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.

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 aced 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.