Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan: Responses

Microsoft

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government or any entity within the U.S. Government. We bear no responsibility for the accuracy, legality, or content of the responses and external links included in this document.
March 4, 2022

Dr. Alondra Nelson, OSTP Deputy Director of Science and Society and Performing the Duties of Director of OSTP  
Dr. Lynne Parker, Assistant Director of OSTP for Artificial Intelligence and Director of the National Artificial Intelligence Initiative Office  
Office of Science and Technology Policy  
The White House  
1600 Pennsylvania Ave NW  
Washington, DC 20500  
Via email to: AI-RFI@nitrd.gov

Dear Dr. Nelson and Dr. Parker,

Microsoft Response to OSTP Request for Information on the Update of the National Artificial Intelligence Research and Development Strategic Plan [Document Number: 2022-02161]

Microsoft appreciates the opportunity to respond to the Office of Science and Technology Policy’s (OSTP) Request for Information on the update of the National Artificial Intelligence Research and Development Strategic Plan. We welcome the review of the National AI R&D Strategic Plan and believe the eight strategic priorities outlined in the 2019 update provide a strong framework to guide ongoing research priorities. We also believe there are a number of research areas across these priorities in which progress could help accelerate the development and adoption of AI in a way that warrants trust and benefits people equitably. We set out our suggestions below.

Developing testing frameworks for AI systems to help identify and mitigate potential risk
Microsoft supports the AI R&D Strategic Plan’s focus on measuring and evaluating technologies through standards and benchmarks (Strategy 6). Work is needed to create standardized testing frameworks and benchmarks that allow for effective evaluation of AI systems to ensure they are performing appropriately for a given use case in a way that is fair, safe, secure, and reliable. This type of testing is an essential part of the risk identification and mitigation process and development of testing frameworks will underpin progress in a number of areas of AI research, including in addressing the ethical, legal, and societal impacts of AI (Strategy 3) and ensuring that systems are performing in a way that is safe and secure (Strategy 4). It will also be important to help organizations implement risk management frameworks, including the AI Risk Management Framework being developed by NIST.\(^1\) Currently, evaluation of AI systems is challenging given the way in which systems can be brittle and difficult to transfer effectively outside of the training environment, with failure modes often not well described.


---

Effective evaluation will therefore require the development of testing frameworks that allow for both benchmark testing of systems during the development phase and local, operational testing of systems post-deployment and across the entirety of a system's lifecycle. As well as allowing for the general evaluation of a system's performance against standardized benchmarks, testing frameworks need to address pockets of failures that could result in potentially significant costs. Developing standardized benchmarks for developmental and operational testing will also allow for direct comparisons of AI systems, including by deployers, in a way that can help drive continued technological advancement and build trust in the technology. While important work is underway in forums like the ISO/IEC JTC 1 committees (including SC42 on AI, SC37 on biometrics, and SC27 on security, privacy and AI) there is currently a gap in relation to these types of testing methodologies and benchmarks. We would encourage the AI R&D Strategic Plan to prioritize further research in this area in a way that advances understanding, creates frameworks that organizations can easily utilize, and aligns with ongoing initiatives.

**Advancing understanding of fairness harms**

Microsoft supports the AI R&D Strategic Plan’s focus on understanding and addressing the ethical, legal, and societal implications of AI (Strategy 3). Work in this area is foundational to ensuring that everyone can realize the benefits of AI technologies, including by ensuring that people view AI technology as trustworthy and that there is legal certainty around how organizations deploy AI systems. Microsoft would encourage further research into the range of fairness harms that AI systems can pose. This includes potential quality of service harms, where an AI system may not work as well for one demographic group as it does for another; allocation harms, where a system may allocate resources or opportunities in essential domains such as education, healthcare, and employment to different demographic groups in a manner that leads to inequitable outcomes; and representation harms, where an AI system may describe, depict, or otherwise represent people, cultures, or society in a way that creates potential for stereotyping, demeaning, or erasing of relevant demographic groups impacted by the system. Important progress has been made in recent years in developing tools and techniques for identifying, measuring, and mitigating fairness harms. This includes work Microsoft has done to build out the Responsible AI Toolbox\(^2\) which helps advance responsible AI use, including by allowing developers to identify fairness challenges in their model performance. Microsoft would encourage the AI R&D Strategic Plan to prioritize research that can further advance understanding of how fairness harms can occur across the development and deployment of AI systems, as well as research into tools and techniques that can help identify and mitigate fairness harms. More work is needed to advance understanding of representation harms in particular, where many important questions around identification, measurement, and mitigation are still at an early research stage.

\(^2\) See [Microsoft Responsible AI (responsibleaitoolbox.ai)](https://responsibleaitoolbox.ai).
Advancing effective human-AI teaming

Microsoft supports the AI R&D Strategic Plan’s prioritization of research into effective methods for human-AI collaboration. This work is critical for ensuring AI can be used in a way that advances human dignity and flourishing and allows for the effective oversight of AI systems. Microsoft believes that the importance of this work is sometimes underappreciated and would recommend prioritizing investments in this area in a way that allows for tangible progress to be made.

Microsoft believes further investment is needed to develop training programs for those using AI systems so that they are able to interpret and act on a system’s output in an appropriate manner. Training is also needed for those overseeing AI systems to identify and respond to any potential risks the system may generate. Investment in training programs for those using and overseeing high-risk systems should be prioritized and training should be appropriately tailored to a given use-case and the type of human-AI interaction, including the level of autonomy with which a system is operating. Training programs should also include proficiency testing and refresher courses that allow for training to keep pace with technological developments.

Further research should be prioritized around questions of how to design AI systems in a way that enables effective use and oversight. Advancing understanding of how to design systems in a way that is explainable and interpretable is important. However, work in this area must extend beyond these issues and identify how best to communicate system information to users, including how to act on outputs in an appropriate manner and how to avoid over-confidence in or over-reliance on system output. Research should also advance understanding of how to design effective human-AI interactions within the context of the wider operational system that the human-AI interaction is a part of, ensuring that actions taken on the basis of an AI system are appropriate. Further research into different types of human-AI interaction, including how best to design AI systems to support human decision making and actions would also be beneficial.

Responsibly integrating AI systems into status quo systems and workflows

Microsoft supports the National AI Initiative Act’s emphasis on updates to the National AI R&D Strategic Plan that encompass AI research, development, and deployment, with a view toward harnessing R&D to successfully field beneficial applications. Once AI technologies are developed, critical challenges remain for integrating and interleaving them into current systems and workflows. We believe this is another critical area of research that warrants more attention and investment. Federal research should advance best practices for integrating both AI technologies and responsible AI methods and practices within existing workflows, in a manner that reduces friction for deep integration with the status quo—and for plans and trajectories to carry operations and programs beyond the status quo via efforts and redesign and reformulation of legacy processes and systems.
Advancing generalizability of systems across time and task
In making long-term investments in AI research (Strategy 1), and to effectively harness shared public datasets (Strategy 5) to benefit populations and institutions across the country, the Federal Government should invest in R&D to improve the generalizability of systems across time and task. Key challenges remain with R&D on the robustness of methods for use in different environments, and for, more generally, transferring knowledge from one task to another, so that models perform well when applied in new contexts or regions that differ in subtle ways. Improvements across task performance would, for instance, enable systems designed for one type of lighting or weather condition to be used in another, or learnings from a hospital system trained on one population to be harnessed to serve a different population. Improving the generalizability of model performance across time is another key challenge, considering shifts in how a task environment changes over time, challenges understanding how model updates will change a system’s operation (including with new errors), and the need for requirements and guidelines for maintaining AI systems over time.

Advancing research and infrastructure development for leading edge R&D with large-scale neural models
In fulfilling the National AI R&D Strategic Plan’s aim to develop shared public datasets and environments for AI training and testing (Strategy 5), 2022 updates should ensure such developments facilitate leading edge R&D with large-scale neural models. AI R&D breakthroughs for national prosperity and security, across domains from healthcare to climate change, will benefit from researchers nationwide having access to the advanced computation and data resources needed to power cutting-edge model development. Training, running, and maintaining large-scale neural models can require vast resources, making them projects requiring robust governance structures, academic consensus-building, and mobilization of special resources and programs. Additional R&D will be needed to drive findings and deliberation on how emerging AI capabilities, including large-scale platform models (also referred to as “foundation” models), should be constructed and applied (e.g., via “fine-tuning” methods) and their risks managed. Microsoft commends the National Artificial Intelligence Research Resource (NAIRR) Task Force for tackling such resource and governance questions as it develops its roadmap and recommendations for Congress; we further support the National AI R&D Strategic Plan including an emphasis on the urgency of adequate infrastructure development and governance in this area.

Using AI to advance science and address major challenges

---

3 For more on current challenges and future research opportunities, see A Study in Transfer Learning: Leveraging Data from Multiple Hospitals to Enhance Hospital-Specific Predictions.
4 See An Empirical Analysis of Backward Compatibility in Machine Learning Systems.
5 Microsoft shares these recommendations given insights from supporting AI R&D breakthroughs via its partnerships with academic research institutions. See for instance Microsoft Turing Academic Program (MS-TAP) - Microsoft Research.
6 Risks with large-scale platform models pertain to safety and the understandability of emergent behaviors, the potential for systems to generate offensive output, and malevolent uses of new capabilities.
Microsoft appreciates OSTP’s request for suggestions of AI R&D focus areas that could create solutions for healthcare and climate change, two areas where we too are committed to harnessing AI. As noted above, continued development of large-scale neural models will improve the nation’s ability to respond rapidly to a range of challenges, including in relation to healthcare and the current and potential future pandemics. There are also great opportunities ahead for harnessing AI advances to enhance the delivery of quality and efficacious healthcare services. R&D investments should be made to address multiple challenges of building, fielding, maintaining, and employing diagnostic and predictive models in ambulatory and in-patient healthcare delivery settings. Moving from clinical practice to the biosciences, there is a great opportunity to build on the progress achieved to date with harnessing AI to assist with protein design and for related tasks such as predicting protein-protein interactions, to enable the faster creation and updating of vaccines and the creation of new kinds of therapeutics.

The National AI R&D Strategic Plan should also prioritize AI R&D to address key knowledge gaps that will be critical to address challenges with climate change, including to help organizations reach net-zero goals. These include leveraging AI 1) to improve carbon accounting (in particular, our ability to directly measure methane and land use emissions, which are areas of great uncertainty); and 2) for materials engineering that can be used to decarbonize the economy (including, for example, to develop improved batteries, sustainable fuels, and sorbents for carbon dioxide removal). Other areas where AI R&D could have significant impact include the development and scaling of tools for climate resilience (e.g., improving high-quality environmental data collection and processing to aid the decision-making of policymakers and local communities), and R&D to advance the sustainability of semiconductor fabrication (e.g., leveraging AI to optimize fabrication with renewable energy and abate or replace greenhouse gases from fabrication). Microsoft notes that as the nation makes investments to advance the U.S. semiconductor industry with national security in mind, a key opportunity also exists to embed sustainability breakthroughs into their design and production processes.

**Advancing understanding of risks around AI systems that can be used for tracking and surveillance**

Microsoft recommends further research into the risks of AI systems that can be used for tracking and surveillance as part of addressing the ethical, legal, and societal implications of AI (Strategy 3). While these systems, including those used for biometric identification, can offer benefits to society, they can also pose potential risks to civil liberties and democratic freedoms if not used responsibly. Microsoft continues to advocate for the urgent development of legal safeguards for facial recognition technology. The need is particularly acute for government and law enforcement use of facial recognition technology given the consequential nature of the decisions these organizations make. Microsoft believes that laws in this area should provide civil liberty protections and advance transparency and accountability, including through benchmark and operational testing of FRT systems. More work is needed to understand the way in which these systems can pose potential risks of harm, including via their interactions with other parts of decision-making systems in higher-risk domains (e.g., use by
government and law enforcement). Work should be accelerated around the development of training programs for users and overseers of such systems to ensure they are using the systems and their outputs appropriately. Urgent work is also needed to build out testing frameworks for these systems that allow for effective benchmark and operational testing in a way that ensures systems are performing to an acceptable standard for a given use case (with a higher standard needed for higher risk use cases). As mentioned above, while important work is underway in ISO/IEC JTC 1/SC37 to define the responsible use of AI and biometrics in passive identification systems (e.g. surveillance), more work is needed to develop methodologies and benchmarks. Robust testing frameworks, when combined with appropriate legal safeguards, will help advance transparency around system performance and demonstrate that the technology is trustworthy.

Developing best practice around privacy-enhancing technologies
To advance OSTP’s priority of maintaining the core values behind America’s scientific leadership, including democratic values, we recommend that the 2022 National AI R&D Strategic Plan prioritize research to advance the adoption of privacy-enhancing technologies (PET) such as homomorphic encryption, secure multi-party computation, and differential privacy. Microsoft applauds the Biden Administration’s recent initiatives to increase PET adoption, including through bilateral innovation prize challenges and federal agency collaboration with the United Nations. Work is already underway in ISO/IEC JTC1 SC27 to assess the impact of AI on privacy, but at the same time, further R&D continues to be needed to address open questions in privacy preserving machine learning\(^7\) and to overcome barriers to achieve widespread adoption.\(^8\)

Advancing privacy protections
Microsoft strongly supports the strategic aim to understand and address the ethical, legal, and societal implications of AI (Strategy 3), and OSTP’s desire to advance research that informs the intersection of AI application with privacy and civil liberties. Novel AI capabilities can reveal personal information (e.g., based on location trails, interactions, and interests) that if used without safeguards could chill freedoms of expression and association. As a result, R&D is needed to better understand legal gaps and gray areas that exist, including with government use of large-scale commercial datasets, and to develop effective technical and policy approaches that limit agency use of (commercial or government) data about U.S. citizens for models and inferences in cases lacking disclosure and consent.

Advancing international R&D cooperation around assessing system performance, including for fairness and privacy

---
\(^7\) See [Privacy Preserving Machine Learning: Maintaining confidentiality and preserving trust - Microsoft Research](#).

\(^8\) These suggestions stem from Microsoft’s research on challenges for PET adoption including: how to make the privacy techniques computationally tractable; how to make them more usable by developers; and how to make them explainable and accountable to stakeholders and wider society. See [Exploring Design and Governance Challenges in the Development of Privacy-Preserving Computation - Microsoft Research](#).
As part of the strategic aim to measure and evaluate AI technologies through standards and benchmarks (Strategy 6), the 2022 Strategic Plan should extend international R&D cooperation with strategic partners to (1) better understand and address AI fairness and privacy challenges, and (2) drive consensus on jointly agreed upon test, evaluation, validation and verification approaches for assessing AI system performance and trustworthiness. Through the strategic plan’s R&D investments, the Federal Government has an opportunity to further the development of common AI measurement and risk mitigation techniques that could support future harmonized standards and/or conformity assessments.

**AI and cybersecurity**
The adoption and use of AI introduces both new "AI attack surfaces" with several well-understood types of attacks and the use of AI itself as a cyberwarrior technology. AI systems can fail in new ways as a result of security attacks which exploit vulnerabilities in AI systems due to the way they are developed and their heavy reliance on data. Although work is already underway in ISO/IEC JTC1 SC27 and via a MITRE and Microsoft partnership to assess the security impact and failure modes of AI, more work is needed to understand how to develop AI systems that are robust and resilient in the face of these new exploits. The use of AI as a cyberwarrior technology received attention from the National Academy of Sciences and the National Security Commission on AI, but requires ongoing study to better understand evolving capabilities and limitations as AI will play an increasingly important role in protecting infrastructure.

Thank you for the opportunity to contribute input on updating the National Artificial Intelligence Research and Development Strategic Plan.

Sincerely,

Eric Horvitz
Chief Scientific Officer
Microsoft Corporation

---

9 See, for example, work on the Adversarial Machine Learning Threat Matrix: [MITRE, Microsoft, and 11 Other Organizations Take on Machine-Learning Threats | The MITRE Corporation; GitHub - mitre/advmthreatmatrix: Adversarial Threat Matrix.](https://github.com/mitre/advmthreatmatrix)

10 See [Implications of Artificial Intelligence for Cybersecurity A Workshop | National Academies.](https://www.nationalacademies.org/content.aspx?id=34973)