March 3, 2022

RFI Response: National Artificial Intelligence Research and Development Strategic Plan—White House Office of Science and Technology Policy
87 FR 5876; Document Number 2022-02161

The Center for Security and Emerging Technology (CSET) offers the following submission for consideration by the Office of Science and Technology Policy; the NSTC Select Committee on Artificial Intelligence; the NSTC Machine Learning and AI Subcommittee; the National AI Initiative Office; and the NITRD National Coordination Office. OSTP, on behalf of its partners, requested input on updating the National Artificial Intelligence Research and Development Strategic Plan.

In this submission, CSET makes 15 recommendations to improve upon the National Artificial Intelligence Research and Development Strategic Plan (Strategic Plan). We recognize that most of our suggested changes would require additional resources to implement. As such, CSET generally recommends that federal agencies account for any new efforts in their budgets, and reallocate resources or request additional appropriations as needed. To ensure these new efforts come to fruition, OSTP must also work with the Office of Management and Budget to provide agencies with the necessary appropriations.

Our recommendations are as follows:

Strategy 1: Make long-term investments in AI research

Recommendation: Adopt caution when pursuing long-term AI research that could generate "general purpose" or "human-like" artificial intelligence.

> In its current state, the Strategic Plan calls for federal research into "general purpose artificial intelligence" and "human-like" AI. As the plan notes, we are likely still decades away from developing these human-level systems. However, other AI research goals articulated within the Strategic Plan—such as the need to ensure appropriate goal specification and alignment of AI systems—should be treated as strong prerequisites before attempting any project that has a realistic chance of producing a highly generalizable, human-like AI system. Ensuring that machine learning systems behave in accordance with their designers' intentions is a challenge that will grow increasingly critical and complex as AI systems are deployed in higher-states and more dynamic settings.[1] While it is sensible to pursue AI that can be more generalizable across specific domains, we must approach the goal of a "general-purpose" AI with caution. The Strategic Plan should not commit to pursuing technologies that may not ultimately serve

[1] Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: Specification in Machine Learning" (Center for Security and Emerging Technology, December 2021). https://doi.org/10.51593/20210031; Stuart J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking Press, 2019).

humanity's best interests, or that may prove to be difficult to usefully and reliably deploy once developed.

Additionally, we recommend editing the Strategic Plan's discussion of general purpose systems to reflect the recent development of systems such as large language models, which are not narrowly targeted to one application. These systems still fall far below what is usually meant by "artificial general intelligence," but they nonetheless are "general purpose" in the sense that they can be used across a wide range of application and topic areas. Any aspects of the Strategic Plan that depend on a clear division between "narrow" and "general purpose" systems should be reconsidered to reflect this development."

Strategy 2: Develop effective methods for human-AI collaboration

Recommendation: Promote the development of tools for "trust calibration" to enable safe and effective human-AI collaboration.

For humans to safely and effectively use AI, it is critical that they understand the specific strengths and limitations of these systems. "Trust calibration" tools allow users to understand how much they should or should not trust in a given AI system under different circumstances. Enabling users to calibrate their trust in AI systems is an important component of human-AI collaboration that is distinct from—but complementary to—building systems that are more reliable, secure, and interpretable. Trust calibration tools are particularly important in the military context, because without understanding the limitations of the autonomous and AI-enabled systems developed by the defense community, it is impossible to ensure these technologies are employed in safe, secure, effective, and ethical ways.[2]

Strategy 3: Understand and address the ethical, legal, and societal implications of AI

Recommendation: NSF and DARPA should fund privacy-preserving computer vision research as an alternative to automated mass facial surveillance.

Today, China relies on widely deployed facial recognition systems to repress broad swaths of its population, and through the export of this technology, it has enabled other authoritarian governments to construct sweeping surveillance programs of their own. To protect human rights and promote AI development that centers around democratic values, the United States and its allies can develop and promote the spread of privacy-preserving computer vision systems.[3] These technical methods—such as differential privacy and homomorphic encryption—would serve as the basis for technologies that would support

[2] Margarita Konaev, Tina Huang and Husanjot Chahal, "Trusted Partners: Human-Machine Teaming and the Future of Military AI," (Center for Security and Emerging Technology: February 2021). DOI: 10.51593/20200024.

[3] Dahlia Peterson, "Designing Alternatives to China's Repressive Surveillance State," (Center for Security and Emerging Technology, October 2020). DOI: 10.51593/20200016; Andrew Imbrie, Ryan Fedasiuk, Catherine Aiken, Tarun Chhabra and Husanjot Chahal, "Agile Alliances: How the United States and Its Allies Can Deliver a Democratic Way of AI" (Center for Security and Emerging Technology, February 2020). https://doi.org/10.51593/20190037.

law enforcement without violating the privacy and civil liberties of those being surveilled, thus providing a viable alternative to the mass automated surveillance systems developed by Chinese companies and the Chinese state.[4]

Strategy 4: Ensure the safety and security of AI systems

Recommendation: Promote research into attacks on AI systems that are more likely to resemble real-world threat scenarios.

Currently, AI security is a small and relatively neglected area of AI research, with some estimates suggesting that less than 1 percent of AI research is dedicated to security topics.[5] As it stands, most AI security research appears to be dedicated to the study of adversarial examples (i.e. injecting inputs into machine learning models that purposely cause them to make mistakes). While some of these efforts attempt to explore the vulnerability of real-world AI systems, much research assumes idealized conditions in which attackers have full access to models. Research on data poisoning similarly tends to focus on idealized attack situations, such as when attackers can easily manipulate all inputs to a model's training data. More research should be done to explore how vulnerable AI systems are to disruption under the less-than-ideal attack conditions that real-world adversaries are likely to face, and into methods for securing AI models from less mathematically sophisticated forms of attack.

Additionally, instead of studying the vulnerabilities of "AI systems" broadly, the federal government should support the development of shared standards for evaluating the impact and severity of a given system's vulnerabilities in different circumstances. These standards may resemble the Common Vulnerability Scoring System used in the field of cybersecurity.[6] Policymakers should also encourage collaboration between cybersecurity professionals and AI experts to promote a better understanding of how information security risks extend to the AI domain.[7]

Recommendation: Dedicate resources to studying and mitigating software vulnerabilities in the AI supply chain.

The AI industry is built upon shared software, shared data, and shared models. However, we know little about how vulnerabilities in one layer of the AI supply chain might propagate to further layers.

[4] Tim Hwang, "Shaping the Terrain of AI Competition" (Center for Security and Emerging Technology, June 2020), cset.georgetown.edu/research/shaping-the-terrain-of-ai-competition/ https://doi.org/10.51593/20190029

[5] Helen Toner and Ashwin Acharya, "Exploring Clusters of Research in Three Areas of AI Safety" (Center for Security and Emerging Technology, February 2022). https://doi.org/10.51593/20210026

[6] "Common Vulnerability Scoring System SIG," *Forum of Incident Response and Security Teams*, 2021, https://www.first.org/cvss/.

[7] Jonathan Spring, "Comments on NIST IR 8269: A Taxonomy and Terminology of Adversarial Machine Learning," *Carnegie Mellon University*, February 13, 2020, https://insights.sei.cmu.edu/blog/comments-on-nist-ir-8269-a-taxonomy-and-terminology-of-adversarial-machine-learning/.

Today, many segments of the AI supply chain rely on a few common software or code resources, such as libraries like TensorFlow and PyTorch, hosting providers like HuggingFace, pre-trained models like BERT and ResNet, or shared datasets.[8] More work is required to understand the vulnerabilities introduced by these shared dependencies. Oftentimes, a decision that is made at one level of the supply chain for the sake of improving efficiency and overall performance ends up introducing a vulnerability in a different, more adversarial operational context. Designers at one level may not understand the potential vulnerabilities they are importing from previous layers of the technical stack.[9] Further research into this area can help practitioners quantify the vulnerabilities of these shared resources; increase visibility into the potential flaws of models, datasets, or code in different operational contexts; and determine whether or not a "software bill of materials" approach for AI models may be appropriate. The government should also consider using competitions and "red teaming" to identify vulnerabilities in their AI systems; funding efforts to secure open source software, datasets, and other public resources; and promoting better risk management practices.[10]

Recommendation: Research and develop AI specific standards and processes for assessing AI maturity.

Maturity standards and processes enable consistent discussions and comparison across different AI systems. With new development tools and increased investment, organizations are now able to create and deploy AI systems faster than ever before. However, speedy deployment—and in fact, deployment itself—does not mean that an AI is mature. The maturity of an AI system can vary depending upon its mathematical complexity, the type of data it uses or produces, the context in which it is used, and the quality of its training processes. With better assessments of AI maturity, practitioners could more effectively determine when a given system is ready for deployment Additionally, these standards could prevent organizations from using immature systems, which are more prone to vulnerabilities and accidents.

Strategy 5: Develop shared public datasets and environments for AI testing and training

Recommendation: Pursue cybersecurity-relevant datasets and testbeds as a special area of focus.

AI tools have a wide range of potential applications in the cybersecurity industry, including intrusion detection, vulnerability discovery, and attack response.[11] However, the

[8] Micah Musser, "Managing the Security of AI Models Across the ML Pipeline," The AI Summit New York, December 9, 2021, https://docs.google.com/presentation/d/1JmTcBDUExXmBtV41GloHHDY8frkHG3z1qKdmvfFK6LA/edit#slide=id.g659fabf128_0_2.

[9] Andrew Lohn, "Poison in the Well: Securing the Shared Resources of Machine Learning" (Center for Security and Emerging Technology, June 2021). https://doi.org/10.51593/2020CA013

[10] Andrew Lohn, "Poison in the Well: Securing the Shared Resources of Machine Learning" (Center for Security and Emerging Technology, June 2021). https://doi.org/10.51593/2020CA013

[11] Micah Musser and Ashton Garriott, "Machine Learning and Cybersecurity: Hype and Reality" (Center for Security and Emerging Technology: June 2021). https://doi.org/10.51593/2020CA004

training data that could support the development of such tools is rarely shared, and there are few common benchmarks for evaluating the performance of AI models in the cyber domain.

The federal government historically played a large role in stimulating research in machine learning-based methods of intrusion detection. In 1998 and 1999, DARPA simulated several large-scale datasets of network data that included various attack signatures; these datasets spurred significant research in the area of intrusion detection and helped lead towards the machine learning methods that many companies use today to detect attacks. However, a lack of other public datasets or benchmarks has meant that, as late as the late 2010s, a substantial majority of research into intrusion detection systems was still using these badly outdated datasets to test their models.[12]

Private industry is unlikely to be motivated to ever publicly release meaningful datasets for fear of inadvertently losing their competitive edge and revealing details about their own networks. It is also possible that large cybersecurity companies now have sufficiently detailed in-house datasets that public datasets would not meaningfully shift the needle for intrusion detection tasks specifically, though this is still an area that should be explored more, along with the potential for public-private partnerships. Nonetheless, simulated cyber environments could be useful for developing new capabilities, especially if gym-like environments could be constructed that could allow for more extensive testing of reinforcement learning approaches in realistic cyber domains. This approach is one that has also been identified as a major research area for China, where PengCheng Laboratories is currently attempting to build one of the largest supercomputers in the world specifically for the purpose of developing a realistic cyber range.[13]

Recommendation: Increase the interoperability of existing data resources.

The AI community would benefit significantly from the curation and publication of government datasets.[14] While the government already maintains and releases data that is critical for research—especially Census and financial data—these datasets are often not highly interoperable. For instance, users who wish to access the vast troves of data collected by the Bureau of Labor Statistics are often required to download and manually merge a vast array of disparate spreadsheets. Other government entities publish data in PDFs or other formats that make analysis difficult. Such idiosyncrasies complicate any analysis project, and they render large-scale data analysis that undergirds much modern machine learning nearly impossible.

[12] Hanan Hindy, David Brosset, Ethan Bayne, Amar Seeam, Christos Tachtatzis, Robert Atkinson, Xavier Bellekens, "A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems," (IEEE, 2020), https://doi.org/10.1109/ACCESS.2020.3000179
[13] Dakota Cary, "Down Range," (Center for Security and Emerging Technology, forthcoming)
[14] We recognize that the process of curating and integrating federal datasets would be expensive. Additional appropriations would likely be required to support this effort.

The government is taking some steps to remedy this problem, for instance, through the Census Bureau's Statistical Data Modernization project.[15] However, other agencies that produce publicly available data should undertake similar efforts. A more interoperable data environment is one that is likely to spur more AI researchers to develop useful insights out of government data.

Strategy 6: Measure and evaluate AI technologies through standards and benchmarks

Recommendation: Encourage the use of operationally relevant metrics and the evaluation of AI in the operational context and condition in which it will be used.

Testing and evaluating an AI in the conditions it will be used provides a better understanding of performance, utility and potential for harm. AI systems often fail because the conditions in which the systems are developed and tested are different from those in which they are used. The evaluation of AIs is often done in clean 'lab-like' conditions and with assumptions about who the user is and how they will interact with the AI. Upon deployment and use in real-world conditions, performance issues are discovered and harm is done.

Testing should also use metrics that are operationally relevant to AI's use context.[16] Frequently AI performance is evaluated solely using a handful of typical metrics (accuracy, precision, etc.). While these metrics are useful for development, they do not provide end-to-end contextual information. For instance, for AIs developed to improve maintenance, metrics associated with repair times and overall system availability will be more informative than how accurately a maintenance action is predicted.

Recommendation: Emphasize characterizing performance across use conditions.

Compared to aggregate metrics (calculations that summarize performance across all conditions), characterizing performance provides information on what conditions an AI does and does not operate well. The results of aggregate metrics like mean, standard deviation, and accuracy are dominated by typical values. This makes them useful for comparing different AIs, but not useful for identifying the atypical conditions in which performance is inadequate. By characterizing performance in relationship to use conditions, an AI system can be deployed with constraints that limit it working in conditions where its performance is degraded or it is more likely to do harm.

Recommendation: The United States should designate developing global facial recognition standards a new priority on its AI standards list, and incentivize U.S. companies' participation in standards bodies.

[15] "Evolving to Meet 21st Century Data Needs," *U.S. Census Bureau*, January 11, 2021, https://www.census.gov/about/what/transformation.html.

[16] We recognize that the process of creating such metrics would be expensive. Additional appropriations would likely be required to support this effort.

To date, China is the only country that has proposed facial recognition standards to the United Nations International Telecommunication Union.[17] These standards—which are often adopted by developing nations in Africa, Asia, and the Middle East—are problematic because they go beyond technical specifications to propose policy recommendations for how and where the technology should be deployed, which allows the technology to be deployed for politically repressive uses. However, no other country has proposed viable alternative standards.[18] The United States should incentivize companies to help create standards for facial recognition, as doing so is often prohibitively expensive. (Work and travel can cost $300,000 per engineer annually.)[19] Twelve of the 15 industry groups that responded to a recent NIST RFI recommended that the U.S. government incentivize companies' participation through grant funding, potentially via industry associations, and revise the R&D tax credit to include standards development work.[20]

Strategy 7: Better understand the national AI R&D workforce needs

Recommendation: Define the AI R&D workforce, in addition to computer research scientists, and compile and publish data on the composition of this workforce.

While the current strategic plan correctly notes that the AI R&D workforce spans multiple disciplines, it does not clearly define who makes up this workforce. Not only is AI R&D conducted on multi-disciplinary teams, but translating that research into applications and responsible use requires an even wider range of talent.[21] Defining and including all these types of talent in the strategic plan will galvanize more federal investment in AI-related workforce development initiatives. Given the dynamic nature of the AI field, this definition should be updated periodically to reflect changes in the AI R&D workforce.

Defining and publishing data on the AI R&D workforce will also facilitate greater diversity in the field by elevating other critical AI-related careers. Data on the demographic composition of a defined AI R&D workforce would highlight the extent and nature of representative gaps, and provide policymakers with measurable objectives for

[17] Dahlia Peterson, "Designing Alternatives to China's Repressive Surveillance State," Center for Security and Emerging Technology, October 2020. https://doi.org/10.51593/20200016.

[18] Meng Jing, "Chinese Tech Companies Are Shaping UN Facial Recognition Standards, according to Leaked Documents," South China Morning Post, December 2, 2019, https://www.scmp.com/tech/policy/article/3040164/chinese-tech-companies-are-shaping-un-facial-recognition-standards.

[19] Jeanne Whalen, "Government Should Take Bigger Role in Promoting U.S. Technology or Risk Losing Ground to China, Commission Says," *The Washington Post*, December 1, 2020, https://www.washingtonpost.com/technology/2020/12/01/us-policy-china-technology/.

[20] Jacob Feldgoise and Matt Sheehan, "How U.S. Businesses View China's Growing Influence in Tech Standards," *Carnegie Endowment for International Peace*, December 23, 2021, https://www.washingtonpost.com/technology/2020/12/01/us-policy-china-technology/; "Study on Chinese Policies and Influence in the Development of International Standards for Emerging Technologies," *National Institute for Standards and Technology*, https://www.regulations.gov/docket/NIST-2021-0006/comments.

[21] Diana Gehlhaus and Santiago Mutis, "The U.S. AI Workforce: Understanding the Supply of AI Talent" (Center for Security and Emerging Technology, January 2021). https://doi.org/10.51593/20200068

targeting investment. Such demographic data could be ascertained from several sources: (1) the American Community Survey (U.S. Census Bureau) for defined occupations; (2) the Survey of Earned Doctorates and Survey of Doctoral Recipients (NSF) for selected fields; (3) annual reports from the NSF Graduate Research Fellowships Program; (4) the Integrated Postsecondary Education Data System (IPEDS) database for selected degrees, awards, and fields of study (NCES), and (5) in coordination with R1 research institutions receiving federal grants, subsetting those grants for AI and AI-related R&D using a predetermined definition.

Recommendation: Identify and remedy inefficiencies in the immigration process that prevent foreign-born AI experts from obtaining residency in the United States.

Today, the United States relies heavily on foreign-born talent to bolster its AI workforce—roughly half of the AI experts in academia and industry were born outside of the United States.[22] Many of these individuals initially enter the country as students and then choose to remain after graduation, founding companies and making valuable contributions to the economy and society at large.[23] However, stay rates among international students would likely be higher if not for certain restrictions in the U.S. immigration system. For example, numerical caps on green cards force individuals from certain countries (namely China and India) onto decades-long waitlists.[24] Policymakers have started taking steps to address some of these roadblocks. For example, the America COMPETES Act, recently approved by the House, would create a new visa category for foreign-born entrepreneurs and lift green card limits for foreign-born STEM PhDs (though it remains unclear whether this measure will be enacted into law).[25] A coordinated effort to identify chokepoints and obstacles in the immigration system that limit the foriegn-born AI experts from moving to the United States and bolstering its AI R&D ecosystem would have long-term benefits for the AI talent pipeline.

Strategy 8: Expand Public–Private Partnerships to accelerate advances in AI

Recommendation: Locate gaps in private sector's AI R&D agenda and forge public-private partnerships to target these areas.

The lion's share of U.S. AI R&D is conducted in the private sector, but few of these companies focus explicitly on the national security and defense applications of AI.[26] A number of defense agencies, such as the Defense Innovation Unit (DIU), have already

[22] Remco Zwetsloot, James Dunham, Zachary Arnold and Tina Huang, "Keeping Top AI Talent in the United States: Findings and Policy Options for International Graduate Student Retention" (Center for Security and Emerging Technology, December 2019). https://doi.org/10.51593/20190007.

[23] Tina Huang, Zachary Arnold and Remco Zwetsloot, "Most of America's 'Most Promising' AI Startups Have Immigrant Founders" (Center for Security and Emerging Technology, October 2020). https://doi.org/10.51593/20200065.

[24] David J. Bier, "Employment-Based Green Card Backlog Hits 1.2 Million in 2020," *Cato Institute*, November 20, 2020, https://www.cato.org/blog/employment-based-green-card-backlog-hits-12-million-2020.

[25] Jackson Lewis, "Bill Passed by House Benefits Immigrants in STEM Fields, Entrepreneurs in Start-Ups," *JDSupra*, February 9, 2022, https://www.jdsupra.com/legalnews/bill-passed-by-house-benefits-8093761/.

[26] Zachary Arnold, Ilya Rahkovsky, Tina Huang, "Tracking AI Investment: Initial Findings from the Private Markets" (Center for Security and Emerging Technology, September 2020). https://doi.org/10.51593/20190011.

created public-private partnerships to fill this gap, but these programs operate in a piecemeal fashion and remain largely disconnected from the broader defense procurement pipeline.[27] More targeted and coordinated efforts across the Defense Department and national security community are required to integrate more AI capabilities in major military platforms and systems. Understanding which private companies lead in defense-relevant AI activities, what real-world problems their technologies may address, and what barriers they face to working with the federal government will lead to better decisions about the development of AI applications that align with market realities and government priorities.

NEW—Strategy 9: Increase transparency in the federal AI R&D ecosystem

Recommendation: Make increasing transparency in the U.S. federal AI R&D ecosystem a new strategic priority.

Today, there is little comprehensive public data on federal AI R&D activities. The data that does exist is related to all federal R&D at select R1 research institutions and is restricted use access.[28] CSET proposes creating a new strategy to increase transparency into this ecosystem. This strategy appears to be missing from the current Strategic Plan, but it is necessary for achieving each of the plan's overarching goals. Such a strategy would align with President Biden's "Memorandum on Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking," the final 2017 report from the Commission on Evidence-Based Policymaking, and in the spirit of former OSTP Director John Marburger's seminal essay "Wanted: Better Benchmarks."[29]

One way to increase transparency would be to create and maintain a dashboard on all federal AI R&D funding, including abstracts. This platform could be modeled on FastLane (NSF) or Federal RePORTER (NIH), both of which include abstract information. Having a publicly accessible database of federally-funded AI research grants would offer insights into the trajectory of existing investments and help guide future federal AI R&D policy. We envision a dashboard that would include information on grants (title, abstracts, awardee, award horizon, and award value), as well as information from grant recipients (personnel, expenditures, and other related AI R&D activities).[30] The database could be populated by funding agencies and departments, or by R1 research institutions, both of which would require new reporting requirements. With public aggregate data, researchers and funding agencies could track awards over time, create

[27] Melissa Flagg and Jack Corrigan, "Ending Innovation Tourism" (Center for Security and Emerging Technology, July 2021). https://doi.org/10.51593/20210030.

[28] Administered by the Institute for Research on Innovation and Science (IRIS). See for more: http://iris.isr.umich.edu/research-data/.

[29] "Memorandum on Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking," *White House*, January 27, 2021, https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/27/memorandum-on-restoring-trust-in-government-through-scientific-integrity-and-evidence-based-policymaking/; "The Promise of Evidence-Based Policymaking," *Commission on Evidence-Based Policymaking*, September 7, 2017, https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf; John H. Marburger III, "Wanted: Better Benchmarks," *Science*, May 20, 2005, https://www.science.org/doi/10.1126/science.1114801.

[30] We recognize this level of proprietary information may not be available in the immediate term.

descriptions of AI award types, and assess portfolio performance similar to that done by the NSF.

Confidential data on research spending on individuals and vendors from all research institutions should be hosted in a secure environment so that independent evaluations can be conducted on the economic, workforce, and social impact of AI spending, including understanding the equity and diversity effects of the spending on underrepresented minorities and institutions.