PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-z23z-tfov
Comments Due: May 28, 2025 **Submission Type:** API

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0288
Comment on FR Doc # 2025-07332

Submitter Information

Name: Jessica Newman

General Comment

Thank you for the opportunity to provide a response to the Request for Information on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan. We are professors and researchers with expertise in AI research and development, policy, and ethics, affiliated with centers at the University of California, Berkeley, including the AI Security Initiative, Center for Long-Term Cybersecurity (CLTC), School of Information, Berkeley AI Research Lab's Responsible AI Initiative, Center for Information Technology Research in the Interest of Society and the Banatao Institute (CITRIS), Berkeley Center for Law & Technology, and the UC Berkeley AI Policy Hub. Please see the attached file.

Attachments

AIRandD2025_BerkeleyCLTCResponse

RFI Response: Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan
White House Office of Science and Technology Policy
Docket ID No. NSF-2025-OGC-0001.

Dear Michael Kratsios, Director of the White House Office of Science and Technology Policy (OSTP),

We thank the OSTP, the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), and the National Science Foundation for the opportunity to submit comments in response to the development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan. We are professors and researchers with expertise in AI research and development, policy, and ethics, affiliated with centers at the University of California, Berkeley, including the AI Security Initiative, Center for Long-Term Cybersecurity, School of Information, Berkeley AI Research Lab's Responsible AI Initiative, Center for Information Technology Research in the Interest of Society and the Banatao Institute (CITRIS), Berkeley Center for Law & Technology, and the UC Berkeley AI Policy Hub.

Some of us have previously submitted a response to the OSTP on the *Development of an AI Action Plan*, in which we emphasized that to be an innovation leader, the U.S. must develop state-of-the-art AI systems that are robust, reliable, and secure.[1] We also emphasized the **importance of supporting the work of the US AI Safety Institute, working with international allies, and facilitating information sharing about AI testing and evaluation results between AI companies and the US government**, among other recommendations.

Some of us have also previously submitted a response to the OSTP regarding the *2023 update to the National AI R&D Strategic Plan*.[2] In our response, we emphasized small modifications to the eight strategies included in the 2019 AI R&D Strategic Plan including the importance of **supporting multidisciplinary AI research, improving detection of malicious uses of AI, and supporting focus on international cooperation and coordination**, among others. We additionally proposed the inclusion of a ninth strategy to encourage support for **research that identifies effective mechanisms for transparency and documentation of AI systems and applications**. We argued that "Improving classification and documentation of AI systems and applications should be a research priority because the current lack of standardization contributes to the dearth of trust in AI development, preventing increased discovery and adoption." While significant advances have been made in AI transparency mechanisms and methodologies since the time of our comment, we believe the increasing capabilities and complexities of AI models (combined with alarming trends toward potential deception, cheating, and "reward hacking"[3]) means that exploring meaningful transparency remains a critical component of a forward-looking national AI R&D strategic plan.

---

[1] Nada Madkour et al, "Comment to the Department of Networking and Information Technology Research and Development (NITRD) and the National Coordination Office (NCO), on behalf of the Office of Science and Technology Policy (OSTP) on the Development of an Artificial Intelligence (AI) Action Plan," March 15, 2025, https://files.nitrd.gov/90-fr-9088/Nada-Madkour-AI-RFI-2025.pdf.
[2] Anthony M. Barrett et al., "RFI Response: National Artificial Intelligence Research and Development Strategic Plan— White House Office of Science and Technology Policy," March 4, 2022, https://cltc.berkeley.edu/wp-content/uploads/2022/03/OSTP-RFI-Draft-Comments.pdf
[3] See for example "AI models can learn to conceal information from their users" *The Economist*, April 23, 2025, https://www.economist.com/science-and-technology/2025/04/23/ai-models-can-learn-to-conceal-information-from-their-users and "Details about METR's preliminary evaluation of o3 and o4-mini", METR's Autonomy Evaluation Resources, 2025, https://metr.github.io/autonomy-evals-guide/openai-o3-report/.

We recommend the inclusion of three critical areas of focus in the development of a new 2025 National AI R&D Strategic Plan.

Our recommendations are intended to help ensure the new National AI R&D Strategic Plan enables the United States to secure its position as the world leader in artificial intelligence by performing R&D to enhance U.S. economic and national security and promote human flourishing. We argue this requires three areas of focus:

1. Expand research and development of AI standards, security, and reliability
2. Understand and address the implications of AI for human flourishing
3. Develop effective methods to preserve human oversight, control, and accountability

All of the nine strategies that were included in the 2023 update have continued relevance to varying degrees. We particularly recognize the critical importance of four of the strategies, including: Strategy 3: Understand and Address the Ethical, Legal, and Societal Implications of AI; Strategy 4: Ensure the Safety and Security of AI Systems; Strategy 6: Measure and Evaluate AI Systems through Standards and Benchmarks; and Strategy 9: Establish a Principled and Coordinated Approach to International Collaboration in AI Research. These four strategies are even more important today than they were two years ago given the increasing scale, scope, and capabilities of available AI models.

The three recommendations highlighted in this document build upon these previous strategies, taking into account important shifts in the AI landscape. They are designed to support the goal of building the world's best AI systems, which we define as not only being powerful, but also reliable, effective, and secure. We believe these recommendations will make America's AI products better, stronger, more competitive than other countries in both the short and long term, and ensure that AI fulfills its promise of economic growth.

Although our recommendations are generally supported by industry and will help foster a thriving innovation ecosystem, they are also likely to be neglected by industry, which has historically failed to prioritize long-term considerations over short-term profits. Additionally, our recommendations will require long-term R&D investment due to their scientific complexity and need for cross-sector and multidisciplinary collaboration. For the success of all our recommendations, it will be critical to ensure that independent academic research is able to continue with support from government funding and adequate access from industry.

Recommended Areas of Focus

1. Expand research and development of AI standards, security, and reliability

Expanding research into AI standards, security, and reliability is a strategic imperative that is critical for both national security and competitiveness and must complement the United States' innovation-first approach to governing AI. Notably, AI companies including Google, Amazon, Meta, OpenAI and many others support U.S. leadership in developing AI standards, testing, and evaluation.[4] **AI adoption will increasingly be contingent on the security of AI offerings as government and industry both look to implement AI at scale across critical processes.** Therefore, excellence in the next era in AI will require (1) advanced testing and evaluation of AI systems (2) innovation in security and reliability and (3) international coordination on standards and best practices for advanced AI security.

Recent disclosures from leading AI labs underscore a rapidly escalating risk landscape. Google's Gemini 2.5 model has already triggered internal alerts for its potential to significantly lower the barrier to high-impact cyberattacks.[5] Similarly, Anthropic's Claude 4 Opus has been deployed with ASL-3 measures due to significant improvements in chemical, biological, radiological, and nuclear (CBRN) weapons development capabilities.[6] Meanwhile, OpenAI has quietly deprioritized manipulation and deception as a critical risk,[7] even as evidence of the persuasive power of frontier models mounts[8] to a degree already deemed unacceptable by many experts.[9] Nonetheless, apart from some state-led efforts,[10] governance frameworks on this front are still nascent. In addition to malicious use concerns, the security of AI systems is a complex challenge.

AI systems are increasingly a mixture of disparate data sources, pre-trained models, and software libraries. Each element in the MLOps life cycle represents a potential attack vector, from data poisoning during training, prompt injection, and hiding malicious code in widely used libraries. A compromise at an early stage in the supply chain can infect numerous downstream AI systems, leading to cascading failures that are difficult to trace and remediate. This highlights the **need for R&D into secure AI supply chain practices, including robust data provenance mechanisms, model bills of materials (MBOMs), continuous monitoring of AI systems, transparency mechanisms, and verification techniques.**

---

[4] "Tech Industry and Safety Groups Push for AI Safety Institute," Americans for Responsible Innovation, October 22, 2024, https://responsibleinnovation.org/wp-content/uploads/2024/10/20241021ARI_ITIOctoberAISIHillLetter.pdf

[5] Google (2025) Gemini 2.5 Pro Preview Model Card. Google, https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf

[6] Although Anthropic has not yet determined if the model has crossed the threshold that would require ASL-3, they also could not clearly rule out ACL-3 risks. See, Anthropic (2025) System Card: Claude Opus 4 & Claude Sonnet 4. Anthropic, https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.

[7] Sharon Goldman and Jeremy Khan (2025) OpenAI updated its safety framework—but no longer sees mass manipulation and disinformation as a critical risk. Fortune, https://fortune.com/2025/04/16/openai-safety-framework-manipulation-deception-critical-risk/

[8] "AI can do a better job of persuading people than we do," *MIT Technology Review*, May 19, 2025, https://www.technologyreview.com/2025/05/19/1116779/ai-can-do-a-better-job-of-persuading-people-than-we-do/

[9] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz (2024) How persuasive is AI-generated propaganda? PNAS Nexus, https://academic.oup.com/pnasnexus/article/3/2/pgae034/7610937

[10] For instance, legislative proposals such as Illinois HB3506 are beginning to call for enforceable safety thresholds

AI failures can lead to financial, reputational, and competitiveness costs that businesses will not want to incur. High-profile failures can disproportionately erode confidence not just in the specific system or developer involved, but also across the broader AI field. This ripple effect suggests a collective responsibility among developers and a pressing need for baseline security that may at present be inconsistent across the industry. **We strongly recommend increased investment in establishing government leadership on AI standards, security, and reliability** to prevent downstream harms and ensure AI systems operate within acceptable bounds of public safety and democratic accountability. While the private sector plays a vital role in advancing AI capabilities, it has limited incentive to invest in the kind of rigorous, reproducible evaluation infrastructure needed to anticipate failures, mitigate systemic risks, and ensure public accountability, especially in high-risk or under-resourced sectors. Public-sector leadership is essential to build the necessary safety and security artifacts, such as developing datasets, benchmarks, and access-restricted testbeds, to help operationalize a robust, reproducible, and verifiable evaluation paradigm.

Traditional software security relies on the ability to analyze source code, predict system behavior under well-understood conditions, and verify compliance with static requirements. However, for AI systems, especially generative AI and adaptive models, behavior can evolve post-deployment, and their internal decision-making processes are often opaque. Building on international efforts such as the UK AI Safety Institute's *Inspect* framework, **the U.S. should fund State-led R&D to support the development of transparent, reproducible evaluation methods for foundation models.** In addition to testing, the U.S. will need complementary research into safety standards and intervention protocols. While recent proposals from industry consortia have made progress in defining voluntary safety commitments, there remains a pressing need for public investment in building reliable safety and security against accident and misuse risks including CBRN and cyber capabilities. **Research is needed to define intolerable risk thresholds for safe deployment**, operationalize off-switch and rollback mechanisms, and develop inter-organizational coordination procedures for emergency response, particularly for the most dangerous scenarios involving CBRN weapons, cyberwarfare, model deception, or loss-of-control risks.

Importantly, such risks from AI do not emerge in a vacuum. The Administration should fund research to support the design and deployment of sociotechnical systems that reflect context specific goals and values, and evaluation methods that focus on the impact of sociotechnical systems in which AI is embedded rather than only model or system outputs. This systems level research is necessary to support the federal government's development and use of AI consistent with M-25-21 Accelerating Federal Use of AI through Innovation, Governance, and Public Trust and to support development and evaluation efforts in the private sector.

Advisory bodies, researchers, advocates and policy makers have emphasized the **importance of taking a sociotechnical systems approach to managing AI risks**.[11]

[11] National Artificial Intelligence Advisory Committee, NAIAC Year 1 Report 2023, May 2023, https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf;
National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework 1.0, Appendix 3, January 26 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf NIST Publications;
Polemi et al., "Challenges and Efforts in Managing AI Trustworthiness Risks: A State of Knowledge," Frontiers in Big Data 7:1381163, May 9 2024, https://doi.org/10.3389/fdata.2024.1381163
Dobbe et al., "System Safety and Artificial Intelligence," arXiv:2202.09292, 2022, https://arxiv.org/abs/2202.09292;
Raji and Dobbe, "Toward Standardized Documentation of AI Systems," ACM FAccT 2020, https://doi.org/10.1145/3351095.3372831; National Academies of Sciences, Engineering, and Medicine, Assessing and Improving AI Trustworthiness: Current Contexts and Concerns, 2021, https://doi.org/10.17226/26208;
Bogen and Winecoff, "Applying Sociotechnical Approaches to AI Governance in Practice," Center for Democracy & Technology / The GovLab, May 23 2024,

Researchers report that a key failing of current risk management methods is an overemphasis on "the technological artefact…in isolation" and absence of attention to "the human factors and systemic structures that influence whether a harm actually manifests."[12] They find that an emphasis on technical components leaves key sources of risk unidentified and unaddressed.

Moving AI risk management beyond models and data is aligned with learnings from safety science, and risk management practices in other high-risk fields. As Dobbe explains, the field of system safety assumes that "systems cannot be safeguarded by technical design choices on the model or algorithm alone" therefore take an "end-to-end" approach to analyzing risks and a sociotechnical systems—including "the context of use, impacted stakeholders…and institutional environment—approach to deploying mitigations.[13] Governance models in other high-risk fields such as transportation, finance, and medicine reflect this holistic, systems approach to assessing and mitigating risk.[14]

As the NAIAC noted, advancing a sociotechnical approach to AI design, use, and evaluation "requires basic research at the intersection of technology, the humanities, and the social sciences that broadens the conception of AI research beyond technocratic frames" and is consistent with the systems framing of safety science. Such research includes funding for values-in-design and participatory methods to design AI systems to align with contextually specific requirements including ethical and legal obligations; boundary objects and methods to support domain experts and impacted communities participation in system designs and mitigation strategies; and as noted by a recent NASEM report, new measurement and assessment methods and tools that take an expanded view of what should be measured and assessed to assure systems are trustworthy and garner public trust.[15] The outputs of this research should be clear, specific, repeatable methods for designing, and metrics and testing methodologies for evaluations.

As a global leader in AI, **the U.S. should invest in identifying incentive mechanisms, such as procurement preferences, certification schemes, and shared evaluation platforms**, to encourage the adoption of robust safety standards across domains and jurisdictions. Cross-border research collaboration will be especially vital to align emerging standards and promote consistent enforcement of safety norms at scale. A coordinated national effort, starting from the creation of evaluation testbeds to the development and promotion of safety standards, along with efforts on the international diplomacy front, will be essential to ensure that AI capabilities evolve in a manner that is secure, reliable, and aligned with the public interest.

---

https://cdt.org/wp-content/uploads/2024/05/2024-05-23-AI-Gov-Lab-applying-sociotechnical-approaches-to-ai-governance-in-practice-final.pdf.
Office of Science and Technology Policy and Global Partnership on AI, AI Governance Framework, April 2024, https://gpai.ai/projects/governance/framework.pdf.

[12] Weidinger et al., "A Preliminary Research Agenda for AI Safety," arXiv:2401.00001, 2024, https://arxiv.org/abs/2401.00001.

[13] Dobbe et al., "System Safety and Artificial Intelligence," arXiv:2202.09292, Section 1, 2022, https://arxiv.org/abs/2202.09292.

[14] National Academies of Sciences, Engineering, and Medicine, Assessing and Improving AI Trustworthiness: Current Contexts and Concerns, 2021, https://doi.org/10.17226/26208.

[15] National Academies of Sciences, Engineering, and Medicine, Proceedings of a Workshop in Brief: Assessing and Improving AI Trustworthiness, 2021, https://doi.org/10.17226/26208.

2. Understand and address the implications of AI for human flourishing

Our recommendations for understanding and addressing the implications of AI for human flourishing largely align with the content of strategy 3 in the 2023 AI R&D Strategic Plan,[16] but build on those recommendations to account for changes in the AI landscape. **Human flourishing requires an environment in which AI supports human dignity and values. This must be a world that is free from rampant cyber and information warfare, harmful discrimination, fraud, or abuse, and AI-driven job displacement**. **We recommend expanding research efforts on these topics in support of ensuring an environment that is supportive of human flourishing.** We further stress the importance of this research as it is unlikely that industry will invest resources in these areas of research.

Rampant cyber warfare can disrupt societal stability, undermine trust, and cause widespread harm. **We recommend increasing government-led and government-supported research and development to understand and mitigate risks of AI-enabled cyber warfare**, particularly in critical contexts such as infrastructure, healthcare, and national security. Actions can include development of advanced defensive technologies, identifying AI system vulnerabilities that can be weaponized, improving strategies for deterrence, and increased international cooperation.[17] Similarly, information warfare and influence operations may lead to destabilized societies, erosion of public trust, and undermined integrity of political systems. We recommend increased government-led and government-supported research into content provenance techniques.[18] Key actions may include developing tools to identify and counter extremist content, and enhancing international collaboration on standards for information integrity.

AI technologies can exhibit performance disparities across different groups, particularly when certain groups are over- or under-represented in training data. Additionally, AI technologies can reflect and reinforce harmful existing societal patterns, potentially leading to undesirable consequences. **We recommend increasing government-supported research to understand how these failures emerge in different technologies and applications, as well as efforts and innovations for mitigation**.

Additionally, excessive prioritization of automation may lead to AI-driven job displacement and unstable job markets. This particular issue is unlikely to be prioritized by industry due to its focus on technological advancement and cost reduction. **We recommend increasing government-led and government-supported research into the socio-economic impacts of excessive automation on the U.S. employment landscape**, with a focus on identifying vulnerable sectors and developing strategies for workforce transition. Key actions may include creating models to predict displacement, identifying areas to include in training programs for affected workers, researching

---

[16] Office of Science and Technology Policy, The National Artificial Intelligence R&D Strategic Plan, May 2023, https://www.nitrd.gov/pubs/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf.

[17] International cooperation is particularly critical in the context of managing AI-aided cyber warfare as the technology transcends borders and international misalignment will likely lead to significantly increased governance difficulty.

[18] See e.g., Marilyn Zhang “Strengthening Information Integrity with Provenance for AI-Generated Text Using ‘Fuzzy Provenance’ Solutions” February 13, 2025, https://fas.org/publication/strengthening-information-integrity-provenance/

potential new jobs, and mandating human oversight in safety critical or high stakes domains.

It is also important to acknowledge the distinct opportunities that AI technologies can present, by **investing in and incentivising the development of AI applications that support human flourishing**. These areas include, but are not limited to AI-driven solutions that improve quality and access to healthcare,[19] AI-driven innovations to enhance education and expand educational access,[20] AI driven-solutions to promote and aid environmental sustainability and US energy independence,[21] AI-driven solutions to safeguard and support human rights,[22] and research that addresses aligning technological advancement with the broader goal of human flourishing. Support for these application areas must also include support that ensures such tools are responsible – including being fair, transparent, and accountable, while also protecting data privacy and having robust security and safety mechanisms.

We also recommend **prioritizing and investing in strategic international collaborations** in the context of understanding and addressing AI implications on human flourishing. Given that AI development and use often transcend borders, it is crucial to prioritize international collaboration to ensure that critical opportunities and risks are addressed globally. We advocate for further research on expanding cooperation with allies to improve information sharing and avoiding harmful "race-to-the-bottom" dynamics. Key actions may include engaging in global AI research consortia, participating in and leading development of international standards, fostering cross-border AI diplomacy,[23] and creating incentives for international collaborative AI projects.

---

[19] See e.g., Mayo Clinic "AI in healthcare: The future of patient care and health management" March 27, 2024, https://mcpress.mayoclinic.org/healthy-aging/ai-in-healthcare-the-future-of-patient-care-and-health-management/

[20] This is supported by the recent executive order: White House "Advancing Artificial Intelligence Education For American Youth" April 23, 2025, https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth/

[21] See e.g., Alan Willie " AI and Environmental Sustainability: Using AI to Improve Energy Efficiency, Waste Management, and Conservation Efforts" September 2024, https://www.researchgate.net/publication/387363808_AI_and_Environmental_Sustainability_Using_AI_to_Improve_Energy_Efficiency_Waste_Management_and_Conservation_Efforts

[22] See e.g., Theresa Adie " Harnessing Technology to Safeguard Human Rights: AI, Big Data, and Accountability" April 8, 2025, https://www.humanrightsresearch.org/post/harnessing-technology-to-safeguard-human-rights-ai-big-data-and-accountability

[23] This can include agreements similar to the no first use (NFU) nuclear policy (see, Center for Arms Control and Non-Proliferation, n.d., https://armscontrolcenter.org/issues/no-first-use/)

3. Develop effective methods to preserve human oversight, control, and accountability

The capabilities of general-purpose AI, including AI agents, have increased rapidly in recent years and months.[24] As AI systems become increasingly autonomous and influential across critical sectors (including for AI-AI collaboration and communication), preserving human oversight, control, and accountability is essential to safeguard public trust, safety, and national security. In the current environment of competitive commercial pressures, industry is unlikely to sufficiently address these considerations, which means that sustained, long-term investment from the federal government will be important.

In particular, **foundational work on model interpretability and explainability should be elevated to a national priority**: by supporting research into a diverse array of probing techniques and mechanistic analyses—such as linear-probe detection of strategic deception, circuit-level interpretability methods, and concept-activation frameworks —we can equip regulators, auditors, and operators with the tools they need to uncover unexpected or covert model behaviors before they manifest in real-world harms.[25] [26] [27] [28] For the most capable AI models, providing open model weights may not be advisable due to safety, security, and competitiveness considerations, and so advancing R&D efforts for alternative transparency mechanisms is also likely to be critical. **It is also essential to invest in organizational transparency mechanisms—standardized incident-reporting frameworks, secure information-sharing platforms, and clear provenance tracking for model training data**—that keep people apprised of emerging risks and preserve a verifiable chain of accountability.

Equally critical is the development of robust AI control paradigms capable of enforcing human oversight even in unexpected scenarios. Building on research into adversarial-resilient intervention protocols—including automated kill-switch architectures and processes, and safely interruptible agent designs—the government should fund work on fail-safe rollback procedures, and continuous monitoring agents that remain effective under adversarial conditions.[29] [30] **These technical safeguards must be paired with parallel investments in workforce training and process redesign**: from up-skilling domain experts to recognize and respond to model misbehavior, to embedding human-in-the-loop review checkpoints in high-stakes domains (e.g., healthcare diagnostics, critical-infrastructure management).

Lastly, the U.S. government should **support AI governance research to ensure our regulatory processes–formal law, institutional staffing, testing, evaluation and oversight processes–are updated** to address gaps that would undermine civil rights, public safety, and competition. Such research should include topics such as

---

[24] "International AI Safety Report The International Scientific Report on the Safety of Advanced AI," AI Action Summit, January 2025, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.
[25] Nicholas Goldowsky-Dill et al., "Detecting Strategic Deception Using Linear Probes", 2025, https://arxiv.org/abs/2502.03407
[26] Samuel Marks et al., "Auditing language models for hidden objectives", 2025, https://arxiv.org/abs/2503.10965
[27] Huben et al., "Sparse Autoencoders Find Highly Interpretable Features in Language Models", 2024, https://openreview.net/forum?id=F76bwRSLeK
[28] Adly Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet", 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
[29] Ryan Greenblatt et al., "AI Control: Improving Safety Despite Intentional Subversion", 2024, https://arxiv.org/abs/2312.06942
[30] Stuart Russell, *Human Compatible: AI and the Problem of Control*, 2020

organizational structures and public-private arrangements that support optimal testing, evaluation, and oversight of AI models and deployed sociotechnical systems.

Combining a spectrum of interpretability and governance research, organizational transparency improvements, and adversarial control methods—and ensuring that these advances are matched by investments in the people and processes that will use them—will keep AI systems operating under U.S. guidance and maintain accountability to the public.

Contact

Thank you for the opportunity to comment on the National Artificial Intelligence Research and Development Strategic Plan. If you need additional information or would like to discuss further, please contact Jessica Newman at

Our best,

Jessica Newman, Director, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; Co-Director, UC Berkeley AI Policy Hub

AnnaLee Saxenian, Professor, School of Information, Faculty Advisor, Center for Long-Term Cybersecurity, UC Berkeley

Deirdre Mulligan, Professor, School of Information and Co-Director Berkeley Center for Law & Technology, UC Berkeley

Camille Crittenden, Executive Director, Center for Information Technology Research in the Interest of Society and the Banatao Institute (CITRIS), University of California

David Evan Harris, Chancellor's Public Scholar, UC Berkeley; Professional Faculty, Haas School of Business; Senior Policy Advisor, California Initiative for Technology and Democracy; Senior Research Fellow, International Computer Science Institute

Genevieve Smith, Founding Director, Responsible AI Initiative, Berkeley AI Research Lab, UC Berkeley

Nada Madkour, Non-Resident Research Fellow, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Evan R. Murphy, Non-Resident Research Fellow, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Deepika Raman, Non-Resident Research Fellow, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Krystal Jackson, Non-Resident Research Fellow, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Oumou Ly, Non-Resident Research Fellow, Center for Long-Term Cybersecurity, UC Berkeley

Sarah Powazek, Program Director, Center for Long-Term Cybersecurity, UC Berkeley

Grace Menna, Fellow, Public Interest Cybersecurity, Center for Long-Term Cybersecurity, UC Berkeley

This submission is not on behalf of the University of California, Berkeley. The views expressed here are the authors' own and do not reflect the views of the University of California, Berkeley.

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the 2025 National AI R&D Strategic Plan and associated documents without attribution.