Formal Response

Recommendations for an AI Action Plan

Response to OSTP's Request for Information

Author
Center for Security and Emerging Technology

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

March 2025

Introduction

Georgetown University's Center for Security and Emerging Technology (CSET) submitted the following recommendations on the Development of an Artificial Intelligence (AI) Action Plan, as directed by a Presidential Executive Order on January 23, 2025. These recommendations were drawn from CSET's wide body of research in the areas of assessment and evaluations, biosecurity, cybersecurity and AI, military AI applications, technology monitoring, U.S.-China competition analysis, and AI workforce research. The recommendations for the Trump administration broadly fall into three categories: 1) steps the United States can take to advance and secure its leadership in developing cutting-edge AI capabilities, 2) initiatives for competition in AI with China, and 3) actions the U.S. government can take to realize the benefits of AI while mitigating its risks.

We begin by laying out several recommendations for advancing American AI leadership through investments in research, innovation, and talent. First, we recommend increasing funding for public sector research into AI applications in the social and life sciences and sharing findings widely to promote new discoveries. Second, we recommend lowering the barriers to entry for emerging AI companies to stimulate innovation and healthy market competition. Promoting competitive markets and open AI models will enable new, diverse approaches to frontier research, offering alternative pathways to realizing the benefits of powerful models. Finally, we recommend that the United States further develop AI talent to strengthen our research ecosystem and create new business opportunities. Increasing funding for workforce training programs and creating an AI scholarship-for-service program will equip private sector workers and public servants with the necessary skills to develop and use AI tools most effectively. Maintaining AI leadership requires a sustained commitment to the research and expertise that have proven a fundamental advantage to the United States.

We next offer recommendations for navigating AI competition between the United States and China. We first recommend the administration take a two-pronged approach to strengthening U.S. technological competitiveness. Preventing illicit technology transfers, coupled with enhancing the effectiveness of U.S. export controls on key AI components, will ultimately weaken the Chinese AI ecosystem. This will require close cooperation with allies and partners to devise joint strategies and garner broad support for export control objectives: cooperation will translate into greater effectiveness. We also recommend that the administration prioritize alleviating information asymmetries between the U.S. government and frontier AI companies to avoid technological surprise and for intelligence sharing with allies. Investments in open source intelligence collection, both related to domestically-created capabilities and developments in China's AI ecosystem, will be critical to informing proactive U.S. AI policy. Finally, we recommend the administration

implement an AI incident reporting regime to aggregate data on AI failure modes and emerging risks, which can better inform safety science research priorities and risk mitigation strategies.

Finally, we recommend actions the U.S. government can take to realize AI's benefits while minimizing risks. Providing public assurances for redressing harms from AI systems and shielding whistleblowers from company retaliation can help the public more confidently engage with AI. We recommend that the administration create standard pathways to contest adverse AI-enabled decisions and establish whistleblower protections for employees at frontier AI firms, which may ultimately improve system performance by disincentivizing dangerous practices. The administration should also proactively guard against AI risks by creating threat profiles for how AI systems could be used to cause harm in areas such as biotechnology, along with tailored model safeguards to address these risks. AI evaluations and standards have a critical role to play in elucidating system capabilities and driving forward AI progress through collaboration with the private sector. The administration should empower federal bodies to advance AI evaluation science and standards and demonstrate how adoption of evaluations can facilitate AI's use for mission success.

RFI Response: The Development of an Artificial Intelligence (AI) Action Plan
Federal Register Document Citation: 90 FR 9088
Agency Name: National Science Foundation, Networking and Information Technology Research and Development National Coordination Office
Organization: Center for Security and Emerging Technology (CSET), Georgetown University
Primary POCs: Mia Hoffmann (mh2171@georgetown.edu), Jack Karsten (jk2497@georgetown.edu), Mina Narayanan (mjn82@georgetown.edu)

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following comments in response to the National Science Foundation Networking and Information Technology Research and Development National Coordination Office's request for comments on **the Development of an Artificial Intelligence Action Plan**. A policy research organization within Georgetown University's Walsh School of Foreign Service, CSET provides decision-makers with data-driven analysis on the security implications of emerging technologies, focusing on artificial intelligence, advanced computing, and biotechnology. We appreciate the opportunity to offer these comments.

Overview

Georgetown University's Center for Security and Emerging Technology (CSET) presents the following recommendations on the Development of an Artificial Intelligence (AI) Action Plan, as directed by a Presidential Executive Order on January 23, 2025. These recommendations are drawn from CSET's wide body of research, and fall largely into three categories: 1) steps the United States can take to advance and secure its leadership in developing cutting-edge AI capabilities, 2) initiatives for competition in AI with China, and 3) actions the U.S. government can take to realize the benefits of AI while mitigating its risks.

Promoting AI Research and Development

Driving AI research and development (R&D) should be a priority in the new AI Action Plan. The federal government fills a critical R&D gap and is uniquely situated to motivate research prioritization and production in specific areas. Given the private sector's dominance in driving AI R&D, the government can play a key role addressing important research gaps where there are no immediate profit motives, like understanding the social, political, and economic implications of AI. To promote U.S. AI R&D leadership, the government should incentivize and award projects that take interdisciplinary approaches, encourage research findings to be disseminated openly and widely, and support public sector research in coordination with private sector innovation. Since AI is a general-purpose technology, basic R&D supports downstream model development for commercial use, application, and, eventually, profits. We must build a thriving, interdisciplinary, and cross-sector AI research ecosystem to enhance America's AI dominance.

- **Couple commercial AI development and innovation with robust, sustained funding for R&D driven by the public sector**. This includes research done at universities, national labs, federally funded research and development centers, and nonprofits, sometimes in collaboration with the private sector, across a range of technical and non-technical fields. Among other things, such research examines how AI interacts with and impacts human behavior, processes, and structures; aims to improve our metrics for understanding and mapping the conversion of research to commercial innovation and real-world use; and creates well-documented and widely usable open data sets.

- **Promote AI-enhanced scientific progress:** The intersection of AI and Biotechnology (AIxBio) holds a great deal of promise—not only for biomedical and industrial innovations, but also as a competitive global battleground for the economic advantages and technological influence that AIxBio technologies will grant to the countries that lead in their development.

A future AI Action Plan must include robust initiatives to advance AI-driven biotechnology in order for the United States to enjoy these benefits. Such an action plan should include:

- **Support for AIxBio funding, capacity, infrastructure, and workforce.** While some of the largest, compute-intensive biological AI models are developed by industry, many of the more specialized, domain-specific AIxBio models come from academic research groups. Providing additional resources to AIxBio researchers will enhance the speed and range of achievable developments and facilitate their transition to real-world applications.
- **Infrastructure for biological data.** At present, there are significant limitations for the types of biological data that could power future AIxBio applications, including the lack of large, standardized, labeled, and annotated training sets for AI systems. The United States should prioritize the development of more robust databases, and explore incentives to encourage researchers to collect biological data in an AI-usable format.

Stimulating Markets, Competition, and Innovation

Foster dynamic and competitive markets

Sustaining long-term U.S. leadership in AI will require a dynamic and diversified domestic market for AI systems. Such an ecosystem—one in which many different developers are empowered to pursue a wide variety of AI techniques and tools, and build their successes into viable businesses—will promote innovation and ensure the United States remains at the forefront of AI. Today, however, the U.S. AI industry is dominated by a small group of incumbent firms that have the power and incentive to thwart rival American AI developers and potentially suppress disruptive AI innovation. To promote the long-term health of the U.S. AI industry, policymakers should lower barriers to entry for new AI developers and ensure that incumbent firms do not wield their power to stifle competition in the AI market. To that end, we offer three recommendations:

- **Promote a more open and competitive cloud computing market.** Compute resources are a crucial input for building and deploying AI systems. Many U.S. AI developers access these resources through cloud services providers (CSPs) like Amazon Web Services, Microsoft Azure, and Google Cloud Platform. Policymakers should crack down on egress fees, restrictive contracting provisions, and other tactics that CSPs use to "lock-in" customers, and implement nondiscrimination and open access rules to prevent CSPs from advantaging large AI developers over smaller AI firms.
- **Maintain open distribution channels for AI products.** In order to build successful businesses, AI developers market and distribute their products to customers. Many of the most prominent "distribution channels" for AI products—mobile devices, software suites, online marketplaces and app stores, and cloud platforms—are owned by incumbent technology companies that develop AI systems in-house or maintain financial ties with prominent third-party developers. Policymakers should prohibit companies from engaging in

self-preferencing, bundling, and other tactics that allow them to exclude rival AI developers from valuable product distribution channels. Such behavior can make it more difficult for new AI developers to gain a foothold in the market, hindering competition and innovation over the long-term.

- **Closely monitor mergers and acquisitions (M&A) and corporate "partnerships" within the AI industry.** M&A transactions can have positive innovation effects, allowing companies to gain economies of scale and acquire new technologies and talent, as well as negative effects, potentially reducing innovation incentives for incumbents and allowing them to prevent disruptive rivals from entering the market. In recent years, incumbent technology companies have also started engaging in "partnerships" with outside AI developers. These arrangements are not subject to the same regulatory scrutiny as traditional M&A transactions, and while their details are often opaque, they likely have similar positive and negative effects on innovation. The Federal Trade Commission and Antitrust Division of the Department of Justice should continue to closely scrutinize M&A transactions and corporate partnerships in the AI sector, and block corporate combinations when necessary for maintaining a competitive and contestable market.

Promote open AI models

Recent advances in AI foundation models have stirred debate over the benefits and risks of freely releasing model weights. We recommend that the Trump administration refrain from inhibiting the release of open models by U.S. companies unless they exceed clear and measurable risk thresholds. Open models feed into America's strengths: they foster innovation and competition, facilitate advances in AI security, and encourage entrepreneurial entry into the AI industry. In particular, we recommend the following:

- **Support the release of open-source AI models, datasets, and tools** that can be used to fuel U.S. AI development, innovation, and economic growth. Open-source models and tools enable greater participation in the AI domain, allowing lower-resource organizations that cannot develop base models themselves to access, experiment, and build upon them. They stimulate economic growth by increasing competition and drawing in more entrepreneurs. Open-source datasets allow for consistent benchmarking of AI progress, enabling a greater understanding of competitive advantages or gaps, and incentivize developers to strive for further successes.

- **Provide resources for the evaluation and analysis of the effects of open models** on advancing AI research and pushing forward private-sector AI growth. Such analyses can contextualize the benefits of open models, which can be compared to the potential risks of open model release from the perspective of AI competition with China.

- **Develop best-practices for the release of frontier open models** to help preemptively identify and minimize risks, but avoid regulations that unnecessarily hinder or disincentivize the opening of models. Work with industry to establish clear and measurable thresholds for intolerable risk that warrant keeping models closed, and avoid over-indexing on hypothetical risks. Ensure these thresholds account for marginal risk, and balance said risks with the benefits of openness.

- **Prioritize, alongside AI capability advancements, the diffusion of American AI models in the U.S. and global AI ecosystem**. Adoption of U.S. open models abroad builds reliance on U.S. technology, thereby endowing the U.S. government with soft power, serving as a foundation for stronger relationships and alliances with partners, and encouraging further paid use of related U.S. AI technologies like enterprise subscription services and cloud platforms. Promotion of U.S. AI technology abroad can also combat the growing influence of Chinese models especially in developing and emerging economies, and prevent China from providing the foundation for large parts of the global digital infrastructure, with implications for the diffusion of Chinese ideologies on the world.

Incentivize multiple approaches to frontier research

- **Incentivize alternative approaches to research on advanced AI in the United States and among our allies.** The United States and western nations regard large generative AI models as the main path to artificial general intelligence despite these models' cost, infrastructure demands, and known limitations. There are likely other ways to advance in AI without investing significant resources into a paradigm that is already approaching the limits of scale. By contrast, China also invests in alternative, brain-inspired projects such as human brain modeling, non-therapeutic brain-computer interfaces, and embodied AI that learns through value-driven interaction with the environment. Large-scale implementation of this last approach is underway now in Chinese cities. If the U.S. is concerned about the emergence of artificial general intelligence, it should make multiple bets on viable pathways, acknowledging the brittleness of a single focus and supporting alternatives, a truism Chinese policymakers recognize.

Developing and Securing Access to Talent

Strengthen the growing AI workforce

Apprenticeships offer a pathway to training, re-training, and upskilling for American workers of all backgrounds across the entire country. This includes apprenticeships in AI-related occupations, which have rapidly increased in number over the last decade. This increase has coincided with greater federal funding and support from successive administrations. Future success will depend on continuing support for these programs. We recommend that the Trump administration:

- **Increase funding for the federal National Apprenticeship system,** with an emphasis on technical occupations and industry intermediaries. The government should also provide funding for data collection and tracking of employment outcomes for Registered Apprenticeship Programs to determine if they lead to well-compensated jobs for apprentices.

Community colleges have enormous potential for training the next wave of AI workers, but require funding and support to succeed. Community colleges are located in every state and have a long history of training workers of all ages in emerging industries. However, they face a number of challenges, including uncertain, complicated, and insufficient funding streams. We recommend that the Trump administration:

- **Fully fund and reauthorize career and technical education programs** like the Strengthening Career and Technical Education for the 21st Century Act (Perkins V), the NSF Advanced Technical Education program, and the Strengthening Community Colleges Training Grants program. Many colleges rely on federal funding from these programs to develop and continue to offer training in emerging technology fields like AI.

Promote a broader scope of AI education

Maintaining U.S. competitiveness in AI requires developing and sustaining the necessary workforce. It is even more imperative that the U.S. government is able to attract, recruit, and retain technical talent for the federal workforce. Among the numerous vehicles for getting top talent into government, scholarship-for-service programs remain a direct talent pipeline into government service. For example, the National Science Foundation's (NSF) CyberCorps scholarship-for-service program is largely considered a success due to its longevity and sustained congressional funding.

- **Support the creation of an AI scholarship-for-service program.** In 2024, the NSF released a report detailing the feasibility of and need for an AI scholarship-for-service program following the CHIPS and Science Act. The NSF's AI Research Institutes offer a promising place to cultivate a potential AI scholarship-for-service program because of the institutes' specific focus on AI applications for a variety of fields and their existing relationships with the federal government. There are 23 institutes across the country with the NSF designation, nine have an active CyberCorps program, and 12 are designated as National Centers of Academic Excellence in Cyber (NCAE).

AI literacy typically dominates education policy conversations. Educators, school systems, and departments of education have mobilized to adapt and respond to the emergence of AI within the education system. However, focusing solely on AI literacy efforts within the classroom excludes many segments of the American citizenry. AI literacy can support citizens by making them aware of AI and its limitations, de-mystifying fears or concerns, and helping individuals take ownership of their creative and original work, thoughts, and ideas.

- **Work with Congress to support AI literacy efforts for the American people.** In 2024, Senators Kelly and Rounds introduced a bill aimed at bolstering consumer awareness and confidence in the use of AI products and services. A companion bill was later introduced by Representative Blunt Rochester. Together, these bills go beyond the classroom and seek to provide American citizens with the necessary education and information to make informed decisions about their AI use and consumption.

Competing with China

Stop illicit technology transfer to the PRC

China's comparative strength, now and historically, has been in commercializing scientific advances made outside China. Although the Chinese diaspora and global data pipelines have obscured China's true capabilities, Chinese scientists still acknowledge their dependency on foreign basic science. This is where China's legal and illegal technology transfer programs increasingly focus. While recognizing China's new capacity for indigenous research, the country continues to reap major advantages from what it acquires through theft, misappropriation, and other one-sided practices. This is especially true in AI, where products are digital and easier to acquire surreptitiously.

- **Create an office or task force within ODNI to track technology transfers to China**. U.S. government efforts to address this problem through research security initiatives (the NSF-backed SECURE project) and criminalizing participation in China's talent programs move the ball forward. Unfortunately, exposed venues and practices are replaced by novel acquisition stratagems. CSET analysts have a suite of practical recommendations to mitigate China's excesses in this arena, but what is lacking is a focal point where monitoring and deliberations can take place within the U.S. government.

Assess and monitor export controls for effectiveness

Export controls are an important economic statecraft tool designed to impose a delay and costs on China and other competitors' technological ambitions, particularly in semiconductor technology and AI. Given the rapidly developing capabilities of AI hardware and both open- and closed-weight AI models, it is critical to ensure that export control policies are adjusted accordingly.

- **The Bureau of Industry and Security (BIS) in the Department of Commerce should institute scenario planning assessments** before implementing new export controls and rigorously monitor the effectiveness of current export control policies.

- **Scenario planning assessments should be clearly articulated.** They should include export control policy objectives, analyses and testing of underlying assumptions, assessments of economic impact on U.S. and allied firms, evaluations of potential Chinese countermeasures and adaptations, and considerations of near- and long-term consequences.

- **BIS should also conduct regular post-implementation assessments** that track progress toward stated control objectives, second-order effects, impact on China's semiconductor manufacturing equipment industry, developments in China's semiconductor fabrication capabilities, and advancements in China's AI sector.

Cooperate with allies and partners to ensure controls remain effective

BIS should continue closely working with allies on a joint export control strategy and improve communication and information sharing about why the controls are needed to protect common interests.

- **Clearly articulate and justify the objectives of the export controls** to allies in order for the broader U.S. export control strategy to work.
- **Avoid overuse of the Foreign Direct Product Rule (FDPR)** to expand the reach of U.S. export controls. Increasing use risks incentivizing foreign companies to design without U.S. technology and components, undermining multilateral efforts and undercutting U.S. strategy in the long-term.

Improving the AI Information Environment

Leverage open source intelligence to avoid technological surprise

In stark contrast to most of the major technological advances of recent decades, AI is being developed, deployed, and used almost entirely outside of the federal government, and to a significant extent in nations outside the United States altogether. Among other impacts, this puts the U.S. government at an inherent informational disadvantage. U.S. policymakers can only exploit AI's potential economic, strategic, and innovation benefits, avoid the risks it poses, and ensure American AI leadership if they have reliable information about the current state of the technology and where it is headed in the coming years. Unfortunately, there is currently no office or agency, inside or outside government, that is able to provide this comprehensive view of the AI landscape. We therefore recommend the following:

- **Significantly expand open-source intelligence (OSINT) gathering and analysis on AI.** This work is particularly neglected in the intelligence community, which remains focused on classified sources. It is critically underdeveloped and under-resourced elsewhere in the federal government. Significant investments are needed in collection, interpretation, and dissemination of AI OSINT, incorporating sources like research publications, supply chain data, market research, patents, capital markets data, and workforce data.
- **Make China's AI ecosystem a special focus for this AI OSINT program.** The lack of a serious program to track China's AI progress undermines federal efforts across policy domains such as export controls, trade policy, research security and industrial policy, and

raises the risk of technological surprise. China itself runs a 100,000-person open source-based monitoring system directed at U.S. military and civilian R&D, fueling its own national development in AI and other critical technologies. The federal government should significantly ramp up efforts to monitor China's AI ecosystem, including the Chinese government itself (at all relevant levels and organizations), related actors such as state-owned enterprises, state research labs, and state-sponsored technology investment funds, and other actors, such as universities and tech companies.

Share intelligence across the government and private sector

Critical information about frontier AI capabilities is siloed in AI companies. Transparency in companies' development practices is necessary for the U.S. government to respond to rapid AI developments and anticipate emerging threats to national security. The U.S. government should also enhance its horizon-scanning abilities by tapping into information from its allies about significant AI developments. Conversely, the U.S. government collects intelligence that could help companies harden their defenses against attacks by lone or state actors. We recommend that the Trump administration exchange critical information about AI capabilities with allied countries and AI companies and remove barriers to participating in such information sharing.

- **Establish reporting programs to gather information on AI development processes from AI companies**. Reporting programs could ask companies to provide detailed documentation on training procedures and environments, unexpected or concerning capabilities found in new models, model specifications (also known as constitutions) that define the behaviors that companies want AI models to have, and evaluations. Detailed documentation on AI development practices would decrease the information gap between AI developers and the government, enabling the government to quickly respond to sudden jumps in AI system capabilities.

- **Partner with companies to share threat intelligence.** The U.S. government should partner with AI companies to share suspicious patterns of user behavior and other types of threat intelligence. In particular, the Intelligence Community and the Department of Homeland Security should partner with AI companies to share cyber threat intelligence, and the Department of Homeland Security should partner with AI companies to prepare for potential emergencies caused by malicious use or loss of control over AI systems. In addition, the Department of Commerce should receive, triage, and distribute reports on CBRN and cyber capabilities of frontier AI models to support classified evaluations of novel AI-enabled threats, building on a 2024 Memorandum of Understanding between the Departments of Energy and Commerce.

- **Contribute to and draw from the collective intelligence of U.S. allies regarding AI capabilities.** The impacts of AI systems transcend national borders, and observations about

AI developments in allied countries may also be relevant at home. The U.S. government should draw on information about frontier AI capabilities from trusted allies like the United Kingdom and its AI Security Institute and share information with them to maintain trust.

Encourage reporting of AI incidents to facilitate technology adoption

The growing deployment of AI systems by the public and private sector has inevitably led to a growing number of failures and harmful incidents involving AI. If the U.S. government continues not to track such AI incidents, it will miss a critical opportunity to boost AI innovation and adoption. AI incident reporting and analysis accelerates learning about AI failures, which surfaces where AI research is most needed and helps developers innovate and improve their models. By preventing repeated failures and enhancing the reliability of AI systems, incident reporting not only reduces the risk of harm to the American public, but also helps to build consumer and user trust in the technology. This promotes widespread AI adoption and, consequently, the realization of economic benefits of AI. We recommend the U.S. government:

- **Implement a mandatory AI incident reporting regime for sensitive applications across federal agencies.** Federal agencies deploy AI systems for a wide range of safety- and rights-impacting use cases, such as using AI to deliver government services or predict criminal recidivism. AI failures, malfunctions, and other incidents in these contexts should be tracked and investigated to determine their root cause, inform risk management practices, and reduce the risk of recurrence. When AI systems are acquired from third parties, vendors should be required to report AI incidents to the agency within 24 hours of detection.
- **Direct agencies overseeing high-risk domains to implement hybrid incident reporting schemes in their respective industries.** High-risk domains include, but aren't limited to, healthcare, transportation, education, employment, finance, housing, insurance, utilities, and critical infrastructure. Criteria for what constitutes an AI incident or malfunction should be determined by each agency. Federal agencies should be authorized to investigate such incidents in order to identify causes, commonalities, and emerging trends, and disseminate lessons learned and updated AI risk management recommendations.

Mitigating Risks from AI

Protect the public from harm caused by AI

Risks from AI systems threaten to undermine U.S. AI policy objectives. If citizens do not have assurances that they can seek remedy from AI harms or that AI systems work properly, then they could be disinclined to adopt AI systems, to the detriment of policy objectives related to realizing AI's benefits.

- **Create standard pathways to contest AI results**. U.S. citizens who are materially impacted by an AI-assisted decision will not trust AI systems if they do not have an efficient and accessible way to contest erroneous decisions. Furthermore, reporting by affected persons is an effective channel for identifying AI mistakes, which is a necessary precursor to remedying these mistakes.

- **Establish whistleblower protections for employees who report dangerous conduct by AI companies.** Whistleblower protections can shield employees from company retaliation and help ensure that AI companies abide by their commitments and the law. The Trump administration should establish a secure line for employees to report problematic company practices, such as failure to report system capabilities that threaten national security.

Protect against AI-enabled biological risks

There are concerns that AI could exacerbate biological risks, for example by making it easier for a non-expert to produce a biological weapon or by enabling the creation of more severe or targeted pathogens and toxins. We recommend the following steps to guard against such outcomes:

- **Build an integrated biosecurity ecosystem.** While a malicious actor could use AI in their plans to cause biological harm, the foundational information and resources necessary are available without AI. A biosecurity strategy focused purely on controlling AI use will fail without defending against both AI-enhanced and AI-agnostic biological agents. Rather, mechanisms targeting AI use should be integrated into broader biosecurity strategies and viewed as just one tool in a more comprehensive governance toolkit.

- **Deploy appropriate model safeguards.** Model safeguards can be deployed to address safety concerns and target various nodes in the AI lifecycle. In a CSET report, we identified potential model governance mechanisms: biosecurity training for developers, training data filtration, access restrictions to certain datasets and computing infrastructure, pre-release assessments, model access controls, usage monitoring, and harm reporting mechanisms.

- **Require future policies to specify whether they apply to models solely trained on biological or chemical data**, particularly when using terms like "foundation model" or "large language model." Some terms relate to both general-purpose and chem-bio AI models (models that can aid in the analysis, prediction, or generation of novel chemical or biological sequences, structures, or functions), but are defined differently depending on the context, creating confusion when they are referenced. This is a particular challenge for existing regulatory and guidance documents, the majority of which are vague about whether they include chem-bio AI models.

- **Define capabilities of concern and support the creation of threat profiles for different types of AI models.** Assessing whether an AI model can output potentially risky biological

information and quantifying that risk on a spectrum is challenging because many pathogens evolve over time and are dangerous in some conditions but harmless in others. Similarly, distinct combinations of users and AI tools impact the potential for harm and the most effective likely policy solutions for evaluation strategies and relevant mitigation measures. A coalition of government agencies should develop frameworks that clearly define risky capabilities, including chem-bio capabilities of concern, so evaluators know what risks to test for. These frameworks could draw upon Appendix D of the National Institute of Standards and Technology's (NIST) draft Managing Misuse Risk for Dual-Use Foundation Models. In addition, government agencies should build threat profiles that consider different combinations of users, AI tools, and intended outcomes, and design targeted policy solutions for these highly variable scenarios.

Advancing AI Evaluation Science and Standards

Advance AI evaluation science to understand model capabilities

Evaluations should inform decision-making about the safety and suitability of AI systems for certain applications and whether AI systems should be used at all. The results of evaluations can provide insight into AI system capabilities and the effectiveness of interventions like implementing technical guardrails or upskilling AI users that shape the future development or adoption of AI. Given the increasing importance of AI systems in many policy domains, it is crucial that AI evaluations are rigorous and reliable and that decision-makers understand how to interpret their results. We recommend that the AI Action Plan deploy evaluations as a foundational tool for driving AI progress.

- **Federal grantmaking bodies such as the National Science Foundation should support basic research related to improving AI evaluation science overall and for "agentic" systems specifically.** Agentic AI systems are capable of independently pursuing complex goals in complex environments and are expected to become increasingly capable in the near future. Basic research into the evaluation of AI agent capabilities and technical mechanisms to control and govern the behavior of AI agents is essential to improve their performance over time.

- **The U.S. AI Safety Institute (AISI) should work with other stakeholders to advance AI evaluation science and de-duplicate efforts to evaluate frontier AI models.** AISI has so far worked effectively with industry, academia, and other partners to improve AI model evaluations. The Trump administration should empower AISI to develop quantitative benchmarks for AI, including benchmarks that test a model's resistance to jailbreaks, usefulness for making CBRN weapons, and capacity for deception. The recent announcement of AISI's collaboration with Scale AI demonstrates how the U.S. government can work with third parties to develop frontier AI evaluations and make efficient use of testing

infrastructure. This type of partnership also gives the U.S. government access to highly capable models that it otherwise would not have, allowing it to build new evaluations of frontier AI models while limiting redundant efforts.

- **Government procurement bodies should implement testing requirements for vendors of AI systems.** Examples of requirements could include listing the benchmarks used to measure model performance, reporting the results of red-teaming exercises intended to discover vulnerabilities or failures in specific contexts, or conducting pre-procurement user testing with government employees. Government procurement bodies may also establish minimum performance requirements for vendors' AI systems based on reported evaluations. Having a clear measure of the risk associated with deploying a given model would enable the U.S. government to effectively harness AI systems and avoid AI failures.

- **Facilitate knowledge and expertise exchange about AI evaluations and risks across the federal government.** Beyond AISI, other parts of the U.S. government are well-situated to conduct evaluations of national security risks posed by AI systems. The National Security Agency has expertise in offensive cyber threat risks and the Department of Energy is well-equipped to test for nuclear and radiological risks. These agencies could build AI evaluation infrastructure that complements other testing infrastructure provided through partnerships such as the one between AISI and Scale AI. The AI Action Plan should encourage these agencies to leverage the tools and knowledge they already have—which may not be readily found outside of government—to ensure that AI systems work as intended.

Develop, adopt, and synchronize standards

Standards can promote smooth market function, interoperability, and consumer safety. However, establishing standards for AI is challenging because of the technology's rapid evolution and explosion of potential use cases across sectors and applications. While many frontier AI companies have published frameworks for managing risks posed by their models, these frameworks often lack sufficient detail about risk mitigation efforts and are prone to being weakened or abandoned altogether in the face of competitive pressure. In order to effectively mitigate the risks associated with AI systems, the U.S. government should develop AI standards in coordination with other stakeholders and model how standards adoption can facilitate the use of AI for mission success.

- **Develop and adopt standards to mitigate risks from AI.** In coordination with stakeholders in academia, civil society, and industry, AISI should develop standards that cover topics including model training, pre-release internal and external security testing, cybersecurity practices, if-then commitments, AI risk assessments, and processes for testing and re-testing systems as they change over time. Standards on when to conduct different types of evaluations, what the best practices are for each, and how model evaluations are reported should be developed to allow fair comparison between models. The U.S. government can

model how to effectively harness AI systems by adopting these standards, incorporating them into procurement requirements, and sharing lessons learned from adoption.

- **Synchronize baseline AI standards across federal agencies**. Companies providing AI tools, or those including AI tools as a part of software solutions, must currently navigate multiple overlapping requirements or standards when selling products to different agencies of the government. For example, NIST's AI Risk Management Framework identifies characteristics of trustworthy AI that differ from the Department of Defense guidance. Within the Department of Defense (DOD), different military services have different generative AI policies. If all federal agencies agree to abide by a unified set of minimum AI standards for purposes of acquisition and deployment, this would greatly reduce the burden on companies offering AI solutions, accelerate the adoption of standard tools and metrics, and reduce inefficiencies caused by the need to repeatedly draft and respond to similar but different requirements in government contracts. By unifying the U.S. government behind common standards for industry contracts, the U.S. government could also drive the adoption of international AI standards favorable to U.S. businesses.

- **Reduce the duplication of effort across military services by empowering the Office of the Secretary of Defense (OSD) to set AI security standards and expand authorization-to-operate (ATO) reciprocity.** OSD has not empowered a DOD-wide entity to set AI policies for the services. This results in duplication of efforts across the military services, with multiple memos guiding efforts across the DOD in different ways. For example, within each service, different commands have different network ATO standards, which require substantial rework by the government and AI vendors to satisfy before deployment. Continuous ATOs and ATO reciprocity must be enforced across OSD and an entity should be empowered to synchronize policies, rapidly certify reliable AI solutions, and act to stop emerging security issues. Where the DOD establishes standards and policies, these should be shared with other government agencies and state agencies to further synchronize standards and accelerate responsible adoption.

Acknowledgments

This response is the product of contributions from all of CSET's staff and builds substantially on previous CSET publications and data insights. We are especially grateful to Mia Hoffmann, Jack Karsten, and Mina Narayanan for their project leadership, and to Owen Daniels, Igor Mikolic-Torreira, and Dewey Murdick for their comprehensive reviews. Special thanks also go to Catherine Aiken, Zachary Arnold, Kendrea Beers, Jack Corrigan, Kathleen Curlee, Jacob Feldgoise, Rebecca Gelles, William Hannas, Jessica Ji, Kyle Miller, Adrian Thinnyun, and Vikram Venkatram for their valuable consolidation of team inputs.