federal acquisition regulation process.

2. In your testimony, you highlight confidential computing, including hardware innovations, as a research area for mitigating risk and securing AI. How can this technology be researched and leveraged to bolster American innovation and dominance in semiconductors and AI hardware?

Defensive cybersecurity techniques for securing frontier models, including confidential computing, would benefit from a portfolio of R&D funding strategies. For instance, federal R&D funders could consider entering into flexible research partnerships with hardware companies in the U.S. to develop sophisticated Trusted Execution Environments (TEE) and tamper-proof electronic systems on cutting-edge AI chips. Ultimately, if the US succeeds in developing a combination of hardware and software innovations that enable strong privacy and cybersecurity assurances, that will benefit the US lead in foundation AI models as well as in global chip exports.

Questions submitted by Rep. Haley Stevens

1. We know some of the harms that happen when AI systems fail—whether that is a generative AI system providing false information or a credit ranking system discriminating against a protected class of people. However, we don't have a sense of how common or rare these failures are. Moreover, most modern evaluations don't account for how these systems affect the people that use them in the real world.

a. In your testimony you describe the need for improved benchmarking and evaluation for real-world capabilities and weaknesses of today's models. What are the barriers to ensuring AI benchmarks and evaluation criteria account for the human in the loop?

Developing benchmarks that allow for open-ended task environments or humans-in-the-loop is an objectively more difficult technical task and requires a different development paradigm than many of the traditional AI benchmarks. Because the capabilities of today's frontier models are so novel, it will take time and investment to develop new benchmarks that truly test these systems. The [paper](#) "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?" presents one intriguing example of the sorts of benchmarks that are appropriately challenging for today's models. However, this also highlights that the most relevant benchmarks (both for AI capabilities and AI harms) are likely to be far more specialized and use-case-specific in the future. The federal government has a vital role to play in facilitating the development and testing of these new benchmarks.

2. One of the major barriers to training trustworthy AI systems is a lack of clean, representative, and privacy-enhanced training data. Last October, researchers discovered that a popular open-source dataset used to train AI systems contained private medical records—violating the privacy of those exposed.

a. In your testimony, one of the priority research priorities you describe is privacy-preserving machine learning. I agree that we need to improve privacy enhancing technologies to address privacy risks with AI systems. Can you describe the state of readiness of privacy-enhancing technologies? What should the Federal government be doing to improve their effectiveness and promote their adoption?

Some techniques, like differential privacy, have been well-studied but have had issues on the implementation side — for example, the initial rollout of differential privacy at the Census Bureau was widely critiqued by social scientists who relied on the data. Other techniques, like fully homomorphic encryption, present enormous promise but are still immature and need more fundamental research efforts to improve the computational efficiency tradeoffs. The federal government should fund additional technical work in this area, with a particular focus on translational research and pilot programs that demonstrate effective implementation strategies that could be scaled in federal dataset contexts.

Questions submitted by Rep. Rick Crawford

1. As the federal government considers this AI Executive Order and the impending OMB guidance, do you think it is important to look to utilizing the NIST AI RMF when considering how best to implement procurement guidelines for federal agencies when purchasing AI products and services?

Utilizing NIST's AI RMF is an important first step for establishing a standardized approach to evaluating and mitigating risks associated with AI technologies in federal

procurement. This framework could provide a consistent methodology for assessing AI systems.

2. Along with the publication of the AI RMF, NIST has published additional “cross-walks” that are intended to provide mapping and guidance for the relationship to other widely recognized specifications for trustworthy AI practices. What are the most important additional standards and/or guidance frameworks that federal agencies can use to augment their adoption of good governance practices for AI systems? Is NIST, or the U.S. government more broadly, active in any of those other standardization activities that are important to the future use of the AI RMF?

Apart from the AI RMF, international standards like ISO/IEC 20546 on big data and ISO/IEC JTC 1/SC 42 for AI are potentially significant. Active participation in international standard-setting bodies is crucial for ensuring that global standards for AI align with American technological values. NIST's involvement in these bodies helps in harmonizing standards and practices globally. Beyond the adoption of standards, red-teaming for potential vulnerabilities and malicious use cases will be an important capability to develop across government.

3. Do you believe NIST is currently well resourced and ready to implement the goals of the AI RMF?

NIST has recently been given a substantially expanded role in AI measurement and standard-setting under the recent Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. However, these additional responsibilities come without a proportional increase in funding which can come only from Congress. Prior to the executive order, NIST would have benefited from additional funding and resources to implement the AI RMF, but now this task has become even more essential. In addition to additional funding and infrastructure, it may be prudent to consider creating a NIST Foundation so that the agency would be enabled to partner with philanthropic communities.

Appendix II

ADDITIONAL MATERIAL FOR THE RECORD

STATEMENT SUBMITTED BY REPRESENTATIVE VALERIE FOUSHEE

Statement for the Record of Rep. Anna G. Eshoo

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

Joint Oversight & Investigations and Research & Technology Subcommittee
Committee on Science, Space, and Technology
October 18, 2023

Artificial intelligence (AI) has extraordinary potential to benefit society. It also has peril. The safe use of AI can lead to life-changing breakthroughs in medicine and science, but unsafe use can result in bias and discrimination, exacerbating existing inequities in society. To promote the research and development of AI that will benefit everyone, I introduced bipartisan, bicameral legislation, the *CREATE AI Act (Creating Resources for Every American to Experiment with Artificial Intelligence)*.

Today, research and development of leading AI is primarily in the hands of a few large technology companies. It's essential that *all* sectors have access to these resources, including startups, small businesses, colleges and universities, nonprofits, and the government. The *CREATE AI Act* establishes a National Artificial Intelligence Research Resource (NAIRR), a shared cyber research infrastructure for researchers across *all* sectors to access the high-powered computing, data sets, and other resources needed to develop cutting-edge AI technology.

Sustained investments in R&D have made the United States the global leader in AI, from the beginning to the present day. American companies lead the world in developing advanced AI systems and identifying cutting-edge AI applications that benefit society and improve our economy, but we cannot take this leadership for granted. We will only maintain our position through investment, competition, and active engagement with the full breadth and diversity of U.S. AI talent. The ability of the U.S. to fully realize the benefits of AI will require U.S. researchers across *all* sectors of society to access the necessary resources, especially researchers who have been historically excluded from researching and developing cutting-edge technology, including AI.

Passing the *CREATE AI Act* and establishing the NAIRR is an important action Congress can take to strengthen U.S. AI R&D. As AI continues to advance, both Democrats and Republicans can agree that we must ensure the U.S. leads the world in responsible AI development by allowing everyone with innovative ideas to leverage this technology. Cultivating a vibrant, competitive, and inclusive AI innovation ecosystem that reflects American values will protect national security, foster scientific progress, and drive economic growth. This will in turn increase America's future technological competitiveness. These outcomes won't be possible through action by any single sector or entity. It will require collaborative action among government, academia, the private sector, and non-profits, and that's exactly what my legislation does by putting the NAIRR in place which will democratize AI R&D and ensure America remains the world leader.

LETTER SUBMITTED BY REPRESENTATIVE HALEY STEVENS



October 18, 2023

Chairman Mike Collins and Ranking Member Valerie Foushee
 Chairman Jay Obernolte and Ranking Member Haley Stevens
 House Committee on Science, Space and Technology
 Subcommittee on Investigations and Oversight and
 Subcommittee on Research and Technology
 2318 Rayburn House Office Building
 Washington DC 20515

Re: CAIDP Statement for the Record: Joint Committee Hearing on *Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence*, October 18, 2023

Dear Chairman Collins and Ranking Member Foushee, Chairman Obernolte and Ranking Member Stevens, and Committee Members,

We write to you, on behalf of the Center for AI and Digital Policy (CAIDP)¹, regarding the October 18, Joint Committee Hearing on *Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence*.²

We previously testified on AI policy before the House Oversight Committee on *Advances in AI: Are We Ready For a Tech Revolution?*³ As CAIDP President Merve Hickok told the Committee: “We do not have the guardrails in place, the laws that we need, the public education, or the expertise in government to manage the consequences of the rapid changes that are now taking place.”⁴

¹ Center for AI and Digital Policy, <https://www.caidp.org>

² Joint Oversight & Investigations and Research and Technology Subcommittee Hearing, *Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence*, Oct. 18, 2023, <https://science.house.gov/2023/10/joint-oversight-investigations-and-research-technology-subcommittee-hearing-balancing-knowledge-and-governance-foundations-for-effective-risk-management-of-artificial-intelligence>

³ Testimony and statement for the record of CAIDP President Merve Hickok, *Advances in AI: Are We Ready For a Tech Revolution?* House Committee on Oversight and Accountability: Subcommittee on Cybersecurity, Information Technology, and Government Innovation (Mar. 8, 2023), https://oversight.house.gov/wp-content/uploads/2023/03/Merve-Hickok_testimony_March-8th-2023.pdf.

⁴ *Id.*



Your hearing on Foundations for Effective Risk Management could help address some of these challenges. In this statement, we respond to several of the questions set out in the hearing charter.⁵ Our recommendations in brief are:

1. Governance of AI systems should be centered on the rule of law, human rights, and democratic values
2. Congress should move forward with comprehensive AI legislation. A voluntary risk management framework is not a substitute for AI governance
3. The FTC should complete its investigation of OpenAI with urgency
4. Congress should promote the use of Privacy Enhancing Technologies (PETs) that reduce the collection and use of personal data
5. The US should lead on global AI policy and push forward the Council of Europe AI treaty

About CAIDP

The Center for AI and Digital Policy (CAIDP) is an independent research organization based in Washington, DC. We advise national governments and international organizations regarding artificial intelligence and digital policy. CAIDP currently serves as an advisor on AI policy to the OECD, the Global Partnership on AI, the European Union, the Council of Europe, UNESCO, and other national and international organizations. In April 2023, we released the third edition of our *Artificial Intelligence and Democratic Values* Index, providing a comprehensive review of AI policies and practices in 75 countries.⁶

1. What types of methods, standards, and tools currently exist for managing risks associated with AI systems?

*Established frameworks such as the Universal Guidelines for AI provide clear, implementable guidance for AI systems based on human centered values.*⁷

⁵ Committee on Science, Space, and Technology, *Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence*, https://republicans-science.house.gov/_cache/files/6/2/62beb7bf-0a1e-45d1-be3f-9735bb992b5d/78C980B22030E727E358E0DECfBE2545.2023-10-18-balancing-knowledge-and-governance-foundations-for-effective-risk-management-of-artificial-intelligence-hearing-charter.pdf

⁶ CAIDP, *Artificial Intelligence and Democratic Values* (2023), <https://www.caidp.org/reports/aidv-2022/>

⁷ Public Voice, *Universal Guidelines for Artificial Intelligence*, Guideline 5, <https://thepublicvoice.org/ai-universal-guidelines/>



The Universal Guidelines on AI (UGAI) were adopted in 2018 and over 330 leading experts and 60 associations (including the AAAS, the ACM, and the IEEE) have endorsed the UGAI. The UGAI sets out 12 principles that are foundational for the governance of AI systems. The aim of the Universal Guidelines is to maximize the benefits and minimize the risk of AI systems.

We advise Congress to implement these principles and move forward with governance frameworks and technical standards based on the UGAI.

2. Where do fundamental knowledge and methodological gaps exist for mitigating risks associated with AI systems?

The single most significant gap for mitigating risks associated with AI systems is the lack of federal legislation for AI. A voluntary risk management framework is not a substitute for regulations that ensure public safety. Congress should establish clear prohibitions on AI techniques that violate human rights such as mass surveillance and social scoring, mandate ex-ante human rights impact assessments for high-risk AI systems, and implement meaningful transparency measures.

We need federal AI legislation that would set standards of liability and accountability of actors in the AI life cycle. Providers of high-risk systems should be obligated to conduct ex-ante “human rights impact assessment.” Accountability mechanisms should also incorporate independent, third-party audits during the lifecycle of AI systems not just pre-deployment. Disclosure obligations should apply to public and private entities deploying AI systems and should also mandate explicit disclosure when content (including text, imagery, and other audio-visual content) is AI-generated.

Developers of foundation models as well as downstream users should be held responsible for ensuring safety by design and implementing specific safeguards on transparency of data practices and disclosure on AI-generated content. “As a part of careful data collection practices, researchers must adopt frameworks to describe the uses for which their models are suited and benchmark evaluations for a variety of conditions. This involves providing thorough documentation on the data used in model building, including the motivations underlying data selection and collection processes. This documentation should reflect and indicate researchers’ goals, values, and motivations in assembling data and creating a given model.”⁸

⁸ Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Margaret Mitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big, FAccT ’21*:



AI systems can replicate and amplify human biases. NIST has also reported on the problems of bias in data sets.⁹ Further, vulnerable populations are the most affected by biased AI outputs. There are examples of datasets created from human biases influencing AI to replicate systemic racism in facial recognition.¹⁰ The overriding objective of AI accountability mechanisms such as certifications, audits, and assessments should be to determine whether an AI system should be deployed. The NIST AI Risk Management Framework is voluntary, meaning that it does not set adequate and appropriate incentives for accountability.¹¹ This is critical given the individual and collective risks posed by unregulated AI systems.

Generative AI systems are built on unlicensed data, do not account for copyrights, and are already harming and marginalizing creators. In focusing investments in AI, Congress must be mindful of frameworks that result in concentration of technology in the hands of few corporations. This would severely restrict diffusion of innovation and be at cross-purposes with the U.S. National AI Research Agenda.¹²

We endorse the Hawley-Blumenthal bipartisan AI Act which offers a comprehensive framework for the governance of AI.

3. What role should the federal government play in the oversight of AI systems?

Companies should not release AI products that are not safe. President Biden has repeatedly stated that tech companies have a responsibility to make sure that their AI products are safe before making them public.¹³ OpenAI has itself acknowledged a dozen risks with ChatGPT. The federal government needs to act.

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (March 2022). Pages 610, 615 (Stochastic Parrots”), <https://doi.org/10.1145/3442188.3445922>

⁹ *More to AI Bias than Biased Data, NIST Report Highlights*, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (Mar. 1, 2022), <https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights>.

¹⁰ *Id.*

¹¹ See U.S. Department of Commerce, National Institute of Standards and Technology, “AI Risk Management Framework,” https://aifc.nist.gov/AI_RM/Knowledge_Base/AI_RM/.

¹² CAIDP, *Statement to OSTP on US National AI Strategic Plan*, (March 4, 2022), <https://www.caidp.org/app/download/8378181763/CAIDP-Statement-OSTP-03042022.pdf?t=1660245988>

¹³ The White House, *Readout of White House Meeting with CEOs on Advancing Responsible Artificial Intelligence Innovation*, (May 4, 2023) (“President Biden dropped by the meeting to underscore that companies have a fundamental responsibility to make sure their products are safe and secure before they are deployed or made public.”), <https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/readout-of-white-house-meeting-with-ceos-on-advancing-responsible-artificial-intelligence-innovation/>



The Federal Trade Commission (FTC) should move forward and conclude its investigation of OpenAI initiated by CAIDP. Pending legislation, the Federal Trade Commission (FTC) has a unique opportunity at this time to establish guardrails for AI. Earlier this year, CAIDP filed a detailed complaint with the FTC regarding OpenAI. The FTC has opened the investigation of OpenAI we requested.¹⁴ This is clearly a positive development, but now the FTC must prioritize this investigation. It took two years from the time we filed similar complaints with the FTC concerning Google and Facebook before there was a settlement.¹⁵ We can't wait that long this time. AI products are evolving rapidly and being deployed downstream in consumer facing services. Before the end of this year the FTC must complete the investigation of ChatGPT and enter into a settlement with OpenAI that ensures the companies will abide by the practices for AI companies the Commission has previously issued.

We acknowledge (and are actively promoting) the recent announcements of several federal agencies, including NTIA, OSTP, and PCAST, to provide opportunities for public comment on AI policies.¹⁶ But even these Requests for Comments have not yet led to notice of rulemaking that would lead to regulation. Meanwhile, the OMB rulemaking to establish regulations for the use of AI across the federal government is more than two years behind schedule.¹⁷ We also note with concern that the National AI Advisory Committee (NAIAC), is lagging in its commitments under the Federal Advisory Committee Act and public comment opportunities at NAIAC have been wanting.¹⁸

[intelligence-innovation](https://www.wsj.com/articles/unite-against-big-tech-abuses-social-media-privacy-competition-antitrust-children-algorithm-11673439411); Joe Biden, *Republicans and Democrats, Unite Against Big Tech Abuses*, Wall Street Journal, (January 11, 2023), <https://www.wsj.com/articles/unite-against-big-tech-abuses-social-media-privacy-competition-antitrust-children-algorithm-11673439411>

¹⁴ Cecilia Kang and Cade Metz, F.T.C. Opens Investigation Into ChatGPT Maker Over Technology's Potential Harms, The New York Times, July 13, 2023, <https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html>; John D. McKinnon and Ryan Tracy, *ChatGPT Comes Under Investigation by Federal Trade Commission*, Wall Street Journal, July 13, 2023, <https://www.wsj.com/articles/chatgpt-under-investigation-by-ftc-21e4b3ef>.

¹⁵ Federal Trade Commission, Facebook Settles FTC Charges That It Deceived Consumers By Failing To Keep Privacy Promises, Nov. 29, 2011, <https://www.ftc.gov/news-events/news/press-releases/2011/11/facebook-settles-ftc-charges-it-deceived-consumers-failing-keep-privacy-promises>

¹⁶ CAIDP, *Public Voice*, <https://www.caidp.org/public-voice/>

¹⁷ CAIDP Letter to Shalanda Young, Director Office of Management and Budget, Apr. 24, 2023, [hKps://www.caidp.org/app/download/8454950563/CAIDP-Statement-OMB-04242023.pdf](https://www.caidp.org/app/download/8454950563/CAIDP-Statement-OMB-04242023.pdf)

¹⁸ CAIDP *Statement to the National AI Advisory Committee* (Oct. 28, 2022), <https://www.caidp.org/app/download/8420970963/CAIDP-NAIAC-FACA-10282022.pdf>; see generally, CAIDP, *National AI Advisory Committee*, <https://www.caidp.org/resources/naiac/>



We also strongly support the notification for a rulemaking petition issued by the Federal Election Commission for reforming campaign finance laws to preserve the democratic process and election integrity from AI generated misinformation and disinformation.¹⁹ We believe that the FEC should act fast on this process and amend its regulation on fraudulent misrepresentation of campaign authority to make clear that the related statutory prohibition applies to deliberately deceptive Artificial Intelligence campaign advertisements and public communications.²⁰

4. Where should the federal government focus investments to promote the development and deployment of trustworthy AI?

The federal government should incentivize the use of Privacy Enhancing Technologies to minimize the risks of AI systems. Many machine learning systems rely on the collection of personal data. These systems can violate privacy and be used for mass surveillance. New AI systems pose new challenges for privacy, dignity, autonomy, and equality. Metrics for explainability, interpretability, and transparency should be established to protect fundamental rights, human well-being, and to increase public trust. These metrics alongside Privacy Enhancing Technologies would help protect privacy.²¹ **Privacy Enhancing Techniques (PETs) minimize or eliminate the collection and use of personal data.**

In December 2021, the Biden administration announced an initiative to encourage development of ‘Democracy-Affirming Technologies’ that support democratic values and governance. Relatedly, the U.S. and UK announced plans to promote Privacy Enhancing Technologies (PETs), including low-data AI, the deletion of unnecessary data, and techniques for robust anonymity.²²

5. How will international approaches to AI governance influence U.S. policies?

The United States should support a comprehensive treaty for the governance of AI. The EU is moving ahead with the EU AI Act and there have been a lot of proposals on AI governance

¹⁹ Federal Election Commission, *Artificial Intelligence in Campaign Ads*, 11 CFR Part 112, Notice 2023-13, Federal Register Vol. 88, No. 157, Aug. 16, 2023, <https://sers.fec.gov/fosers/showpdf.htm?docid=423639>.

²⁰ 11 C.F.R. §110.11, §11 C.F.R. 110.16. *See also*, 52 U.S.C. §30124

²¹ *Comments of CAIDP to OSTP on National Artificial Intelligence Research and Development Strategic Plan* at 4, <https://www.caidp.org/statements/> (March 4, 2022)

²² CAIDP, *Artificial Intelligence and Democratic Values Index (2022)*, pg. 1070, <https://www.caidp.org/reports/aidv-2022/>; *See also*, The White House, *US and UK to Partner on Prize Challenges to Advance Privacy-Enhancing Technologies* (December 8, 2021) <https://www.whitehouse.gov/ostp/news-updates/2021/12/08/us-and-uk-to-partner-on-a-prize-challenges-to-advance-privacy-enhancing-technologies/>



Center for AI and
Digital Policy

in Congress—many of which are bi-partisan.²³ The US and EU can move ahead with a transatlantic accord on AI governance which would harmonize and align the governance of high-risk AI systems.

The Council of Europe AI Treaty is “a standard-setting instrument developed through a global approach, based on international law, focusing on human dignity and human rights, as well as gender equality, social and economic justice and development, physical and mental well-being, diversity, interconnectedness, inclusiveness, and environmental and ecosystem protection can guide AI technologies in a responsible direction.”²⁴ The United States should support a comprehensive international treaty such as the Council of Europe AI Treaty.

The United States has recently rejoined UNESCO, citing the important work of UNESCO on artificial intelligence.²⁵ The U.S. should also take a leadership role in the G7 to establish common guardrails and international technical standards based on human-centered values.

Thank you for your consideration of our views. We would be pleased to provide you and your staff with additional information. We ask that this statement be included in the hearing record.

Sincerely,

Merve Hickok
CAIDP President

Christabel Randolph
CAIDP Law Fellow

Marc Rotenberg
CAIDP Executive Director

Brianna Rodriguez
CAIDP Law Fellow

²³ Marc Rotenberg, *Everything. Everywhere. All at Once: AI Policy When Congress Returns*, ACM Blog, Aug. 24, 2023, <https://cacm.acm.org/blogs/blog-cacm/275742-everything-everywhere-all-at-once-ai-policy-when-congress-returns/fulltext>

²⁴ CAIDP, Council of Europe AI Treaty, <https://www.caidp.org/resources/coe-ai-treaty/>.

²⁵ The American Independent, *Biden leads US back to membership in UNESCO*, Jun. 30, 2023, <https://americanindependent.com/joe-biden-donald-trump-china-unesco-blinken-ai/>.



Center for AI and
Digital Policy





Center for AI and
Digital Policy



ARTICLE SUBMITTED BY REPRESENTATIVE SUZANNE BONAMICI

10/18/23, 5:32 PM

How Flawed Data Aggravates Inequality in Credit

Stanford University



Open

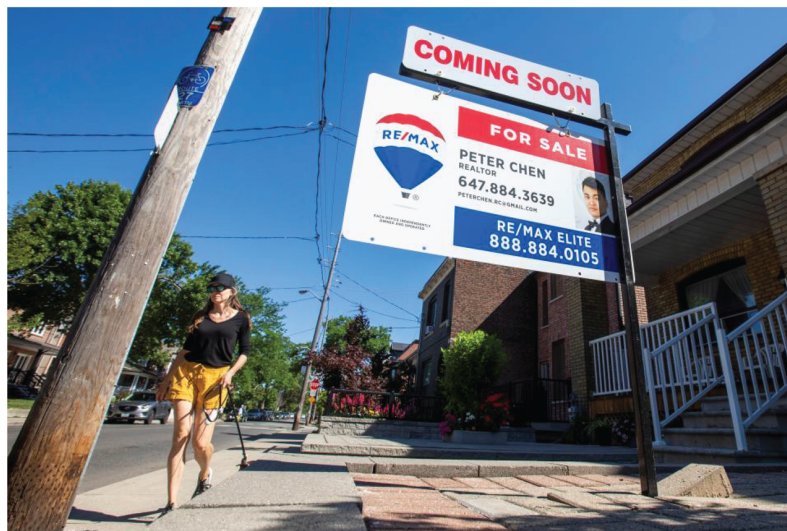
Search this site

Ethics and Justice, Machine Learning

How Flawed Data Aggravates Inequality in Credit

AI offers new tools for calculating credit risk. But it can be tripped up by noisy data, leading to disadvantages for low-income and minority borrowers.

Aug 6, 2021 | Edmund L. Andrews



10/18/23, 5:32 PM

How Flawed Data Aggravates Inequality in Credit

Credit scoring models can be between 5 and 10 percent less accurate for lower-income and minority homebuyers, new research shows. | Carlos Osorio

or aspiring home buyers, getting a mortgage often comes down to one talismanic number: the credit score.

F Banks and other lenders are turning to artificial intelligence to develop increasingly sophisticated models for scoring credit risk. But even though credit-scoring companies are legally prohibited from considering factors like race or ethnicity, critics have long worried that the models contain hidden biases against disadvantaged communities, limiting their access to credit.

Now a preprint study in which researchers used artificial intelligence to test alternative credit-scoring models finds that there is indeed a problem for lower-income families and minority borrowers: The predictive tools are between 5 and 10 percent less accurate for these groups than for higher-income and non-minority groups.

It's not that the credit score algorithms themselves are biased against disadvantaged borrowers. Rather, it's that the underlying data is less accurate in predicting creditworthiness for those groups, often because those borrowers have limited credit histories.

A "thin" credit history will in itself lower a person's score, because lenders prefer more data than less. But it also means that one or two small dings, such as a delinquent payment many years in the past, can cause outsized damage to a person's score.

"We're working with data that's flawed for all sorts of historical reasons," says Laura Blattner, an assistant professor of finance at the **Stanford Graduate School of Business**, who co-authored the new study with Scott Nelson of the University of Chicago Booth School of Business. "If you have only one credit card and never had a mortgage, there's much less information to predict whether you're going to default. If you defaulted one time several years ago, that may not tell much about the future."

The Root of the Problem

In analyzing the issue, the researchers used artificial intelligence and huge volumes of consumer data to test different credit-scoring models.

The first step was to figure out if the standard credit-score approaches were equally accurate across different demographic groups. They used AI to crunch anonymized consumer credit data on 50 million people that was provided by one of the major credit-score companies. They also worked with a huge marketing dataset that makes it possible to identify borrowers by income, race, and ethnicity.

Part of the challenge was figuring out whether people who were rejected for home loans would have been likely to default on mortgages if they had been approved. To do that, the AI models looked at how rejected mortgage applicants had kept up with other kinds of loans, such as car loans, which correlate very closely with how likely a person is to default on a mortgage.

Sure enough, the credit scores turned out to be less accurate for low-income and minority borrowers than others.

Overall, Blattner and Nelson estimate that there is substantially more "noise" or misleading data in the credit scores of people in minority and low-income households. They found that scores for minorities are about 5 percent less accurate in predicting default risk than the scores of non-minority borrowers. Likewise, the scores for people in the bottom fifth of income are about 10 percent less predictive than those for higher-income borrowers.

But why? Were the algorithms biased because they couldn't pick up on the distinctive patterns of particular demographic groups?

To find out, Blattner and Nelson tested alternative scoring models that had been fine-tuned to minority and low-income borrowers. The results didn't change much: The scores were still less accurate for them.

The real problem was in the data. People with very limited credit files, who had taken out few loans and held few if any credit cards, were harder to assess for creditworthiness. That was especially the case for people who had one or two blemishes on their records. Because minority borrowers and low-income earners were more likely to have thin or spotty credit records, their credit scores were less accurate.

<https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit#:~:text=AI offers new tools for,low-income and minority borrowers.&text=F...> 2/4

10/18/23, 5:32 PM

How Flawed Data Aggravates Inequality in Credit

By making the existing credit report data more informative, such as by compiling thicker and more diverse files for minorities, Blattner and Nelson estimate that it's possible to eliminate half the disparity in accuracy.

A Misallocation of Credit

The findings have a number of important implications. The study suggests that people from disadvantaged communities are being turned down for mortgages more often than they probably should be, though some of those borrowers may also get approved when they probably shouldn't be. That creates a misallocation of credit. It can also perpetuate the inequalities because people who can't get mortgages have less opportunity to build up solid records. They also then lose out on a crucial avenue for building wealth.

Some experts have argued that the algorithms run into trouble because financial institutions are legally prohibited from incorporating factors like ethnicity, gender, or race into their models. Those prohibitions are intended to prevent discrimination, but they can also block lenders from recognizing differences between groups that might actually elevate some borrowers' credit scores.

The study finds that is not a problem. Even when Blattner and Nelson created scoring models that were fine-tuned to minority and low-income borrowers, the scores were still less accurate for those groups.

There is no simple solution, Blattner says. But one possible strategy would be for financial companies to run experiments in which they approve loans to people with relatively low credit scores.

"If you're a bank, you could give loans to people and see who pays," says Blattner. "That's exactly what some fin-tech companies are doing: giving loans and then learning."

Stanford HAI's mission is to advance AI research, education, policy and practice to improve the human condition. [Learn more.](#)

More News Topics

Ethics and Justice, Machine Learning
Related Content



How Zillow, Census Data, and Machine Learning Could Improve Our Water Use

by Michelle Horton
November 19th, 2020

Stanford researchers develop a new approach to help cities better understand water use and design water-efficient...



HAI Fellow Shazeda Ahmed: Understanding China's Social Credit System

by Mary Duan
August 31st, 2020

Does a tracking system making laws more enforceable actually improve society?