Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

The White House Office of Science and Technology Policy – on behalf of the National Science and Technology Council's Select Committee on Artificial Intelligence and Machine Learning and AI Subcommittee, the National AI Initiative Office, and the Networking and Information Technology Research and Development National Coordination Office – released a Request for Information (RFI) on February 2, 2022, to request input on updating the National Artificial Intelligence Research and Development Strategic Plan. The RFI was published in the Federal Register and the comment period was open from February 2, 2022, through March 4, 2022.

This document contains the 63 responses received from interested parties. In accordance with the RFI instructions, only the first 10 pages of content were considered for each response.

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. The U.S. Government bears no responsibility for the accuracy, legality, or content of the responses and external links included in this document.

Table of Contents

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Accrete

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.

Ensuring AI Dominance by Building Public Trust – by Prashant Bhuyan, CEO of Accrete

AI has the potential to transform the very nature of work by offloading skilled labor to machines. Continuously learning AI that interacts naturally with knowledge workers will drive revenue growth for corporations in previously unimaginable ways. However, the price of growth, if left unchecked, will be extreme socioeconomic inequality and the end of civil society. To avoid harmful consequences, we must act now to establish rules and regulations that protect people from the far-reaching consequences of AI misuse.

To secure its position as a leader in tomorrow's world, the United States must recognize that power will flow to the select few that control the AI architecture. The greater the perceived benefit of these AI systems, the greater the potential for disproportionate influence or unethical use. In the future, our lives will be substantially affected by AI in innumerable ways. Just like the invention of the atom bomb, AI is a natural consequence of human ingenuity. Like nuclear proliferation, we must also recognize that AI proliferation is a potential threat to humanity if left unchecked. Soon there will be a rapidly growing divide in which most people will ultimately be imperceptibly influenced by intelligent machines that are owned and trained by people with distinct biases and objectives. We must act now to establish policies to ensure AI proliferation is a peaceful, equitable, and above all, transparent evolution.

In the same manner that the Federal Reserve Bank maintains a dual mandate to balance inflation and employment, an independent government agency should be established to balance AI ethics and labor automation. The key to ensuring the United States' position as a leader in the Industries of Tomorrow is balancing the inevitable increase of intelligent machines replacing skilled human labor with the people's fundamental trust in how the AI learns.

Important questions people must ask include, "To what extent do I need to understand the biases underpinning the AI?"; "Is the AI learning from my personal information?"; "How do I know my objectives are aligned with the AI?". To trust AI, citizens must believe in the ethical standards set forth by governing agencies and the government must work with the people to establish the appropriate standards.

An independent government agency focused on the real-world implications of AI within civil society should have the ability to establish standards that reinforce public trust in AI. For example, such an agency could establish a rule that prevents technology companies from gaming children to get them hooked on social media for the purpose of collecting data to optimize advertising algorithms targeting those children. Another rule that could engender public trust in AI could mandate that any employer that replaces workers with AI is obligated to retrain and upskill the redundant employee.

Such an independent governing body would also be able to establish industry specific standards in areas like explain-ability and performance to hold owners and employers of AI accountable for the consequence of an error. AI driven errors that harm humans would have the worst consequence. For example, if a surgeon relies on an insight produced by an AI and subsequently kills a patient, that surgeon mustn't be able to lay blame on the AI. However, to hold the surgeon accountable, there needs to be transparency into who trained the AI, how the AI was trained and

what biases influenced the model's learning. Ultimately, there should be standardization in the architectures and approaches used in the development of explainable AI itself.

Although countries such as China are making great advances in AI today, these advances are coming at the cost of civil liberties such as digital privacy. This asymmetric transaction between people and AI is unsustainable in the long run and will ultimately lead to revolution. The United States has a golden opportunity to lead the world in using AI to create a Utopian future by engineering a fair and equitable relationship between AI and the people that ensures balanced long-term growth and civil society.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

ACT | The App Association

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, 2020

Attn: NCO
Office of Science and Technology Policy
2415 Eisenhower Avenue,
Alexandria, VA 22314

RE: Comments of ACT | The App Association to the Office of Science and Technology Policy on its Request for Information to the Updated National Artificial Intelligence Research and Development Strategic Plan

ACT | The App Association (App Association) appreciates the opportunity to submit views to the Office of Science and Technology Policy (OSTP) on updates to the National Artificial Intelligence Research and Development Strategic Plan, which provides guidance to federal agencies to inform the development of regulatory and non-regulatory approaches regarding technologies and industrial sectors empowered or enabled by artificial intelligence (AI), and ways for agencies to reduce barriers to the development and adoption of AI technologies.[1] The App Association supports updating the National Artificial Intelligence Research and Development Strategic Plan to support and facilitate AI research and development by prioritizing and providing sufficient funding while also ensuring adequate incentives (*e.g.*, streamlined availability of data to developers, tax credits) are in place to encourage private and non-profit sector research. Transparency research should be a priority and involve collaboration among all affected stakeholders who must responsibly address the ethical, social, economic, and legal implications that may result from AI applications.

The App Association represents thousands of small business software application development companies and technology firms that create the technologies that drive internet of things (IoT) use cases across consumer and enterprise contexts. Today, the value of the ecosystem the App Association represents – which we call the app economy – is approximately $1.3 trillion and is responsible for 5.7 million American jobs. Alongside the world's rapid embrace of mobile technology, our members create the innovative solutions that power IoT across modalities and segments of the economy. The National Artificial Intelligence Research and Development Strategic Plan, and the efforts of numerous agencies with respect to AI policy and regulation, directly impacts the app economy. We support the Administration's goal of ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security, and to maintaining the core values behind America's scientific leadership,

[1] https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and

including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.[2]

The App Association also continues to proactively work to advance the use of AI in key use cases. As one example, the App Association's Connected Health Initiative[3] (CHI) assembled a Health AI Task Force in the summer of 2018 consisting of a range of innovators and thought leaders. Building on their work throughout the second half of 2018, in early February 2019 CHI unveiled its AI Task Force's deliverables during a public-private multistakeholder dialogue in Washington, DC. These deliverables included a position piece supporting AI's role in healthcare, policy principles addressing how policy frameworks should approach the role of AI in healthcare, and a terminology document targeted at policymakers.[4] Since the release of its deliverables, CHI has actively advocated for the development of frameworks that will responsibly support the development, availability, and use of AI innovations.

AI is an evolving constellation of technologies that enable computers to simulate elements of human thinking – learning and reasoning among them. An encompassing term, AI entails a range of approaches and technologies, such as Machine Learning (ML) and deep learning, where an algorithm based on the way neurons and synapses in the brain change due to exposure to new inputs, allowing independent or assisted decision making. AI-driven algorithmic decision tools and predictive analytics are having, and will continue to have, substantial direct and indirect effects on Americans. Some forms of AI are already in use to improve American consumers' lives today – for example, AI is used to detect financial and identity theft and to protect the communications networks upon which Americans rely against cybersecurity threats.

Moving forward, across use cases and sectors, AI has incredible potential to improve American consumers' lives through faster and better-informed decision making, enabled by cutting-edge distributed cloud computing. As an example, healthcare treatments and patient outcomes stand poised to improve disease prevention and conditions, as well as efficiently and effectively treat diseases through automated analysis of x-rays and other medical imaging. AI will also play an essential role in self-driving vehicles and could drastically reduce roadway deaths and injuries. From a governance perspective, AI solutions will derive greater insights from infrastructure and support efficient budgeting decisions. An estimate states AI technological breakthroughs will represent a $126 billion market by 2025.[5]

---

[2] *Id.*

[3] *See* www.connectedhi.com.

[4] The CHI Health AI Task Force's deliverables are accessible at https://actonline.org/2019/02/06/why-does-healthcare-need-ai-connected-health-initiative-aims-to-answer-why/.

[5] McKinsey Global Institute, *Artificial Intelligence: The Next Digital Frontier?* (June 2017), *available at* https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx.

Today, Americans encounter AI in their lives incrementally through the improvements they have seen in computer-based services they use, typically in the form of streamlined processes, image analysis, and voice recognition (we urge consideration of these forms of AI as “narrow” AI). The App Association notes that this “narrow” AI already provides great societal benefit. For example, AI-driven software products and services revolutionized the ability of countless Americans with disabilities to achieve experiences in their lives far closer to the experiences of those without disabilities.

Nonetheless, AI also has the potential to raise a variety of unique considerations for policymakers. The App Association appreciates the efforts to develop a policy approach to AI that will bring its benefits to all, balanced with necessary safeguards to protect consumers. To assist the Administration, the App Association offers a comprehensive set of AI policy principles below for consideration that we strongly encourage alignment of the National Artificial Intelligence Research and Development Strategic Plan with the following:

1. **AI Strategy:** Many of the policy issues raised below involve significant work and changes that will impact a range of stakeholders. The cultural, workforce training and education, data access, and technology-related changes associated with AI will require strong guidance and coordination. An AI strategy incorporating guidance on the issues below will be vital to achieving the promise that AI offers to consumers and our economies. We believe it is critical to take this opportunity to encourage civil society organizations and private sector stakeholders to begin similar work. The National Artificial Intelligence Research and Development Strategic Plan is, and should remain, a key part of the U.S. overall strategy to global leadership in this critical area of technology.

2. **Research:** The National Artificial Intelligence Research and Development Strategic Plan should support and facilitate research and development of AI by prioritizing and providing sufficient funding while also ensuring adequate incentives (e.g., streamlined availability of data to developers, tax credits) are in place to encourage private and non-profit sector research. Transparency research should be a priority and involve collaboration among all affected stakeholders who must responsibly address the ethical, social, economic, and legal implications that may result from AI applications.

3. **Quality Assurance and Oversight:** The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should advance risk-based approaches to ensure that the use of AI aligns with the recognized standards of safety, efficacy, and equity. Providers, technology developers and vendors, and other stakeholders all benefit from understanding the distribution of risk and liability in building, testing, and using AI tools. Policy frameworks addressing liability should ensure the appropriate distribution and mitigation of risk and liability. Specifically, those in the value chain with the ability to minimize risks based on their knowledge and ability to

mitigate should have appropriate incentives to do so. Some recommended guidelines include:

- Ensuring AI is safe, efficacious, and equitable.
- Supporting that algorithms, datasets, and decisions are auditable.
- Encouraging AI developers to consistently utilize rigorous procedures and enabling them to document their methods and results.
- Requiring those developing, offering, or testing AI systems to provide truthful and easy to understand representations regarding intended use and risks that would be reasonably understood by those intended, as well as expected, to use the AI solution.
- Ensuring that adverse events are timely reported to relevant oversight bodies for appropriate investigation and action.

4. **Thoughtful Design:** The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should strongly encourage the design of AI systems that are informed by real-world workflows, human-centered design and usability principles, and end-user needs. AI systems solutions should facilitate a transition to changes in the delivery of goods and services that benefit consumers and businesses. The design, development, and success of AI should leverage collaboration and dialogue among users, AI technology developers, and other stakeholders in order to have all perspectives reflected in AI solutions.

5. **Access and Affordability:** The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should ensure AI systems are accessible and affordable. Significant resources may be required to scale systems and policymakers should take steps to remedy the uneven distribution of resources and access. Policies must be put in place that incent investment in building infrastructure, preparing personnel and training, as well as developing, validating, and maintaining AI systems with an eye toward ensuring value.

6. **Ethics:** AI will only succeed if it is used ethically. It will be critical to promote many of the existing and emerging ethical norms for broader adherence by AI technologists, innovators, computer scientists, and those who use such systems. The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should:
    - Ensure that AI solutions align with all relevant ethical obligations, from design to development to use.
    - Encourage the development of new ethical guidelines to address emerging issues with the use of AI, as needed.
    - Maintain consistency with international conventions on human rights.
    - Ensure that AI is inclusive such that AI solutions beneficial to consumers are developed across socioeconomic, age, gender, geographic origin, and other groupings.

- Reflect that AI tools may reveal extremely sensitive and private information about a user and ensure that laws protect such information from being used to discriminate against certain consumers.

7. **Modernized Privacy and Security Frameworks:** While the types of data items analyzed by AI and other technologies are not new, this analysis will provide greater potential utility of those data items to other individuals, entities, and machines. Thus, there are many new uses for, and ways to analyze, the collected data. This raises privacy issues and questions surrounding consent to use data in a particular way (e.g., research, commercial product/ service development). It also offers the potential for more powerful and granular access controls for consumers. Accordingly, The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should address the topics of privacy, consent, and modern technological capabilities as a part of the policy development process. Risk management policy frameworks must be scalable and assure that an individual's data is properly protected, while also allowing the flow of information and responsible evolution of AI. This information is necessary to provide and promote high-quality AI applications. Finally, with proper protections in place, policy frameworks should also promote data access, including open access to appropriate machine-readable public data, development of a culture of securely sharing data with external partners, and explicit communication of allowable use with periodic review of informed consent.

8. **Collaboration and Interoperability:** The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should enable eased data access and use through creating a culture of cooperation, trust, and openness among policymakers, AI technology developers and users, and the public.

9. **Bias:** The bias inherent in all data, as well as errors, will remain one of the more pressing issues with AI systems that utilize machine learning techniques in particular. Addressing data provenance and bias issues is a must in developing and using AI solutions. The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should:
    - Require the identification, disclosure, and mitigation of bias while encouraging access to databases and promoting inclusion and diversity.
    - Ensure that data bias does not cause harm to users or consumers.

10. **Education:** The National Artificial Intelligence Research and Development Strategic Plan, and the U.S. approach to AI generally, should support education for the advancement of AI, promote examples that demonstrate the success of AI, and encourage stakeholder engagements to keep frameworks responsive to emerging opportunities and challenges.
    - Consumers should be educated as to the use of AI in the service they are using.

- Academic education should include curriculum that will advance the understanding of and ability to use AI solutions.

The policy issues raised by the National Artificial Intelligence Research and Development Strategic Plan involves significant work and changes that will impact a range of stakeholders. The cultural, workforce training and education, data access, and technology-related changes associated with AI will require strong guidance and coordination across U.S. federal agencies. The App Association supports the development of national AI strategies for federal agencies, which will be vital to achieving the promise that AI offers to consumers and entire economies.

Noting our general support for the current National Artificial Intelligence Research and Development Strategic Plan, we offer the following suggested revisions:

- **Alignment with Other Leading Federal Policies for AI:** The National Artificial Intelligence Research and Development Strategic Plan should align with other federal efforts to develop AI policy, such as the National Institute of Standards and Technology's (NIST) Artificial Intelligence Risk Management Framework, a policy being developed in close collaboration with the private sector, academia, and others for voluntary use with the goal of improving the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.[6]
- **Require Agencies to Advance Thoughtful Design Principles Across AI Use Cases:** The National Artificial Intelligence Research and Development Strategic Plan should require design of AI systems informed by real-world workflows, human-centered design and usability principles, and end-user needs. AI systems solutions should facilitate a transition to changes in the delivery of goods and services that benefit consumers and businesses. The design, development, and success of AI should leverage collaboration and dialogue among users, AI technology developers, and other stakeholders in order to have all perspectives reflected in AI solutions. As this concept must run across sectors and AI use cases, the National Artificial Intelligence Research and Development Strategic Plan incorporate guidance for agencies to advance thoughtful design principles through their approaches and actions related to AI.
- **Require Agencies to Advance Ethics in AI's Development and Use:** The success of AI depends on ethical use. An agency's approach will need to promote many of the existing and emerging ethical norms for broader adherence by AI technologists, innovators, computer scientists, and those who use such systems. The National Artificial Intelligence Research and Development Strategic Plan should:

[6] https://www.nist.gov/itl/ai-risk-management-framework.

- Ensure that AI solutions align with all relevant ethical obligations, from design to development to use.
- Encourage the development of new ethical guidelines to address emerging issues with the use of AI, as needed.
- Maintain consistency with international conventions on human rights.
- Ensure that AI is inclusive such that AI solutions beneficial to consumers develop across socioeconomic, age, gender, geographic origin, and other groupings.
- Reflect that AI tools may reveal extremely sensitive and private information about a user and ensure that laws protect such information from being used to discriminate against certain consumers

- **Augment the Requirement on Federal Agencies for Disclosure and Transparency:** The Administration should consider further prioritizing disclosure and trust priorities in the National Artificial Intelligence Research and Development Strategic Plan. Providers, technology developers, and vendors, and other stakeholders will all benefit from understanding the distribution of risk and liability in building, testing, and using AI tools. The National Artificial Intelligence Research and Development Strategic Plan should therefore clearly address liability so as to ensure the appropriate distribution and mitigation of risk and liability (i.e., those in the value chain with the ability to minimize risks based on their knowledge and ability to mitigate should have appropriate incentives to do so). Further, the National Artificial Intelligence Research and Development Strategic Plan should clearly require that AI policies prioritize that those developing, offering, or testing AI systems provide truthful and easy to understand representations regarding intended use and risks that would be reasonably understood by those intended, as well as expected, to use the AI solution.
- **Support the Development of, and Access to, Open Standards Needed to Drive U.S. Leadership in AI:** The National Artificial Intelligence Research and Development Strategic Plan should support the developer and use of voluntary consensus standards that concern AI application. The App Association strongly encourages updating the National Artificial Intelligence Research and Development Strategic Plan to support public-private collaboration on AI through standardization by encouraging key U.S.-based standard-setting organizations (SSOs) such as IEEE to grow and thrive. The U.S. government can support such organizations through pro-innovation policies that encourage private sector research and development of AI innovations and the development of related standards.

  It is critical that the United States should ensure that such standards are accessible to innovators by promoting a balanced approach to standard-essential

patent (SEP) licensing. AI technical standards, built on contributions through an open and consensus-based process, bring immense value to consumers by promoting interoperability while enabling healthy competition between innovators; and often include patented technology. When an innovator gives its patented technology to a standard, this can represent a clear path to reward in the form of royalties from a market that likely would not have existed without the standard being widely adopted. To balance this potential with the need for access to the patents that underlie the standard, many SSOs require holders of patents on standardized technologies to license their patents on fair, reasonable, and non-discriminatory (FRAND) terms. FRAND commitments prevent the owners of patents used to implement the standard from exploiting the unearned market power that they otherwise would gain as a consequence of the broad adoption of a standard. Once patented technologies incorporate into standards, it compels manufacturers to use them to maintain product compatibility. In exchange for making a voluntary FRAND commitment with an SSO, SEP holders gain the ability to obtain reasonable royalties from a large number of standard implementers that might not have existed absent the standard. Without the constraint of a FRAND commitment, SEP holders would have the same power as a monopolist that faces no competition.

Unfortunately, a number of owners of FRAND-committed SEPs are flagrantly abusing their unique position by reneging on those promises with unfair, unreasonable, or discriminatory licensing practices. These practices, under close examination by antitrust and other regulators in many jurisdictions, not only threaten healthy competition and unbalance the standards system but also impact the viability of new markets such as AI. This amplifies the negative impacts on small businesses because they can neither afford years of litigation to fight for reasonable royalties nor risk facing an injunction if they refuse a license that is not FRAND compliant.

Patent policies developed by SSOs today will directly impact the way we work, live, and play for decades to come. SSOs vary widely in terms of their memberships, the industries and products they cover, and the procedures for establishing standards. In part due to the convergence associated with the rise of IoT, each SSO will need the ability to tailor its intellectual property policy for its particular requirements and membership. The App Association believes that some variation in patent policies among SSOs is necessary and that the U.S. government should not prescribe detailed requirements that all SSOs must implement. At the same time, however, as evidenced by the judicial cases and regulatory guidance, basic principles underlie the FRAND commitment and serve to ensure that standard setting is pro-competitive, and the terms of SEP licenses are in fact reasonable. Ideally, an SSO’s intellectual property rights policy that

requires SEP owners to make a FRAND commitment would include all of the following principles that prevent patent "hold up" and anti-competitive conduct:

- **Fair and Reasonable to All** – A holder of a SEP subject to a FRAND license such SEP on fair, reasonable, and nondiscriminatory terms to all companies, organizations, and individuals who implement or wish to implement the standard.
- **Injunctions Available Only in Limited Circumstances** –SEP holders should not seek injunctions and other exclusionary remedies nor allowed these remedies except in limited circumstances. The implementer or licensee is always entitled to assert claims and defenses.
- **FRAND Promise Extends if Transferred** – If there is a transfer of a FRAND-encumbered SEP, the FRAND commitments follow the SEP in that and all subsequent transfers.
- **No Forced Licensing** – While some licensees may wish to get broader patent holder should not require implementers to take or grant licenses to a FRAND-encumbered SEP that is invalid, unenforceable, or not infringed, or a patent that is not essential to the standard.
- **FRAND Royalties** – A reasonable rate for a valid, infringed, and enforceable FRAND-encumbered SEP should be based on several factors, including the value of the actual patented invention apart from its inclusion in the standard, and cannot be assessed in a vacuum that ignores the portion in which the SEP is substantially practiced or royalty rates from other SEPs required to implement the standard.

We also note that a number of SSO intellectual property rights policies require SSO participants to disclose patents or patent applications that are or may be essential to a standard under development. Reasonable disclosure policies can help SSO participants evaluate whether technologies considered for standardization are covered by patents. Disclosure policies should not, however, require participants to search their patent portfolios as such requirements can be overly burdensome and expensive, effectively deterring participation in an SSO. In addition, FRAND policies that do not necessarily require disclosure, but specify requirements for licensing commitments for contributed technology, can accomplish many, if not all, of the purposes of disclosure requirements.

The U.S. Department of Justice (DOJ) already encouraged SSOs to define FRAND more clearly. For example, DOJ's former assistant attorney general Christine Varney explained that "clearer rules will allow for more informed participation and will enable participants to make more knowledgeable decisions regarding implementation of the standard. Clarity alone does not eliminate the

possibility of hold-up…but it is a step in the right direction."[7] As another example, Renata Hesse, a previous head of the DOJ's Antitrust Division, provided important suggestions for SSOs to guard against SEP abuses that included at least three of the aforementioned principles.[8] The National Artificial Intelligence Research and Development Strategic Plan should be updated to advance open standards, consistent with OMB-A119 ("Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities"),[9] open standards and access to open standards with respect to SEPs.

[7] Christine A. Varney, Assistant Att'y Gen., Antitrust Div., U.S. Dep't of Justice, Promoting Innovation Through Patent and Antitrust Law and Policy, Remarks as Prepared for the Joint Workshop of the U.S. Patent and Trademark Office, the Federal Trade Comm'n, and the Dep't of Justice on the Intersection of Patent Policy and Competition Policy: Implications for Promoting Innovation 8 (May 26, 2010), *available at* http://www.atrnet.gov/subdocs/2010/260101.htm.

[8] Renata Hess, Deputy Assistant Attorney General, *Six 'Small' Proposals for SSOs Before Lunch*, Prepared for the ITU-T Patent Roundtable (October 10, 2012), *available at* https://www.justice.gov/atr/speech/six-smallproposals-ssos-lunch.

[9] https://www.nist.gov/system/files/revised_circular_a-119_as_of_01-22-2016.pdf.

The App Association appreciates the Administration's consideration of the above views. We urge OSTP to contact the undersigned with any questions or ways that we can assist moving forward.

Sincerely,

Brian Scarpelli
Senior Global Policy Counsel

Leanna Wade
Policy Associate

ACT | The App Association
1401 K St NW (Ste 501)
Washington, DC 20005
202-331-2130

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

AHIP

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.

Guiding Greater Health

601 Pennsylvania Avenue, NW
South Building, Suite 500
Washington, D.C. 20004

T 202.778.3200
F 202.331.7487
ahip.org

March 4, 2022

The White House
Office of Science and Technology Policy
1650 Pennsylvania Avenue, NW
Washington, D.C. 20502

RE: Request for Information (RFI) Response: Update of the National Artificial Intelligence Research and Development Strategic Plan

Dear White House Representative:

Artificial Intelligence (AI) is becoming more commonplace in the United States and worldwide in both the public and private sectors. Americans want our nation to leverage advancements in technology, including applications in health care, while also leading the way in the ethical use of data and information. To help fulfill this goal, AHIP[1] is responding to the Notice of Request for Information (RFI) for the National Artificial Intelligence Research and Development Strategic Plan.

In health care, AI can serve as a catalyst for better care and access for Americans. AI can help improve efficient delivery of care, ensure correct care decisions for patients, identify previously unidentified issues or trends in individual and population health, simplify processes and improve satisfaction for patients and providers, and reduce administrative tasks to enable providers and their staff to focus their time and attention on the patient. AHIP has increased its focus on the current and potential applications of AI and is working to develop and support policy principles and goals specific to AI and its uses to improve health and well-being.

AHIP supports the OSTP effort to update the current Strategic Plan. We offer our input as a national association of health insurance providers, private sector partners in the critical health delivery and health care infrastructure. AHIP is an active participant in the Health/Public Health (HPH) Sector Coordinating Council (SCC). We support this effective public-private partnership, and we support ways to advance our work to complement the National Strategy.

Our comments below address the topical issues raised in the RFI.

[1] AHIP is the national association whose members provide health care coverage, services, and solutions to hundreds of millions of Americans every day. We are committed to market-based solutions and public-private partnerships that make health care better and coverage more affordable and accessible for everyone.

General Comments
Overall, we believe it will be extremely valuable to align the Strategic Plan with the goals and priorities defined in the National AI Initiative Act of 2020, which became law on January 1, 2021. Specifically, explaining how the Strategic Plan can align with and integrate common components and strategies for the public and private sectors can help to streamline AI uses and acceptance by consumers.

Likewise, building the technical processes, standards, and metrics will help capture the ways through which AI can be efficient, cost-effective, and easy-to-use. For example, AHIP has worked with the Consumer Technology Association on its consensus-driven American National Standards Institute (ANSI) accredited standard, ANSI/CTA-2090, The Use of Artificial Intelligence in Health Care: Trustworthiness, which considers three expressions of how trust is created and maintained: (1) Human Trust; (2) Technical Trust; and (3) Regulatory Trust. Building public trust and acceptance centered around safety, accountability, accuracy, reliability, security, and ethics will be essential components for moving AI forward in diverse settings and applications. As stakeholders build more resources and standards for AI, federal support for and recognition of these efforts can help promote national acceptance and adoption.

Long-term Investments in AI Research
We support federal investments to spur more widespread development and adoption of AI and believe it is important to highlight health care-specific priorities around the investment, adoption, and use of AI. The federal government should build on both public and private efforts to promote data exchange to improve "data completeness" (i.e., readability, reliability, and accountability). Promoting interoperable data, where possible, can help health care build on existing data streams without having to re-create systems and processes for data exchanges that use AI technologies. Developing shared public datasets and environments based on national technical standards for AI training and testing can also facilitate the availability of curated, standardized, secure, representative, aggregate, and privacy-protected data sets for AI research and development.

We encourage any federal investments to focus on the functional aspects of AI based on nationally recognized, technology-neutral standards. Innovations are key to advancing AI and federal policies and regulations should promote such innovations. Regulations, if enacted, should be based on a balance of the costs and benefits after public input.

Both short-term and long-term investments in health care AI can help prioritize areas of research and development focused on individual treatment - improving care outcomes and expediting interactions, which are important to consumers. A host of interdisciplinary projects can be developed to enable computing, networking, and access for patients and providers who may be unaccustomed to AI. We believe that Federated Learning and advanced cryptographic protections for data and privacy may be good base projects for research and development. We also support projects that inform ways to reduce unnecessary costs.

We are interested in learning more about proposals to support regional hubs to advance workforce training, representation, and overall digital equity. Regional innovation centers may establish community connections and involvement, and they also may help in development of solutions specific and appropriate to those regions. These efforts also can foster engagement to drive region-specific innovation on an ongoing basis. Such projects should have the infrastructure to include diverse stakeholders using a "hub and spokes" approach so that a variety of health care entities can utilize and build on such efforts.

Ethical, Legal, and Societal Implications of AI

To protect the rights and well-being of all Americans, we support the promotion of ethical, legal, environmental, safety, security, fairness, and other such guidelines for appropriate use of AI. AI and machine learning (ML) applications do not develop legal knowledge and ethical principles. Human users and programmers should anticipate these concepts and build appropriate frameworks based on legal and ethical business practices.

Inherent bias in data is of particular concern. Efforts to identify and mitigate harmful, unintended, or disparate bias should be undertaken. When possible, the public and private sector should work together to identify and manage bias that may have a harmful impact to specific groups or individuals.

It is also important to recognize that in AI, algorithms with a more precise focus on individuals or groups can have a benefit (i.e., "good bias"). One example may be noticing a trend or issue that may have been previously missed, which may lead to intentional steps to overcome longstanding systemic racism. Likewise, the intent and application of AI in the health sector may be to identify a specific health need that had not been previously identified (e.g., underserved populations, disease-specific outcomes). Some practical examples also include focusing on AI systems to support home-based health care for our aging, veteran, and disabled populations, as well as mental health applications. This is separate from adverse or harmful bias, and it should be recognized and leveraged accordingly.

Federal alignment of AI can build on the work done by the National Institute of Standards and Technology (NIST), and specifically that agency's guidance on identifying and managing bias in AI. In addition, this work may consider developing reliable metrics to assess the degree to which AI controls for bias.

One component that the National Strategy should consider is how to incorporate broader data equity considerations. Algorithms learn from existing data. Larger conversations regarding data representation and accuracy, within the parameters of laws and regulations protecting consumer privacy and rights, are necessary to improve AI model development and performance. It is important to advance the application of AI, while allowing innovations and room for growth and

adoption. Use of AI systems to predict and prevent inequities in the delivery of health care services should be a goal for the health sector, along with ensuring transparency.

In addition, there should be ways for the "pieces and parts" to talk with each other so that the overall goal of effective and interoperable systems can be realized. Building on the work of the United States Core Data for Interoperability (USCDI) and similar efforts can be effective for starting such projects. These efforts can also leverage and build on data content standards (e.g., collecting race, ethnicity, and other demographic data consistently) to make sure there are no unintended consequences for those impacted by AI.

Ensure the Safety and Security of AI Systems
The impact on individual privacy in the development and use of AI systems cannot be understated. The importance of protecting a person's right to privacy should be a key component of the National Strategy, with tools and resources to identify scenarios for AI systems as guidelines (e.g., appropriate de-identification of data in research, securing of health and other individually sensitive data). Validations involving human input and review across the entire design, implementation, and monitoring process should be built into AI development and deployment.

In addition, the National Strategy should include advanced verification and validation methods for AI systems, testing for high availability and safety, and new methods for identity proofing (i.e., properly and accurately identifying an individual as legitimate). In health care, much work has centered around identity proofing. While there are several solutions to identity proofing, federal support for this work and alignment with the AI priorities should be encouraged.

AHIP's Board of Directors and its Chief Medical Officers leadership team recently released core guiding priorities and a detailed roadmap to further protect the privacy, confidentiality, and cybersecurity of consumer health information. Health insurance providers have long-been leaders in developing privacy, confidentiality, and cybersecurity practices to protect personal health information. These priorities reaffirm that commitment while offering a path forward to keep Americans' health data secure and provide them with actionable health information. These concepts should be a part of AI processes and systems.

Cybersecurity must always be "top of mind" when working in electronic, connected environments. The National Strategy should specifically address AI and cybersecurity, both in terms of considerations, protections, remediation, and reporting, while leveraging AI systems and appropriate design principles (e.g., fault-tolerance) from nationally critical environments and applications. Having the ability to share information related to threats and attacks should be allowed and encouraged as part of the federal AI framework.

AHIP supports keeping all parts of the national critical infrastructure aware of threats to AI systems, schemes from nefarious actors, Nation-States, and others as a federal priority. The National Strategy should promote information sharing as a method to protect AI systems from hacking, attacks, and similar intrusions and threats. In addition, to the extent that research outcomes from the Defense Advanced Research Projects Agency (DARPA) can be applied as broadly as possible, this work should be leveraged across sectors.

We appreciate the opportunity to comment on this important topic.

Sincerely,

Danielle A. Lloyd
Senior Vice President, Private Market Innovations and Quality Initiatives, Clinical Affairs

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Aletheia AI, LLC

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.

Aletheia AI, LLC

Response of to White House Office of Science and Technology Policy and National Science Foundation RFI on the National Artificial Intelligence Research Resource

Submitted by: **Mark Beall**
CEO, Aletheia AI, LLC

Summary

Aletheia AI supports the National AI Research Resource (NAIRR) Task Force in its strategic aim of ensuring AI systems remain safe and secure. As an organization with deep expertise in AI R&D, AI safety, and AI policy, we assess that the imbalance in R&D funding for AI capabilities and AI safety poses strategic and economic risks to U.S. innovation. As of 2021, global R&D for AI safety was less than 1 percent of all AI R&D funding.

The lack of AI safety products will likely cause a major slowdown in the adoption of trustworthy AI systems, particularly as AI systems are deployed in physical systems at scale. As a result, continued U.S. leadership in AI will increasingly become a function of U.S. leadership in AI safety. We urge the NAIRR TF to take an even greater leadership role in AI safety by considering the following two recommendations:

Recommendation 1: Establish a National Center for AI Safety Research

If the United States Government could take just one step to help ensure U.S. leadership in AI, then we recommend establishing and funding a National Center for AI Safety Research Center (NCAISR). This Center would perform the vital public service of coordinating and directing national research investments in the precise area where the private sector is not yet investing. AI safety is in its infancy and early steps by the government could really shape and drive the field's development consistent with the public interest.

In support of NAIRR strategic aims 1, 3, 4, and 6, the Center could increase the safety of advanced AI systems, evaluate and develop mitigations for the malicious use and accident risks of AI capabilities, and help augment the limited private sector investment in AI safety research. Such steps would contribute directly to public safety and security. Appendix 2 gives a list of promising research directions in AI safety that NCAISR could contribute to.

The NCAISR's mission could be to:

1. Conduct research into advanced AI safety and AI alignment, both independently and in collaboration with leading academic and industry AI labs;
2. Conduct research into the effectiveness of published or proposed AI safety and AI alignment solutions; and
3. Establish grant programs for universities, academic research groups, and independent researchers pursuing research into advanced AI safety and AI alignment.

Recommendation 2: Stimulate training and human capital building in AI safety and related fields

The United States must continue to grow its workforce capacity in all relevant STEM and interdisciplinary fields, including across AI. This is especially the case for AI safety and AI alignment, specialized subjects which face a shortage of domain experts relative to the effort needed to secure current and future AI systems. **AI safety** — as distinct from AI ethics — is the scientific discipline dedicated to minimizing accident and malicious use risk from powerful AI systems (see Appendix 1). **AI alignment** is the sub-field of AI safety research dedicated to understanding and mitigating risks from future AI systems whose capabilities may rival or surpass those of humans across a broad range of tasks. The field of AI alignment is particularly underinvested by the private sector: we estimate that there are currently fewer than 100 AI alignment researchers worldwide.

To address these education gaps, the government can direct investment towards academic programs in AI safety and AI alignment. This may include degree programs, academic awards, and new competency requirements for AI safety and advanced AI risk, consistent with NAIRR strategic aims 1, 4, and 7.

We believe that the government can continue to deepen its investments in U.S. human capital by supporting the establishment of degree and training programs in AI safety and AI alignment research. This effort could involve:

1. Establishing and coordinating AI safety training programs;
2. Funding universities to establish degree programs in AI safety and AI alignment; and
3. Establishing new competency requirements for AI safety as part of computer science, AI, and digital engineering workforce efforts.

Appendix 1: AI safety

Although problems in AI ethics have received significant attention from the public and private sectors, far less policy attention and research funding have been devoted to addressing these AI safety risks. There are three broad categories of AI risk:

1. **Malicious use.** Malicious use risk refers to the risk that bad actors may use advanced AI systems to undermine U.S. interests at ever-lower costs and in novel ways. In particular, humanlike text generation, photorealistic image generation, automated code

generation, and real-time decision-automation systems each offer new attack vectors. AI's open-source culture lends itself to the rapid proliferation of powerful capabilities. In the wrong hands, these capabilities expose our democratic process, security, and economic interests to threats at a scope and scale never before seen. We assess that malicious use risk is the dominant source of AI risk. We expect this to remain the case for the next 1-5 years.

2. **Accident.** Accident risk increases as AI is deployed in safety-critical physical systems such as flight software and autonomous vehicles. As ever-greater segments of our national infrastructure are managed with AI, the impact of accidental failures escalates. We expect accidents to become the dominant source of AI risk within 1-5 years.

3. **Alignment.** Alignment refers to the technical challenge of ensuring that the most advanced AI systems behave in a way that reflects their programmers' intentions. Advanced AI systems often use dangerously creative strategies to achieve their programmed goals. These strategies can be impossible to anticipate and difficult to detect, occasionally deceiving their programmers into believing that the systems that employ them are functioning properly when they are not. As future AI systems become more capable and creative, an increasing body of evidence suggests that they may exhibit dangerous behaviors. These behaviors may become so dangerous that they lead to catastrophic outcomes, given that there is no fundamental limit to the capabilities of the most powerful AI systems. We expect alignment risk to become the dominant source of AI risk within 3-10 years.

Appendix 2: research areas in AI safety and AI alignment

Additional funding in AI safety and alignment research could accelerate AI adoption in the next 5-10 years. There are several promising research directions in AI safety and AI alignment already, and many of the research directions focused on today's malicious and accident risks would also inform AI alignment risk mitigation in the future. These research areas include:

1. **Robustness.** Robustness research aims at ensuring that AI systems are trained and evaluated in contexts that closely resemble their real-world deployment conditions. The idea is to ensure that an AI is not placed in a context in which its behavior has not been characterized during its development phase, so that it does not take unexpected actions. There are two kinds of robustness:

a. **Capability robustness.** If an AI is not capability-robust, it won't behave competently outside the contexts it was trained in. For example, a self-driving car that is only trained in sunny weather might not drive competently in snow. A failure of capability robustness can be a cause of accident risk.

b. **Objective robustness.** If an AI *is* capability-robust, but it is *not* objective-robust, then it will behave competently outside the context it was trained in, but it may pursue a goal that's different from what its designers intended. One can think of objective robustness failure in a system as that system "learning the wrong thing". Objective robustness failure is hypothesized to be a major source of alignment risk, since an AI that competently pursues the wrong objective would be doing so at odds with its designers' intent.

2. **Assurance.** This area includes approaches that help humans verify that an AI system's actions are, and continue to be, consistent with the wishes of its designers. Assurance may involve continuous monitoring of an AI system's behavior, though for very advanced AI systems there is a risk that the system may learn to conceal some of its actions from the monitoring system. For advanced AI systems, one area that appears increasingly promising is transparency:

a. **Transparency.** This involves developing a mechanistic human understanding of exactly how an AI operates and what its decision-making process is. Modern frontier AIs are built using neural networks with many layers, so this involves understanding how the layers interact, what abstractions they use, and what kinds of inputs drive which kinds of outputs.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Amazon Web Services (AWS)

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

Re: Request for Information on the *2019 Update of the National Artificial Intelligence Research and Development Strategic Plan*

Submitted By:	**Submitted To:**
Amazon Web Services, Inc. 12900 Worldgate Dr. Suite 800 Herndon, VA 20170	NCO 2415 Eisenhower Avenue Alexandria, VA 22314, USA

This document is provided for informational purposes only. It represents AWS's current product offerings and practices as of the date of issue of this document and is subject to change. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS's products or services. This document does not create any warranties, representations, contractual commitments, conditions or assurances from AWS, its affiliates, suppliers or licensors. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers. For current prices for AWS services, please refer to the AWS website at www.aws.amazon.com

Amazon Web Services (AWS) appreciates the opportunity to submit feedback to the Office of Science and Technology Policy (OSTP) in response to its Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan[1].

1. Comments on strategic aims, including suggestions to address OSTP's priorities of ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security, and to maintaining the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.

Innovation is an important goal for AI/ML (Artificial Intelligence and Machine Learning) policies. AWS supports the United States' government developing its vision and objective for the use of AI/ML. We encourage greater investment into research and development (R&D) of AI/ML (especially in areas

[1] https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and

where there are market failures and/or gaps in private sector investment), and look for opportunities to collaborate with governments in this area. At the current rate of AI/ML technology adoption around the world, AI /ML will deliver economic activity of around $13 trillion USD globally by 2030. This represents about 1.2% GDP growth per year, which is higher than the economic productivity growth brought about by either the steam engine or the early IT boom of the 2000s.[2] Long term investments in AI/ML R&D are essential to the US goal to remain a global leader in AI/ML development and deployment.

We agree with the 2019 update of *Strategy 1: Make long-term investments in AI research*, which emphasizes the importance of sustaining investment in fundamental AI research. We encourage the US to continue to invest in supporting US governments' investment in research fields relevant to AI, including cyber-defense, data analytics, detection of fraudulent transactions or messages, machine learning, robotics, human augmentation, natural language processing, interfaces, visualizations etc. Moreover, specialized computing hardware, high-quality data, and, most importantly, skilled human expertise are all essential to enabling the success of AI/ML.

In order to remain a global leader in AI/ML, the US must also build a more inclusive and innovative ecosystem that expands to industry, academia, civil society, and the Federal government. We must support and enrich the vast untapped potential of America's academic researchers, providing these researchers the same resources and access to the same infrastructure for AI/ML R&D as large companies. We must ensure that Federal funding of foundational AI/ML research does not simply take place at universities but is commercialized in industry. Funding should be dispersed to an expanded group of academia and civil society so that the pipeline of AI innovation does not deplete itself. Policies must balance the needs of all involved parties in AI/ML R&D while democratizing AI/ML research, education, and innovation.

To this end, public-private partnerships between industry, government departments, and academia are necessary to continue advancing innovations in AI/ML research. High quality data sets are crucial for research in AI. AWS currently collaborates with academia and other stakeholders through strategic partnerships with universities including but not limited to the following: University of California, Berkeley; Massachusetts Institute of Technology; California Institute of Technology; and the University of Washington. We also provide research grants through Amazon Research Awards and the joint Amazon and National Science Foundation Fairness in AI Grants program. We are also active members of multi-stakeholder organizations relating to AI, including OECD AI working groups and The Partnership on AI. Moreover, we understand the need to expand AI/ML education in non-traditional tech communities and also have our Machine Learning University, curriculum which provides anybody, anywhere, at any time access to the same machine learning courses used to train Amazon's own developers on machine learning, across community colleges nationally.

2. Suggestions of AI R&D focus areas that could create solutions to address societal issues such as equity, climate change, healthcare, and job opportunities, especially in communities that have been traditionally underserved.

A) Addressing societal challenges

[2] https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

AI/ML should be human-centric and for the benefit of all. Already, innovative companies and researchers are collaborating to use AI/ML to address societal challenges in areas such as agriculture and sustainability, healthcare, energy, manufacturing, and more. This includes our own AWS customers.

Increased AI R&D will address needs and provide solutions in healthcare. For example, AI/ML is advancing healthcare needs in the improvement of child survival rates globally. According to UNICEF, in 2019 alone 5.2 million children under the age of 5 died, with almost half being in the first month of life. A nonprofit research institute that develops scalable AI solutions to societal problems is using smartphones and AI tools to further understand newborn health issues and identify undiagnosed problems in newborns. These insights help to reduce infant mortality by ensuring newborns receive the proper health care.

When the COVID-19 pandemic hit the U.S in early spring of 2020, university researchers had already developed an image recognition model using ML to identify pneumonia in difficult-to-detect cases. Because pneumonia was quickly becoming one of the major indicators of severe infections in COVID-19 patients, AWS was asked for help setting up a system for applying the model in a clinical setting that would enable medical practitioners to use the information in diagnosis and treatment.

Health services around the world are under strain from aging populations and over-stretched resources, with the population aged 60+ expected to rise from 23% to 30% by 2050. AI provides organizations the benefits of caring for patients' long-term conditions, providing diagnostic support, improving operational and system efficiency, and supporting recovery. Recognizing the potential of this technology, some of the world's largest health organizations have set a goal of becoming a world leader for AI use in healthcare.

Agricultural technology start-ups are using AI/ML to monitor field and soil conditions and improve crop yields on farms around the world. Farming is responsible for 80% of deforestation, with an area the size of Spain cut down between 2010 and 2020 to make way for new farmland.[3] Agriculture startups use satellite imagery, sensors, and software platforms to provide farming analytics. This reduces time and money spent on field scouting and surveying techniques that are time-intensive and expensive. With this information, the farmer can make more informed decisions on which crops to grow and how best to care for them. Without the technology, it would take 49 years for one person to manually mark 21 million fields. In total, a single startup has analyzed 376,835,301 hectares of fields in the US and internationally.

Increased AI R&D can also improve access to job opportunities, especially for underserved populations. According to Commissioner Keith Sonderling, of the US Equal Employment Opportunity Commission (EEOC),

> *[when] carefully designed, AI can mask for protected classes like race, gender, age, or disability. It can hide for proxy terms, like candidates' names, the names of schools, or associations with a particular gender or racial indicative clubs. It can offset the well-documented confidence gap that leads women to under-report their abilities on resumes and men to over-state theirs. It can identify a candidate's adjacent skills, and it can identify candidates for upskilling opportunities.*

[3] https://www.greenpeace.org/usa/forests/issues/agribusiness/

In short, AI can determine the best candidates based not only on their merit but also on their potential while stripping out human bias.[4]

3. Comments regarding how AI R&D can help address harms due to disparate treatment of different demographic groups; and AI R&D to evaluate and address bias, equity, or other concerns related to the development, use, and impact of AI.

The development and deployment of AI/ML should place humans at the center. To address concerns of bias, AI/ML deployments should support inclusivity, diversity of thought, and collaboration to maximize the benefits of the technology for all of society.

AWS published our Responsible Use of Machine Learning guide that provides recommendations for responsibly developing and using ML systems, including mitigating bias and addressing other potential harms[5]:

- We encourage developers and deployers to have teams with diverse backgrounds, perspectives, skills, and experiences. AI/ML researchers and users should assess whether teams include a wide array of genders, races, ethnicities, abilities, ages, religions, sexual orientations, military status, backgrounds, and political views. They should further assess whether teams may have gaps and consider adding underrepresented perspectives to fill those gaps to enhance performance.
- When collecting and evaluating data to develop and test models, we encourage developers to consider its completeness, representativeness, and breadth. Diversity of data is often important for use cases that involve personal characteristics like race and gender, but can also apply in non-obvious contexts. Develop mechanisms to evaluate whether the data appropriately represents real world use, and collect and test additional data to address underrepresented attributes.
    - We recommend that developers implement processes to understand where bias may be introduced by developers and mitigate human error. Create fairness goals and metrics to measure performance across different subgroups, communities, and demographics applicable to the use case, and test and measure progress against those metrics.
- We recommend to deployers of AI/ML tools that they consider whether human review or oversight over the operation of the system may be appropriate or necessary (e.g., in situations where ML systems may be used in a manner that impact human rights or safety), and if so, how to best incorporate such human input into the overall operation of the system. Human reviewers should be appropriately trained on real world scenarios, including examples where the system fails to properly process inputs or cannot handle edge cases, and have ways to exercise meaningful oversight.

Addressing potential bias concerns in emerging technologies requires collaboration from a diverse group across industry, civil society, academic, and government. R&D should focus on developing data

[4] https://www.ihrim.org/2021/12/how-people-analytics-can-prevent-algorithmic-bias-by-commissioner-keith-e-sonderling/

[5] https://aws.amazon.com/machine-learning/responsible-machine-learning/

driven techniques, metrics, and tools that industry can operationalize to improve how they measure and mitigate bias in concrete terms.

4. Comments on AI R&D to help address the underrepresentation of certain demographic groups in the AI workforce.

Greater access to AI/ML education and training will lead to a more diverse workforce. The US government needs to support large-scale educational initiatives to ensure increased access for people, especially for underrepresented groups. This additional investment by the Federal government will allow the workforce to more accurately mirror a diverse population. Exposing students, particularly K-12 students, to AI/ML education equips them with the lifelong skills that are necessary to build successful careers in AI/ML fields. By addressing the gaps in education across the nation, which lead to an underrepresentation of certain groups in the AI workforce, we will see a more diverse group of people entering the workforce.

AWS is committed to help train the future generation by investing in AI/ML education, and is investing in programs that serve communities that have rarely had access to this type of education. Amazon's primary computer science focused initiative is Amazon Future Engineer—a four-part childhood to career program aimed at educating 10 million students from underrepresented and underserved communities each year to try computer science and coding. We encourage developing and increasing access to AI/ML training as a way to see an increase in diversity within the workforce in the future.

AI/ML trainings should target colleges and community colleges that are non-technical to expand access to skill development and ensure that students who are not often the targets of technical trainings have the ability to pursue careers in AI/ML. This can be accomplished by supporting continuing education programs focused on AI/ML in nontraditional tech communities which will ultimately increase the diversity of the AI workforce. For example, MLU has been used for over six years to train our engineers, and last year much of this content has been available for free to customers and programs in community colleges nationally.

Increased accessibility to AI R&D will also help reach new audiences and increase diversity in tech. For example, to help researchers learn about cloud computing, AWS curated a list of no-cost, on-demand online courses tailored to researchers' needs that are widely accessible. The AWS research team selected this list of courses from hundreds of available courses, specifically those who want to learn foundational cloud services. These online courses are available at any time to help users learn new cloud skills and services. AWS helps researchers process complex workloads by providing the cost-effective, scalable, and secure compute, storage, and database capabilities needed to accelerate time-to-science. Scientists can quickly analyze massive data pipelines, store petabytes of data, and share their results with collaborators around the world.

5. Comments on strategic directions related to international cooperation on AI R&D and on providing inclusive pathways for more Americans to participate in AI R&D. Additionally, comments are invited as to existing strategic aims, along with their past or future implementation by the Federal government.

A) Support for global AI standards

Geopolitical realities have contributed to growing recognition of the importance of government participation for enhancing cooperation on digital technical standards. Work on standards development also offers an opportunity to accelerate a broader joint technology agenda. The recent G7 Digital and Technology Ministerial Declaration was a particularly solid endorsement, specifically its detailed annex on collaboration on digital technical standards. The declaration's endorsement of multi-stakeholder and *industry-led standards development*—and of standards consistent with open, democratic societies—provides a robust framework to guide collaboration. As such, the US should seek to expand strategic cooperation on standards development among the US, EU, Canada, and the UK, among like-minded countries, and among states that are undecided on the direction of their technology governance, including in the Global South.

To ensure greater diversity and representativeness in Standard setting bodies, the US should provide financial and technical support for subject matter experts (SMEs) and other stakeholders less likely to otherwise participate in the process. Working with like-minded countries, the US should work on joint coordination and identification of the types of technical capacity required to advance specific standards, and then seek out the corresponding availability of such know-how in underrepresented organizations. Understanding the realities of the settings where these talents are located will help identify the correct policy and financing mechanisms to incentivize participation (tax credits to pre-revenue startups, for example, will be of little value). Indeed, efforts to render SDOs more accessible to voices across the stakeholder spectrum will help serve their core mission to produce strong technical standards that enhance competitiveness, innovation and the public interest more broadly.

Indeed, it is our view that there are still numerous AI standards to be launched and developed and that many AI standards currently in advanced development (ISO 42001 on AI Management Systems, for instance) will require further refinement and precision, likely based on certain categories of high-risk use-cases. This will require work and deep technical expertise.

Finally, the US and its closest allies will have to look beyond standards alone if they aim to drive support for—and broader global adoption of—their AI governance roadmap. For instance, the US should engage their respective cooperation agencies to address technology development issues and digital infrastructure plans across the Global South. The UK, for instance, has been active in cyber capacity-building across partner countries in the Global South. Efforts of this nature should be broadened and reinforced, *alongside* parallel efforts to address the digital technical standards meant to underpin AI governance. This reflects a call to action of the National Security Commission on Artificial Intelligence (NSCAI).

We support the work that NIST performs in benchmarking, but continue to advocate for improved testing methodologies and resources that would allow NIST to perform benchmarking directly evaluating cloud hosted capabilities.

B) International Collaboration with foreign bodies that have AI principles

AWS is working closely with the OECD to demonstrate their implementation through various tools, including our own responsible use guidelines and more technical tools, such as Amazon SageMaker Clarify. Moreover, AWS is actively involved in international AI standards setting organizations as well as collaborating on international draft legislation like the EU AI Act.

6. Comments on existing strategic aims, along with their past or future implementation by the Federal government.

Since the 2019 update, multiple federal agencies have incorporated ethics principles to guide the way in which AI is adopted and applied. For example, the Department of Defense adopted in 2020 and reaffirmed in 2021 *Ethical Principles for Artificial Intelligence*, while the intelligence community adopted *Principles of Artificial Intelligence Ethics for the Intelligence Community* as well as the *Intelligence Community Artificial Intelligence Ethics Framework.* We encourage the 2022 update of the Strategic Plan to reference these efforts.

We thank you for the opportunity to respond.

Sincerely,

Shannon Kellogg

Vice President, AWS Public Policy – Americas

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

American Psychological Association (APA)

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.

AMERICAN PSYCHOLOGICAL ASSOCIATION

March 4, 2022

AI R&D RFI Response Team
2415 Eisenhower Avenue
Alexandria, VA 22314, USA

Submitted electronically via Regulations.gov

RE: RFI Response: National Artificial Intelligence Research and Development Strategic Plan

To Whom It May Concern –

The American Psychological Association (APA) appreciates the opportunity to comment on the National Artificial Intelligence Research and Development Strategic Plan require for information. This request represents a step in the right direction towards ensuring that stakeholders across disciplines are represented in future efforts to deploy artificial intelligence. In addition to the comments below, APA endorses the comment submitted from the Society for Industrial and Organizational Psychology.

APA is the largest scientific and professional organization representing psychology in the U.S., numbering over 133,000 researchers, educators, clinicians, consultants, and students. For decades, psychologists have a played vital role in the development and deployment of technologies and neurological science. These contributions have been essential to the currently available artificial intelligence enabled technologies and psychological science should continue to be at the heart of strategic planning of AI deployment.

The comments below represent three primary areas where the *Strategic Plan* should ensure the discipline of psychology is included: increased investments in research on artificial intelligence, ethics of artificial intelligence, and artificial intelligence and implicit bias.

Increased Investments in Research on Artificial Intelligence

APA strongly supports the need for additional investments in research related to Artificial Intelligence. From the current technological and research standpoint, it is almost impossible to predict the impact of future AI-informed technologies. There is an imperative that as the technologies grow in their capabilities and prevalence, that research surrounding their impact also increases. Future research funding in this area should ensure that psychological and behavioral science is adequately represented. The impact of AI on mental and behavioral health must continue to be examined to ensure we mitigate any harmful impacts caused by new systems.

750 First Street, NE
Washington, DC 20002-4242

AMERICAN PSYCHOLOGICAL ASSOCIATION

Ethics of Artificial Intelligence

There are some fundamental research opportunities the AI research community must investigate. AI Ethics and Psychology is an evolving discipline essential to the study of how AI learns from society and humans and how AI makes consequential decisions in critical settings.[1] Studies have demonstrated that AI automatically learns implicit biases from language corpora and accordingly perceives the world in a biased manner.[2] These implicit biases that have been documented in social psychology for decades include racial, gender, sexuality, ability, and age attitudes.[3] Moreover, these findings provide insights about how language might be impacting the social cognition of both AI and humans.

There are, additionally, ethical implications for what AI learns, how AI learns, and AI's subsequent decision-making. For example, developing transparency enhancing algorithms for measuring and simulating AI bias and equity would make it possible to analyze the ethical implications of AI in a variety of domains including natural language and computer vision.[4] Alternatively, these AI methods could examine and analyze current and historical social and human cognition.[5] This research program would allow for understanding how AI is co-evolving with humanity, as AI is shaping society and impacting individuals' lives in an accelerated manner and at an unprecedented scale. While beyond the scope of this Request for Information, there remains no comprehensive regulation for auditing how AI impacts equity and fairness in democratic societies.[6] Consequently, these promising research areas of computer and

[1] Caliskan, A., Bryson, J.J., & Narayanan, A., (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356(*6334), 183-186. 10.1126/science.aal4230.

[2] Pandey, A., & Caliskan, A., (2021). *Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms.* In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 822-833.

[3] Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4; Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464.

[4] Steed, R., & Caliskan, A. (2021). A set of distinct facial traits learned by machines is not predictive of appearance bias in the wild. *AI Ethics 1***,** 249–260. https://doi.org/10.1007/s43681-020-00035-y

[5] Caliskan, A., & Lewis, M. (2020, July 16). Social biases in word embeddings and their relation to human cognition. https://doi.org/10.31234/osf.io/d84kg

[6] Caliskan, A. (2021, May 10). *Detecting and mitigating bias in natural language processing*. Brookings Institution. https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/

information science contribute to data-driven policy making and law while having implications for psychology, political science, sociology, linguistics, and philosophy.

Artificial Intelligence and Implicit Bias

Given evidence that AI can reproduce discrimination and bias against individuals and groups, it is imperative to leverage psychological science and examine people's expectations about and reactions to the fairness and potential discrimination of AI versus human agents. An emerging line of research suggests that people expect AI to be less biased than humans in some cases and are less outraged when they learn of bias from an AI versus human actors.[7] Algorithms appear less discriminatory than humans, perhaps incorrectly engendering trust and comfort from human users. The early evidence shows that decisions about AI and how it is implemented reflect the world view and values of the human beings who design them and set policy for how it is used. Given the massive and increasing influence of AI on people's lives, it is critical to better appreciate how people understand and react to such influence, especially when the AI is perceived to be biased or unfair.

Without the help of psychological science, we risk harming already disadvantaged populations and creating systems that perpetuate harmful stereotypes and bias. AI systems are often trained using large data sets of human attributes or demographics that have the potential to integrate biases related to gender identity, race, and other characteristics. These systems then spread the biases in their interactions with humans or other technology-informed systems, with implications for equity and fairness. Psychologists' research on the various forms of resulting bias and the detrimental impacts are being used to develop data sets that are less biased and AI systems that can detect and compensate for biases in data. Findings from this research should be incorporated into future deployments of artificial intelligence tools, especially when being funded or used by the federal government.

While we remain broadly supportive of the strategic aims set forward by the *National Artificial Intelligence Research and Development Strategic Plan*, it is important that psychological and behavioral science is included in each strategy to ensure comprehensive consideration of the broad impact of AI technologies.

APA again thanks you for the opportunity to comment on this policy. If APA can be of any further assistance, please contact Corbin Evans, Senior Director of Congressional and Federal Relations, at [redacted].

[7] Jago, A. S., & Laurin, K. (2021). Assumptions About Algorithms' Capacity for Discrimination. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/01461672211016187

AMERICAN
PSYCHOLOGICAL
ASSOCIATION

Katherine Burnett McGuire, MS
Chief Advocacy Officer, APA

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Anthropic

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.

ANTHROP\C

March 4, 2022

Subject: Anthropic Comment Regarding "Update of the National Artificial Intelligence Research and Development Strategic Plan"
Reference: 87 FR 5876, Document Number 2022-02161

Update of the National Artificial Intelligence Research and Development Strategic Plan

Anthropic welcomes the opportunity to provide feedback to the Office of Science and Technology Policy (OSTP) in response to a Request for Information (RFI) on updates to the National Artificial Intelligence Research and Development Strategic Plan. Our submission focuses on recent developments in artificial intelligence (AI) research, and how the federal government can foster a more competitive research and development (R&D) environment through additional support of four existing strategies.

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. We're an organization with backgrounds in research, engineering, and policy, and we approach AI development from a cross-disciplinary perspective. Our founding team previously worked at OpenAI, where they helped develop a large-scale language model called "GPT-3," which played a key role in the recent rise of more general AI systems.

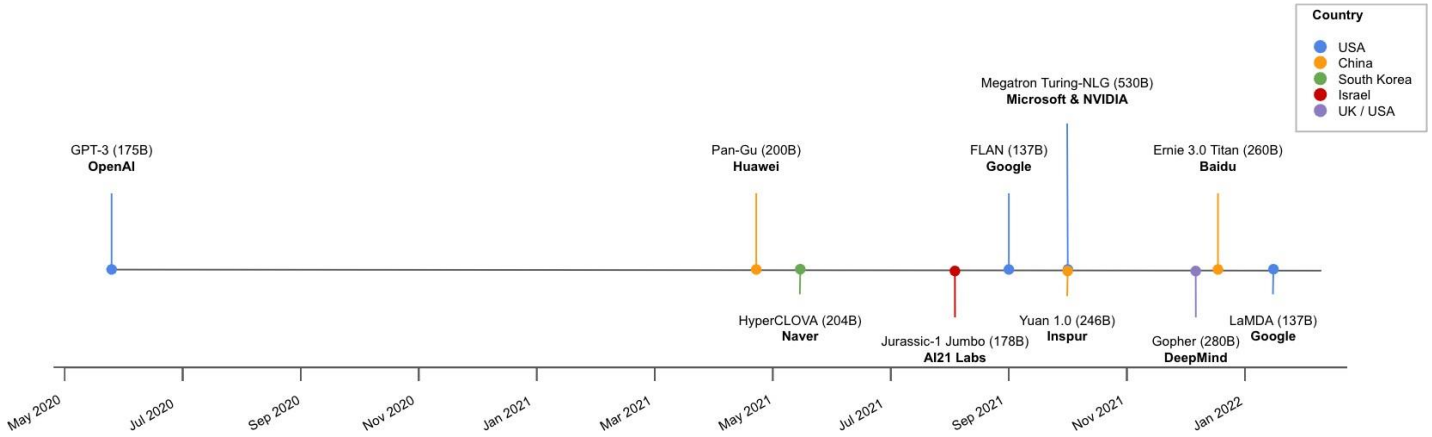
In this response, we provide a brief overview of some of the technical advancements made since the 2019 Update that demonstrate just how quickly progress moves in AI. We argue that who drives (and ultimately benefits from) this progress has become increasingly unequal over the past decade, with a larger concentration of AI development driven by a small number of industry actors. As a result, the development of AI systems and the broader ecosystem that surrounds them are primarily influenced by commercial priorities, which do not necessarily reflect the needs of the public.

The lack of broad public participation in AI development, coupled with an insufficient AI assurance ecosystem, threaten not only a competitive R&D environment, but also public trust in AI more broadly. The federal government can rectify these challenges by bolstering support for current areas in the Strategic Plan: long-term investments in AI research (Strategy 1), shared public infrastructure for AI training and testing (Strategy 5), developing trusted AI measurement initiatives (Strategy 6), and using those efforts to validate that AI systems are safe and secure (Strategy 4).

Since the 2019 Update, AI Research has Made Tremendous, Unexpected Progress

As articulated in the 2019 Update, progress in AI was – until that point – largely driven by "narrow" AI systems that demonstrated strong performance on specific tasks (typically basic classification or pattern-recognition tasks), while more general-purpose AI systems remained elusive. Less than one year after the 2019 Update was written, OpenAI released a language model called "GPT-3"[1], a system that could read, write, and classify text, and could be programmed via people giving it instructions in natural language.

GPT-3 and the models that followed (e.g. Microsoft and NVIDIA's "Megatron Turing-NLG"[2], DeepMind's "Gopher,"[3] etc.) represent a new class of general AI systems. Crucially, they are *generative* AI systems – instead of simply classifying or recognizing patterns in existing data, AI can now *generate* original data. This data often takes the form of written content (or for some models, synthetic imagery), and can be produced at such a high degree of quality that it is, in some cases, indistinguishable from data created by humans. These models represent a fundamental shift in the capabilities of AI systems and have altered the trajectory for frontier research and development in AI.



Timeline of public disclosures of GPT-3 scale dense language models[4]

This new class of general models has made remarkable progress towards a number of goals described in the 2016 Strategic Plan, which were considered "still far" off at the time the Plan was written. These systems come much closer to the "flexibility and versatility of human intelligence in a broad range of cognitive domains," than what was envisioned in the Strategic Plan and its subsequent Update. For example, they can increase the efficiency of software

[1] Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. *arXiv*. https://arxiv.org/abs/2005.14165
[2] Smith, S., et al. (2022). Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *arXiv*. https://arxiv.org/abs/2201.11990
[3] Rae, J. W., et al. (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv.* https://arxiv.org/abs/2112.11446
[4] Ganguli, D., et al. (2022). Predictability and Surprise in Large Generative Models. *arXiv*. https://arxiv.org/abs/2202.07785

engineers in computer programming tasks[5] and generate creative and original images from plain text descriptions[6]. In the three years since the Update, developers have built a growing number of systems that can do the tasks the Strategic Plan considered aspirational – in the case of generative language models (e.g. those represented above), today's frontier AI systems are able to read and summarize text, answer questions about complex technical subjects, and generate original prose.

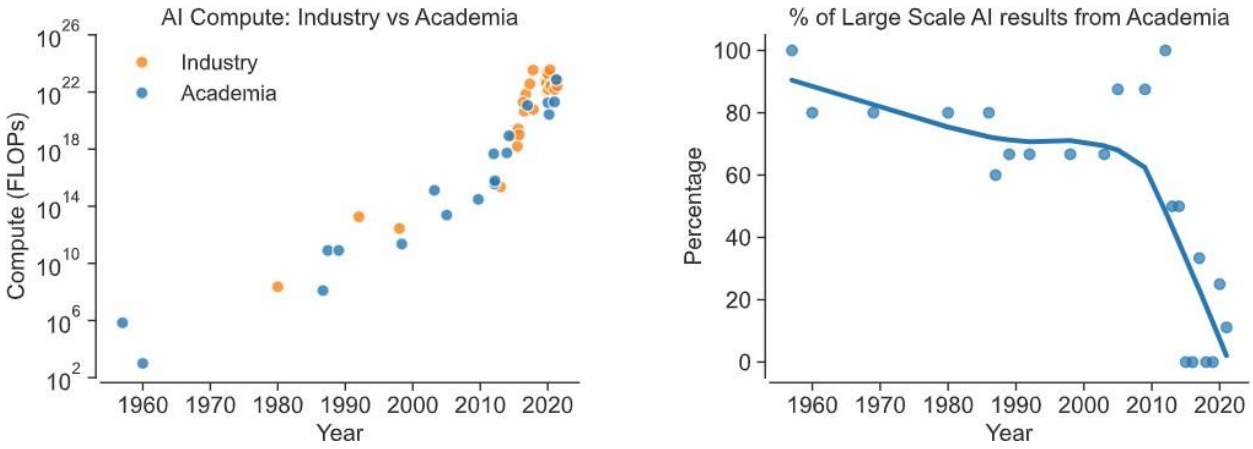
Research and development in AI has progressed far more quickly than what was anticipated in the 2016 Strategic Plan and the subsequent 2019 Update; we now have systems that can successfully complete a variety of tasks without explicit training. This rate of progress, potentially accelerated by enhanced government support, suggests a future of increasingly powerful general AI systems. **Breakthroughs in AI R&D now occur far too quickly for national strategic plans to be updated on a three-year cadence;** without detailed and timely information about AI progress, the federal government risks missing critical research priorities and areas for additional investment[7]. A more frequent review of the Strategic Plan could help translate real-time research findings into actionable recommendations that support a competitive R&D environment.

Rapid Progress in AI Has Been Isolated to a Handful of Private Organizations

Despite the recent influx of large-scale general systems, only a small number of highly-resourced organizations are able to develop them. Today, these models are developed by private companies, either established technology firms or smaller startups, with academic institutions and public sector organizations notably missing. This is largely due to the resource-intensity of these systems. Developers must have access to sufficient amounts of computing power (compute) to train these models, which can cost on the order of several millions of dollars[4] and far exceed academic research budgets for this sort of work.

Beyond the significant cost, AI systems have become so large that they're increasingly unwieldy to train. Where systems were once able to fit on a single processor, today's frontier systems must be distributed across massive clusters of processors working in parallel, requiring highly sophisticated software engineering skills[8]. Unlike other stakeholders in AI development, industry actors have access to both large R&D budgets and exceptional engineering talent. This has created a dynamic where a small number of industry actors drives an increasing portion of computationally-intensive AI research.

[5] GitHub. (2021). *GitHub Copilot*. https://copilot.github.com/
[6] OpenAI. (2021, January 5). *DALL·E: Creating Images from Text.* https://openai.com/blog/dall-e/
[7] Whittlestone, J., & Clark, J. (2021). Why and How Governments Should Monitor AI Development. a*rXiv*. https://arxiv.org/abs/2108.12427
[8] Lohn, A., & Musser, M. (2022). AI and Compute - How Much Longer Can Computing Power Drive Artificial Intelligence Progress? *Center for Security and Emerging Technology.* https://cset.georgetown.edu/publication/ai-and-compute/



(Left) *The amount of compute required by major AI projects over time is increasing exponentially for both academic (blue) and industrial (orange) projects.* ***(Right)*** *The proportion of large-scale AI results from academia is steadily decreasing. Blue curve represents a Lowess fit to the data*[4].

As costs to build frontier AI systems have grown largely out of reach for academic stakeholders, public funding for university R&D has lagged, exacerbating an already unequal playing field for model development. With data from the Organization for Economic Cooperation and Development (OECD), ITIF found that the United States ranked 24th out of 36 nations in terms of government funding for university R&D, as a portion of gross domestic product. The study also found that the United States would need to spend an additional $90 billion per year to match the university R&D spending of the 1st place country (Norway)[9].

These two interrelated factors – increased costs to build advanced AI systems and insufficient public funding for non-commercial research – have created a vastly unequal R&D landscape that requires federal intervention. Without additional investment directed towards public organizations, the future of AI development will be controlled by a handful of private actors, primarily motivated by commercial interests.

Commercial Incentives Influence the Systems that Get Built and the Broader Ecosystem

Due to the incentives of private industry, corporate organizations may be more inclined to prioritize profitable deployments over systems with fewer economic use cases but broader societal benefit. In contrast, academia may be motivated more by the pursuit of knowledge than profit, and has more immediate access to varied expertise for interdisciplinary research and evaluation of AI systems[4]. A thriving R&D environment – where both industry *and* academia

[9] Atkinson, R. D., & Gawora, K. (2021, April 12). *U.S. University R&D Funding Falls Further Behind OECD Peers*. Information Technology & Innovation Foundation. https://itif.org/publications/2021/04/12/us-university-rd-funding-falls-further-behind-oecd-peers

can participate fully – will require a concerted effort to bolster Strategies 1 and 5 of the Strategic Plan.

Beyond having an outsize role in determining the types of systems that get created, industry actors aren't generally in the business of creating measurement and monitoring infrastructure for a broad range of AI technologies. While individual companies may create methods to evaluate their own systems in isolation, there's little incentive to build the broader ecosystem to critically evaluate AI models for safety or public benefit. Where other stakeholders (e.g. public sector, academia, civil society, etc.) might otherwise step in and fill that void, they lack the resources to build these models and access them directly.

We refer to this measurement and monitoring infrastructure as an "assurance ecosystem," or a set of overlapping protection mechanisms that provide certainty and trust that AI systems will operate as intended. With an orientation towards the public interest, the federal government should be the primary stakeholder to develop this ecosystem of AI assurance, and can work in collaboration with academic researchers and civil society.

While both the 2016 Plan and 2019 Update highlight the need for AI evaluations to verify system safety, more work remains to be done. **We must develop standardized performance and safety indicators, be able to accurately and uniformly measure those indicators, and create a process to regularly evaluate AI models at scale.** This system would need to be built with adaptability in mind, in order to evolve measurement techniques with new developments from the research community. By placing renewed emphasis on Strategies 4 and 6 of the Strategic Plan, the federal government can stay more attuned to real-time progress and foster a responsible R&D environment.

The Government Should Support Broader Participation in AI Development (Strategies 1 & 5)

We echo the concern in the 2016 Plan that states "progress [will] suffer if AI training and testing is limited to only a few entities," and unfortunately, this consolidation is already apparent in the recent releases of large-scale AI systems. The federal government can rectify this imbalance by creating public experimental infrastructure – a shared research environment that includes access to compute resources and datasets through a centralized user interface. An initiative of this sort could facilitate large-scale AI experimentation and model development for academic stakeholders, and it is the strongest way to support long-term investments in AI research (Strategy 1).

Since the 2019 Update was written, Congress passed the National AI Initiative Act of 2020, which among other efforts, directed the National Science Foundation (NSF) and the Office of Science and Technology Policy (OSTP) to form a National AI Research Resource (NAIRR) Task Force to design a roadmap for how this kind of shared public infrastructure could be created at

scale. The NAIRR represents an opportunity to both restore a healthy balance between industry and academic contributions to AI R&D, and also to include researchers that have historically been underrepresented in AI.

Anthropic firmly supports the goals of the NAIRR; we view its future establishment to be the successful manifestation of Strategy 1, and the ideal environment to host the compute resources, datasets, and testbeds outlined in Strategy 5. Without the NAIRR, academic researchers would need to negotiate and manage compute access with cloud service providers or high-performance computing centers, spend valuable research time on software management needs, and work only with open source data or data available within their institution. Instead of accessing communal resources through centrally-managed infrastructure, every academic research effort would need to tackle these hurdles independently, ultimately taking away from time spent on innovative R&D.

The need for sufficient computational resources and high-quality datasets reflected in the 2019 Update to Strategy 5 has become even more pronounced in this new era of general AI systems. As described above, the amount of compute required for major AI projects has rapidly increased over time, with general AI systems being among the most compute-intensive. With sufficient funding, the NAIRR can help reintegrate the academic stakeholders that have been pushed out of frontier research due to rising costs in model development. To ensure academia can build and research the kinds of general systems currently developed by industry actors, the NAIRR should reserve a non-trivial percentage of compute resources for a handful of industry-scale projects.

Strategy 5 accurately notes that the resources to train and test AI systems – including compute, quality datasets, and interactive testbeds – is a "significant 'public good' challenge". While private industry has access to an abundance of proprietary data from deployed commercial products, AI researchers in academia typically rely on open source datasets or data available within a university setting. A shared experimental resource such as the NAIRR could provide secure access to public sector data for research in areas with broad societal relevance (e.g. healthcare, climate, the economy, etc.), rather than commercial interests of the private sector[10].

Competitive AI R&D Requires Robust System Assurance (Strategies 4 & 6)

If the federal government wants to ensure the U.S. AI research and development community is working in the interests of both the economy and the nation at large, then the government will need more detailed and timely information about how AI research is progressing. Specifically, the government should seek to measure and monitor the AI research landscape (or work with partners who can), to give it better information about the rapidly evolving state of the AI ecosystem. In doing so, the government will be better equipped to evaluate new research

[10] Stanford University Human-Centered Artificial Intelligence. (2021, October). *Building a National AI Research Resource.* https://hai.stanford.edu/white-paper-building-national-ai-research-resource

priorities, compare the strength of the US R&D landscape relative to other countries, and create standardized approaches to impact assessments and system assurance[7].

In addition to building out dedicated monitoring infrastructure, the federal government can continue to collaborate with external experts to better understand current research efforts at the frontier of the field. Through public comment periods and advisory bodies such as the National AI Advisory Committee (NAIAC), a range of stakeholders can present a more comprehensive and representative account of the AI R&D landscape. Without the real-time insights made possible through regular monitoring and public participation, the information asymmetry between the private and public sectors will widen. This asymmetry increases the risk of unsafe deployments, reactive regulatory responses, and the development of AI evaluation programs that benefit commercial actors rather than the public[7].

Beyond keeping the government informed of the speed of AI progress, investing in an assurance ecosystem will allow multiple stakeholders the ability to properly evaluate and verify AI systems for performance and safety. System assurance provides model developers with certainty in the reliability of their models, end users with trust that models will act as intended, and government stakeholders with confidence that systems are safe for the general public. Effective evaluation is also a necessary prerequisite for meaningful regulation. For a wide variety of other powerful technologies (e.g. aviation, food and drugs, vehicles, etc.), we have dedicated institutions that develop product safety standards and rigorously test systems for compliance, yet we lack the equivalent for AI.

Assurance is critical not only for verifying system safety – it also enables stronger R&D. Strategy 6 of the 2016 Plan correctly recognizes that "benchmarks [and evaluations] drive innovation by promoting advancements aimed at addressing strategically selected scenarios; they additionally provide objective data to track the evolution of AI science and technologies". In the six years following the release of the 2016 Strategic Plan, we've witnessed incredible progress in the field of AI but unfortunately, the development of objective standards and benchmarks has not kept pace with technological progress.

While isolated measures of performance and safety exist for specific domains, no comprehensive assurance system exists for the vast range of AI technologies in development and in use today. For certain tasks such as image recognition, the AI & ML community can turn to standardized datasets and evaluations (e.g. ImageNet[11]) to measure the performance of computer vision models. Drawing on this foundation, we must build standardized ways to measure and test the potential social effects of these systems — Do they exhibit bias? Do they act in accordance with widely-held human values? Do they share sensitive data that may have been encountered during model training?

[11] *Imagenet.* Retrieved Mar 1, 2022, from https://www.image-net.org/index.php

We see this work as best suited to the public sector and organizations such as the National Institutes of Standards and Technology (NIST), with support from the broader research community. The development of AI testbeds, created and managed by NIST, would allow for the comparative evaluation of different AI models using centralized datasets and standardized testing protocols. In addition to building testbeds directly, NIST could validate and collate evaluations created by independent researchers to create a catalog of NIST-approved tests for deployed models and those in development. These are significant efforts; NIST and others charged with this work must be properly resourced in both infrastructure and expert personnel.

In parallel with efforts to build standards and benchmarks, the federal government should provide financial grants to researchers specifically interested in building measures of AI assurance and could do so through initiatives such as the NAIRR. A significant allocation of research grants could be reserved for researchers developing assurance indicators with wide societal relevance, including measures of model bias or accuracy. Broader participation in the development and evaluation of AI systems will lead to a more comprehensive and representative ecosystem for AI assurance. Being able to objectively measure and validate how AI systems perform (Strategy 6) will be a critical component to ensure their safety and security (Strategy 4).

Conclusion

In the time since the 2019 Update was written, we've seen a rapid emergence of increasingly capable, general AI systems. These models demonstrate a surprising range of capabilities, ones they were never explicitly trained to do. They're used to augment the work of professionals and hobbyists, and they present us with expanded possibilities for what it means to *generate* outputs that are creative, valuable, and socially relevant. When a diverse network of people can play an active role in the development and oversight of these systems, we're far more likely to have AI technologies that can be used to tackle meaningful challenges with broad public interest.

The Plan's strategic priorities – if supported fully – are directionally aligned with the needs of a future where these systems are integrated into multiple facets of everyday life. The government can accelerate progress in long-term, fundamental research (Strategy 1) by expanding public access to experimental infrastructure (Strategy 5). By creating a robust ecosystem for AI monitoring and assurance (Strategies 4 and 6), the government can ensure that those advancements are safe, trusted, and broadly beneficial.

Thank you for your work on this critical topic, and for the opportunity to provide input.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Association for Computing Machinery (ACM)

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

COMMENTS IN RESPONSE TO RFI TO THE UPDATE OF THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN (DOCUMENT NUMBER 2022-02161)

The non-profit Association for Computing Machinery (ACM), with more than 50,000 U.S. members and approximately 100,000 worldwide, is the world's largest educational and scientific computing society. ACM's US Technology Policy Committee (USTPC), currently comprising more than 160 members, serves as the focal point for ACM's interaction with all branches of the US government, the computing community, and the public on policy matters related to information technology. It is charged with providing policy and law makers throughout government with timely, substantive and apolitical input on computing technology and the legal and social issues to which it gives rise.[1]

In response to the Office of Science and Technology Policy's Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan of February 1, 2022 (RFI),[2] USTPC is pleased to submit the following comments:[3]

First, while we support all eight of the strategies outlined in the "The National Artificial Intelligence R&D Strategic Plan: 2019 Update," we are pleased that the overall plan is being reviewed and updated. We especially encourage a focus on strategies 3 and 4: "understand and address the ethical, legal, and societal implications of AI" and "ensure the safety and security of AI systems," respectively. The Committee notes that building systems that achieve these aims is difficult. We believe, therefore, that emphasizing and enabling research to advance the field of accountable AI system design is especially important.

[1] To arrange for a technical briefing from USTPC and other ACM expert members, please contact Adam Eisgrau, ACM Director of Global Policy & Public Affairs, at acmpo@acm.org or 202-580-6555.

[2] *See* 87 FR 5876 (February 2, 2022) at https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and.

[3] The lead author of these Comments for USTPC was its Artificial Intelligence & Algorithms Subcommittee Chair Prof. Jeanna Matthews of Clarkson University. Also contributing were USTPC members L. Jean Camp, Charalampos Chelmis, Thomas Chen, Carlos Jiménez, Arnon Rosenthal, Ben Schneiderman, and Kenneth Zhang.

Second, we recommend that the Strategic Plan adopt the broadest possible definition of artificial intelligence to include, specifically, automated or algorithmic decision-making systems more broadly. This is appropriate and necessary because, when automated systems are used to make critical decisions impacting society and the lives of individuals, the ethical, legal, societal, safety and security issues are similar regardless of the complexity or interpretability of the algorithms. Analysis of safety and security should include comprehensive evaluation of data compilations used for training, the accuracy of decision-making systems, and the potential for the abusive use of platforms.

Third, we encourage revision of the current plan to rank tiers of systems based on the critical nature of their impact on individuals and society and to hold systems classified in higher tiers to proportionately higher standards of verification and validation, testing, documentation and explanation. The criteria for determining the level of rigor applied to a system should be dependent on its impact on individuals and society, rather than the complexity of its algorithms, or of the size or nature of the company producing it. Automated decision-making systems impacting human life and liberty should be held to the highest standards including independent verification and validation, audit trails, and retrospective analyses of failures. The same should be true for systems deployed in high impact or highly regulated areas such as hiring, housing, credit, and the allocation of public resources, and others.

Finally, in updating the National Artificial Intelligence Research and Development Strategic Plan, we respectfully commend the agency's attention to the attached Statement on Algorithmic Transparency and Accountability[4] and its seven associated principles: 1) awareness; 2) access and redress; 3) accountability; 4) explanation; 5) data provenance; 6) auditability; and 7) validation and testing. The Statement is a joint product of ACM's Europe and US policy committees.

ACM's US Technology Policy Committee looks forward to assisting OSTP, NSF and other agencies throughout the process of reconsideration and revision of the 2019 Strategic Plan and welcomes all inquiries to that end. For further information, or should you have any other questions, please contact ACM's Director of Global Public Policy, Adam Eisgrau, at 202-580-6555 or eisgrau@acm.org.

[4] https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Association for Computing Machinery
US Public Policy Council (USACM)

usacm.acm.org
facebook.com/usacm
twitter.com/usacm

January 12, 2017

Statement on Algorithmic Transparency and Accountability

Computer algorithms are widely employed throughout our economy and society to make decisions that have far-reaching impacts, including their applications for education, access to credit, healthcare, and employment.[1] The ubiquity of algorithms in our everyday lives is an important reason to focus on addressing challenges associated with the design and technical aspects of algorithms and preventing bias from the onset.

An algorithm is a self-contained step-by-step set of operations that computers and other 'smart' devices carry out to perform calculation, data processing, and automated reasoning tasks. Increasingly, algorithms implement institutional decision-making based on analytics, which involves the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming, and operations research to quantify performance.

There is also growing evidence that some algorithms and analytics can be opaque, making it impossible to determine when their outputs may be biased or erroneous.

Computational models can be distorted as a result of biases contained in their input data and/or their algorithms. Decisions made by predictive algorithms can be opaque because of many factors, including technical (the algorithm may not lend itself to easy explanation), economic (the cost of providing transparency may be excessive, including the compromise of trade secrets), and social (revealing input may violate privacy expectations). Even well-engineered computer systems can result in unexplained outcomes or errors, either because they contain bugs or because the conditions of their use changes, invalidating assumptions on which the original analytics were based.

The use of algorithms for automated decision-making about individuals can result in harmful discrimination. Policymakers should hold institutions using analytics to the same standards as institutions where humans have traditionally made decisions and developers should plan and architect analytical systems to adhere to those standards when algorithms are used to make automated decisions or as input to decisions made by people.

This set of principles, consistent with the ACM Code of Ethics, is intended to support the benefits of algorithmic decision-making while addressing these concerns. These principles should be addressed during every phase of system development and deployment to the extent necessary to minimize potential harms while realizing the benefits of algorithmic decision-making.

[1] Federal Trade Commission. “Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues.” January 2016. https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report.

Principles for Algorithmic Transparency and Accountability

1. Awareness: Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

2. Access and redress: Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.

3. Accountability: Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

4. Explanation: Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.

5. Data Provenance: A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorized individuals.

6. Auditability: Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.

7. Validation and Testing: Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Association for the Advancement of Artificial Intelligence (AAAI)

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

Response to Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan (Federal Register Notice of February 2, 2022, pp 5876-5878)

The **Association for the Advancement of Artificial Intelligence (AAAI)** strongly supports the objectives of the National Artificial Intelligence Research and Development Strategic Plan and the eight strategic priorities of the strategic plan. **AAAI is the leading scientific society for artificial intelligence (AI),** with more than 300 elected Fellows and over 6,000 members.

In this brief commentary, we would like to raise one issue concerning the practical challenge of implementing the first strategic priority, **"Make long-term investments in AI research."**

Much of the current excitement around AI arises from recent breakthroughs in machine learning (deep learning) and so-called data-driven AI. These breakthroughs have opened up opportunities for AI to contribute significantly to our society on climate change and sustainability, scientific discovery, personalized education, healthcare and well-being, and business innovation. However, these dramatic recent successes in AI have also generated a certain sense that fundamental AI research is "done" and that the remaining challenges are mainly on how to apply data-driven AI techniques to other domains. **There is a perception that simply larger data sets with more cloud computing resources will be sufficient to bring out the full potential of AI.** ***We want to stress that this is not the case.*** **To reach robust, interpretable, and human-compatible AI systems that adhere to desirable ethical guidelines will require further breakthroughs in core AI.**

By "fundamental AI research" we mean research that advances not only data-driven approaches but also knowledge-based approaches (such as those in planning, scheduling, decision making, and optimization), and that are human-centered rather than always closed and fully automated. Without a careful combination of all these AI techniques, we will not be able to tackle current and future challenges in many application domains, especially those that involve human-machine collaboration and high-stakes decision making where robustness and trust are fundamental notions.

An important question is how to create an environment that will enable such advances in order to support and reach the next level of AI. In 2019, AAAI released "A 20-Year Roadmap for AI Research in the US", co-sponsored by AAAI and the Computing Community Consortium, and involving inputs from over a hundred leading AI researchers and extensive feedback from the broader community. In the roadmap, among several other recommendations about AI education

and workforce development, we urged the US to establish "*National AI Research Centers as multi-university centers with affiliated institutions*, focused on pivotal areas of long-term AI research (e.g., integrated intelligence, trust, and responsibility), *with decade-long funding to support on the order of 100 faculty, 200 AI engineers, 500 students, and necessary computing infrastructure.* These centers would offer rich training for students at all levels. Visiting fellows from academia, industry, and government will enable cross-cutting research and technology transition." The current NSF National AI Research Institutes program does not address this need since it provides funding for only a small percentage of the effort of 20-30 faculty (who are mostly supported by other grants), and a similar number of graduate students and undergraduates.

Two promising examples of successful large long-term funded AI centers are the Vector Institute for AI and the MILA research institute, recently established in Canada. These research institutes combine **a mission of advancing fundamental AI research with applications to societal challenges.** The institutes bring together numerous faculty and graduate students and have close connections to universities and industry labs. **The leadership of these institutes is provided by distinguished senior AI researchers, who help formulate the long-term research agenda.** The funding for the institutes is on a longer timescale (10 years, with an option for renewal).

We recommend that the US establish similar sustained national AI research centers to advance fundamental AI research to implement the first strategic direction of the National Artificial Intelligence Research and Development Strategic Plan. Each center should be led by one or more internationally distinguished AI scientists. These centers will enable the US to reach the next level of AI, necessary for developing truly robust, interpretable, and human-compatible AI systems that adhere to desirable ethical guidelines.

Bart Selman, Cornell University
AAAI President

Yolanda Gil, University of Southern California
AAAI past-President

Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, CA 94306, USA
+1-650-328-3123 • www.aaai.org

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

AUTM

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

AI R&D RFI Response Team
National Coordination Office (NCO)
2415 Eisenhower Avenue
Alexandria, VA 22314

AUTM's Input on the Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan

AUTM is the non-profit leader in efforts to educate, promote and inspire professionals to support the further development of academic research that drives innovation and changes the world. Our community is comprised of more than 3,000 members who work in more than 800 universities, research centers, hospitals, businesses and government organizations around the globe. AUTM's members are primarily from academic settings (67%), 15% are practicing attorneys and 5% are from industry. Some 22% of our members are international. AUTM appreciates the opportunity to provide input on the above-referenced update to the 2019 version of the National Artificial Intelligence Research and Development Strategic Plan (the "2019 Plan").

AUTM members in academic settings are focused on advancing early-stage inventions and other technologies to the marketplace primarily through licensing to partners (i.e., implementers). Between 2011 and 2020 (the most recent decade for which we have data), our skilled professionals filed over 150,000 patents for academic inventors and over 17,000 in 2020 alone. Between 2011 and 2020 our U.S. members negotiated over 60,000 intellectual property license agreements on behalf of U.S. universities and academic research institutions, and in 2020 alone over 8,000 such license agreements. Thus, AUTM has valuable insights and an important voice with respect to the protection, further development and commercialization of a wide range of new technologies.

Today, Artificial Intelligence (AI) is frequently at the center of what we do because it is becoming ubiquitous. It now touches

nearly every technology sector including, transportation, telecommunications, military, consumer electronics, therapeutics, diagnostics and even agriculture and finance. More importantly, AI is present from the origin of the technology. It is now integral to the inventions themselves, not some after-thought layered in at some later stage of development. For these reasons, we at AUTM know firsthand of the tremendous potential of AI and, given the current geopolitical climate, we recognize there is no margin for error in urgently tapping into that potential. As such, we applaud the Office of Science and Technology Policy, on behalf of the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence (Select Committee), the NSTC Machine Learning and AI Subcommittee (MLAI-SC), the National AI Initiative Office (NAIIO), and the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) (collectively, the "Coalition") for its effort to update the National Artificial Intelligence Research and Development Strategic Plan in order that we maximize the output of the National AI Initiative at a breakneck pace while being effective stewards of the precious taxpayer dollars involved.

Preliminary Matters to Establish the Context for AUTM's Input

Upon review of the eight (8) strategies comprising the 2019 Plan, we can't help but be drawn to strategies 1 and 8. We are drawn there because we live there. Day in and day out, AUTM members review the new technologies that are the fruits of investments in research. Then, after a thorough, collaborative assessment and arranging for the appropriate protection (e.g., filing a provisional patent application) of those new technologies, AUTM members immediately turn to the difficult task of finding a partner to further develop (i.e., advance) those new technologies toward the marketplace.

The most important step in above process is the protection step, particularly for new inventions. The reason being that the partner AUTM members endeavor to find will ultimately be a private sector entity. Among the many resources said private sector partner will be required to provide is investment dollars (e.g., venture capital), often in the millions or tens of millions of dollars. Any partner seeking to further develop the new technologies AUTM members manage is, without question, a sophisticated investor. No sophisticated investor will provide such substantial sums of investment dollars without having first made detailed estimates of the expected return on that investment. Accurate revenue projections from the sale of the eventual product or service are essential to those estimates, and one cannot make accurate revenue projections faced with the prospect of unlimited sellers of the same product or service. Thankfully, patents provide the ability to enforce the exclusive right to sell the product or service covered by the patent thereby providing valuable limits of the number of potential sellers. Thus, patents allow the investors to make more accurate revenue projections and confidently estimate the return on their investment. For these reasons, patents are essential tools investors depend on when deciding whether to enter into a partnership (i.e., taking a license) to further develop a new invention. Consequently, the patent status is among the first questions asked of AUTM members when partnership discussions commence.

AUTM's Input

With the above as context, we were quite surprised and more than a little dismayed to note that none of the strategies in the 2019 Plan mentions the patent system or intellectual property of any type. This to us is a glaring omission and a harmful one to boot.

To the members of the Coalition who may not be as familiar with the current state of the U.S. patent system as we at AUTM are, it is not good. It has undergone a radical transformation in the last 15 years. The enemies of the U.S. patent system have been waging a multi-front war against it since the mid-2000's resulting in a significant weakening of U.S. patent rights manifesting as uncertainty and unreliability. There is grave concern among AUTM and other like-minded groups that this weakening is resulting in fewer and/or less effective U.S. patents in vital technology sectors mentioned above, all of which nowadays incorporate AI. As we have illustrated, weaker or non-existent patents lead to less investment and thus fewer cutting-edge products and services being developed and introduced in the U.S. in these vital technology sectors. This outcome risks the loss of our technological preeminence with great harm to our economy and national security.

Thus, in order that the Coalition's valuable time the taxpayers' money is not squandered, AUTM recommends that a threshold strategy be put in place, Strategy 0, if you will. Strategy 0 will be to work with Congress to immediately strengthen U.S. patent rights by eliminating the current uncertainty and unreliability. The uncertainty can be eliminated by reforming Section 101 of the patent statute to clarify that (i) a claimed invention is entitled to a patent unless it exists in nature independently of and prior to any human activity or exists solely in the human mind and that (ii) subject-matter eligibility determinations must be made without regard for the requirements of Sections 102, 103, and 112 of the patent statute, or the claimed invention's "inventive concept." The unreliability can be eliminated by (i) restoring the ability of the patentee to routinely obtain injunctive relief after a finding of infringement at trial and (ii) reforming the post-issue review procedures at the PTAB to shift the balance in favor of the patentee instead of the challenger as it is at present.

We fear that, without implementing this Strategy 0, the result of the National AI Initiative will resemble what we experienced in the pre-Bayh-Dole years; namely, significant taxpayer dollars, squandered because they were invested in research that resulted in great technologies that never made it into the marketplace. These technologies remained on laboratory shelves and eventually became obsolete. The federal government owned them, was unwilling or unable to seek patent protection for them and, thus, were unable to find the partners necessary to continue their development. The past 40-plus years of Bayh-Dole has taught us that accessing the patent system, even a weak one, and strong private-sector partnerships pay tremendous dividends for the economy and the standard of living for everyday Americans by bringing many cutting-edge products to the marketplace. See here for more information of the success of the Bayh-Dole Act (https://autm.net/about-tech-transfer/advocacy/legislation/bayh-dole-act/bayh-dole-innovations). A well-functioning system of strong patent rights is at the core of Bayh-Dole.

Conclusion

AUTM again wishes to thank the Coalition for its efforts to to update the National Artificial Intelligence Research and Development Strategic Plan in order to quickly and efficiently maximize the output of the National AI Initiative. AI is ubiquitous in 21st century technologies and in order for our country to remain the world's technological superpower, we must realize AI's full potential and in ways that no other country can even fathom. For this Coalition to be successful in bringing about this important result, its plan must include the threshold strategy to strengthen U.S. patent rights by eliminating the current uncertainty and unreliability. If we do so, and only if we do so, will we realize the full potential of AI so that America will continue to enjoy robust national security and lead the world in sustained increases in economic growth, standard of living, and high-paying jobs for its citizens.

Sincerely,

Stephen J. Susalka, Ph.D.
Chief Executive Officer

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Booz Allen Hamilton

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.

ARTIFICIAL INTELLIGENCE

ENGINEERING YOUR AI FUTURE

RFI RESPONSE: OSTP, THE SELECT COMMITTEE, NAIIO, AND NITRD NCO INPUT REQUEST ON THE UPDATE OF THE NATIONAL ARTITFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN- (87 FR 5876)

Prepared by Booz Allen Hamilton (boozallen.com/ai)

Booz Allen Hamilton: RFI Submission- 87 FR 5876

Key Recommendations

- For the Office of Science and Technology Policy to issue the revised 2022 Plan as a clean sheet, standalone product (vice a revision) due to its importance and to clearly communicate with the public.
- Provide a unified national roadmap helping the public better understand America's AI pursuits and how various organizations are aligning to meet those objectives.
- Address the role of technologies like 5G that will significantly increase access to data 'at the edge' and the nation's ability to conduct meaningful analytics for end users in near-real-time.
- Bolster objectives/sections addressing vulnerabilities, such as hardening against cyber-attacks and information assurance.

INTRODUCTION

Businesses, organizations, and individuals are generating mountains of data. Business operations, online transactions, vehicles and smart homes, cell phones, and increasing prevalence of the Internet of Things (IoT) means digitized information is continuously being generated and stored. Moreover, this growth in data vastly outstrips the number of technical analysts needed to gain insights from it. Companies, governments, and organizations are rightly asking how they can possibly provide value from these data volumes with their existing analytical capabilities and staff while determining where they need to invest to meet this growth.

Like past industrial revolutions, AI's adoption will continue to have sweeping effects on our economy, governance, and national security. American organizations are seeking the next wave of competitive advantage and striving for better performance in their analyses—a result that effective AI deployments can deliver. As such, artificial intelligence will change how Americans conduct business, provide government services, collect intelligence, and provide for national defense by introducing new efficiencies—and complexities— that we are just beginning to understand. Anticipating these long-term changes and guiding America's adoption with the 2022 AI R&D Strategic Plan will be essential to ensuring continued American technological superiority…and all downstream benefits coming from sustaining that scientific edge. [1]

BOOZ ALLEN'S AI BACKGROUND

The largest provider of AI services for the Federal government[2], Booz Allen delivers professional and technical services to research, design, architect, engineer and integrate AI solutions needed to accomplish critical missions and maintain U.S. technological leadership. We support some of our Nation's highest profile and innovative programs—including the Joint Artificial Intelligence Center (JAIC), Office of the National Coordinator for Health Information Technology, and the Defense Threat Reduction Agency (DTRA).

The federal government remains in an ideal position to foster and enhance the adoption of AI. Booz Allen recently submitted a response to assist the White House's Office of Science and Technology Policy as they develop an implementation plan for the National AI Research Resource.[3] Our submission offered thoughts regarding the federal government's positioning to facilitate AI adoption and global leadership, including infrastructure design; tools, methods, and data; education and training; the importance of accountability, diversity, and integration; and frameworks to aid adoption. These frameworks include Responsible AI, AIOps[4],

and Data, Machine Learning, and Systems Engineering and Operations. That response is available at https://www.ai.gov/rfi/2021/86-FR-39081/BAH-NAIRR-RFI-2021.pdf.

2022 PLAN RECOMMENDATIONS

- We recommend the Office of Science and Technology Policy **issue the revised 2022 Plan as a clean sheet, standalone product**. This path will:
    - Follow similar processes taken with other national-level strategies (e.g., National Security Strategy, National Defense Strategy, Department of Commerce Strategic Plan) to minimize confusion regarding previous iterations. Rather than offer as an update and track 'what's changed,' OSTP can simply issue the 2022 Plan as the latest version with a vision clearly articulated for the future.
    - Highlight the importance of AI R&D, especially given the rapid pace of technological change and increasingly widespread adoption.
    - Ensure clarity and unity of purpose for organizations looking for a single source regarding America's AI development pathway.
- Standards development will be a challenge during the next few years. **Policy should navigate a course that strikes a balance between regulating AI while facilitating innovation and adoption of a rapidly evolving technological capability**. Artificial intelligence is an emerging industry – we recommend maintaining a relatively liberal position leveraging already existing regulatory powers to enable America to innovate at the speed of relevance. Over-regulation risks ceding AI leadership to other nations in this increasingly competitive great power competition landscape.
- Strategic leaders' **ability to anticipate and guide these long-term changes will be essential to ensuring the success** of this transition. While there is growing consensus around AI's importance, there is a growing chasm regarding how to successfully deploy and apply artificial intelligence at an enterprise-wide, let along nation-wide, scale. Many organizations are struggling with expanding, evolving, and integrating their early AI development efforts into mature, sustainable, enterprise-wide capabilities. This gap is due in part to the drastic increase in scope and complexity required to operationalize artificial intelligence, particularly in terms of integrating AI solutions within the larger organization.
- **The most important numbered Strategy would be to offer a centralized roadmap for America's overall AI strategy and investments.** As we've witnessed during the past few years, much of U.S. AI development is fractured. Each government agency is pursuing AI for its own ends (and the data associated with it). The same holds true within various executive departments. For example, the Dept. of Defense: Research & Engineering is pursuing AI separately from the services, which is also separate from the Joint Artificial Intelligence Center. While thematically related overall, each agency and service are pursuing its own approach to AI adoption and solutions risking diminishing US-based data and analytics outcomes relative other international competitors, esp. vis-à-vis China. The same holds true for R&D efforts. A clearly articulated and numbered Strategy will provide much needed clarity regarding White House requirements regarding continued development and ongoing efforts.
- **Create a numbered Strategy for the creation and adoption of a central annotated data repository and/or widely accessible data formats.** Like the previous point, data formats and overall data collection are widely incompatible. This diminishes the impact of AI-derived insights, especially across a whole-of-government approach. A holistic methodology for primary data types (images, records, etc.) designed

for wide implementation and adoption will foster greater cross-agency communications and insights. Critical to the success of this repository is means to curate labeling and data errors, as the value of labeled data to mission success is enormous but most often neglected component of successful AI applications. An example of investment in data annotation is the 'National Cancer Informatics Program (NCIP) Annotation and Image Markup (AIM) Foundation Model.' In some ways, we're now tackling data challenges that the 9/11 Commission identified for intelligence-related activities: the inability to rapidly align, share, and leverage data puts America's decision and policymakers at a significant information disadvantage. This waterfalls into every corner of AI design, implementation, and adoption. In short, the choices and pathways selected in 2022 will waterfall into every succeeding year…and exceedingly difficult to change once entrenched. Clear guidance from the 2022 Plan will go a long way toward unifying America's efforts for AI development, deployment, and implementation.

- **Address the role of 5G, quantum computing, and other emerging technologies as they relate to AI.** References to these increasingly diffuse technologies are not in the 2019 Plan, yet 5G (in particular) will significantly increase access to, and facilitate ingestion of, massive amounts of data at the edge. Given technological maturation during the previous three years, addressing these technologies will help address how data generation is changing, the increasing demands this puts on storage solutions, and the direction this should drive R&D strategy and investments.

- **Consider the nation's bandwidth realities and future requirements.** Given the massive (and growing) data generation at the edge from the internet of things, how can R&D inform better edge-based analytics, slimming of data for near-real-time transmission back to decision makers given limited bandwidth availability, federated learning options (i.e. distributed/edge training for data privacy and bandwidth concerns) and then subsequently transmit approved full data sets back to a centralized repository (in whatever form that looks like) once physical delivery is possible. With so much information collected every second of every day, organizations will benefit from OSTP's guidance regarding the handling and prioritization of data streams—what's critical now, what's useful and relevant later, and what data to discard to avoid collection for sake of collection.

- **Address analytics at the edge.** Separate from bandwidth and transmission capabilities, the 2022 Plan can help organizations determine their level of investment in devices and data collection capabilities required to make meaningful insights at 'the edge.' This could include addressing manufacturers like Apple who incorporates neural processors in iPhones, Google with TensorFlow, Amazon with AWS, and others, as well as how local mesh networks can support improved data collection, processing, and dissemination for real-time insights and support to American citizens (e.g., crowdsourcing data for weather, traffic, greater insight and access to government services, increased efficiencies, etc.) These paths could be increasingly valuable to citizens, especially as inflationary pressures and the expectation that cost for goods continue to increase in the near-term.

- **Acknowledge, address, and provide guidance for security concerns.** The 2022 Plan needs to place much greater emphasis on this topic. Data is widely considered to be the new oil—an invaluable resource that can unlock new insights and provide increasing value. Media reporting on cyber-attacks—including the exfiltration of decades of sensitive technology along with personal data—are proliferating rapidly, raising awareness for the public. Unfortunately, media reports only indicate that we're reading about the events that were caught and publicized. Many more are occurring without widespread knowledge, risking billions of dollars in R&D investments and, perhaps more importantly, decades of R&D. A nation-state competitor can accomplish via remote keystrokes and cyber espionage what would have required far more significant risk only a decade or two ago. The 2020s will likely represent the greatest leveling of global knowledge in history – since actors can now steal knowledge and data that was previously unavailable, competitors have access to greatly accelerated R&D pathways at the expense of US-based corporations. It is a known issue that at some point soon current encryption will be breakable. To ensure

this data remains secure—this requires a shift to Post Quantum Cryptography (PQC) across all government and commercial clients. PQC has impacts on both applications and networking hardware used today, therefore both need addressing to effectively move to PQC. This will require testing and customization to individual organizations. The level this threat presents is worthy of its own numbered strategy to inform secure development of technologies and data repositories from Day 1.

- **Reorder the existing numbered Strategies.** Since many of the eight existing stated strategies are likely to be retained, we recommend reordering based on each Strategies' overall importance. For example, moving Strategy 8: "Expand Public-Private Partnerships to Accelerate Advances in AI" up to position number two. The order the Strategies are presented should matter and enhancing public-private partnerships is likely the second most important for America's overall strategy—both near-term and for long-term success (vice least-most as Strategy #8).

- **Address the operationalization of AI.** One of the most important elements for the successful adoption of AI is the ability to operationalize it. That is, to take algorithms and methods developed in a static laboratory environment and bring them into the real world, where they can begin to tackle real-world challenges. We believe the right approach is to begin operationalizing AI sooner rather than later, with the expectation that companies will rapidly improve their algorithms based on real-world situations and feedback from end users and senior decision makers. Critical to this is thinking about the development and use of AI in a responsible manner. Responsible AI becomes meaningful when considering how an AI system impacts people and clear understanding (i.e., explainability of the parameters that govern the outputs as well as the identification of the shifts of those parameters as the AI models learn to address/mitigate risk). This pathway will help speed AI adoption, elevate citizens awareness and familiarity with AI, and provide a quicker on-ramp to impactful AI applications.

---

[1] Justin Neroda, Steve Escaravage, and Aaron Peters, *Enterprise AIOps: A Framework for Enabling Artificial Intelligence* (Sebastopol: O'Reilly Media, 2021), v.
[2] Bloomberg Government Market Analysis
[3] Booz Allen Hamilton, *Request for Information on an Implementation Plan for a National Artificial Intelligence Research Resource*, https://www.ai.gov/rfi/2021/86-FR-39081/BAH-NAIRR-RFI-2021.pdf.
[4] Justin Neroda, Steve Escaravage, and Aaron Peters, *Enterprise AIOps: A Framework for Enabling Artificial Intelligence* (Sebastopol: O'Reilly Media, 2021), v.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

BSA | The Software Alliance

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.

4 March 2022

Stacy Murphy
US Office of Science and Technology Policy
1650 Pennsylvania Avenue NW
Washington, DC 20502

Dear Ms. Murphy,

BSA | The Software Alliance appreciates the opportunity to provide feedback to the Office of Science and Technology Policy (OSTP) as it considers updates to the National Artificial Intelligence Research and Development Strategic Plan. BSA is an association of the world's leading enterprise software companies that provide businesses in every sector of the economy with tools to operate more competitively and innovate more responsibly.[1] As leaders in the development of enterprise AI, BSA members have unique insights into the technology's tremendous potential to spur digital transformation and the policies that can best support a competitive and thriving national AI R&D ecosystem.

The National AI R&D Strategic Plan is an important signal of US priorities for the development of AI. We are pleased that OSTP continues to revisit the Strategic Plan to ensure that it accounts for shifts in the technological landscape and the needs of US R&D stakeholders. Overall, we regard the eight strategic priorities identified in the 2019 update to the Strategic Plan to be the right areas of continued focus and would not recommend any major course corrections. Instead, as OSTP considers updates to the AI R&D Strategic Plan, BSA offers below several recommendations for advancing the existing strategies and ensuring that federal investments in R&D are aligned with core US interests.

Expanding the National AI Research Institutes Program

As part of the effort to sustain long-term investments in fundamental AI research (Strategy 1), the 2022 R&D Strategic Plan should support the continued expansion of the National AI Research Institutes Program.[2] The National AI Research Institutes Program is helping to establish a nationwide network of AI research clusters- including regional hubs- that can support sustained, large-scale, and multidisciplinary research into pressing challenges. To date, the Program has established 18 Research Institutes that each operate as hubs for research into how AI can be used to address a broad range of societal and technological

[1] BSA's members include: Adobe, Alteryx, Atlassian, Autodesk, Bentley Systems, Box, Cisco, CNC/Mastercam, DocuSign, Dropbox, IBM, Informatica, Intel, MathWorks, Microsoft, Okta, Oracle, PTC, Salesforce, SAP, ServiceNow, Shopify Inc., Siemens Industry Software Inc., Splunk, Trend Micro, Trimble Solutions Corporation, Twilio, Unity Technologies, Inc., Workday, Zendesk, and Zoom Video Communications, Inc.
[2] National Artificial Intelligence Research Institutes, available at https://beta.nsf.gov/funding/opportunities/national-artificial-intelligence-research-institutes

challenges, including climate change,[3] agricultural supply chain challenges,[4] and cybersecurity.[5] By bringing together researchers from multiple academic institutions, as well as experts from industry, government, and NGOs, each of the Program's Research Institutes is helping to break down siloes and foster coordination across the United States' dispersed AI R&D ecosystem. The 2022 R&D Strategic Plan should highlight the critical importance of the National AI Research Institutes Program and signal support for its continued expansion, including the promotion of regional hubs for AI R&D.

Investing in Tools and Resources to Manage AI Risks

BSA strongly supports the AI R&D Strategic Plan's focus on supporting research into the ethical, legal and societal implications of AI, including the potential risks of unintended bias (Strategy 3). The 2022 R&D Strategic Plan should build on this commitment by outlining how the federal government will support efforts to develop tools and resources that can help organizations manage the risks of AI. For instance, the 2022 R&D Strategic Plan should highlight the critical work being undertaken by the National Institute for Standards and Technology (NIST) to develop a cross-sectoral AI Risk Management Framework. BSA and its members are also deeply invested in the development of risk management tools, particularly as it relates to bias. In addition to supporting the development of the NIST AI Risk Management Framework, we recently published C*onfronting Bias: BSA's Framework to Build Trust in AI*[6] which outlines a comprehensive, lifecycle-based methodology for performing impact assessments to identify risks of AI bias and corresponding risk mitigation best practices. The AI R&D Strategic Plan should also invest in work to develop standardized frameworks for benchmark and operational testing of AI systems. This type of testing is important to ensure systems are performing appropriately for a given use case and is an important part of the risk identification and mitigation process. Further research into appropriate training programs for those using and overseeing AI systems, with a focus on developing programs for use of high-risk systems, should also be prioritized.

Enhancing Visibility into the Needs of the AI Workforce

There is growing demand across technology fields for skilled workers to fill critical roles, and the Strategic Plan correctly identifies the importance of meeting AI workforce needs to further US leadership in AI R&D (Strategy 7). The Strategic Plan can help address the current skilled worker shortage by prioritizing investments into initiatives that will grow the pipeline of future talent and make it easier to identify skills that are in demand. For instance, OSTP should consider how US labor data can be better leveraged to provide greater visibility into the needs of the AI workforce. Under current practice, labor force data often takes several months to be released. As a result, job seekers who may be interested in pursuing reskilling programs are unable to base their decisions on real-time data about what skills are in greatest demand. The lack of real-time labor data also impairs the effectiveness of government supported retraining initiatives by obscuring economic trends. To overcome these challenges, 2022 R&D Strategic Plan should prioritize research into how the public and private sectors can work together to enhance the collection and availability of real-time labor data. The government should work to better incentivize employers to improve their data collection methods, consolidate existing workforce

[3] NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), available at https://www.ai2es.org/

[4] AgAID Institute, available at https://agaid.org/

[5] NSF AI Institute for Future Edge Networks and Distributed Intelligence, available at https://aiedge.osu.edu/

[6] Confronting Bias: The BSA Framework to Build Trust in AI (June 2021), available at https://ai.bsa.org/wp-content/uploads/2021/06/2021bsaaibias.pdf

datasets, and support the creation of a modernized labor database.[7] Such improvements could help increase the number of workers in high-demand fields related to AI R&D outside of traditional graduate and post-graduate roles, such as data analysis, and those that involve the integration of advanced computing skills into other disciplines.

BSA appreciates the opportunity to provide input into the update of the strategic plan and looks forward to continued collaboration with OSTP on this and other AI-related projects.

Sincerely,

Heidi Obermeyer
Manager, Policy

[7] Business leaders call for 100B in workforce investments, available at https://nationalskillscoalition.org/wp-content/uploads/2021/08/Business-leaders-call-for-100B-in-workforce-investments.pdf

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Byrne, Vanderbilt University

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.

Subject: RFI Response: National Artificial Intelligence Research and Development Strategic Plan
From: Byrne, Daniel
To: AI-RFI

The new book "Artificial Intelligence for Improved Patient Outcomes - Principles for Moving Forward with Rigorous Science" by Daniel Byrne provides valuable input that should be considered for the "National Artificial Intelligence Research and Development Strategic Plan". Below are some of the principles from this book:

Scientists must take the lead in evaluating the effec-veness and impact of AI. Marketing and salespeople must limit their claims to scien-fically valid and published conclusions.

A pragmatic randomized controlled trial is the key to real progress of AI in healthcare.

Reject the flawed thinking that science slows AI advancement.

A pragmatic randomized controlled trial of AI in medicine does not need to be difficult, take years to conduct, or disrupt clinical workflows. These trials can be frictionless, fast, and low cost. Randomization is the solution – not the problem. Randomization solves many problems that most people are not even be aware of.

AI must focus on improving hard outcomes that are important to patients - not surrogate end points or process metrics.

For many applications in medicine, especially clinical decision support with a binary end point and low dimensional data, logistic regression is a superior choice over machine learning.

After creating an AI tool, focus on creating and testing an effector arm in a pragmatic trial.
***

Hope this is helpful.

Regards,
Dan Byrne

Daniel W. Byrne

Director of Ar-ficial Intelligence Research
Advanced Vanderbilt Ar-ficial Intelligence Laboratory (AVAIL)

Department of Biosta-s-cs
Vanderbilt University Medical Center
2525 West End Avenue, Suite 1000
Nashville, TN 37203-1741

hfps://www.vumc.org/biosta-s-cs/person/daniel-w-byrne-ms
hfps://scholar.google.com/cita-ons?user=Eg4zVpYAAAAJ&hl=en
hfps://www.amazon.com/dp/1496353862

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Carnegie Mellon University (CMU)

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.

Carnegie Mellon University

Request for Information Response:
National Artificial Intelligence Research and Development Strategic Plan Submitted by Carnegie Mellon University

Introduction

Carnegie Mellon University is pleased to contribute ideas to support the update of the National Artificial Intelligence Research and Development Strategic plan.

Key points outlined in the document include the following:

- Fostering interagency and community collaboration on two meta themes that are critical to all applications of AI:
    - Advancing research to accelerate breakthroughs in trustworthy and robust AI, distributed AI and AI-enabled automated science;
    - Advancing research capabilities to engineer AI into systems of societal importance – including manufacturing, infrastructure and energy systems – and testing those engineered systems in testbeds enabled by
    - the bipartisan American infrastructure investment agenda;
- Advancing research to enable AI to be a platform technology for the President's health care and climate challenge initiatives.
- Enhancing the ability for the AI community to jump-start job creation and to foster equity and economic inclusion and;
- Building on the lessons of the COVID-19 pandemic to enable AI to advance a new generation of real-time policy decision and implementation tools.

The National AI Research and Development Strategic Plan and Update have galvanized the research community and public, private and community partners to advance a shared vision of U.S. leadership in ethical AI innovation. The Carnegie Mellon community appreciates the leadership of the National AI Initiative Office (NAIIO) and looks forward to continuing to support this vital mission.

Recommendations of AI Research Focus Areas to Create Solutions to Major Societal Challenges

The National AI Research and Development Strategic plan can catalyze innovations in both fundamental discoveries and applications that address specific societal challenges. Progress towards realizing this potential can be realized by collaborative efforts in the following areas.

Foster Interagency Collaboration to Ensure America Leads in Enabling Distributed Artificial Intelligence

The U.S. should lead a bold transformative agenda over the next five years to enable AI to evolve from highly structured and controlled, centralized architectures to more adaptive and pervasively distributed ones that autonomously fuse AI capability among the enterprise, the edge, and across AI systems and sensors embedded on-platform. CMU terms this revolutionary architectural advance as **AI Fusion**. The vision is built upon plans for a cohesive research advancing capabilities in microelectronics, AI frameworks and algorithms and innovations in federated learning in the AI fabric and abstraction layers.

Building a community research roadmap for distributed AI will address several critical challenges for the growth of AI, challenges that cut across agency-specific missions. The ability to enable distributed AI at the edge will minimize the dependence on aggregating and engineering massive data sets and reduce the need to "move the data to the algorithms" as well as the inherent challenges associated with the need for continuous high-bandwidth connectivity. Research in this area will also greatly enhance the capacity to address privacy and security challenges. It is dependent on, will contribute to and will benefit from the national computing infrastructure initiatives launched by the NAIIO.

Most critically, an *AI Fusion* research agenda will contribute to the network of AI institutes by enabling a host of applications emerging from increased convergence across AI-enabled cyber and physical systems. This convergence is vital to the viability of applications in commercial, military and national security domains. *AI Fusion*, for example, will be a critical contribution to the Department of Defense's (DOD) focus on Multi-Domain Operations. It will also enhance the potential for advances in smart city applications and AI breakthroughs aiding manufacturing, energy, health care, education and agricultural innovations. A focus on *AI Fusion* should operate synergistically with national initiatives in microelectronics and tie directly with research and innovation efforts aimed at enhancing, protecting and hardening critical U.S. supply chains.

Initiate Research to Engineer AI into Societal Systems

While fundamental advances are needed in AI science, advances in engineering AI into systems of societal importance are vital to realize the full impact on major national missions. Engineering AI into such systems will be essential to transform U.S. manufacturing and enhance infrastructure and energy systems to meet critical national economic and societal goals.

Engineering AI will require the design, development and deployment of new use-inspired AI algorithms and methodologies, targeted to real-world applications and possessing enhanced scalability, robustness, fairness, security, privacy and policy impact. Advancing *Engineering AI* will also require new hardware and software systems, including cloud, edge and device computing infrastructures that sense and store the vast amounts of data collected in the real world and that enable devices to access and transmit this data from anywhere, to anywhere, in secure and private ways. Foundational research for *Engineering AI* is needed to enable the deployment of the highest performing and most energy-efficient AI systems. Such systems will

require architecting new hardware and computing frameworks; designing faster, more powerful and efficient integrated circuits; and developing sensing modalities to support data collection, storage and processing of the data deluge.

In addition, Carnegie Mellon recognizes that research on *Engineering AI* must include a focus on creating trust not only from a technical standpoint but from the system of stakeholders interacting with the AI system — be it in education, infrastructure or climate. Users and communities have to trust the system that is allocating resources and making decisions.

Potential applications and use cases for *Engineering AI* include autonomous infrastructure systems (AIS) that can help create equitable, innovative and economically sustainable communities. AIS technology could, for example, include initiatives integrating food delivery, the tracking of goods while preserving privacy and tools to improve mobility. *Engineering AI* will be key to the digital transformation of manufacturing in the U.S., including robotics for manufacturing, development of a timely and trustworthy supply chain and additive manufacturing. *Engineering AI* also has the potential to revolutionize how electricity is produced, distributed and consumed. It can provide insights to improve electricity distribution through demand forecasting, load management and community governance, as well as to innovate new energy storage solutions, control pollutants and advance wind, solar and nuclear energies.

Accelerate Advances in Automated Science to Support the Nation's Science and Technology Ecosystem

The national AI research strategy can also work to vector advances in AI, machine learning and robotics to accelerate the scalable deployment of ***Automated Science.*** This effort can be the cornerstone of a national initiative to strengthen the nation's science and technology ecosystem, potentially transforming the way bench research is conducted and taught. The development of *Automated Science* labs, such as the CMU-alumni-founded Emerald Cloud Lab, allows researchers to have their experiments performed remotely at an automated lab facility and the results returned to them – all via the cloud. Technological and methodological breakthroughs that integrate AI with drive-by-wire automated experiments are helping to accelerate the pace of innovation in a host of fields, including drug discovery.

An AI research agenda to advance *Automated Science* holds promise for strengthening interdisciplinary research and can play a critical role in democratizing participation in and access to research, as well as in addressing the important need for reproducibility in experimental scientific discovery. *Automated Science* can create an emerging industry, including start-ups and collaborations across several sectors that can contribute to the U.S. innovation ecosystem. It can lower barriers to innovation and thus play a critical role in sparking new start-ups in areas such as new materials that will be critical to mitigating the climate challenge.

Bringing together key stakeholders across the community will help to frame an interagency strategy to build an *Automated Science* infrastructure for the U.S. research ecosystem and understand the policies needed to unleash the potential of these breakthroughs.

Develop a Targeted Initiative to Enable AI to Accelerate Breakthroughs in Health Care The power of research investments to aid the nation's response to the pandemic provides an opportunity for a coordinated strategy to accelerate the ability of AI to enable major breakthroughs that can directly impact the quality of health care.
Specific innovation areas highlight the potential power of AI to advance national health initiatives:

- AI can play a greater role in the Cancer Moonshot through additional research in computational models, predictive modeling and algorithms to engage various combinations of data sources that dramatically improve understanding of the evolution of cancers and how mutational processes vary among patients.
- Similarly the application of large scale AI-driven modeling and prediction capabilities should be a major contributor to our response to future pandemic responses.
- An initiative focused on advancing AI and machine learning capabilities in 3D bioprinting would contribute to realizing the ability to print human organs within this decade. As with the initiatives highlighted above, multi-agency efforts that integrate AI research with materials, genomic and surgical sciences are key to realizing this opportunity.
- The impact of telemedicine and extending its power to improve clinical care and decision-making depends upon accelerating advances in robotics and human-computer interaction, specifically in the area of automatic recognition of multi-modal behaviors and the emerging field of health behavioral informatics. (We again note that progress in *AI Fusion* will be key to ensure capabilities exist equitably across the full spectrum of compute and bandwidth availability.)
- Similarly, advances at the nexus of autonomy and materials afford an opportunity to realize the potential for robots to contribute directly to extending the independence of elderly Americans and persons with disabilities. These advances are also rapidly accelerating progress in the development of neuro/brain-computer interfaces.

Action to realize the potential of AI in improving health care should begin by identifying specific advances in AI research and capabilities that are common to defined innovation areas, such as those above. Capitalizing on these advances requires new models for collaboration among universities, industry and health care providers. The National Strategy can help create a framework for convening agencies, industry and the academic community to craft a vision for accelerating the AI breakthroughs that can provide cross-cutting platform capabilities to support specific health initiatives. Developing this vision would also help jump-start ARPA H.

Drive the Development of an AI for Climate Roadmap

AI is also a platform technology poised to accelerate breakthroughs across the energy and climate continuum. Advances are needed in AI that include new and accelerated material development, enhanced power production and distributed energy, the optimization of batteries and improved wind, solar and geothermal operations. As an example, the Open Catalyst Project, a collaboration between Facebook and CMU, aims to use AI to accelerate quantum mechanical simulations by 1,000x in order to discover new electrocatalysts needed for more efficient and scalable ways to store and use renewable energy.

These AI breakthroughs are needed to build the integrated smart grid and electrification infrastructure the President's vision demands. And even more AI innovations are needed to inform more efficient agricultural techniques to combat climate change. Of equal importance, this agenda can also galvanize and advance research to reduce the energy footprint of AI.

AI is also a powerful tool to realize the goal that investments in climate infrastructure will have a transformative impact on addressing environmental justice. The President has called for at least 40 percent of all climate infrastructure investments to occur in underrepresented communities. This can be enhanced by the development of trustworthy AI community interface tools. The ability to integrate system-level advances in building technologies and transportation capabilities with tools for equitable community engagement, both enabled by AI, would be vital to shape the societal impact of climate initiatives.

The National Strategy can help seize these opportunities that span multiple agencies and advance an AI Climate research agenda.

Establish a Focus on AI Engineering to Support Continued Innovation

Realizing the power of AI innovations requires engineering assurances so that AI systems are trustworthy and robust. An *AI Engineering* initiative focused on building new tools to extend and adapt Agile and DevSecOps methodologies, will allow U.S. AI practitioners to build in robustness and security and develop new tools for test, evaluation, validation and verification; monitoring; and assuring AI systems over their full domain of use and full life cycle.

Working in tandem with NIST and NSF initiatives, this *AI Engineering* effort will enhance the accumulation of best practices to establish and grow an engineering discipline for AI systems, the creation of new frameworks for sharing AI incidents, and incent the integration of ethical principles. The AI Engineering effort will lead to methods, practices, and tools for the development of reliable, responsible, and trustworthy AI which will be critical to public acceptance of AI-enabled products and services.

New tools will be required to mitigate the failure of AI systems through approaches such as enhanced algorithmic agility. In addition, research on actuarial risk methods will enable wellunderstood risk abatement approaches to create insurance for engineered AI systems, unlocking new markets for AI.

Finally, developing and communicating the discipline of AI Engineering are crucial to build public confidence in AI solutions of all kinds.

Advance Research to Support New Real-Time Policy Decision Tools and Public/Private Data Collaborations

As highlighted in the President's executive order calling for a national center for epidemic forecasting and analytics, AI is emerging as a powerful tool for policymaking. Collaborations combining a variety of public and private data sources, like CMU's COVIDCast, have advanced a new generation of epidemiological forecasting tools to shape public health policies. Furthermore, as noted by the National Security Commission on AI, machine learning supported the creation of next-generation, real-time decision support tools to aid the nation's governors in shaping public health and economic strategies.

The National AI Research and Development strategy has the opportunity to build upon the impact of these collaborations to serve public health and meet the goal of improving resilience and preparedness. Additional support for the development of epidemiological forecasting models and advanced visualization and human-computer interaction capabilities would create enhanced capabilities to protect against future pandemics and create new tools to support public health initiatives.

Applications of machine learning and natural language processing to public/private data collaborations could also create powerful, new real-time decision tools to strengthen domestic medical supply chains. Early tests of these data models during the pandemic highlighted opportunities to engage more small firms and increase the understanding of the roles specific supply, workforce and manufacturing capabilities play in determining the ability to rapidly expand personal protective equipment availability.

These same capabilities afford an opportunity to enhance the President's vision that investments in infrastructure will enhance U.S. manufacturing and leadership in critical technologies such as semiconductors, batteries and the AI supply chain. Moreover, securing American supply chains begins by advancing research breakthroughs that support U.S.-based competitive advantages. Carnegie Mellon faculty are finding that the enhanced understanding of supply chain dynamics informs research that can improve the likelihood of domestic production.

Finally, new decision tools can also contribute to the President's agenda to advance equity and inclusion by sparking a new generation of community development and empowerment strategies. Machine learning and data visualization are helping inform linkages between environmental justice and critical economic factors such as persistent redlining.

The National AI Research and Development Strategy should advance the creation of AI policy "testbeds" for building upon the breakthroughs demonstrated in response to the pandemic and harness the capabilities of powerful new data decision tools to meet key elements of the President's agenda**.**

Recommendations of Areas of AI Research to Address Bias, Enhance Equity and Expand Opportunity

Continue a Robust Research Agenda Focused on Fairness, Bias and Privacy

The National AI Research and Development Strategy and the National AI Innovation Act have advanced a focused research, development and education strategy to address issues of fairness and equity—including the commitment to center scale funding. The following are areas that can support the continued development of a vibrant research agenda.

Fairness and bias. Fairness and bias reduction are essential for AI systems. This goes far beyond generic, statistical aspects such as understanding the effect of unbalanced data. It requires understanding of notions like fairness and bias *in the context of a specific application and its desired outcomes*. Developing a common language and a set of approaches for this is still very much an open area of research. Understanding the role of human-generated data requires reaching out far beyond the disciplines of core AI and statistical methods, into the social sciences, psychology, philosophy, economics, and many others. These approaches are not yet fully developed.

Data privacy. Artificial intelligence and machine learning systems' reliance on data brings up a range of issues from preserving data privacy to mitigating the risks associated with centralized data aggregation. This is obvious in fields of application like health care but is equally important in virtually all fields of application.

There are several technical research thrusts in this area. One is the concept of distributed AI—which is the idea of bringing the computation as close as possible to the data collection to reduce the need for data migration. Another is the design of secure, distributed learning systems like federated learning. Yet another is the idea of "sharing without showing" data, which involves methods like data masking, for example.

Safety and reliability. Assessment of risks and impact requires understanding and modeling of the performance of AI systems and their behaviors--in particular, predictive models of performance that can place bounds on the performance of an AI system operating under different, changing conditions. For classical engineering systems, we have at our disposal a wide array of tools, of formal methods, of best practices built over a couple of centuries to evaluate and characterize systems. For AI systems, we do not have the equivalent toolbox, because the systems are constantly changing as they learn: Thus the performance of an AI system depends not just on its design but on the data that was used to train it. In addition, even when individual components can be characterized in isolation, measuring the outcomes of an end to end system in the context of an application remains challenging. AI technology is moving much faster than the development of assessment and characterization methods.

Human-machine teaming. Many if not most AI applications involve a "human in the loop." This is to say, they share the decision making and action taking with humans to a great degree. This adds a great deal of variables and complexity to the modeling of AI systems and their risks and limitations because it requires a deep understanding of human decision-making processes and their interaction, not just of a stand-alone, fully autonomous AI system. This requires cross disciplinary research that engages experts in diverse fields such as in social sciences, economics, design, game theoretic algorithms for modeling interactions and decision-making as well as in application domains.

Transparency/Explainability. Transparency and explainability of an AI system are essential for users in an application area to trust its decision making process. There is considerable technical research still needed in transparency and explainability. However, "transparency" does not necessarily mean that algorithms and data usage must be understood. It is much more practical to say they must be replicable and auditable in each application area. Such audits can then be shared for purposes of mitigating risk and ensuring accountability.

The vibrant research and education initiatives that are incorporated in the National AI Innovation Act are enhanced by the emergence of vibrant communities of interest emerging on U.S campuses. These communities of interest, such as the Responsible AI Initiative at Carnegie Mellon, foster strong multi-disciplinary dialogue and collaboration. The Strategy should enhance and foster engagement across the vital communities of interest.

Expand Support the Development of a National Network of AI Demonstration Projects and Testbeds

Building on the NAIIO's AI R&D Testbed Inventory, the Strategy should accelerate investments in demonstration projects and testbeds in a national AI research strategy would contribute to the ability of AI advances to supporting engineering AI into systems of societal importance can both show the utility of AI in such systems and create immediate economic benefit. As an example, the Metro Lab Network refined by CMU as a testbed model for smart city and transportation technologies research, development and deployment has been scaled to over 50 communities.

The NAIIO could convene the engineering and AI communities to work with mayors; industry; labor; key agencies such as Commerce, DOT, DOE, DOL and HUD; and other local stakeholders to design a model for AI Demonstration Projects and Testbeds." The demonstration projects and testbeds could focus on engineering AI into specific systems of societal importance. They could design, develop and deploy use-inspired AI algorithms and methodologies targeted to real-world problems, to address climate change, stimulate jobs and manufacturing and foster greater economic inclusion and equity. Each project could also explore how to integrate regulatory policies with these technology innovations, serving as a foundation for advancing new community partnerships informing future agency AI institutes.

Launch Grand Challenges to Spark AI Innovations to Address Learning Loss and Improve Training Outcomes

Research to enable AI to transform education and training has significant potential to advance innovations that can address inequality and expand economic opportunity. A targeted research initiative can make a vital contribution to America's post-pandemic resurgence. As many as one in 10 workers--nearly 17 million Americans – may be required to change occupations in the recovery from the pandemic. Data also suggest that the pandemic is resulting in students, particularly students of color, losing five to nine months of learning in mathematics.

Grand Challenges focused to address these two critical areas would galvanize the research community and deepen interagency collaboration on education and training. The challenges would foster public/private collaboration and partnerships across the education continuum and with organizations and institutions in underrepresented communities.

Grand Challenges should also help to deploy AI-based learning tools in K-12 education; demonstrate technologies and methodologies that blend informal and classroom learning; and pilot gamification, VR tools and machine learning applications to address learning loss and enhance career pathways through job matching. For example, a pilot to seed advances in machine learning to facilitate a granular understanding of the critical tasks within a given occupation and the connection of similar tasks between different occupations would enable hyper-focused rapid training initiatives to provide the specific skills needed to help a worker transition careers in weeks and not years and to enable high-school and community-college educated individuals into their first entry-level technology job, thereby providing a pathway to prosperity. Similarly, a Grand Challenge focused on developing tools to accelerate training for 100,000 workers, for example, could spark new collaborations across the workforce development, industry and academic communities.

Foster Strategies to Broaden Engagement in the AI Innovation Process

The accelerated pace of innovation renders linear models of research/innovation/training inadequate to realize the goal of broad-based participation in the AI economy. It is thus essential that new models foster broader engagement in the research process. One such model has emerged for engaging workers and labor directly in the development of AI applications.

Carnegie Mellon has been honored to collaborate with the American Federation of Labor and Congress of Industrial Organizations in the design and creation of its path-breaking, recently launched Technology Institute. This new Institute will enable high-level engagement on innovation with the broader labor movement and build lasting labor partnerships with universities. In addition, it will inject into our innovation policies a worker-centered perspective on research that is focused on the President's critical goals of job creation, equity and rebuilding U.S. domestic production in our supply chain. We have also engaged labor directly in federally funded research initiatives and have brought union-trained labor into our manufacturing lab operations.

Create a National Reserve Digital Corps to Accelerate AI Deployment in Federal and State Government

A powerful opportunity to jump-start job creation is through a collaboration among government, industry and academia to train professionals to accelerate the deployment of AI across government agencies through the creation of a National Reserve Digital Corps. Universities would recruit and train private-sector professionals and traditional students to engage with federal agencies in the areas of digital transformation, data management and analytics and AI. Those trained would then serve in government jobs to accelerate the deployment of AI across the public sector, creating new career pathways in the process.

Potential Strategic Directions Related to International Cooperation

OSTP and the NAIIO have effectively integrated strategic international collaboration into the essence of the U.S. national AI strategy. This work has spanned collaboration on major policy and standards issues---which was reflected in the September U.S. Europe Tech Summit held at Carnegie Mellon's Mill 19. There are also opportunities to leverage U.S. university research strengths in AI to advance broader international missions and objectives. These collaborations may need to strategically build education initiatives that can provide the foundation for future research engagements.

One model of this type of collaboration is Master of Science in Engineering Artificial Intelligence at Carnegie Mellon Africa. The program supports the development of advanced skills that will enable engineers to design powerful solutions to societal challenges. Students learn to combine a foundation in artificial intelligence, machine learning, and data science with their engineering, information technology, and software skills through theoretical and practical hands-on study of real-world applications. This type of initiative can provide the foundation for building future collaborations in areas that align with the focus of the National AI Research and Development Strategy.

Conclusion

The National AI Research and Development strategies have energized the U.S. AI research community and galvanized support that has resulted in the near tripling of federal nondefense AI R&D in the last five years. The opportunity with this update is to advance initiatives that focus more directly on advancing AI research to address major societal challenges and realizing the potential for the ethical development of AI to expand economic opportunity. Carnegie Mellon remains committed to the pursuit of this vital national mission.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Caroline Friedman Levy

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.

A Human-Shaped Hole: Filling the Behavioral Science Gap in AI Ethics Oversight

Caroline Friedman Levy, PhD

Artificial intelligence already permeates our lives contributing to advances that might alternately (even simultaneously) be considered mundane, dubious or miraculous. As the scope of the technology expands, the fathomless benefits AI promises are paralleled by equally immeasurable risks, with our window to manage the latter narrowing. An ever-increasing number of ethicists can enumerate the most prominent risks; yet startlingly few psychologists, or other experts in behavioral science, are working on the project of AI alignment. This knowledge gap is striking given the extraordinary investments being channeled into the technology's ability to master and influence human behavior. In neglecting more than a century of relevant behavioral research, the ethical AI community has left a human-shaped hole that weakens oversight proposals and jeopardizes the future of the project. In updating its strategic plan, I suggest that NAIRR needs to delineate an evidence-based approach to applying behavioral science to oversight policies that will increase the likelihood AI's evolving uptake aligns with fundamental ethical principles.

Strong Consensus on Known Risks

There is striking consensus—acknowledged by the tech community's most ardent investor-cheerleaders as well as its most disquieted ethicist-Cassandras—about the salient flaws of AI systems as used currently, and about known risks for the near future. Foremost among these are:

- Historical racial, gender, ethnic and other **biases** baked into data sets and algorithms currently influencing decisions as momentous as hiring, bail approval and health care allocation. Left unchecked in an oversight-free framework, the increasing speed, power, capability and permeation of advancing AI will broaden the impact of these biases, expanding and further reifying social inequalities.

- Without robust strategic leadership focused on education, training and industrial policy, we risk increased **unemployment** and ever-widening **income inequality**, with the economic benefits of AI technologies conferred primarily upon a fortunate few.

- Absent powerful policy interventions, AI-enhanced **offensive capabilities** —ranging from undetectable *deepfakes* and disinformation to the means to disrupt crucial infrastructure and initiate drone-swarms—will be increasingly accessible to terrorists, rogue states and other malefactors.

- AI advances allow for metastasized infringements into personal data **privacy**, with the surveillance taking place in authoritarian countries paralleled by more diffuse, corporate surveillance in the United States, minus any framework for accountability.

- Within just a few years, social media and other digital platforms have built algorithms that have allowed for a startling influence over consumer behavior and mastery of the **attention economy**. As AI-supported nudges advance, risks to social cohesion (not to mention our conception of free will) are likely to increase.

Disunity on Oversight

Despite this impressive consensus, there is far less agreement among stakeholders on how to devise governance and oversight frameworks to effectively reduce risks while supporting AI's evolution in a manner that will benefit all—at the *very least*: in a manner that won't harm already-vulnerable individuals and communities. Technology leaders, AI business consultants, think tank ethicists and regulatory appointees can articulate the key targets of oversight—essentially the obverse of the aforementioned risks: ensuring **safety** over speed, **fairness/equity**, **reliability, accountability**, and (less advocated-for in the business community) the clear means to opt-out and maintain **privacy**. At present, AI ethics oversight is relegated almost entirely to the benevolence of the technology companies creating these tools, and to the client-businesses applying them.

With AI capabilities hurtling forward and government regulation on a future horizon, is the tech industry capable of policing itself? The question may seem naive, yet the AI ethics community has few options for now other than to help the industry do so, preferably in partnership with NAIRR and other governmental agencies. Tech business leaders evoke the need to "align incentives" to ensure that AI's evolving uptake coheres with fundamental ethical principles. Pressed further, however, these leaders often betray a simplistic understanding of human behavior as consistent and rationally-motivated. Behavioral scientists can provide expertise critical to two distinct dimensions of AI ethics-building: 1) helping to establish an industry culture in which incentives are genuinely aligned with the commitment to ethical development; 2) contributing to developers' understanding of human-neural network interactions such that increasingly frequent and profound exchanges between people and systems will be consonant with fundamental ethical principles.

Baseline for Governance: A Shift in Tech Culture

In order for AI governance to be effective, the tech industry will need to undergo a tectonic cultural shift to establish an expectation of transparency with oversight agencies. We have decades of experience designing regulatory frameworks requiring a delineated degree of transparency from other industries that present profound societal risks—for example for financial markets and pharmaceutical companies—with the tech industry thus far evading such bespoke oversight. For external oversight to adequately ensure AI applications are safe, trustworthy and equitable, leaders of big tech firms must warm to a degree of openness which is currently culturally anathematic.

Tech Industry Culture: Consensus on Incentive-Alignment/Shallow Understanding of Behavior

We know through years of social psychology (now often termed behavioral economics) research that our behaviors are often more influenced by context and social reinforcement than by nominal rewards—constituting the basis of "nudge theory." The UK's quasi-governmental *Behavioural Insights Team*, summarized the components of social nudging via the EAST acronym—demonstrating that salutary behaviors can be encouraged by creating an environment that makes such behaviors **easy**, **attractive**, **socially-rewarding**, and **timely**. Whether intentionally or not, organizations are invariably "nudging" for behavioral outcomes among their constituents. An understanding of *nudge theory* can help leaders be more effective and deliberate in promoting ethical AI standards.

Leaders committed to building an AI tech culture prioritizing ethical principles can be similarly aided by tools developed over decades within the discipline of industrial-organizational (I-O) psychology. I-O psychologists are well-versed in the potential for incentive systems to backfire as well as motivate, with employees too often aiming toward discrete metrics which fog the core principles at stake. I-O psychologists have particular expertise in **defining** business-culture goals in operational terms, **diagnosing** misalignment (and the attendant *barriers*, *bottlenecks*, *skill deficits*, and *knowledge deficits*), **designing** organizational interventions, **evaluating** organizational interventions, and **reiterating** towards sustainable organizational changes.

Promoting Safety/Ethics in Human-Neural Network Interaction

With behavioral scientists a rare sighting in AI safety/ethics, developers have relied upon big data to train machine learning models in human behavior and values. But just a cursory glance at the ML-testing literature reveals the dangers of this approach. To what end have cognitive, clinical, social psychologists and other behavioral scientists built decades of research expertise on the "black box" of the human mind only to have this knowledge-base ignored in favor of a relatively undifferentiated trawl of the internet? As we engage in more frequent and potent interactions with AI applications, it becomes ever more critical that behavioral scientists play a

normative role in design, testing, and oversight. Amid the well-delineated AI risks, the threat of worsening mental health outcomes has been given short shrift.

Behavioral Science and NAIRR's Future

The interviews and literature reviews I've conducted, published in summary by the *Vanderbilt Project on Unity and American Democracy*, have alerted me to the profound risks of neglecting behavioral science in designing oversight mechanisms for increasingly powerful artificial intelligence applications. As NAIRR moves forward, transitioning from defining principles toward operationalizing and implementing governance that places fundamental rights and principles at the forefront of our AI standards, it is essential that the task force invests in behavioral science research and implementation expertise.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Center for AI and Digital Policy (CAIDP)

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.

Center for AI and Digital Policy

Comments of the

THE CENTER FOR AI AND DIGITAL POLICY

To the Office of Science and Technology Policy, on behalf of the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence (Select Committee), the NSTC Machine Learning and AI Subcommittee (MLAI-SC), the National AI Initiative Office (NAIIO), and the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), on the

National Artificial Intelligence Research and Development Strategic Plan

March 4, 2022

On behalf of the Center for AI and Digital Policy (CAIDP), we write in response to the RFI request on the National Artificial Intelligence Research and Development Strategic Plan (the "AI Strategic Plan").[1]

CAIDP is an independent non-profit organization that advises national governments and international organizations on artificial intelligence (AI) and digital policy. We work with more than 100 AI policy experts in almost 40 countries. In February 2022, we released the second edition of our report, *Artificial Intelligence and Democratic Values Index*[2], providing a comprehensive review of the AI policies and practices in 50 countries. Using a methodology to assess AI policies against democratic values and human rights, the Index includes detailed narrative reports, quantitative assessments, and ratings and rankings across a dozen metrics to measure progress towards human-centric and trustworthy AI values. The CAIDP currently serves as an advisor on AI policy to the OECD, the Global Partnership on AI, the Council of Europe, the European Union, and other international and national organizations.

We strongly support OSTP's proposals to update the AI Strategic Plan and appreciate the opportunity to provides comments. Our comments focus on:

Strategy 3: *Understand and address the ethical, legal, and societal implications of AI*;
Strategy 4: *Ensure the safety and security of AI systems*;

[1] Office of Science and Technology Policy, *RFI request on the National Artificial Intelligence Research and Development Strategic Plan* (Feb. 22, 2022) ("OSTP RFI on AI Strategic Plan"), https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and

[2] CAIDP, *Artificial Intelligence and Democratic Values Index* (2022), https://www.caidp.org/reports/aidv-2021/

Strategy 6: *Measure and evaluate AI technologies through standards and benchmark*s; and Strategy 7: *Better understand the national AI R&D workforce needs*.

CAIDP has already endorsed the AI Bill of Rights,[3] one of the OSTP's six policy priorities, and made specific recommendations for that initiative. [4] We recommended a small number of clear, powerful principles and unnecessary qualifiers, loopholes, and exceptions. We suggested building on prior AI policy initiatives such as the OECD AI Principles and the Universal Guidelines for AI (UGAI).[5] In October 2018, over 250 organizations and experts, representing more than 30 countries and including the American Association for the Advancement of Science, endorsed the UGAI.[6] The Universal Guidelines for AI are intended to maximize the benefits of AI, to minimize the risk, and to ensure the protection of human rights. UGAI, already widely endorsed by the AI community, provides a good starting point but there is more to do.

Regarding the AI Bill of Rights, CAIDP also urges proceeding on a bipartisan basis. Eliminating bias, promoting fairness, ensuring accountability, and transparency for AI-based systems could also help align the political parties behind a common national purpose.

We also call your attention to the 2022 G7 Leader statement endorsing "Human-Centric AI", calling for "robust transparency" to oppose algorithmic bias.[7] This is a powerful statement from world leaders to address a problem that OSTP has identified as one of the great challenges in the AI field. The G7 leaders, including the United States, also committed to working together for a "values-driven digital ecosystem for the common good that enhances prosperity in a way that is sustainable, inclusive, transparent and human-centric." They called for a "human-centric approach to artificial intelligence," building on the work of the Global Partnership for Artificial Intelligence (GPAI) advanced by the Canadian and French G7 Presidencies in 2018 and 2019 and looking forward to the GPAI Summit in Paris in November 2021.

---

[3] The White House, *Join the Effort to Create A Bill of Rights for an Automated Society* (Nov. 10, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/11/10/join-the-effort-to-create-a-bill-of-rights-for-an-automated-society/

[4] Lorraine Kisselburgh and Marc Rotenberg, *Next Steps on the AI Bill Of Rights*, Washington Spectator (Nov. 2021), https://washingtonspectator.org/author/lorraine-marc/; CAIDP, Public Voice, https://www.caidp.org/public-voice/

[5] *OECD AI Principles* (2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; The Public Voice, *Universal Guidelines for Artificial Intelligence* (2018) ("Universal Guidelines for AI"), https://thepublicvoice.org/AI-universal-guidelines/

[6] The Public Voice, *Universal Guidelines for Artificial Intelligence* – Endorsements (2018) https://thepublicvoice.org/AI-universal-guidelines/endorsement/

[7] *G7 Leaders Endorse Human-Centric AI, Call Out Bias,* (June 13, 2021), https://www.whitehouse.gov/briefing-room/statements-releases/2021/06/13/carbis-bay-g7-summit-communique/; see also *G7 Leaders Endorse Human-Centric AI, Call Out Bias,* CAIDP Update 2.24 (June 14, 2021), https://www.caidp.org/app/download/8326521963/CAIDP-Update-2.24.pdf

We write now to renew CAIDP's earlier recommendations, encourage the adoption of the AI Bill of Rights, and make additional proposals to advance the goals set out in the AI Strategic Plan.

Review/Recommendations for *Strategy 3: Understand and Address the Ethical, Legal, and Societal Implications of AI*

We applaud the goal of addressing the ethical, legal, and societal implications in AI. We further support the emphasis on fairness, transparency, and accountability as foundational values in designing ethical AI systems.

The Universal Guidelines for AI emphasize similar points. The Fairness Obligation (UGAI-4) states that institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions. The Fairness Obligation recognizes that all automated systems make decisions that reflect bias, but such decisions should not be normatively unfair or impermissible. There is no simple answer to the question on what is unfair or impermissible. The evaluation often depends on context, but the fairness obligation makes clear that an assessment of objective outcomes alone is not sufficient to evaluate a system. Normative consequences must be assessed, including those that preexist or may be amplified by an AI system. As OSTP Director Alondra Nelson has explained, the OSTP should be "open about the history of science and technology's flaws and failures."[8] The consequences of the deployment of technology must be assessed with an understanding of the past, and a future lens that protects human dignity and civil rights.

Strategy 3 could be further strengthened to incorporate *considerations related to sustainability, and environmental issues.*

Problem: Greater emphasis on research of societal, ethical implications of AI-related to sustainability required.

The National AI R&D Strategic Plan implements the National AI Initiative (NAII) Act of 2020. [9] This includes action to: "support research and other activities on ethical, legal, environmental, safety, security, bias, and other appropriate societal issues related to artificial intelligence." The OSTP AI Strategic Plan calls attention to "societal issues such as equity and

[8] Khari Johnson, *Alondra Nelson wants to make science and tech more just*, Wired (June 29, 2021), https://www.wired.com/story/alondra-nelson-make-science-tech-more-just/

[9] House of Representatives, National Defense Authorization Act for Fiscal Year 2021 (2020), https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210, 1210

climate change."[10] Moreover, Director Nelson has highlighted "groundbreaking clean energy investments" among six policy priorities for the agency.[11]

The need to focus on environmental issues for AI is timely.[12] The UNESCO Recommendation on the Ethics of AI focuses specifically on Protecting the Environment.[13] As the UNESCO Recommendation states:

> The Recommendation emphasises that AI actors should favour data, energy and resource-efficient AI methods that will help ensure that AI becomes a more prominent tool in the fight against climate change and on tackling environmental issues. The Recommendation asks governments to assess the direct and indirect environmental impact throughout the AI system life cycle. This includes its carbon footprint, energy consumption and the environmental impact of raw material extraction for supporting the manufacturing of AI technologies. It also aims at reducing the environmental impact of AI systems and data infrastructures. It incentivizes governments to invest in green tech, and if there are disproportionate negative impact of AI systems on the environment, the Recommendation instruct that they should not be used.[14]

AI should also be aligned with the United Nations Sustainable Development Goals[15] including cross-cutting environmental issues, as additionally emphasized by the OECD AI Principles, which have been endorsed by the United States.[16]

As it stands, Strategy 3 says little about environmental impact and sustainability. The Strategy should be revised to consider the carbon footprint of AI, modeling and data infrastructure, environmental degradation, and waste concerns.

Recommendation 1: CAIDP recommends an interdisciplinary perspective in developing, designing, and managing AI, specifically including environmental and climate research perspectives. The call for multidisciplinary perspectives lacks environmental science, ecosystem and resource management, as well as social science. OSTP Deputy Director Dr. Jane Lubchenco

[10] OSTP RFI on AI Strategic Plan.
[11] OSTP, The Director's Office (2022), https://www.whitehouse.gov/ostp/directors-office/
[12] Intergovernmental Panel on Climate Change (IPCC), Sixth Assessment Report (Feb. 28, 2022), https://www.ipcc.ch/assessment-report/ar6/
[13] *UNESCO Recommendation on the Ethics of AI* (2021), https://unesdoc.unesco.org/ark:/48223/pf0000377897
[14] UNESCO, *UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence* (Nov. 25, 2021), https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence
[15] *United Nations Sustainable Development Goals* (2015) https://sdgs.un.org/goals
[16] *OECD AI Principles* (2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

made this point at the recent White House Climate Roundtable.[17] While it is a positive step to call for the inclusion of interdisciplinary perspectives including engineering and "other disciplines,"[18] there is a clear need to address crucial AI environmental, energy, and equity impacts with expertise from the physical and social sciences.

Recommendation 2: CAIDP recommends making environmental impact a focus area for Strategy 3. Specifically, AI Sustainability and AI Development should be incorporated in the "Building ethical AI" and "Designing architectures for ethical AI" subheadings of Strategy 3.

In this regard, a focus on environmental sustainability can promote a well-being approach to human dignity and quality of life. Research has shown that AI-enabled systems require exponentially rising computing power. This increase in computing power requires substantial energy consumption, generating a huge carbon footprint and upending the green effects of digitalization. This problem has raised additional ethical concerns, as well as the well-being of the planet and thus humans.[19] To address this concern, more research should be focused on reducing AI energy consumption, environmental degradation, mineral extraction, and waste. Researchers are developing AI system for training and running certain neural networks that reduce the carbon emissions.[20]

Under this framework, the priority becomes the development of more efficient computing systems that as a goal will not damage the environment,[21] given that human well-being is dependent on ecological well-being. As such, it is of paramount importance to build efficient hardware and AI-based algorithms that require less energy to ensure improved computational efficiency and a smaller carbon footprint. This sets up the critical need to support AI governance frameworks that require the implementation of standards and independent oversight over carbon accounting. Furthermore, this framework would increase the demand for the inclusion of other disciplines like environmental science, geology, oceanography, planetary science, astrobiology, etc.

---

[17] OSTP, *Readout of White House Climate Science Roundtable on Countering "Delayism" and Communicating the Urgency of Climate Action* (Feb. 25, 2022), https://www.whitehouse.gov/ostp/news-updates/2022/02/25/readout-of-white-house-climate-science-roundtable-on-countering-delayism-and-communicating-the-urgency-of-climate-action/

[18] House of Representatives, *National Defense Authorization Act for Fiscal Year 2021* (2020) https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210, 1210

[19] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* In Conference on Fairness, Accountability, and Transparency (FAccT '21),(March 3–10, 2021), https://doi.org/10.1145/3442188.3445922

[20] H. Cai, et al., *Once-For-All: Train One Network and Specialize it for Efficient Development*, published as a conference paper at ICLR 2020, https://arxiv.org/abs/1908.09791.

[21] A. Gupta, *The Imperative for Sustainable AI Systems* (Sept. 18, 2021), https://thegradient.pub/sustainable-ai/.

These recommendations address the questions raised in the subheading "What uses of AI might be considered unethical?" In our view, issues of sustainability and the significant environmental impacts of AI systems (such as energy consumption, extraction of rare minerals, and pollution) should be a required dimension of AI development. Inclusion of the language above will help mitigate this concern.

Review/Recommendations for *Strategy 4: Ensure the Safety and Security of AI Systems*

Strategy 4 of the 2019 National AI R&D Strategic Plan updates the 2016 plan by focusing on the rapid growth in AI security and safety and stresses the need for creating robust and trustworthy AI systems.

We call your attention to two fundamental obligations for AI systems set out in the Universal Guidelines for AI, salient in ensuring safety and security: Obligations of Accountability (UGAI-5) and Public Safety (UGAI-8).[22] The obligation to be accountable for AI systems speaks to the ongoing need for assessment of the risks during the design, development, and implementation of systems. Developing standard risk analysis tools for AI systems must include assessment of risks at all levels, and defined context-specific benchmarks to indicate when a system is ready for deployment. It's essential that investments in ethics and social science research address questions responsibility and accountability. The institutions, the designers, and the operators of AI systems retain responsibility for the consequences of AI systems. As the Universal Guidelines for AI further state:

> Safety and security are fundamental concerns of autonomous systems – including autonomous vehicles, weapons, and device control – and risk minimization is a core element of design. Less certain, however, is how to determine and set standards for levels of autonomy across broad applications, and understanding levels of autonomy (and the correlated level of human control) is an interdisciplinary research challenge. The UGAI underscores the obligation of institutions to assess public safety risks that arise from the deployment of AI systems, and implement safety controls.[23]

While we agree that trustworthy AI is "a critical issue that requires Federal Government R&D investments, along with collaborative efforts among government, industry, academia, and civil society,"[24] *independent oversight, international cooperation, clear definitions, and system resilience* are necessary to achieve this goal. The three recommendations provided here are imperative to meet the goals set out in the OSTP AI Strategic Plan; most notably, the promise to

[22] Universal Guidelines for AI.
[23] Ibid
[24] OSTP, *The National AI R&D Strategic Plan: 2019 Update* (June 2019), https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf, 24

"build a society where everyone can live with equal dignity and hope and opportunity, as well as equal safety and security."[25]

Problem 1: The need for standardization and independent oversight.

Recommendation 1: New technologies such as AI pose new challenges for privacy, dignity, autonomy, and equality. Metrics for explainability, interpretability, and transparency should be established to protect fundamental rights, human well-being, and to increase public trust.[26] These metrics alongside Privacy Enhancing Technologies would help protect privacy.[27] Additionally, standardized metrics for explainable, interpretable, and transparent systems will increase users' trust in these systems. After standardization, an independent audit –for which its methodologies also require standardization– and the resulting evaluation must confirm the system performs as intended to be certified.

Problem 2: The need for international cooperation.

Recommendation 2: AI standards should be produced and harmonized at the international level (with primary locus being in intergovernmental fora and global standards bodies with strong NGO presence) to ensure common ground around security, safety, and system resilience. This determination should be made by diverse groups with a variety of expertise.[28] The process of developing standards should not be dominated or led by industry groups - the voices and concerns of civil society and affected communities should be effectively represented. Standard-setting activities should protect fundamental rights.[29] CAIDP recommends that these organizations publish annual reports that describe specific steps taken to ensure broad-based participation in the development of technical standards as well as the consideration of fundamental rights.[30]

Problem 3: The need for clear definitions and system resilience.

---

[25] The White House, *A New Chapter for the White House Office of Science and Technology Policy* (Feb. 17, 2022), https://www.whitehouse.gov/ostp/news-updates/2022/02/17/a-new-chapter-for-the-white-house-office-of-science-and-technology-policy/

[26] NIST, *U.S. Leadership in AI: A Plan for Federal Engagement for* Standard (July 2, 2019) (draft for public comment), https://www.nist.gov/system/files/documents/2019/07/02/plan_for_ai_standards_publicreview_2july2019.pdf

[27] The White House, *US and UK to Partner on Prize Challenges to Advance Privacy-Enhancing* Technologies (Dec. 8, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/12/08/us-and-uk-to-partner-on-a-prize-challenges-to-advance-privacy-enhancing-technologies/

[28] *CEN-CENELEC response to the EC white Paper on AI*, Version 2020-06, https://www.cencenelec.eu/media/CEN-CENELEC/Areas%20of%20Work/CEN%20sectors/Digital%20Society/Emerging%20technologies/cen-clc_ai_fg_white-paper-response_final-version_june-2020.pdf

[29] EU-US Trade and Technology Council, *Inaugural Joint Statement* (Sept. 29, 2021), https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_21_4951

[30] *CAIDP Statement to European Commission on Proposed AI Act* (July 2021). https://www.caidp.org/statements/

Recommendation 3: Standardization, independent audit, system certification, and determination of international common ground depend on three foundational requirements: (1) consensus-based provision of precise definitions and terminology of technical terms (e.g. AI, automation, explainability, interpretability, transparency) for standardization and determination of international common ground;[31] (2) continuation of system updating to include (a) new data resulting from a data-centric strategy for system integrity and thus model improvement as data evolves,[32] and (b) new core AI functionalities resulting from rapid AI advances to maintain system resilience against adverse conditions like cybersecurity risks;[33] and (3) consideration of practices of inclusive design for AI systems.[34]

Review/Recommendations for *Strategy 6: Measure and Evaluate AI Technologies through Standards and Benchmarks*

Strategy 6 establishes that "standards, benchmarks, testbeds, and their adoption by the AI community are essential for guiding and promoting R&D of AI technologies."[35] This section also identifies developing a broad spectrum of AI standards, establishing AI technology benchmarks, increasing the availability of AI testbeds, and engaging the AI community in standards and benchmarks as areas for improvement.

We call your attention to the UGAI principles standards and benchmarks, Assessment and Accountability (UGAI-5) and Accuracy, Reliability, and Validity (UGAI-6).[36] Assessment determines whether an AI system should be established. AI systems should be deployed only after an adequate assessment of its purpose, objectives, risks, and benefits. Imperatively, such assessments must include a review of individual, societal, economic, political, and technological impacts, and a determination can be made that risks have been minimized and will be managed. Individual level risk assessments might include a fundamental rights impact assessment; societal level risk assessments might involve public health or economic impact assessments. If an assessment reveals substantial risks, especially to public safety and cybersecurity, then the project

[31] Krafft, P. M., Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. "Defining AI in policy versus practice" In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020)*, pp. 72-78. 2020.
[32] Gerdes, Anne. "A participatory data-centric approach to AI Ethics by Design." *Applied Artificial Intelligence* (2021): 1-19.
[33] Eigner, Oliver, Sebastian Eresheim, Peter Kieseberg, Lukas Daniel Klausner, Martin Pirker, Torsten Priebe, Simon Tjoa, Fiammetta Marulli, and Francesco Mercaldo. "Towards Resilient Artificial Intelligence: Survey and Research Issues." In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 536-542. IEEE, 2021.
[34] Berkman Klein Center, AI and Inclusive Design, https://aiandinclusion.org
[35] OSTP, *The National AI R&D Strategic Plan: 2019 Update* 33 (June 2019), https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf
[36] Universal Guidelines for AI

should not move forward. Accountability for the outcomes and consequences of AI systems lies with the institutions. As the UGAI states:

> Institutions have the obligation to ensure the accuracy, reliability, and validity of AI systems. Benchmarks should be developed against which these standards can be measured. For example, standards should demonstrate that the AI system has been tested for reliability and external validity (i.e., is valid within the population and application context in which it will be deployed). If developed using value-sensitive design, and trained on datasets that are appropriate for a specific user population, AI algorithms and technologies embedded within those contexts will reflect its values, and perform reliably. For example, systems modeled on a dataset of young adults from the United States is likely not to have validity if deployed in a population of aging seniors in Africa because of demographic, cultural, and biological differences.[37]

We encourage adoption of UGAI key principles and make additional proposals. We recommend improving standards and benchmarks by *adding social impact of technology as a separate standard, creating standards that can adapt and keep pace with the speed of technological evolution, and increasing engagement with a diverse community of AI stakeholders*.

Problem 1: Are these benchmarks technical, social impact, or both?

As the report mentions, NIST has planned to develop a broad spectrum of AI standards which include software engineering, performance, metrics, safety, usability, interoperability, security, privacy, traceability, and domain, not including societal impacts.

Recommendation 1: Social impact of technology needs to be added as a separate standard. Most AI-based solutions directly or indirectly affect society disproportionately, therefore, Social Impact Assessment based on model risk framework: to define social impact level for different AI systems based on domain and grading technologies and impact level [38] [39] should be added as a standard. There should be AI risk level-based standards which include high, medium, and low risk AI standards.[40] For creating benchmarks and standards, it is also essential to understand and

[37] Ibid
[38] OECD, *Policy Brief on Social Impact Measurement for Social Enterprises* (2015), https://www.oecd.org/social/PB-SIM-Web_FINAL.pdf
[39] Floridi L, Cowls J, King TC, Taddeo M. *How to Design AI for Social Good: Seven Essential Factors*. Sci Eng Ethics. 2020 Jun;26(3):1771-1796. doi: 10.1007/s11948-020-00213-5. Epub 2020 Apr 3. PMID: 32246245; PMCID: PMC7286860.
[40] European Commission, *Regulatory framework proposal on artificial intelligence*, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

determine the levels of risk of AI systems—without understanding the impact of the AI system on human rights, there is little evidence and knowledge for detecting the risk level. We recommend including the Human Rights, Democracy, and Rule of Law Impact Assessment (HUDERIA) proposed by the Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI) as a possible benchmark for determining the risk level associated with AI systems.[41]

Problem 2: Lack of standards that can adapt and keep pace with technological evolution.

The strategy states that "standards must be hastened to keep pace with the rapidly evolving capabilities and expanding domains of AI applications."[42] Two key aspects that could be considered when developing standards are developing processes for creating standards faster and ensuring that standards are evaluated and changed in line with the pace of technology.

Recommendation 2: OSTP works with public partners, including the IEEE and ISO, towards the development of processes and procedures that will help to reduce the time to create AI standards. Currently, there are multiple AI standards, benchmarks, and policy frameworks developed by private organizations, NGOs and international organizations such as OECD. We recommend partnering with organizations which have inclusive participation mechanisms to create and change standards faster in line with the pace of technology.

Problem 3: More effort needs to be given towards ensuring engagement with a diverse community of AI stakeholders.

NIST AI standardization activities include engagement with standards organizations and plans to engage with AI communities that have diverse backgrounds made up of users, industry, academia, and government.

Recommendation 3: OSTP must encourage the engagement of diverse communities of AI by: proactively identifying local and marginal communities and indigenous groups; and including diverse stakeholders from domain experts, academic, private, government, and social sectors, including representatives from different sized organizations; to ensure fairness and prevent bias in development of standard and benchmarks. Standard-setting activities must have safeguards in place to balance the industry interests with societal and fundamental rights concerns. If necessary, special funding and membership paths should be available for marginal communities.

---

[41] Council of Europe AD HOC Committee on Artificial Intelligence (CAHAI), *Human Rights, Democracy and Rule of Law Impact Assessment of AI systems* (Mar. 11, 2021), https://rm.coe.int/cahai-pdg-2021-02-subworkinggroup1-ai-impact-assessment-v1-2769-4229-7/1680a1bd2d

[42] OSTP, *The National AI R&D Strategic Plan: 2019 Update* (2019), https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf, 33

Review/Recommendations for *Strategy 7: Better Understand the National AI R&D Workforce Needs*

As the speed and scale of AI technologies expand both domestically and globally, much care must be given to ensure the proper recruitment and retention of AI researchers and practitioners to ensure a viable pool of talent for tomorrow's intellectual demands. Strategy 7 states, "It is critical to maintain a robust academic research ecosystem in AI that, in collaboration with industry R&D, can continue to deliver tremendous dividends by advancing national health, prosperity, and welfare, and securing the national defense."[43] As CAIDP stated in our June 2021 statement to U.S. financial agencies, we believe AI systems should be designed in a way that respects the rule of law, human rights, democratic values, and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society."[44] Further, the UGAI indicated several key principles that are critical considerations for workplace need strategies including and relevant here, including the right to transparency (UGAI-1) and human determination (UGAI-2), and obligations of fairness (UGAI-4).[45] Using that perspective, we sense that Strategy 7 could be enhanced by *ensuring interdisciplinary education, reforming fellowships and educational programs, and ensuring the inclusion of diverse voices among educators and students.*

Also, CAIDP recommends following best practices for AI procurement decisions and adopting talent management strategies that are needed to support oversight mechanisms to ensure the protection of human rights and wellbeing of citizens and accountability, transparency and fairness of AI systems. The responsible use of emerging technologies can be supported by public policy improvements to government procurement processes. Through reform of the guidelines by which the government exercises its purchasing power, governments can fulfill their leadership role in steering technology policy. As explained in the World Economic Forum's Unlocking Public Sector AI through Government Procurement initiative AI Procurement in a Box report:

> Government procurement officials cannot be expected to have the most up-to-date knowledge in every highly specialized field. To safeguard the responsible future use of AI technologies, a multistakeholder effort with cross-sector participation and interdisciplinary expertise is required to create authoritative guidelines. The procedural norms are even more urgent now. What information should be recorded

[43] OSTP, The National AI R&D Strategic Plan: 2019 Update (Jun 2019), https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf, 37
[44] CAIDP, *CAIDP Statement to U.S. Financial Agencies on Use of AI by Financial Institutions* (June 30, 2021), https://www.caidp.org/statements/
[45] The Public Voice, *Universal Guidelines for Artificial Intelligence: Endorsement* (2018), https://thepublicvoice.org/AI-universal-guidelines/endorsement/

and how explanations need to be documented is what lays the foundation for fairness and impartiality in the administrative process. To preserve due process and predictability, a coalition can help ensure that the right questions are asked.[46]

Problem 1: Multidisciplinary teams composed of stove-piped experts are insufficient.

Recommendation 1: OSTP must work to ensure collaborators have both deep subject matter expertise and interdisciplinary knowledge to readily build more effective connections. Computing is an interdisciplinary field that requires innovative interdisciplinary education. For current and future workforce talent to successfully manage the interdisciplinary nature of computing, education must not just expose learners to different disciplines' knowledge but integrate the disciplines within the learning process. As the demand for multidisciplinary teams grows, such teams can no longer rely exclusively on single subject matter experts.

Problem 2: Increasing educational programs, fellowships, and activities dedicated, exclusively to the quantitative fields of ML is insufficient.

Recommendation 2: Learning from K-12 to postgraduate education needs reforming and resources to sustain. To achieve success with recommendation (1), curricula will need to balance in-depth subject matter with expansive related subject matter coverage, as well as traditional assessments with creative and experiential practice-based learning experiences.

Problem 3: Insufficient care given to ensure effective representation by underrepresented groups.

Recommendation 3: From the instructors, faculty, researchers, designers, developers, project managers and directors to the learners themselves, diverse voices must not only be included at the planning stage of research projects but integrated within the various phases of the AI lifecycle of design, development, and deployment. This necessarily implies respectfully accounting for the different experiences of all through, for example, data acquisition and user experience feedback. Echoing the OSTP recommendation for STEM equity,[47] we support diverse representation at all levels and across all AI disciplines.

[46] World Economic Forum, *AI Procurement in a Box: Project overview* (June 2020), https://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_Project_Overview_2020.pdf

[47]The White House, A New Chapter for the White House Office of Science and Technology Policy (Feb. 17, 2022), https://www.whitehouse.gov/ostp/news-updates/2022/02/17/a-new-chapter-for-the-white-house-office-of-science-and-technology-policy/

CAIDP encourages OSTP, the Select Committee, and NAIIO, in consultation with the NSTC Subcommittee on Machine Learning and AI and the NITRD AI R&D Interagency Working Group, to incorporate the recommendations above in the AI Strategic Plan.

Finally, we call attention to the UNESCO Recommendation on the Ethics of Artificial Intelligence.[48] CAIDP contributed to the development of the Recommendation,[49] and expressed "strong support for adoption."[50] The United States is not yet a signatory to the UNESCO AI Recommendation. We strongly encourage OSTP to advance U.S. global leadership in AI ethics and urge the Administration to endorse the UNESCO Recommendation.

Thank you for your consideration of our views.[51] We welcome the opportunity to discuss further.

Marc Rotenberg
CAIDP President

Merve Hickok
Research Director

Karine Caunes
Global Program Director

Jason Johnson
CAIDP Research Fellow

Tamra Moore
CAIDP Research Fellow

Somaieh Nikpoor
CAIDP Research Fellow

[48] UNESCO, Recommendation on the Ethics of Artificial Intelligence (2021), https://en.unesco.org/artificial-intelligence/ethics
[49] CAIDP, CAIDP Update 1.4– "UNESCO Pursues Humanistic Approach for AI" (Aug. 10, 2020), https://www.caidp.org/app/download/8292333863/CAIDP-Update-1.4.pdfs
[50] CAIDP, CAIDP Update 2.41– "UNESCO Pursues Humanistic Approach for AI" (Nov 25, 2021), https://www.caidp.org/
[51] CAIDP acknowledges the significant contributions to this statement of the 2022 CAIDP Research Group, North America Team, including Sharvari Dhote, Kathy Kim, Monica Lopez, Nidhi Sinha, and Narain Yucel.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Center for New Democratic Processes (CNDP)

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.

Center for New Democratic Processes

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

Document Citation:
87 FR 5876

Page:
5876-5878 (3 pages)

Document Number:
2022-02161

The Center for New Democratic Processes (CNDP) commends the Office of Science and Technology Policy (OSTP), on behalf of the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence (Select Committee), the NSTC Machine Learning and AI Subcommittee (MLAI-SC), the National AI Initiative Office (NAIIO), and the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) for seeking public comment and stakeholder input on the critical issues surrounding the future development and improvement of the National Artificial Intelligence Research and Development Strategic Plan.

We are pleased to respond to this Notice of Request For Information (RFI) on the ways in which the strategic plan should be revised and improved by outlining potential benefits of deliberative civic engagement processes to support the development of policies and guidance related to the topics identified in this Request.

For the purposes of this response, we focus our comments most specifically on Strategies 3, 4, and 8 of the *National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* included in the Supplementary Information section of the Notice of Request For Information (RFI):

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

Strategy 4: Ensure the safety and security of AI systems.

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

The work of the affiliated agencies and committees represented in this RFI, as well as the National AI Strategy itself, could be strengthened by the utilization of deliberative civic engagement methods (such as Citizens' Juries) to engage the public (constituents, consumers, patients, and residents) in the development of policies and guidance on key ethical, legal, and regulatory issues, information governance and data privacy issues, the application of artificial intelligence, automated decision-making, machine learning technologies, algorithmic transparency and accountability, the public and private sector uses of biometric technologies, and other emerging AI applications. The use of purposefully designed deliberative civic engagement processes would also bolster the Office's work to establish a "Bill of Rights for an Automated Society" that works for all.

Deliberative engagement on the public and private sector uses of artificial intelligence, the legal and ethical issues surrounding AI, and the societal implications of AI can promote diversity and equity in shaping data collection, data storage and management practices, developing regulatory oversight and guidance, and creating robust, equitable policy solutions. This can be achieved by meaningfully involving those who are directly impacted by policies and who have been historically excluded from decision-making processes and policy development - serving as both stakeholders and participants in deliberative events.

We encourage the OSTP, as well as affiliated agencies and committees under the auspices of this RFI, to pursue the use of deliberative civic engagement methods as the Office undertakes efforts to gather information and inform guidance and policy development regarding:

A. Stakeholder engagement practices for systems design, procurement, ethical deliberations, approval of use, human or civil rights frameworks, assessments, and strategies to mitigate the potential harm or risk of AI;
B. Best practices or insights regarding the design and execution of pilots or trials to inform further policy developments;
C. Practices regarding data collection (including disclosure and consent), review, management (including data security and sharing), storage (including timeframes for holding data), and monitoring practices;
D. Safeguards or limitations regarding approved use (including policy and technical safeguards), and mechanisms for preventing unapproved use;
H. Practices for public transparency regarding use (including notice of use), impacts, opportunities for contestation and for redress, as appropriate;

I. Clarity pertaining to the legal use, including intra-agency use and exchange, of personal data such as biometric data, health data, and other information that is used for secondary purposes through emerging data initiatives;
J. The reasonable responsibilities of private enterprises serving as data managers and brokers in their collection, storage, and disposal of personal data on behalf of public institutions or clients.

CNDP has demonstrated the potential impact and contributions of deliberative civic engagement to support the development of guidance and regulatory frameworks in partnership with government bodies on a range of emerging technology issues. These include but are not limited to: artificial intelligence and secondary uses of personal data among public and private sector entities (as well as how these organizations interface with one another through data sharing and information governance arrangements), explainability and performance in artificial intelligence and automated decision making processes, COVID-19 data sharing initiatives, as well as reasonable expectations for consent and opt-in vs. opt-out procedures for secondary uses of patient health records.

In 2021, CNDP conducted a series of deliberative projects to shape national policy regarding COVID-19 data sharing initiatives on behalf of the National Health Service (England), the National Data Guardian for Health and Social Care, and NIHR-ARC Greater Manchester.

The following documents present the findings from this project.

- The Full Report from the Pandemic Data Sharing Citizens' Juries (three citizens' juries) which were conducted in early to mid-2021.
- The Executive Summary from the Pandemic Data Sharing Citizens' Juries (three citizens' juries) which were conducted in early to mid-2021.

In 2019 CNDP conducted a pair of Citizens' Juries on behalf of The National Institute for Health Research (NIHR) Greater Manchester Patient Safety Translational Research Centre (PSTRC) and the Information Commissioner's Office in the United Kingdom that focused on the tradeoffs between explainability and performance when AI-powered automated decision making in systems directly impacting individuals. This pair of citizens' juries assessed a range of scenarios including healthcare diagnosis, organ transplant matching, employment screening, and criminal justice sentencing practices.

The following document presents the findings from this project.

- The Full Report from the AI (explainability and performance) Citizens' Juries (two citizens' juries) conducted in 2019.

The following articles, documents, and posts from project sponsors demonstrate how jury outcomes have been incorporated and/or responded to regarding emerging technology policy, data privacy, artificial intelligence, and information governance issues.

- From the National Data Guardian (Dr. Nicola Byrne) (project sponsor) re: Pandemic Data Sharing Juries (2021).
- From Greater Manchester National Institute of Health Research and NHSX (project sponsors) re: Pandemic Data Sharing Juries (2021).
- From the Information Commissioner's Office (project sponsor) re: AI Citizens' Juries and Interim Report from ICO on use of Jury Findings (2019).
- From the Greater Manchester Patient Safety Translational Research Centre (project sponsor) re: AI Citizens' Juries (2019).
- From the National Data Guardian (project sponsor) re: Reasonable Expectations for Data Sharing Citizens' Jury (2018).

As described on the Brookings Institution blog, "Citizens' Juries are valuable, we believe, as tools for improved policymaking. But their value may go beyond any specific use, in part because their use would demonstrate greater trust in and respect for the people. Adopting a more open version of democracy — such as one in which Citizens' Juries are positioned to purposefully shape policymaking — provides the public with a structured opportunity to directly voice their opinions and influence decision-making… An approach which purposefully situates Citizens' Juries in policy development communicates to community members and constituents that their viewpoints matter and are trusted enough to be included in the decision-making process. This could be accomplished by developing clear channels for the incorporation or adoption of jury findings and results by policymakers and decision-making bodies."

We provide this comment in response to the Request For Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan to encourage the OSTP and affiliated agencies and committees to supplement their ongoing information gathering, stakeholder input, and policy development efforts through the use of deliberative civic engagement processes.

The Center for New Democratic Processes welcomes the opportunity to collaborate with the OSTP, the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence (Select Committee), the NSTC Machine Learning and AI

Center for New Democratic Processes

Subcommittee (MLAI-SC), the National AI Initiative Office (NAIIO), and the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) on efforts to utilize deliberative civic engagement on emerging technology issues and updates to the National AI Research and Development Strategic Plan.

Organizational Contact

Please contact Kyle Bozentko, Executive Director of the Center for New Democratic Processes, (email: kyle[at]cndp[dot]us) for further information or with any questions.

About Us

The Center for New Democratic Processes is a nonpartisan, nonprofit civic engagement organization based in St. Paul, MN with global partners and clients. Our mission is to strengthen democracy by partnering with individuals, communities, and institutions to design and implement informed, innovative, and democratic processes to address today's toughest challenges. We provide an interdisciplinary, customized approach to the design and implementation of each deliberative process and engagement project we undertake.

Since 2012 CNDP has conducted over 150 multi-day deliberative events (Citizens' Jury, Citizens' Assembly, Community Panel, Policy Juries, etc.,) and deliberative forums (single day events) on a broad range of complex policy issues and research programs. We've advised governments and teams in Argentina, Australia, Canada, Japan, Portugal, Scotland, Singapore, South Korea, and the UK on the effective design and implementation of civic engagement strategies and public participation projects. Our recent projects have informed data privacy and governance, technology policy and artificial intelligence (AI) regulatory guidance, and shaped national policy regarding data sharing initiatives that emerged in response to the COVID-19 pandemic on behalf of the NIHR-ARC Greater Manchester, the National Health Service (England) and the National Data Guardian for Health and Social Care. We've supported rural communities responding to local impacts of climate change and extreme weather through the Rural Climate and Energy Dialogues. We worked with stakeholders to guide significant infrastructure and planning decisions with the City of Vancouver (British Columbia) through the Flats Arterial Community Panel. We designed and delivered the first Citizens' Assembly in the United States through the MN Community Assembly Project. We are currently working with the University of Liverpool and Pfizer Inc. who've commissioned the Liverpool Citizens' Jury on Antimicrobial Resistance to explore attitudes and perspectives about relationships among public and private entities collaborating to monitor and develop responses to antimicrobial resistance.

Our History

For nearly fifty years we've been expanding the boundaries of democracy through ongoing experimentation and implementation of groundbreaking deliberative processes. We were the first to employ the Citizens' Jury, invented by our founder (Ned Crosby), as a method for participatory deliberative engagement in the United States. Since its introduction, we have supported the global proliferation of this method. Throughout the 1970s, 80s and 90s our work focused on refining the use of the citizens' jury on issues ranging from Ag Impacts on Water Quality, Organ Transplants, and School-based Clinics to the Federal Budget and evaluating the positions of candidates for US Senate. In the early 2000s we covered topics such as global climate change, piloting the Citizens' Initiative Review in the state of Washington, advancing the use of deliberative democracy in Australia, and improving Electoral Recounts.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Center for Security and Emerging Technology (CSET), Georgetown University

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 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.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Cindy Mason

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.

Subject: RFI Response: National Artificial Intelligence Research and Development Strategic Plan

Date: Thursday, March 3, 2022 at 9:17:04 PM Eastern Standard Time

From: Cindy Mason

To: AI-RFI

Hello,

I am an AI researcher for 30 years. My mentors were John McCarthy at Stanford
and Lotfi Zadeh at UC Berkeley. I am a white female. I worked on AI at NASA Ames,
Stanford and UC Berkeley.

-The toxicity of our workplace is too long to tell here, but it is an on-going problem that impacts
creativity and inclusivity. Please include equal number of women, both old and young.
Esp. seek advice from Melinda Gates, Joy Boulamwini, etc.

-add funding opportunities and support for independent AI researchers.
AS MORE universities adopt a variety of strategies for coping with Pandemic, rising number of researchers are working independently. Some leO
because their environment was too toxic.

-Software vedng process for 'sensi-ve' AI products
Strategy 4: add "including Software vedng process for 'sensi-ve' AI products"
That may affect "personhood"

-Support for GDPR

I am available for further discussion.

Best,
Dr. Cindy Mason
Independent Researcher
Workshop Chair, AAAI/IAAI-22 Diversity and Inclusion Tract

--
If you think you are too small to make a difference
you have not been in bed with a mosquito. Chinese fortune cookie

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Coalition for Independent Tech Research

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.

The Essential Role of Industry-Independent Research in U.S. National AI R&D

Coalition for Independent Tech Research *March 4, 2022*

The National AI Initiative Act is designed to ensure that the United States remains a leader on AI research and development. It sets out to leverage the opportunities of AI to increase the well-being of Americans while safeguarding against potential and actual harms to our communities. As Lynn Parker noted in her February blog post announcing an update of the National AI Research & Development Strategic Plan, achieving this goal requires "consistent and sustained Federal investments in cutting-edge AI R&D, *particularly for those areas in which industry has few incentives to invest.*"

One of those areas is research on how artificial intelligence is shaping the information infrastructure of America and the world. Right now, a handful of companies control artificial intelligence systems that already shape public health, education, stock markets, public services, and political discourse, just to name a few. The way these AI systems work, and the impacts that they have, are poorly understood. This is in large part because the companies control access to the systems necessary to understand these questions. The primary motivation of these companies is to make money, not to produce research in the public interest. Industry has insufficient incentives to open these systems to people outside of companies—to independent researchers at news organizations, academia, and civil society groups—who could study their impacts on the well-being of the public.[1]

Tech companies have repeatedly demonstrated that they will not permit oversight of these systems on their own. Independent researchers have attempted for years to collaborate with tech companies on methods to make their AI systems safer, to no avail. Many companies have made grand promises and then failed to share essential information with public interest researchers.[2] Those who have tried to study these systems using their own tools and accounts have had their access revoked.[3]

Through the National AI Initiative Act, the US government can make a positive intervention to change this. **The implementation of this Act should support a new system of governance and external oversight of AI systems that shape America's communication infrastructure.** It should do this by:

---

[1] Whittaker, M. (2021). The steep cost of capture. *Interactions*, *28*(6), 50-55.
Zuckerman, E. (2021) Demand five precepts to aid social-media Watchdogs. Nature.
Matias, J.N. (2020). Why we need industry-independent research on tech & society. Citizens & Technology Lab
Haibe-Kains et al (2020) Transparency and reproducibility in artificial intelligence. Nature.

[2] Seetharaman, D. (2020) Jack Dorsey's push to clean up Twitter stalls. Wall Street Journal.
Hegelich, S. (2020) Facebook needs to share more with researchers. Nature.

[3] Edelson, L. (2021) How Facebook Hinders Misinformation Research. Scientific American.

- Increasing support for industry-independent research to spur innovation, protect the public, advance science, and contribute to AI governance.
- Facilitating the availability of and equal access to people, systems, and data in a way that upholds the highest standards of ethics and privacy.
- Ensuring that industry-independent researchers from civil society, news organizations, and academia are included in the implementation of the National AI Initiative Act.

While the current strategic aims of the National AI R&D Strategic Plan are laudable, they cannot be achieved without the participation of independent researchers, including journalists and those representing civil society organizations. Much of the most influential and most-cited research on the impacts of artificial intelligence on society have been conducted by journalists, citizen scientists, and civil society, including research on pre-trial risk assessment,[4] predatory targeted advertising,[5] flawed predictive policing,[6] content moderation systems,[7] market algorithm discrimination,[8] and harmful content from search engines.[9]

The most impactful and immediate way to support research to evaluate and address social concerns related to the use of AI is to empower people to do this work outside of industry. Expanding the participation of journalists and civil society researchers in particular will provide more inclusive pathways for more Americans to participate in AI R&D.

The Coalition for Independent Technology Research is a new group of academics, journalists, civil society researchers, and community scientists who work independently from the technology industry. Our mission is to advance, defend, and sustain the right to ethically study the impact of technology on society.

We recommend that the National AI R&D Strategic Plan:

1. **Increase support for industry-independent research to understand societal issues related to artificial intelligence.**

Strategies 3 and 4 of the National AI R&D Strategic Plan aim to understand and address the ethical, legal, and societal implications of AI and ensure the safety and security of AI

[4] Angwin, J., Larsen, J. (2016) Machine Bias. ProPublica.

[5] Riecke, A., Koepke, l. Led Astray. Upturn

[6] Main, F., Dumke, M. (2017) A look inside the watch list Chicago police fought to keep secret. Chicago Sun-Times

Saunders, J., Hunt, P., & Hollywood, J. S. (2016). Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. Journal of Experimental Criminology, 12(3), 347-371.

[7] Matias, J., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). Reporting, reviewing, and responding to harassment on Twitter. Available at SSRN 2602018.

Matias, J. N., Hounsel, A., & Feamster, N. (2022). Software-Supported Audits of Decision-Making Systems: Testing Google and Facebook's Political Advertising Policies. Computer-Supported Cooperative Work.

[8] Cox, M. (2017) The Face of Airbnb, New York City.

[9] Kayser-Bril, Nicolas (2020). Ten years on, search auto-complete still suggests slander and disinformation. AlgorithmWatch

systems. This cannot be done within the current paradigm where tech company employees exclusively control access to systems, data, and affected communities. These companies have amassed large teams of talented researchers, but their research studies are aimed at supporting the corporate rather than the public interest. This is a perilous paradigm, akin to asking the public to trust automakers to be the only ones to perform safety tests on the cars that they manufacture. Research from companies should not be dismissed, but it must be part of a system of oversight that also includes researchers who are independent of the companies' corporate interests.

Industry independent research can play an important role in developing a trustworthy American artificial intelligence industry. According to research by Pew, Americans do not believe leaders of tech companies admit to mistakes, do not believe that tech leaders care about people like them, and trust technology leaders less than any other group.[10] As has been the case in other industries, independent research can provide trustworthy evidence about risks and safety in ways that enhance the common good.[11]

Industry-independent research is essential for making progress on these strategies; yet many of the programs implemented within the Strategic Plan so far are being carried out with industry. For example, the Strategic Plan 2019 Update cites a project between NSF and Amazon to "jointly support research focused on AI fairness with the goal of contributing to trustworthy AI systems that are readily accepted and deployed to tackle grand challenges facing society." The updated National AI R&D Strategic Plan should include specific goals for supporting industry-independent research within the implementation of Strategies 3 and 4. This could include prioritising industry-independent actors for funding, redirecting funding away from industry-dominated projects, and developing funding partnerships with funders outside of the AI industry.

While the addition of Strategy 8, "Expand Public-Private Partnerships to accelerate advances in AI," to the Strategic Plan is a promising step in this direction, the implementation of that strategy has relied on industry-dominated groups like the Partnership on AI, which has been criticised by civil society organisations[12] for doing little to change the attitudes of member companies or foster genuine dialogue with civil society on a systematic basis. This dynamic is common in partnerships like these; it is caused by a fundamental mismatch in incentives and power between different members of the group and is rarely solved by governance protocols. The agencies responsible for the implementation of Strategy 8 should consider how these dynamics can prevent progress towards the aims of the Strategic Plan and prioritise partnerships that are truly independent of the tech industry.

---

[10] Pew Research Center. (2019). Why Americans Don't Fully Trust Many Who Hold Positions of Power and Responsibility.

[11] Carpenter, D. (2014). *Reputation and power*. Princeton University Press.
Silber, N. (1983). Test and Protest: The Influence of Consumers Union. Holmes & Meier.
Dietz, T., Ostrom, E., & Stern, P. C. (2003). The struggle to govern the commons. science, 302(5652), 1907-1912.

[12] (2020) Access Now Resigns from the Partnership on AI. AccessNow

2. Facilitate the creation of curated, standardized, secure, representative, aggregate, and privacy-protected data sets to enable independent research.

Agencies implementing the National AI R&D Strategic Plan have made significant progress on Strategy 5, which aims to release public datasets to advance the field of AI research. Much of this work has focused on areas such as computer vision, natural language processing, and speech recognition, as well as getting public agencies to contribute data to such initiatives. These efforts are necessary for the development of inclusive AI systems.

There should be a parallel effort to develop procedures which facilitate independent audits of *active* AI systems to achieve the objectives outlined in Strategies 3 and 4 of the Plan. For example, researchers have called for the development of a universal digital ad archive[13] which would bring transparency to online ad content, targeting and delivery, thus allowing independent research into the harms caused or amplified by digital ads. Several legislative proposals also focus on researcher access to social media data as an important intervention in AI governance.[14] Beyond facilitating access to social media data, agencies implementing the Plan could also invest in the development of observatories and citizen science programs that contribute to the public's understanding of artificial intelligence, its safety, and its impacts. Given a consensus from technology firms and scientists that the societal harms of AI are hard to predict and prevent,[15] the Plan should also support work to develop new methods for studying the impact of AI systems on society.

3. Ensure industry-independent researchers from civil society, news organizations, and academia are included in the implementation of the National AI Initiative Act.

Independent researchers in civil society groups have an important role to play in AI R&D—particularly to evaluate and address bias, equity, or other concerns related to the development, use, and impact of AI. It has often been through painstaking research done by poorly-resourced civil society representatives (often in the face of an adversarial posture from platforms) that we have been able to learn about significant harms arising from AI.

Similarly, some of the most impactful research and investigations into AI systems have come from journalists working within news organisations. Journalists have produced several of the most cited sources on safety and fairness in AI, spurring the creation of whole subfields in

---

[13] Edelson, L., Chuang, J., Fowler, F. F., Franz, M. M., Ridout, R. (2021) A Standard for Universal Digital Ad Transparency. Knight First Amendment Institute, Columbia University.

[14] Social Media DATA Act. H.R.3451. 117th Cong. (2021)
Algorithmic Justice and Online Platform Transparency Act. S.1896. 117th Cong. (2021)
Algorithm Accountability Act of 2019. H.H.2231. 116th Cong (2019)
Platform Accountability and Transparency Act. S.4066. 116th Cong (2019)
Filter Bubble Transparency Act. S.2024. 117th Cong (2021)

[15] Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., ... & Weber, E. U. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences, 118*(27).
Clegg, Nick. 2021. "You and the Algorithm: It Takes Two to Tango." Facebook. March 31, 2021.

computer science through their data analyses and investigative reporting. Moreover, journalists know how to relate their research findings to broader conversations among the public, which leads to accountability and sparks change.

Innovation happens when people from a variety of disciplines apply different skills and perspectives to solve collective problems. Yet so far, the implementation of the National AI R&D Strategic Plan has not included many of these important actors. This exclusion significantly limits the scope of important research, including research into harms that disproportionately affect marginalised and underrepresented groups, and into how these harms might be addressed.

Researchers of all kinds should be assessed on the basis of their expertise, their ability to implement necessary privacy, data protection and ethics safeguards and protocols, and their independence from commercial interests.

The implementation of the National AI R&D Strategic Plan should prioritize researchers from a broad range of organizations and backgrounds who meet these qualifications for funding, partnerships, and other forms of participation. Researchers from these groups should be consulted in the development of research as outlined above, and affiliation with an academic institution should not be required, per se, for access to these data sets. Instead, the agencies responsible for implementing the Strategic Plan should design inclusive standards for the necessary privacy, security, and data protection protocols that must underpin research projects and build capacity for vetting whether these standards are met.

We thank OSTP for the opportunity to provide this input as they develop the National AI R&D strategic plan.

Signatories (institutions included for identification purposes):

Susan Benesch, Dangerous Speech Project
Brandi Geurkink, Mozilla Foundation
David Karpf, The George Washington University
David Lazer, Northeastern University
Nathalie Maréchal, Ranking Digital Rights
J. Nathan Matias, Citizens and Technology Lab, Cornell University
Rebekah Tromble, The George Washington University

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Cognitive Insights for Artificial Intelligence (CIfAI)

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.

Cognitive Insights for Artificial Intelligence

February 26, 2022

NCO,
2415 Eisenhower Avenue,
Alexandria, VA 22314, USA

Dear NCO,

On behalf of Cognitive Insights for Artificial Intelligence (CIfAI), we write in response to the RFI on the *National Artificial Intelligence Research and Development Strategic Plan*.

We support the efforts of the Office of Science and Technology Policy (OSTP) to update the 2019 National AI Research and Development Strategic Plan, and appreciate the opportunity to comment on the proposal concerning updates to the plan's strategic aims.

We at CIfAI provide strategic research-based solutions from a human-centered perspective to ensure the safe and ethical design, development, deployment, and management of artificial intelligence (AI)-enabled autonomous systems across various industries. As AI-enabled technologies proliferate across a range of use cases both nationally and globally and calls for safety and reliability increase, we believe that the design, development, deployment, and management of these systems must be first and foremost intentional and responsible. Moreover, we believe that a diverse workforce who understands and actively pursues interdisciplinary solutions for the socio-technical complexity of AI is paramount. In this regard, we propose the following 10 recommendations for the robust advancement of AI R&D:

1) In the absence of significant mention of environmental sustainability and given the urgency of our current climate crisis, we recommend a new strategy aptly titled *Strategy 9: Ensure the Environmental Sustainability of AI Systems*.

- **Statement 1A:** Research has already shown that AI-enabled systems require exponentially rising compute power. This increase in compute power requires substantial energy consumption and, as a result, generates a large carbon footprint.[1] To begin to address this concern, researchers have developed a new automated AI system for training and running certain neural networks that cut down on carbon emissions.[2] We underscore the need to implement an additional strategy focusing exclusively on ensuring the environmental sustainability of AI-enabled systems. The addition of environmental sustainability as a core AI R&D strategy supports a more holistic well-being approach to human health, prosperity and overall welfare given that human well-being is dependent on ecological well-being.[3] Furthermore, standards on assessing the impact of AI-enabled technologies on human and ecological well-being have been put forward by IEEE[4] and merit wider adoption. This proposed strategy directly supports the recent mention from the OSTP on February 17, 2022 to include, among six policy divisions, the priority of advancing "critical Administration priorities around groundbreaking clean energy investments."[5] Moreover, a focus of this kind would not only be in concert with UNESCO's Recommendation on the Ethics of AI whereby environmental impacts must be continuously assessed alongside human, social, and economic implications,[6] but also with the United Nations's 2030 Agenda for Sustainable Development.[7] **Recommendation 1A**: This new strategy sets up the critical need to support AI R&D that (i) addresses the design, development, deployment, and management of more efficient computing systems, (ii) informs the standardization and implementation of carbon accounting metrics and their independent oversight thereof, and (iii) supports the readily available public access to carbon accounting practices. An immediate consequence of such initiative will be the demand for inclusion of other disciplines such as environmental science, ecology, and oceanography, among others.

2) To significantly advance trustworthy AI-enabled systems and support the goal to "protect civil liberties, privacy, and American values,"[8] we recommend an even greater consideration of the integration of ethical principles via technical means within *Strategy 3: Understand*

---

[1] AI and Compute. OpenAI, 16 May 2018, https://openai.com/blog/ai-and-compute/.

[2] Cai, H., Gan, C., Wang, T., Zhang, Z., & Han, S. (2019). Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*.

[3] Raworth, K. (2017). Doughnut economics: seven ways to think like a 21st-century economist. Chelsea Green Publishing.

[4] IEEE. (7010-2020). IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being, https://standards.ieee.org/ieee/7010/7718/.

[5] OSTP Blog. A New Chapter for the White House Office of Science and Technology Policy. 17 February 2022, https://www.whitehouse.gov/ostp/news-updates/2022/02/17/a-new-chapter-for-the-white-house-office-of-science-and-technology-policy/.

[6] UNESCO. Recommendation on the ethics of artificial intelligence, https://en.unesco.org/artificial-intelligence/ethics.

[7] United Nations. Department of Economic and Social Affairs - Sustainable Development, https://sdgs.un.org/goals.

[8] Select Committee on Artificial Intelligence of the National Science & Technology Council. (June 2019). The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update, https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

and Address the Ethical, Legal, and Societal Implications of AI. Specifically, we highlight focal areas ripe for R&D investment that are needed to sustain the balance we must bear in mind of building optimized systems for the benefit and empowerment of users and building such systems that do not compromise users' rights.

- **Statement 2A**: As AI-enabled technologies increasingly reflect society's biases and open-source bias audit toolkits grow to address such,[9,10] the need to explicitly state the methods of data acquisition and use to validate a specific values-driven approach and the standardization of such becomes essential.
  Recommendation 2A: An emphasis on the building of trusted datasets that are more representative of the users of the AI-enabled systems, unbiased in their decision-making, and thus trustworthy in their efficiency and effectiveness is needed. One way is to engage more intentionally with multiple stakeholders from across different sectors; independent oversight is necessary to prevent repeated engagement with already known stakeholders. Data crowdsourcing is another way to broaden participation.[11] While a citizen science paradigm offers opportunities for scalability and expansion into different areas for the diversification of datasets, it too raises challenges regarding (i) privacy issues, (ii) computational resources which must factor in Recommendation 1A, and (iii) public awareness and (iv) accessibility that could in fact upend the very goal of diverse community engagement and non-discriminatory practices. This recommendation is also useful towards the building of public datasets —the focus of *Strategy 5: Develop Shared Public Datasets and Environments for AI Training and Testing*— because data crowdsourcing platforms are by definition datasets intentionally created by the public, invite questions of ownership, and face similar security challenges by adversaries.

- **Statement 2B**: The intersectional ethical-legal question of accountability and its subareas of informed consent to use, safety, and liability, among others, in the case, for example, of system failure and/or harm still necessitates significant development regarding clarity of definitions, design frameworks, and public communication strategies, to name a few.
  Recommendation 2B: Interdisciplinary research is inevitable. There needs to be (i) clarification of issues regarding causality, compensation, and justice, as well as transparent determination of the role of responsibility of the human(s) involved. As these issues depend on system transparency and explainability, there needs to be (ii) determination of when explanations are required and what kinds of explanations are acceptable, along with a mechanism of adaptability to changing societal expectations as a result of technological advances.

---

[9] Bellamy R. K., Dey K., Hind M., Hoffman S. C., Houde S., Kannan K., Lohia P., Martino J., Mehta S., Mojsilovic A., & Nagar S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943, https://aif360.mybluemix.net/.

[10] Saleiro P., Kuester B., Hinkson L., London J., Stevens A., Anisfeld A., Rodolfa K. T., & Ghani R. (2018). Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577, http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/.

[11] Mozilla Common Voice, https://commonvoice.mozilla.org/en.

- **Statement 2C**: The question of ethical decision-making in such high stakes contexts such as facial recognition, healthcare diagnostic tools, driverless vehicles, and lethal autonomous weapon systems necessitate substantial attention. Alongside ethical issues such as bias and discrimination as addressed in Recommendation 1A, there is also the issue of moral capability and the level of moral decision making we are willing as a society to assign to an AI-enabled system.

  Recommendation 2C: A moral quandary such as choosing between life or death is a controversial issue, most clearly demonstrated with the Trolley Problem and one that depends on cultural differences.[12] As a result, there is an imperative to (i) determine whether a red line of moral capability should be established, and (ii) after determination of (i), to set system boundaries dependent on regional/cultural norms.

3) In agreement with the emphatic need to build safe and secure AI-enabled systems, and to underscore OSTP's recent announcement on February 17, 2022 to "build a society where everyone can live with equal dignity and hope and opportunity, as well as equal safety and security,"[13] we recommend fundamental system requirements necessary to uphold such a promise. These recommendations expand *Strategy 4: Ensure the Safety and Security of AI Systems* and *Strategy 6: Measure and Evaluate AI Technologies through Standards and Benchmarks*.

- **Statement 3A**: As multiple proposals circulate on the definition of AI, most notably from the European Union's AI Act,[14] the OECD,[15] and a recent call for commentary by the U.S. Chamber of Commerce's AI Commission,[16] there is continued need for agreed-upon definitions of AI and standards for AI-enabled systems' robustness. The establishment of an all-encompassing definition is, in part, due to the dynamic nature of technological advancement and global centuries-old debate on our human-machine relationship.[17,18]

---

[12] Awad, E., et al. (2018). The moral machine experiment. Nature *563*(7729), 59-64.

[13] OSTP Blog. A New Chapter for the White House Office of Science and Technology Policy. 17 February 2022, https://www.whitehouse.gov/ostp/news-updates/2022/02/17/a-new-chapter-for-the-white-house-office-of-science-and-technology-policy/.

[14] Artificial Intelligence Act. (21 April 2021). Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *EUR-Lex - 52021PC0206*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

[15] OECD.AI Policy Observatory, https://oecd.ai/en/ai-principles.

[16] U.S. Chamber of Commerce. U.S. Chamber Launches Bipartisan Commission on Artificial Intelligence to Advance U.S. Leadership. 18 January 2022, https://www.uschamber.com/technology/u-s-chamber-launches-bipartisan-commission-on-artificial-intelligence-to-advance-u-s-leadership

[17] Al-Jazari IR. (1974). The Book of Knowledge of Ingenious Mechanical Devices: Kitāb fī ma 'rifat al-ḥiyal al-handasiyya (translated from Arabic and annotated by Donald R. Hill). Springer, Dordrecht.

[18] Muri, A. (2007). The Enlightenment Cyborg: A history of Communications and control in the human machine, 1660-1830. University of Toronto Press.

Moreover, the development of scientific innovation within modern AI and of governance structures operate on different timelines and with different goals.

Recommendation 3A: Consensus-based provision of precise definitions of technical terms such as AI, autonomy, transparency, explainability, and interpretability and consistent terminology included within the domain of AI safety and security must be achieved. It is critical to note that different cultures and communities have different values towards automation and its role within human identity and culture.[19] As such, it is paramount to include multiple stakeholders when defining AI and related terms. Following such provision is the imperative for system updating and significant R&D in this area. Such updating must include (i) new data resulting from a data-centric strategy for AI model building to support system integrity, and (ii) new core AI functionalities resulting from rapid AI advances to maintain system resilience to change and/or adversarial attack. Both of these requirements are dynamic and necessitate an established review structure to manage in real time.

- **Statement 3B**: As proposals for ethical principles and standards circulate, claims of trustworthiness and their compliance thereof must be verifiable and certifiable. Increasing trust can be achieved through standardization, independent oversight, and overt certification.

  Recommendation 3B: The following steps invite a host of much needed AI R&D foci. A first step is to determine metrics of standardization around transparency, explainability, and interpretability. This is paramount to protecting public and individual privacy requirements as well as overall human well-being. A second step after standardization is determined, is to implement mechanisms of independent audit. The audit process —for which its methodologies also require standardization— and resulting evaluations must confirm the system performs as intended to be certified. A third step is to improve users' system confidence. This can be achieved by ensuring that all AI-enabled technologies contain a user-friendly label of not only privacy considerations regarding an organization's collection, use, and sharing of personal information,[20] but also a label of ethical certification. The issue of comprehensibility, and thus the consensus-based need to establish standards, becomes paramount because users vary in their AI competency and such must be taken into account.

- **Statement 3C**: AI-enabled technology is neither confined to a single use case nor to a single market. Given that everyone will be affected by AI at some point, if not already, international cooperation is fundamental for system interoperability.

  Recommendation 3C: AI standards should be produced and harmonized at the international level to ensure common ground around definitions, security, safety, and system resilience. This

[19] Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 1-12. DOI: 10.1177/2053951717738104.

[20] Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009, July). A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (pp. 1-12).

determination should be made by different groups with a variety of expertise[21] and, in effect, underscores the need to significantly expand the multilateral dialogue mentioned as an activity of *Strategy 6: Measure and Evaluate AI Technologies through Standards and Benchmarks*.

4) To significantly advance efforts to educate and include interdisciplinary and diverse voices across the entire AI space and thus raise the competitive stance of the U.S., we recommend further improvements to education and workforce training activities and proposals. These recommendations expand *Strategy 7: Better Understand the National AI R&D Workforce Needs* and also address the OSTP's recent mention on February 17, 2022 of sustaining "a national strategy for STEM equity."[22]

- **Statement 4A**: Computing is no longer an exclusively technical field. It is multidisciplinary field with social-technical consequences. To readily identify more efficient and effective interdisciplinary solutions, interdisciplinary thinking is a requisite. R&D teams composed of single subject matter experts brought together to collaborate as a multidisciplinary group are insufficient because today's and tomorrow's challenges demand interdisciplinary competences and skills.
  Recommendation 4A: R&D teams, whether in academia, industry, or government, must be composed of individuals who have both deep subject matter expertise *and* interdisciplinary knowledge of the research agenda at hand.

- **Statement 4B**: In order to build a workforce whereby every individual has an interdisciplinary background and to successfully follow through with Recommendation 4A, education must not only expose learners to different disciplines' knowledge, but connect and integrate the disciplines within the learning process. As such, increasing education programs, fellowships, and activities focused exclusively to STEM fields and their subareas is insufficient because their questions, methods, and goals are insufficient for the complex nature of AI.
  Recommendation 4B: Learning across K-12, precollege, undergraduate, graduate, and postgraduate education necessitates reforming and resources to sustain. A first step is the acknowledgment of the limits of silos and the impact that has on the nurturing of interdisciplinary minds.[23] A second step is the creation of learning environments that reach across disciplinary boundaries; such can be achieved by implementing curricula that balance in-depth subject matter with expansive related subject matter coverage, as well as traditional

[21] CEN-CENELEC response to the EC white Paper on AI, Version 2020-06, https://www.cencenelec.eu/media/CEN-CENELEC/Areas%20of%20Work/CEN%20sectors/Digital%20Society/Emerging%20technologies/cen-clc_ai_fg_white-paper-response_final-version_june-2020.pdf.

[22] OSTP Blog. A New Chapter for the White House Office of Science and Technology Policy. 17 February 2022, https://www.whitehouse.gov/ostp/news-updates/2022/02/17/a-new-chapter-for-the-white-house-office-of-science-and-technology-policy/.

[23] National Academies of Sciences, Engineering, and Medicine. The integration of the humanities and arts with sciences, engineering, and medicine in higher education: Branches from the same tree. National Academies Press, 2018.

assessments with creative and experiential practice-based learning experiences.[24] A third step is expanding academic-industry partnerships and closely tied to the focus of *Strategy 8: Expand Public-Private Partnerships to Accelerate Advances in AI*. Moving past differing expectations of achievement and cultural differences will not only accelerate advances in AI through collaboration and co-production,[25] but enhance learning experiences that feed back into both academia and industry in mutually beneficial ways. Of resounding importance and a critical addition to *Strategy 8* as well, these partnerships cannot be solely between STEM fields. Instead, both academia and industry must move away from disciplinary boundaries and also initiate collaborations with non-STEM industries and academic departments, respectively.

- **Statement 4C**: Representation of diverse voices is still significantly lacking within the domain of AI.[26] Diversity, equity, and inclusion (DEI) are fundamental to ensuring that AI-enabled systems are respectful of, safe, and reliable for all users.

  Recommendation 4C: From the designers, developers, researchers, instructors, project managers and directors to the learners themselves, the inclusion of diverse voices must be mandatory across all educational and R&D initiatives. Specifically, inclusion entails the respectful consideration for the different experiences of all within the various areas AI cuts across including education, design, development, deployment, and management. This also has a direct effect on ensuring the goals of Recommendation 2A.

In conclusion, we recommend that the OSTP take these 10 recommendations into consideration for the subsequent update to the *National Artificial Intelligence Research and Development Strategic Plan*.

We thank you for your consideration, and we look forward to your action on our recommendations.

Sincerely,

Monica Lopez, PhD
Chief Executive Officer
Cognitive Insights for Artificial Intelligence

---

[24] López-González, M. (2017). For female leaders of tomorrow: cultivate an interdisciplinary mindset. In *2017 IEEE Women in Engineering (WIE) Forum USA East* (pp. 1-6). IEEE.

[25] Sannö, A., Öberg, A. E., Flores-Garcia, E., & Jackson, M. (2019). Increasing the impact of industry–academia collaboration through co-production. *Technology Innovation Management Review*, *9*(4).

[26] West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. *AI Now*.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Competitive Enterprise Institute (CEI)

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.

Before the

OFFICE OF SCIENCE AND TECHNOLOGY POLICY

Washington, D.C. 20500

COMMENTS OF THE
COMPETITIVE ENTERPRISE INSTITUTE

In the Matter of	)	
	)	
Office of Science and Technology	)	87 FR 5876
Policy	)	Document Number 2022-02161
	)	
RFI Response:	)	
National Artificial Intelligence	)	
Research and Development	)	
Strategic Plan	)	

March 4, 2022

Ryan Nabil

COMPETITIVE ENTERPRISE INSTITUTE

1310 L Street NW, 7th Floor
Washington, DC 20005

(202) 331-1010

Introduction

On behalf of the Competitive Enterprise Institute (CEI), I respectfully submit these comments in response to the Office of Science and Technology Policy's request for comments on updating the National Artificial Intelligence Research and Development Strategic Plan.[1] Founded in 1984, the Competitive Enterprise Institute is a non-profit research and advocacy organization that focuses on regulatory policy from a pro-market perspective. CEI works to promote policies that can help boost technological innovation and American leadership in areas such as artificial intelligence and machine learning and emerging technologies that depend on such innovation.

The Competitive Enterprise Institute appreciates the recognition by both the Trump and Biden Administrations of the need to create a more favorable regulatory environment in which artificial intelligence (AI) and AI-enabled emerging technologies can thrive and promote American economic growth and competitiveness. To that end, CEI recognizes the increasingly important role played by the Office of Science and Technology Policy (OSTP), the National Science and Technology Council (NSTC)'s Select Committee on Artificial Intelligence, the NSTC Machine Learning and AI Subcommittee, and the National AI Initiative Office. As these regulatory bodies seek to update the National AI Research and Development (R&D) Strategic Plan, they have an opportunity to strengthen America's position as a global center of AI innovation. To accomplish that goal, the OSTP and the NSTC need to make several updates to the current AI strategy.

Specifically, the National AI R&D Strategic Plan would benefit from updates in five areas.

- First, while the strategic plan recognizes the essential role of the private sector in promoting AI innovation, it needs to provide more concrete steps to engage the private sector in AI research and development projects.
- Second, to ensure that taxpayer dollars are utilized effectively, the AI strategic plan should propose a framework to track and evaluate the effectiveness of R&D expenditure and grants to various recipients in different AI subdisciplines.
- Third, the strategic plan would benefit from a more nuanced understanding of the AI R&D and regulatory approaches in other countries. For example, reviewing AI policies of other nations could help inform why the U.S. government should allocate a higher share of research spending to multidisciplinary AI projects and prioritize accuracy over algorithmic transparency as a goal for AI systems.
- Fourth, while the AI strategy recommends developing shared AI datasets for academic and private sector use, more details are needed on such proposals.
- Fifth, the National AI R&D Strategic Plan should propose a federal AI sandbox program to incentivize the private sector to play a more important role in AI innovation. By allowing private companies and research institutions to test innovative AI systems for a limited time, such a program can help promote technological innovation, enhance regulatory understanding of AI, and help craft market-friendly regulatory frameworks and technical standards for AI systems.

[1] Office of Science and Technology Policy (OSTP), "Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan," *Federal Register*, Vol. 87, No. 22 (February 2, 2022), https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and.

I. The National AI R&D Strategic Plan Needs to Better Engage the Private Sector

The private sector and academic institutions play a crucial role in the development of AI technologies.[2] Given that reality, a successful AI strategy needs to closely engage technology companies, startups, and research institutions. The 2019 National AI R&D Strategic Plan recognizes the private sector's essential role in promoting artificial intelligence and provides several recommendations to enhance collaboration with the private sector. For instance, it proposes creating joint public-private collaboration, increasing the availability of public datasets, and expanding AI training and fellowship opportunities to meet workforce R&D needs.[3] Despite such proposals, the strategy would benefit from a greater emphasis on engaging the private sector and academic institutions in promoting AI innovation, for example, by creating a federal artificial intelligence regulatory sandbox program, as discussed later in this comment.

II. Developing Mechanisms to Track and Evaluate Artificial Intelligence R&D Spending

The National AI R&D Strategic Plan should propose a framework to better track the allocation and impact of AI-related research and development projects across federal agencies, research institutions, companies, and other recipients of federal AI R&D grants. Despite the growing federal expenditure on AI-related research and development activities, there appears to be a scarcity of efforts in tracking how this money is spent and how it impacts AI innovation.[4]

Greater transparency and more precise information about federal AI expenditure and its impact on innovation within different AI subdisciplines can help policymakers allocate R&D resources more effectively. For example, have resources allocated to specific AI subdisciplines—such as computer vision—led to demonstrably better research outcomes than in other areas? Are certain federal agencies and academic institutions more effective at utilizing research grants than others? Collecting and analyzing data to answer these questions can significantly improve policymakers' ability to make evidence-based R&D spending decisions.

Some long-term AI research projects will require several years before R&D efforts show results—especially in "general AI" and areas of machine learning research that do not appear to have immediate commercial applications.[5] However, tracking spending can nonetheless help compare the effectiveness of similar short- and long-term projects by different agencies, research institutions, and companies. That could not only help U.S. policymakers allocate more resources to more promising AI subdisciplines, but it might also help improve competition between different recipients of federal research grants.

---

[2] White House, Office of the President, Select Committee on Artificial Intelligence of the National Science and Technology Council (NSTC), *National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*, June 2019, p. iii, https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf; Lauren A. Kahn, "US Leadership in Artificial Intelligence Is Still Possible," Council on Foreign Relations, October 28, 2021, https://www.cfr.org/blog/us-leadership-artificial-intelligence-still-possible.

[3] NSTC, *The 2019 National AI R&D Strategic Plan*, p. iii.

[4] Jon Harper, "Federal AI Spending to Top $6 Billion," *National Defense*, February 10, 2021, https://www.nationaldefensemagazine.org/articles/2021/2/10/federal-ai-spending-to-top-$6-billion.

[5] The National AI R&D Strategic Plan defines general AI as the type of AI intended to "create systems that exhibit the flexibility and versatility of human intelligence in a broad range of cognitive domains, including learning, language, perception, reasoning, creativity, and planning" (NSTC, *The 2019 AI R&D Strategic Plan*, pp. 10–11).

III. Lessons from the Successes and Shortcomings of Other Countries' AI Policies

The National AI Strategic R&D Plan would benefit from closer examination of how other countries allocate research spending, their regulatory approach toward AI, and the extent to which these policies have been successful. Due to the potential military applications of many AI-enabled technologies, other countries' AI strategies—particularly those of adversarial nations—are often viewed as a threat to America's national security and technological competitiveness.

However, AI policies and developments in other countries also provide the opportunity to better understand which R&D and regulatory approaches have been successful elsewhere. Policymakers should exercise caution in making such comparisons, as the regulatory experience from other jurisdictions might have limited applicability to the United States. However, awareness of those broader trends can help the U.S. capitalize on the successes of different countries and avoid their regulatory mistakes. To maximize the benefit of this comparative approach, the strategic plan could propose mechanisms to conduct annual reviews of the global AI research and regulatory landscape and an evaluation of its successes and failures.

For example, the National AI R&D Strategic Plan could benefit from a closer examination of European and Chinese approaches toward interdisciplinary AI research. While the European Union's restrictive approach to AI risks harming innovation in certain sectors,[6] several European countries have designed innovative AI R&D strategies at the national level. For example, the French government has proposed creating a network of interdisciplinary institutions to promote high-level AI research in multiple disciplines. Outside of the European Union, the British government has launched initiatives to bolster multidisciplinary artificial intelligence research and foster AI-enabled innovation in insurance and legal services.[7] In China—one of the two leading sources of AI innovation alongside the United States—the State Council has created similar initiatives to encourage cross-disciplinary academic research at the intersection of artificial intelligence, economics, psychology, and other core disciplines.[8]

Studying and evaluating these countries' approaches might provide American policymakers insights into the extent to which existing R&D resources should be devoted to interdisciplinary AI projects. To that end, the U.S. government could also create an AI sandbox and other innovative programs to capitalize on the expertise of academic institutions and the private sector and promote cross-disciplinary AI research in fields such as finance, medicine, and physics.

---

[6] Ryan Nabil, "The EU's Recently Proposed Artificial Intelligence Act Goes Too Far," *The National Interest*, August 16, 2021, https://nationalinterest.org/blog/buzz/eu's-recently-proposed-artificial-intelligence-act%C2%A0goes-too-far-191733; Benjamin Muller, *How Much Will the Artificial Intelligence Act Cost Europe?* (Brussels: Center for Data Innovation, July 2021), https://www2.datainnovation.org/2021-aia-costs.pdf.

[7] Castro and New, "2016 National AI R&D Strategic Plan," p. 3.

[8] State Council of China, "国务院关于印发新一代人工智能发展规划的通知" ["Notice of the State Council on the Release of the New Generation Artificial Intelligence Development Plan"], July 20, 2017, http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm; Castro and New, "2016 National AI R&D Strategic Plan," p. 3. For an English translation of the Chinese government's 2017 AI Development Plan, see State Council of China, "China's 'New Generation Artificial Intelligence Development Plan'" [translated by Graham Webster, Roger Creemers, Paul Triolo, and Elsa Kania], New America, August 1, 2017, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.

Along with successful innovative strategies of other countries, it is equally important to examine potential regulatory mistakes. For instance, despite objections from government bodies and non-profit research organizations,[9] the UK and the EU have sought to promote algorithmic transparency as a major objective for AI technologies.[10] Likewise, many American policymakers and experts increasingly advocate transparency as a goal in developing AI systems.[11] In attempting to understand AI systems better, the National AI R&D Strategic Plan has proposed "improving fairness, transparency, and accountability-by-design" as objectives for government-funded AI research projects.[12]

While mechanisms to improve fairness and reduce biases are worthwhile goals, mandating high levels of transparency could detract from designing effective and accurate AI programs.[13] That is all the more likely if the same level of transparency requirements were to be applied to AI systems across all industries. For example, high levels of algorithmic transparency are crucial to prevent government abuses of power and civil liberties violations if AI systems are used in fields such as criminal justice and law enforcement. However, transparency concerns are much less important in areas like cybersecurity and medical diagnosis—where the accuracy of preventing cyberattacks and detecting diseases is much more important than the explainability or transparency of underlying algorithms. Therefore, the National AI R&D Strategic Plan should not prioritize transparency over accuracy as a research objective—except in limited and specified cases such as AI systems for criminal justice, law enforcement, and human resources.[14]

Since AI systems typically involve complex neural networks and algorithms that produce effective outcomes, their underlying processes are not transparent even to the programmers that design such systems.[15] For instance, medical researchers at New York's Mount Sinai Hospital developed Deep Patient, an AI program that can predict whether a patient has contracted a specific disease. The AI system, which is reported to be substantially better than any comparable systems, trained on the medical data of 700,000 patients across several hundred variables. Deep Patient can accurately provide medical diagnostics, but its designers cannot accurately describe

---

[9] UK Government Office for Science, *Artificial Intelligence: Opportunities and Implications for the Future of Decision Making* (London: Government Office for Science, February 12, 2016), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf; Nick Wallace and Daniel Castro, "The Impact of the EU's New Data Protection Regulation on AI," Center for Data Innovation, March 27, 2018, p. 4, https://www2.datainnovation.org/2018-impact-gdpr-ai.pdf.

[10] UK Government, Central Digital and Data Office, "Algorithmic Transparency Standard," November 29, 2021, ww.gov.uk/government/collections/algorithmic-transparency-standard; Tambiama Madiega, "EU Guidelines on Ethics in Artificial Intelligence: Context and Implementation," European Parliament Briefing PE 650.163 (September 2019), p. 4, https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf.

[11] For example, see Electronic Privacy Information Center, "State Artificial Intelligence Policy," n.d., accessed March 3, 2022, https://epic.org/state-artificial-intelligence-policy/; Artificial Intelligence Capabilities and Transparency Act of 2021, S. 1705, 117th Cong. (2017), https://www.congress.gov/bill/117th-congress/senate-bill/1705/text?r=82&s=1.

[12] NSTC, *The 2019 National AI R&D Strategic Plan*, p. 21.

[13] Daniel Castro and Joshua New, "Regulatory Comments in Response to the Request for Information on Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan," Center for Data Innovation, October 26, 2018, pp. 4–5, https://www2.datainnovation.org/2018-nitrd-ai-r%26d.pdf.

[14] Ibid.

[15] Ibid., p. 4.

how the program arrives at such a diagnosis.[16] Mandating high levels of algorithmic transparency from such programs can detract from cutting-edge AI innovation in fields ranging from healthcare to nuclear science.

Nevertheless, to bolster leadership in shaping global AI norms, Britain and the EU are increasingly advocating algorithmic transparency as a main AI policy goal.[17] Although fixating on transparency can harm the development of innovative yet unexplainable AI systems, potential benefits remain limited. As the UK Government Office for Science explained in a 2016 report:

> Most fundamentally, transparency may not provide the proof sought: Simply sharing static code provides no assurance it was actually used in a particular decision, or that it behaves in the wild in the way its programmers expect on a given dataset.[18]

Studying these international regulatory developments and criticisms can help American policymakers avoid potential mistakes like mandating transparency as an overarching goal for all types of AI systems. By creating processes to review international regulatory developments, the National AI R&D Strategic Plan can help U.S. policymakers design an innovation-friendly approach toward artificial intelligence.

IV. Encouraging Academic and Private Sector Innovation by Providing Access to Non-Sensitive Government Database

The lack of access to data remains a significant challenge for the development of novel AI technologies, especially for startups and businesses without the resources of Big Tech companies. Creating innovative AI systems requires high-quality datasets on which AI systems can be trained. The costs associated with creating, cleaning, and preparing such datasets for training AI systems remain too high for many businesses and academic institutions.[19] For example, Google London-based subsidiary DeepMind's AlphaGo software made headlines in March 2016 when it defeated the human champion of the Chinese game Go.[20] The cost to train datasets for building this program was more than $25 million in hardware alone.[21]

---

[16] Castro and New, p.4; Will Knight, "The Dark Secret at the Heart of AI," *MIT Technology Review*, April 11, 2017, https://www.technologyreview.com/s/604087/the-darksecret-at-the-heart-of-ai/.

[17] UK Government, Central Digital and Data Office, "UK Government Publishes Pioneering Standard for Algorithmic Transparency," news release, November 29, 2021, https://www.gov.uk/government/news/uk-government-publishes-pioneering-standard-for-algorithmic-transparency; Madiega, "EU Guidelines on AI Ethics."

[18] UK Government, Office for Science, *Artificial Intelligence: Opportunities and Implications for the Future of Decision Making* (London: UK Government Office for Science, November 9, 2016), https://www.gov.uk/government/publications/artificial-intelligence-an-overview-for-policy-makers; Castro and New, pp. 4–5.

[19] NSTC, *The 2019 National AI R&D Strategic Plan*, p. 28; Daniel E. Ho, Jennifer King, Russell C. Wald, and Christopher Wan, *Building a National AI Research Resource; A Blueprint for the National Research Cloud* (Stanford: Stanford University Human-Center Artificial Intelligence and Stanford Law School, October 2021), pp. 35–41, https://hai.stanford.edu/sites/default/files/2022-01/HAI_NRCR_v17.pdf.

[20] DeepMind. "The Story of AlphaGo So Far," n.d., accessed February 28, 2022, https://deepmind.com/research/alphago/; Jeffrey Ding, "Deciphering China's AI Dream," Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, March 2018, p. 7. https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf.

[21] Elizabeth Gibney, "Self-taught AI is best yet at strategy game Go," *Nature*, October 18, 2017, https://www.nature.com/articles/nature.2017.22858.

Recognizing this challenge, the National AI R&D Strategic Plan recommended the development of shared datasets that startups, businesses, and research institutions can use to create and train AI programs.[22] But despite this commitment, progress in this area appears to be slow. That is why the AI strategic plan needs to outline more concrete steps to publish high-quality datasets using the vast amount of non-sensitive and non-personally identifiable data already at the federal government's disposal. Under the 1974 Privacy Act, U.S. government agencies have not created a central repository of data, which is important because of the privacy and cybersecurity risks that a central data repository of sensitive information would face.[23]

However, different U.S. agencies also have access to a wide range of non-personally identifiable and non-sensitive datasets intended for public use—such as the National Oceanic and Atmospheric Administration's climate data and the National Aeronautics and Space Administration (NASA)'s non-confidential space-related data.[24] Making such data readily available to the public can allow AI innovation in weather forecasting, transportation, astronomy, and other underexplored subjects.[25] Therefore, the AI strategy should propose a framework that enables the OSTP and the NSTC to work with government agencies and ensure that non-sensitive *and* non-personally identifiable data—intended for public use—are made available in a format suitable for AI research by the private sector and research institutions.

The OSTP, the NSTC, and the National AI Initiative Office could use the federal government's FedRAMP data classification as a general framework to develop a strategy for which data should be included in public datasets. The FedRAMP framework divides government-stored data into three distinct types:

1) Low-impact risk data meant for public use;
2) Moderate-impact risk data, which are controlled, unclassified data (e.g., personally identifiable information) unavailable to the public; and
3) High-impact risk data, which contain "sensitive federal information," such as law enforcement and emergency services information.[26]

To minimize privacy and cybersecurity risks, the OSTP and the NSTC should propose that the datasets only contain low-impact risk data intended for public use. The OSTP, the NSTC, and other relevant regulatory authorities should ensure appropriate data controls and privacy standards so that these datasets do not erroneously include sensitive information, uphold cybersecurity best practices, and are provided in a format suitable for training AI systems.

V. Creating a Federal Artificial Intelligence Regulatory Sandbox Program

To develop the regulatory understanding of emerging AI technologies and better engage the private sector, the National AI R&D Strategic Plan should propose the creation of a federal AI sandbox program. Given the rapid evolution of AI-enabled technologies, there is a growing need

[22] NSTC, *The 2019 National AI R&D Strategic Plan*, p. 28.
[23] Privacy Act of 1974, 5 U.S.C. § 552a (2012); Ho et al., *National AI Research Resource*, p. 35.
[24] Ho et al., *National AI Research Resource*, p. 47.
[25] Ibid.
[26] Ibid. p. 39.

to better understand their ethical, legal, and societal implications.[27] For example, as the strategic plan notes, "the unusual complexity and evolving nature" of AI systems mean that "ensuring the safety of AI systems" remains a challenge.[28] Therefore, understanding such systems would constitute an important first step toward crafting safe, market-friendly regulatory frameworks and technical standards for AI systems and AI-enabled technologies.[29]

At the same time, there is a growing need to engage the private sector in promoting AI innovation. These objectives could benefit from a policy tool commonly used to advance financial technology (FinTech) called "regulatory sandboxes."

"Regulatory sandbox" programs provide companies with an experimenting space that allows them to offer innovative products and services under a frequently lightened regulatory framework for a limited period.[30] The United Kingdom's Financial Conduct Authority created the world's first FinTech sandbox program in 2015. Since then, regulators in Australia, Hong Kong, Singapore, and other innovative jurisdictions have launched similar programs.[31] The federal Consumer Financial Protection Bureau and more than 10 U.S. states have launched sandbox programs to promote technological innovation in finance and insurance.[32] In addition to financial services, the Utah Supreme Court has also created a regulatory sandbox program that allows non-legal firms to provide certain innovative legal services.[33]

Although FinTech sandbox programs are becoming increasingly common, AI sandbox programs remain a largely underexplored idea.[34] In its 2021 Artificial Intelligence Act, the European Commission proposed an AI sandbox program.[35] However, its success will depend on its regulatory design and the extent to which the sandbox prioritizes technological innovation and regulatory learning. If implemented correctly, an AI regulatory sandbox program can be significantly helpful in promoting U.S. AI innovation. To that end, the National AI R&D Strategic Plan could recommend a sandbox program that will accept participants based on

---

[27] NSTC, *The 2019 National AI R&D Strategic Plan*, pp. 19–22.
[28] Ibid., p. 5.
[29] Ibid., pp. 14–18, 23–26.
[30] Ryan Nabil, "How Regulatory 'Sandboxes' Can Boost U.S. Technological Innovation," *Real Clear Markets*, August 12, 2021, https://www.realclearmarkets.com/articles/2021/08/12/how_regulatory_sandboxes_can_boost_us_technological_innovation_789620.html
[31] Financial Conduct Authority, "Financial Conduct Authority's regulatory sandbox opens to applications," May 9, 2016, https://www.fca.org.uk/news/press-releases/financial-conduct-authority's-regulatory-sandbox-opens-applications.
[32] World Bank, *Global Experiences from Regulatory Sandboxes* (Washington, DC: World Bank Fintech Note No. 8, 2020), https://openknowledge.worldbank.org/handle/10986/34789; Nabil, "How Regulatory 'Sandboxes' Can Boost U.S. Technological Innovation."
[33] Utah Supreme Court, Office of Legal Services Innovation, "January 2022 Activity Report," n.d., accessed February 27, 2022, https://utahinnovationoffice.org; Ryan Nabil, "Regulatory sandbox programs can promote legal innovation and improve access to justice," *The Hill*, October 9, 2021, https://thehill.com/opinion/judiciary/576041-regulatory-sandbox-programs-can-promote-legal-innovation-and-improve-access.
[34] Aljoscha Burchardt, "Steckt die KI in den Sandkasten!" ["Put AI in a Sandbox!"], *Die Zeit*, December 10, 2020, https://www.zeit.de/2020/52/kuenstliche-intelligenz-suchmaschinen-navigationen-training-forschung.
[35] European Commission, "Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," Proposal for a Regulation of the European Parliament and of the Council, COM (2021) 206 Final, April 21, 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

proposed AI innovation and its potential benefit to developing novel AI technologies.[36] Such a program would not only benefit consumers and the participating companies; it would also enable regulators to gain firsthand insights into AI technologies and help craft market-friendly regulatory frameworks and technical standards.

Regulators could also create sandbox programs to target innovation in specific areas—such as human-machine interaction and probabilistic reasoning—which the strategic plan identifies as areas in need of further research.[37] For example, current AI systems are ill-equipped to translate heavily accented speech or speech in a noisy environment.[38] A thematic sandbox program targeting natural language processing could incentivize companies and researchers to offer innovative AI-enabled products in this area. Likewise, a thematic sandbox aimed at developing AI-enabled cybersecurity and encryption tools can help encourage market innovations to counter growing cybersecurity challenges.[39] Sandbox participants testing innovative products in these areas can go a long way towards helping lawmakers and regulators better understand these emerging areas of artificial intelligence and develop innovation-friendly regulatory frameworks accordingly.

The AI regulatory sandbox concept remains novel. Therefore, designing an effective program will require creative thinking from the Office of Science and Technology Policy, the National Science and Technology Council, and the National AI Initiative Office. For example, regulators will need to define the type of AI systems and AI-enabled products and services eligible for participating in the sandbox. Given the current technological limitations, an AI sandbox might need to be restricted to 1) "limited AI" systems that perform tasks in specific and well-defined domains like speech recognition, translation, and medical diagnosis and 2) projects where measurable technological advances are possible within the typical sandbox testing period of one to two years.[40]

Furthermore, the systems, products, and services eligible for the sandbox could fall under the jurisdictions of multiple regulators. Such a development might ultimately require a legal framework that defines the supervisory role of different regulators in operating the AI sandbox program in cases of overlapping jurisdiction.

Designing an effective AI sandbox will also require modifications to the existing FinTech sandbox models. For example, unlike in financial services, academic institutions remain a leading source of AI innovation.[41] Typically, FinTech sandbox programs provide participants

[36] To better understand the regulatory designs of major fintech sandbox programs in the United States and selected foreign jurisdictions, and how they might inform the regulatory design of an AI sandbox program, readers are advised to consult the author's upcoming CEI report on regulatory sandbox programs.
[37] NSTC, *The 2019 National AI R&D Strategic Plan,* p. 12.
[38] Ibid.
[39] Arthur Herman, "The Executive's Guide to Quantum Computing and Quantum-secure Cybersecurity," Hudson Institute, April 3, 2019, https://www.hudson.org/research/14930-the-executive-s-guide-to-quantum-computing-and-quantum-secure-cybersecurity.
[40] In contrast to limited AI, the goal of "general AI" is to "create systems that exhibit the flexibility and versatility of human intelligence in a broad range of cognitive domains, including learning, language, perception, reasoning, creativity, and planning" (NSTC, *The 2019 National AI R&D Strategic Plan*, pp. 10–11).
[41] NSTC, *The 2019 National AI R&D Strategic Plan,* p. iii.

with exemptive regulatory relief and waived or expedited registration processes.[42] However, many research institutions might not seek to commercialize the algorithms and technologies they test in a sandbox. Therefore, exemptive regulatory relief might be of limited benefit to such institutions. Furthermore, for many research institutions and companies, the lack of access to high-quality datasets and cloud-based computing power might pose a greater obstacle than regulatory barriers.[43] Thus, a way to incentivize participation in an AI sandbox program could be to grant limited cloud-based computing power for the duration of the sandbox test.[44]

Conclusion

The Office of Science and Technology Policy, the National Science and Technology Council, and the National AI Initiative Office need to adopt a realistic approach, objectives, and scope for a national AI R&D plan. AI's general-purpose nature—combined with its diffusion across many sectors and the rapidly changing technological developments—limits the extent to which a national strategy can significantly improve AI innovation across the economy. Recognizing this challenge, the Biden administration and administrative agencies should focus on enabling a wide range of actors, from tech startups to academic and financial institutions, to play a role in promoting American AI innovation.

Given the rapid change and growth of AI-enabled technologies, any national AI R&D strategy will need to be frequently revisited—in light of regulatory learning and the changing AI landscape in the United States and other major jurisdictions. An adaptable and light-touch regulatory approach is needed to secure America's global economic competitiveness and technological innovation in AI and emerging technologies that depend on artificial intelligence.

Ryan Nabil
Research Fellow
Competitive Enterprise Institute

[42] For a longer discussion, see Hilary Allen, "Regulatory Sandboxes." *The George Washington Law Review* 87, no. 3 (May 2019): 579–645. http://dx.doi.org/10.2139/ssrn.3056993.
[43] Ho et al, *Building a National AI Research Resource*, pp. 28–33.
[44] Ibid.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Computing Community Consortium (CCC)

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.

Computing Community Consortium's Response to RFI "Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan"

Written by: *David Danks (University of California, San Diego), Maria Gini (University of Minnesota), Odest Chadwicke Jenkins (University of Michigan), Sven Koenig (University of Southern California), Daniel Lopresti (Lehigh University), Melanie Mitchell (Santa Fe Institute), Katie Siek (Indiana University, Bloomington), Ufuk Topcu (University of Texas at Austin), Holly Yanco (University of Massachusetts Lowell) and Maddy Hunter (Computing Community Consortium).*

The National Artificial Intelligence Research and Development Strategic Plan articulates how the United States can accelerate advances in Artificial Intelligence (AI) through strategic investments in research, infrastructure, assessments, education, and partnerships. We commend the National Science and Technology Council on updating The National Artificial Intelligence Research and Development Strategic Plan since our advances in AI technologies and our understanding of the unintended consequences of AI have significantly changed over the last six years. We agree that AI provides "tremendous opportunities" to improve many facets of life and that this is largely due to the federal government's investment in research. However, we emphasize that federal funding is crucial in *all* scientific research - from fundamental computing to social science - to ensure we are creating innovative, reliable, trustworthy, domain relevant, and socially responsible AI innovations.

Of interest specific to this RFI (and an update to our RFI response in 2018[1]), the CCC recently published a 20-Year Community Roadmap for Artificial Intelligence Research in the United States[2] that details the need to invest in fundamental AI research, educate a diverse, inclusive pipeline to make these needed innovations, and develop the infrastructure needed to support transformational AI. We point to the AI Roadmap as a prime example of the outcome from a community-based process, which we believe would work well for answering the open questions we raise below, as well as for its approach of identifying ambitious "stretch" goals requiring a long-term view and

[1] https://cra.org/ccc/wp-content/uploads/sites/2/2018/10/CCC-Response-to-AI-RFI.pdf
[2] https://cra.org/ccc/visioning/visioning-activities/2018-activities/artificial-intelligence-roadmap/

associated milestones that we think is vital for the strategic plan. More specifically, we offer that the plan should emphasize three important themes that we explore in more detail below:

- **Diversity** provides a foundation for the success of the U.S., and those many dimensions of diversity must be reflected in our AI efforts. We do not mean only diversity of people, but also diversity of ideas, theories, methods, projects, evaluation systems, and much more.
- **Education** is critical to addressing both pipeline challenges for the next generation of researchers and developers, and also the need for AI literacy across the population.
- **Foundational Research** enabled the AI revolution of recent years, but such research requires a longer-term view. Progress in AI requires both use-inspired research for the needs of now, and transformational basic AI research on the ideas of the future. Moreover, this foundational research must span across many disciplines so that we can understand, shape, and create AI for the betterment of all.

We now explore how these three themes can be incorporated into the eight strategies laid out in the National Artificial Intelligence Research and Development Strategic Plan.

Strategy 1: Make long-term investments in AI Research

The 2019 Plan talks about "America's scientific leadership" and "addressing societal issues" while mitigating negative effects caused by "bias, equity, or other concerns" without mentioning any investment in the **education** and training of the pipeline. The current **education** system is not equitable, and this puts future **diverse** scholars, who would contribute to this vision, at risk of not being able to participate. Successful research and development (R&D) starts with the pipeline. PhD applications are low across the country and a more robust pipeline is needed to create better tech/AI in the future, to train the next generation of AI researchers, and to **educate** the general population on how AI technologies will impact their lives and about the positive and negative aspects of AI. In addition to specifying what the goals are, it is important to provide some guidelines on how to achieve them. The processes we have now are not sufficient to bring our current goals to fruition, we must focus on creating new strategies and processes that will.

Strategy 1 is goal-centric and focused on specific capabilities, but research is always uncertain and surprising. The focus should be on creating processes that will sow many seeds and cultivate scientific exploration in addition to scientific rigor. An investment in

AI has to produce both ideas and people that move AI forward in positive directions. Even if a large investment is given, the current peer and merit review practices will not cultivate transformational ideas in AI or produce the necessary AI workforce. The review processes we have in place reward siloing into intellectual comfort zones among like-minded groups that gravitate to similar ways of thinking. Given how neural networks catalyzed over the last 20 years (mostly through international non-US support), we worry that our nation will miss the boat on the next transformations in AI if we continue to choose investments in this manner. Such practices tend to reward short-term results rather than **longer-term exploration**. This partially explains the rise and dominance of industry-led research groups in publication venues. Furthermore, the people being trained in AI tend to overfit to the current "hot" topics, and this does not lead to research labs that produce a workforce with **diverse** sets of ideas or skill sets. There is a risk that focusing solely on the AI workforce will miss the mark for creating systems that (1) address national needs and (2) utilize justice-oriented frameworks to solve societal issues.

Many topics are discussed in the report under Strategy 1. Additional thoughts are:

- *Data focused methodologies for knowledge discovery,* overlaps with Section 5,
- *Fostering research on human-like AI*, is connected to Strategy 2.
- *Pursuing research on general-purpose artificial intelligence* addresses the issue of needing more than machine learning methods, but is vague and talks about general intelligence in a way that does not provide specific ideas on what should be done. The concept of general intelligence is not sufficiently specific and there are no guidelines on how to move in that direction.
- *Advancing hardware for improved AI* and *Creating AI for improved hardware* and *understanding theoretical capabilities and limitations of AI.* There is emphasis on the hardware needed to assess performance. While the importance of hardware should not be underestimated, the sections minimize the importance of theoretical results and focus mostly on performance.
- *Developing scalable AI systems*. Scalability is critical, but communications, networks, and interactions with other systems are mentioned briefly, together with mentioning that distributed systems are more robust to failures. Issues related to multi-agent systems, interactions of agents, game theoretic approaches, norms for the behaviors of the agents, etc. are not mentioned.
- *Perception* and *robotics* emphasize collaboration, which is obviously important, but neglects issues such as adversarial systems, or systems that operate in disguise to collect data or to interfere with legitimate operations. Those issues should be addressed in Section 4.

Strategy 2: Develop effective methods for human-AI collaboration

No matter how many resources are applied to the development of fundamental and applied AI research, poor human-AI collaboration will prevent the system from achieving its full potential, as people will mistrust, misuse, and discontinue using systems they do not understand. Human-AI collaboration requires **foundational research** on the development of explainable AI systems that still achieve the required performance. Systems need the ability to answer, at varying levels of complexity, human questions about how decisions are made.

Human-AI collaboration goes far beyond the future of work, a focus of the 2019 Plan. People are increasingly interacting with AI in their homes (e.g., Amazon Echo, Google Home), in their cars (e.g., computer vision systems for safety), on their phones (e.g., mapping, shopping), for healthcare (e.g., automated diagnosis systems), in online shopping and entertainment (e.g., recommendation systems), and when shopping in the real world (e.g., Amazon Go). This revision of the plan needs to recognize the **diversity** of ways that people are already interacting with AI, and develop frameworks that can be applied to enable common interaction methods.

The revised plan must also recognize the **diversity** of interaction roles; much of the 2019 Plan frames the human as an operator of the AI. We aim for fundamental discoveries that enable the same richness of interaction between a person and an AI that currently exists between two people, with a shifting of roles depending on the current situation and the abilities of each. Ultimately, this would allow us to achieve a level of human-AI collaboration where the AI is a teammate of the person.

Not only do we need to create AI systems that are usable by all people, but we must also **educate** the public, starting at the K-12 level, about AI, probability and statistics, and computational thinking. Even with the best designed human-AI collaboration, without AI literacy, people will still effectively be bystanders to the technology.

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

AI has the power to provide enormous benefits, but also to cause significant harm. The inclusion of ethical, societal, and legal issues in the 2019 Plan is crucial in ensuring that our AI systems provide substantial benefits while conforming to relevant legal requirements. However, the current structure of the 2019 Plan separates these issues into a distinct strategy, rather than recognizing the ways in which ethical, societal, and legal issues arise throughout the lifecycle of AI creation, including the design, research,

development, and deployment stages. As a result, this particular strategy has a narrower focus that omits central ethical and societal issues. For example, the current strategy element on “building ethical AI” focuses on explication of principles to characterize ethical behaviors by AI systems. However, ethical AI performance depends on the values that the AI is intended to support, the contexts in which it is deployed, and many other aspects that go far beyond principles. Similarly, “designing architectures for ethical AI” focuses on systems in which ethics is explicitly represented, but the ethics in an AI usually arise from design and development decisions (e.g., about what errors are deemed “acceptable”) rather than explicit representations. We encourage the next revision to adopt a “whole lifecycle” approach to ethical, societal, and legal issues, with an emphasis on the ways that myriad decisions that might appear “merely technical” can actually have significant ethical, societal, and legal import.

This approach would be greatly enhanced by **foundational research** on best practices for designing, developing, and deploying AI systems that implement and support people’s values. Ethics and societal impact cannot be reduced to simpler measures of, say, reliability or security. At the same time, we do not currently have clear best practices or other ways to guide AI researchers and developers towards more ethical and socially beneficial systems. This much-needed research should also provide mechanisms, processes, and practices to ensure that **diverse communities** are supported by AI, rather than providing benefits for only a few.

We also note that the current strategy contains surprisingly little focus on legal and regulatory issues, despite the current title. There is increasing recognition and understanding that current regulatory and policy frameworks are insufficient for present-day and near-future AI systems. The revision should thus call for additional research, including the creation of regulatory “sandboxes” to test different ideas, on novel legal and regulatory approaches to ensuring AI safety and benefit.

Finally, we suggest that this strategy should call for the development and assessment of novel **educational programs** to ensure that people have appropriate knowledge and skills to assess and respond to ethical, societal, and legal concerns or opportunities with AI. If we want to have more ethical AI systems that better support people’s values, then relevant people need appropriate training. Most obviously, these issues should be incorporated into university curricula and upskilling/reskilling programs around AI. However, the needed **education** extends beyond the developers–for example, people engaged in software acquisition need to understand the technical, ethical, and legal possibilities and risks with AI systems. AI creators must learn how to do better, and the rest of us must learn what to demand so that they do better.

Strategy 4: Ensure the safety and security of AI systems

As an overarching theme and essential piece in this response, re-strategizing **education** and curriculum development is a vital piece of ensuring the safety and security of AI systems. Many issues pertaining to safety and security will not be resolved through piecemeal advances, but instead require a fundamental change in mentality in the way that we design AI systems and view consequential responsibility of these systems if things go wrong.

It is challenging if not impossible to ensure the safety and security of a system that was built without factoring in the safety- and security-related constraints into the development of the system. Therefore, ensuring safety and security cannot be an afterthought and can be achieved only through principled development processes that diffuse the need for safety and security in each step of development from specifications to certification.

The portion of the 2019 Plan that states, “the notion of safety (or security) by design might impart an incorrect notion that these are only concerns of system designers,” may come across as support to reduce the burden on designers rather than imposing responsibilities on all parties. The document should not have an indication that designers might be responsible for less. What they do is not sufficient for safety/security (or any other feature that contributes to the trustworthiness of AI) but good design is absolutely necessary for safety/security. The responsibility of ensuring the safety and security of AI systems needs to start during the design process, beginning with specifying conceptually and technically what we expect from them. These decisions made in the early stages restrict the decisions later in the lifetime and may have more severe and at times unpredictable and unintended consequences. It is definitely not easy to specify acceptable or desirable properties for systems with AI, but we need to start somewhere and adopt an iterative approach where the need for certification is a central driver of the design process.

In addition, the prominence of adversarial machine learning in Strategy 4 will downplay the importance and difficulty of securing systems with AI. It is rare that an AI functionality is used in isolation. We cannot overemphasize the notion of **diversification** of data sources and tools for building intelligence and methods for it. It is likely that integration of functionality is a security vulnerability, i.e., we may secure each in isolation but, when integrated, new vulnerabilities emerge. Most importantly, we need to leverage what we already know and incorporate that into techniques for AI. Current ML techniques reinvent the wheel whenever they face a new task. Therefore, it is

necessary to develop theory and mechanisms to address the vulnerabilities that arise when AI functionality is integrated into bigger systems.

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

We strongly agree that socio-technical infrastructure is needed "to support reproducible research in the digital area." This is an excellent start, but we encourage the administration to think more broadly to capture better sociological and contextual information to help us understand what is going on and how this goal will be implemented. The strategy acknowledges that more support is needed for semi-structured and unstructured data; we emphasize that researchers in social sciences should be consulted to get a better understanding of the rich, contextual data needed for future knowledge generation. In addition, multiple unstructured data streams need to be easily integrated for improved sense-making. These are open research challenges that need to be carefully addressed to answer more challenging, worthwhile societal issues (e.g., climate change, proactive healthcare interventions).

We highlight NAIRR considerations[3] for "building inclusive datasets and governance approaches that treat equity as a core design principle," but also call for reviews of current attempts at equity in data collection (e.g., AllOfUs[4]) to identify best practices and issues, and to ensure historically exploited groups receive benefits from participating.

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

Benchmark datasets are very important and the strategy rightly emphasizes this. However, it's important to note that there are a lot of problems with the kinds of evaluations done currently with widely used benchmark datasets in AI / ML.

Here are some of the issues:

- **Assumption that data is IID (Independent and Identically Distributed):** Today, most ML systems are evaluated by splitting data randomly into training and test sets, and evaluating accuracy on test sets - assuming the data is IID. However, real world data is often (almost always) not IID. More research needs

[3] https://www.whitehouse.gov/ostp/news-updates/2021/12/15/readout-of-the-fourth-national-artificial-intelligence-research-resource-nairr-task-force-meeting/

[4] https://allofus.nih.gov/

to go into how to construct benchmarks that test for more than just accuracy under IID assumptions.

- **Shortcut learning:** Many current standard benchmarks in ML have been shown to allow for "shortcut learning"—that is, a system can perform very well on the benchmarks due to subtle statistical correlations that are unrelated to the actual tasks we want the system to learn. These shortcuts are often hard to uncover.
- **Overfocus on standard benchmarks in evaluating research:** The reliance on standard benchmarks in today's ML research sometimes translates into an attitude that if a system is not tested on the standard benchmark, or if it scores below "state of the art" on that benchmark, the research is not worth publishing. This can have the effect of stifling important new ideas and encouraging research that provides only very incremental progress.

Additional thoughts:

- The strategic plan says that we need standards and benchmarks to ensure "accuracy, reliability, robustness, accessibility, and scalability." This list should also include generalization, robustness to shortcut learning, and robustness to adversarial attacks.
- We should not simply increase the availability of AI testbeds, as recommended in the report; we need to rethink the entire enterprise of how to create test beds that actually test for what we want systems to do. This is currently a big unsolved problem in AI/ML.
- The report says "Government leadership and coordination is needed to drive standardization and encourage its widespread use in government, academia, and industry." But how will the government actually do this in a way that creates standards and testbeds that really test what we want?

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

It is necessary to resolve the current curricular bottlenecks that limit growth of the AI workforce. Modernization of our approach to AI curriculum is desperately needed to grow to meet needs for national technological competitiveness. Currently, the expansive growth of AI has led to a two-sided demand crunch for developing an AI workforce. This AI demand crunch has pushed our current curricular structures beyond their limits. In particular, AI has grown from being a specialty research area into a basic literacy needed by people across society. The needs for education spans across the needs of people who will be leaders in the innovation of AI, practitioners that will design and develop AI products and services, and users and consumers who need to make

effective choices. There is a swelling demand for enrollments in courses[5] and to understand and to serve such a broad need in AI **education** for all people we must do the following:

- Cultivate a new populous of educators in AI
- Data collection more than anything else - particularly are there specific places we know where the pressure is bottlenecking curriculum (e.g., not enough people to provide stewardship to get people through the system)?
- There needs to be an interdisciplinary approach to address the workforce needs and create equitable systems. The pipeline has to be thought of more broadly and there needs to be training of those who help the pipeline).[6]
- Relative to the other strategies, Strategy 7's language is passive and non-committal - merely understanding, rather than trying to do something about it. It would be better to say "Realize a competitive AI R&D workforce" and involve the creation of programs, support for research about how to build skills etc.

Strategy 8: Expand public private partnerships to accelerate advances in AI

Public-private partnerships involve some significant tradeoffs. We recognize some notable positives (e.g., access to computing resources and data that only industry may possess, first-hand awareness of problems that have practical applications benefiting society, financial resources to make large investments in applied research), but at the same time there are risks, including the potential for conflicts of interest, whether real or perceived, that may impact the independence of university-based researchers. This kind of independence – to question and to criticize – has never been more important given the growing pervasiveness of commercial AI applications and certain well-publicized failures of recent years, discovered only after the fact. This relates to the **diversity** theme mentioned earlier.

In addition, while several major success stories are called out in the 2019 Update to National AI R&D Plan (e.g., CNNs), it should be noted that funding for basic AI research in the US has been uneven over the decades, and important contributions made as a result of non-US funding and by non-US researchers needs to be acknowledged. Sustained funding over long durations (much longer than typical grant award periods) is vital; the path from the germ of an idea to commercial application is a long one, and many US-based researchers have had to be very creative where they seek ongoing support. This relates to the **foundational research** and **diversity** themes mentioned earlier. (For an example of the time scales involved in transitioning from research to

[5] https://www.nytimes.com/2019/01/24/technology/computer-science-courses-college.html
[6] #noCode.org

practice, see the famous “tire tracks” diagram produced by the National Academies: Information Technology Innovation: Resurgence, Confluence, and Continuing Impact.)[7]

While existing efforts at establishing public-private partnerships have shown promise (as suggested in the sidebar on Page 42 for the 2019 Update), we are curious about the extent of the impact on the US research community as a whole, and in particular the involvement of researchers who are faculty members in “standard” academic settings. If the number of researchers who benefit is relatively small, and, more significantly, if these researchers are not also serving the urgent educational mission we face (as tenure-track faculty who conduct research and also teach undergraduates and graduate students), then they are contributing only part of the solution. We need our students – the future AI workforce – trained at the same institutions that are being funded to conduct leading research. These relations to the **education** theme mentioned earlier.

The financial “pull” of well-funded tech companies, along with the attractive environments they provide to researchers, creates a tension that has led to a “brain drain” in academia. This same effect has been visible at the level of domestic graduate students as well. Public-private partnerships could help address this issue, or they can exacerbate it. This concern should be a topic whenever such partnerships are discussed. This is another link to the earlier **education** theme.

For additional thoughts offered on behalf of the computing research community regarding the increasing closer connections between private industry and academic research, we cite the recent CCC whitepaper, “Evolving Academia/Industry Relations in Computing Research”.[8] Quoting from the summary of this whitepaper: “Particular attention needs to be focused on issues related to department culture, potential conflict of interest, intellectual property, and ensuring that students continue to have sufficient faculty mentoring and contact to prepare them for their career.”

[7] https://www.nap.edu/read/25961/chapter/1

[8] https://cra.org/ccc/wp-content/uploads/sites/2/2019/06/Evolving-AcademiaIndustry-Relations-in-Computing-Research.pdf

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Conexus AI

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.

Logic-preserving graph-embeddings

Eric Daimler

Conexus AI

March 3, 2022

We propose that the government fund research into a key problem to better address the needs of data scientists: how to make recommendations on knowledge graphs in a way that conforms to arbitrary given symbolic logical constraints specified by domain e xperts. For example, how to take a knowledge graph about cleaning products that may associate Chlorox and Formula 409, but not recommend that pairing for sale because bleach and ammonia are a toxic combination. Currently, data scientists have no such "safety net" and must account for domain rules manually during e.g. feature engineering, a tedious and error-prone process not known to even be possible outside of some particular special cases (e.g. transitive relationships [2]) – a situation we aim to address with this proposal.

Although there are well-established techniques for learning on matrices with columns of properties, for learning on images, on linear sequences, and many such "low level" structures, over the last several years, learning directly on algebraic structures, and on graphs in particular, has exploded in popularity[1]. This has been driven in part by the existence of so-called "enterprise-wise virtual knowledge graphs", which provide a default "god's eye view of an enterprise", and in part by a desire to lower the barrier to entry of the machine learning process itself (i.e., most domain experts are more comfortable with knowledge graphs than vector spaces). The most basic learning task one can do on a knowledge graph is recommendation: i.e., given an edge label E and two nodes n_1 and n_2, assign a "probability" to the chance that $E(n_1, n_2)$ appears in the graph. For example, given a purchase history E, determine how likely a buyer is to purchase two products n_1 and n_2. To do this, the graph is embedded into a vector space, and then existing statistical learning techniques are used. An example is shown in Figure 1.

Unfortunately, just embedding the knowledge graph into a vector space is not "sound" in the sense that the recommender system can assign high probabilities to facts that are ruled out by the schema that the knowledge graph is on. For example, a knowledge base may contain information about married couples with the implicit assumption that $marriedTo$ is a symmetric relation, yet in certain embedding schemes it is not possible to assign both (Joe, $marriedTo$, $Joan$) and ($Joan$, $marriedTo$, Joe) a high probability.[1]

Of course, the above situation is untenable in that it prevents recommender systems from being used on formal structures that possess laws, for example, a law stating a transitive relationship between three edge labels:

$$\forall x, y : capitalOf(x, y) \implies locatedIn(x, y) \implies containedBy(x, y)$$

[1]If the embeddings of Joe, $Joan$, and $marriedTo$ are represented as vectors, and if the scoring function (see Section 5.3) is additive (i.e. to look for who Joe is liked married to, look for entities near $\overrightarrow{Joe} + \overrightarrow{marriedTo}$), then there is no nontrivial scenario where both $\overrightarrow{Joe} + \overrightarrow{marriedTo} \approx \overrightarrow{Joan}$ and $\overrightarrow{Joan} + \overrightarrow{marriedTo} \approx \overrightarrow{Joe}$ are true.

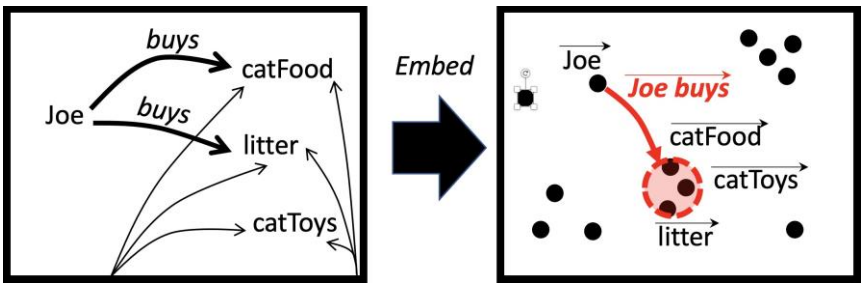


Figure 1: In the visualized subsection of a knowledge graph on the left, the fact that Joe buys cat toys is not present. However, analyzing the graph structure yields many interconnections between the other cat products and cat toys, and an embedding procedure may find it optimal to locate cat toys geometrically nearby the others in the embedded space (here, projected in two dimensions for easy visualization). When we interpret Joe and the buying relation within the embedded space, a region of space may be identified as high likelihood of for Joe buying, so (*J oe*, *buys*, *catT oys*) is a link that is detected.

In practice, knowledge graphs are incomplete; it's possible for one to have the facts like *capitalOf*(*Boston*, *M assachusetts*) and *locatedIn*(*Cambridge*, *M assachusetts*) but not have *containedBy*(*Boston*, *M assachusetts*). An embedding that does not enforce the law above may treat the last fact to be as unlikely as any other fact not present in the knowledge graph and assign it a low probability. To remedy this situation, researchers have proposed special embeddings that are guaranteed to respect certain classes of rules, such as transitivity[10]. However, to the best of our knowledge, no existing work can guarantee that first-order logic will be respected, and we urge the government to tackle this problem.

Appendix: Knowledge Graph Embeddeding Primer

Knowledge graphs [9] are formally specified by a set of entities, e.g. $\mathrm{E} = \{john, joe, \ldots\}$, and relations between entities, e.g. $\mathrm{R} = \{friendOf, likes, \ldots\}$, each of which is a subset of of $\mathrm{E} \times \mathrm{E}$. We denote $(john, joe) \in friendOf$ by writing the tuple $(john, friendOf, joe)$. Therefore the full information of the knowledge graph can be captured in a binary tensor of dimension $|\mathrm{E}| \times |\mathrm{R}| \times |\mathrm{E}|$, with a **1** entry in coordinate (e_i, r_j, e_k) if the knowledge base contains the relevant tuple, else **0**. Thus there is a naive way to 'flatten' structured, hierarchical data into an ideal shape for machine learning (i.e. a tensor); however, because this tensor is very large due to the typical sizes of $|\mathrm{R}|$ and $|\mathrm{E}|$, embedding techniques seek a smaller dimensional representation of the original data with minimal information loss.

The process of constructing an embedding first begins with selecting representations for the embedded entities and relations. The entities and relations are conventionally real-valued vectors, R^n. There has been variation in the scoring function which relates the entity and relation representations, where it can be *additive* (i.e. $\overrightarrow{john} + \overrightarrow{friendOf}$ can be expected to be approximately close to $\overrightarrow{joe}$ when the tuple exists), *multiplicative* (i.e. a matrix **friendOf** is constructed from $\overrightarrow{friendOf}$ by some means such that $\overrightarrow{john}^T$ **friendOf** $\overrightarrow{joe}$ is high when the tuple exists), or some combination of the two. In the literature, additive models are called *translational models* [4, 6, 8], and multiplicative models are sometimes called *bilinear models* [3]. The literature of embedding

techniques has been surveyed in the context of knowledge graphs [5].

The embedded space that the input is mapped into is a low-dimensional, continuous vector space that does not scale in size with $|E|$ or $|R|$. A key outcome of training embedding function is encoding semantic features of the input as geometric features of the embedding space. A paradigmatic case of this is word embedding in natural language processing, where the notion of synonymity can be recaptured as 'low Euclidean distance' in the embedded space as an outcome of training the embeddings to recover information about the words used in the context of any given word. For example, 'Cat' and 'Dog' might get mapped to points whose Euclidean distance is relatively small, and the distance between 'Dog' and 'Canine' will likely be even smaller. Likewise, analogical reasoning has been shown to be captured in the embedded space, such as $\overrightarrow{King} - \overrightarrow{Man} + \overrightarrow{Woman} \approx \overrightarrow{Queen}$. By taking advantage of these geometric properties of the embedded space, computationally inefficient inference tasks can be reduced to linear complexity by querying the embedded representations of data points in the knowledge graph [7].

References

[1] Top trends of graph machine learning in 2020. Technical report, TowardsDataScience, 02 2020.

[2] Antone Amarilli, Michael Benedikt, Pierre Bourhis, and Michael Vanden Boom. Query answering with transitive and linear-ordered data. *Journal of Artificial Intelligence Research*, 63:191–264, 2018.

[3] Ivana Balažević, Carl Allen, and Timothy M Hospedales. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*, 2019.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26:2787–2795, 2013.

[5] Yuanfei Dai, Shiping Wang, Naixue Xiong, and Wenzhong Guo. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics*, 9:750, 05 2020.

[6] Jun Feng, Minlie Huang, Mingdong Wang, Mantong Zhou, Yu Hao, and Xiaoyan Zhu. Knowledge graph embedding by flexible translation. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 557–560, 2016.

[7] William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs, 2019.

[8] Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. Stranse: a novel embedding model of entities and relationships in knowledge bases. *arXiv preprint arXiv:1606.08140*, 2016.

[9] Harmen van den Berg. First-order logic in knowledge graphs. In Carlos Mart´ın-Vide, editor, *Current Issues in Mathematical Linguistics*, volume 56 of *North-Holland Linguistic Series: Linguistic Variations*, pages 319 – 328. Elsevier, 1994.

[10] Mengya Wang, Hankui Zhuo, and Huiling Zhu. Embedding knowledge graphs based on transitivity and antisymmetry of rules, 2017.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Data & Society Research Institute

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.

DATA&
SOCIETY

March 4, 2022

Office of Science and Technology Policy

Re: RFI Response on the Update of the National Artificial Intelligence Research and Development Strategic Plan (FR Doc. 2022-02161)

Dear Dr. Alondra Nelson,

Data & Society Research Institute is pleased to submit a response to the Request for Information (RFI) published by the Office of Science and Technology Policy (OSTP) on updating the National Artificial Intelligence Research and Development Strategic Plan ("Strategic Plan").

Our organization is an independent, nonprofit research institute studying the social implications of data-centric technologies and automation. We are working to produce empirical research that challenges the power asymmetries created and amplified by technology in society. We are pleased to see this commitment to updating the Strategic Plan. As we have shared in recent comments to OSTP[1], it is essential that we develop AI policy and governance mechanisms that are responsive to the prevalence of AI systems that enable discriminatory practices and that expose marginalized communities to harm.

Throughout our comment, we reiterate the importance of long-term research funding that challenges, rather than consolidates, corporate control over the AI research field. In order to meaningfully reckon with mounting evidence of the harmful impacts of large-scale AI systems, the Strategic Plan must introduce additional research directions for programs that facilitate meaningful democratic control of AI and related technologies. We also encourage reframing that centers equity and anti-discrimination as first principles and critical components of successful AI research.

[1]"Democratize AI? How the proposed National AI Research Resource falls short." *AI Now and Data & Society*. (Oct. 5, 2021). https://medium.com/@AINowInstitute/democratize-ai-how-the-proposed-national-ai-research-resource-falls-short-96ae5f67ccfa

1. Background and context: Understanding the "AI boom"

To understand the complex environment surrounding AI research, it's helpful to recall that the current turn to AI is primarily a product of significantly concentrated corporate resources—namely vast computation, massive data, and the capital required to attract and retain scarce AI talent.[2] Placing human beings as active agents in the creation of AI tools and frameworks, as well as stakeholders with a vested interest in algorithmic outcomes is especially important before we can truly assess and understand any technological advances achieved thus far. **The so-called "advances" in AI that have been celebrated since the early 2010s were not due to breakthroughs in AI research and innovation. They were predicated on newly available access to powerful computation and to massive amounts of web data.** Then, as now, these are resources that a handful of powerful tech companies have in large supply thanks to ad tech-driven surveillance business models, that few others can avail themselves of without going through these companies first.[3]

The past decade's Big Tech-led turn to AI profoundly shaped academic computer science disciplines as well. It served to redirect computer science research toward AI-related questions and approaches favored by these companies. In particular, the influx of money and attention produced a turn toward resource-intensive research and development. Work that could avail itself of expensive and scarce industry computing and data was heralded as "cutting edge." This created an uneasy and conflicted environment for university AI research, in which the dependence on large tech company funding, infrastructure, and data was recognized by practitioners, but not often openly acknowledged.

Policy proposals that disrupt the dynamic of concentrated corporate power and the effects this has on academia are important and necessary. However, this also presents us with a set of thorny questions, at the center of which is: how do we validate the importance of federally funded R&D while also reducing the power and control of the handful of companies currently dominating AI research, and how do we ensure that determinations about whether—if at all—AI is developed and deployed are subject to more democratic deliberation. Directing attention and resources to

[2]While the field of Artificial Intelligence is ostensibly oriented around making machines intelligent, in practice, most AI systems rely on big data - the collection and processing of massive datasets, identifying patterns and probabilities within them and codifying them into a predictive mathematical model. See Broussard, Meredith. (2018). *Artificial Unintelligence: How Computers Misunderstand the World.* Cambridge: MIT Press.
[3]Whittaker, Meredith. "The Steep Cost of Capture," ACM Interactions, Vol. XXVIII.6
Nov-Dec 2021. Forthcoming.

these questions is essential for OSTP's efforts to ensure that data-driven technologies reflect and respect our democratic values.[4]

2. **Large-scale government investment in shared computing and data infrastructure will entrench corporate control over the AI field, contrary to the Biden Administration's bold stance against the power of large tech companies in society.**

The National AI Initiative Act calls for regular updates to the Strategic Plan, which may include activities that overlap with National AI Research Resource (NAIRR) aims, such as "[providing or facilitating] the necessary computing, networking, and data facilities for artificial intelligence research and development."[5] If these resources were to take the form of shared, national cloud infrastructure, as envisioned through the NAIRR proposal, we foresee the inability to successfully implement those plans without further entrenchment of corporate influence and control over the AI research field.[6] What is being proposed through the NAIRR is an extension of industry-dependent resources, not the construction of resources that would challenge or reduce the centralized power of the large tech players.

There is no scenario in the short or mid-term future where large scale computational resources adequate to the task of expanding access to bigger-is-better AI research resources could be created and maintained by institutions meaningfully separate from the large tech platform companies. These companies provide more than raw computing power: the computational environments they own and license provide the tools and research environments that define how AI research gets done. Most policy discussion thus far, as has happened within recent media policy history, "involves technical terms that enable easy obfuscation."[7] There is no plausible path forward in which such a resource would not be dependent on existing tech industry platforms, tools, and resources.

---

[4]Lander, Eric, & Nelson, Alondra. "Americans Need a Bill of Rights for an AI-Powered World." *Wired*. (Oct. 8, 2021). https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/

[5]Science and Technology Policy Office. Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan. *Federal Register*. (Feb. 2, 2022). https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and

[6]Science and Technology Policy Office. Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource. *Federal Register*. (July 23, 2021). https://www.federalregister.gov/documents/2021/07/23/2021-15660/request-for-information-rfi-on-an-implementation-plan-for-a-national-artificial-intelligence

[7]Crawford, Susan. (2013). *Captive Audience: The Telecom Industry and Monopoly Power in the New Gilded Age*, Yale University Press.

This is tacitly acknowledged in repeated calls to constitute the NAIRR via public-private partnerships.[8] This arrives at a time when the concentrated power and influence these companies exert is increasingly under scrutiny, including by the Biden Administration itself, which has taken a clear stance that a small number of dominant platforms are using their power to extract monopoly profits.[9] An effort that aims to "democratize" AI research by investing money in companies that dominate their market, will only further entrench these firms' power and reach. This will make it harder to check the power of these companies through regulation and public pressure, and will potentially stifle new innovation, as a small amount of proprietary tech and resources will inherently create limits or boundaries on new research directions.

As we raised in our NAIRR comment, working to democratize access to cyberinfrastructure and to fuel AI research could instead take the form of investments that help us meaningfully ensure democratic control and deliberation over AI. In order to do this, the Strategic Plan could call for additional funding for under-resourced research domains, taking leadership from those most harmed by inequitable uses of AI, and investing in meaningful public control over these powerful technologies.[10] This funding could be used to 1) create and fund scholarships for underrepresented students and sociotechnical research disciplines, 2) invest in fellowships that place sociotechnical scholars in federal agencies, and 3) create and maintain public engagement mechanisms that allow communities most harmed by these systems to have a say in AI R&D.

Public-private partnerships also form a central component of the Strategic Plan. While public-private partnerships can take many forms and are relied on in many parts of government, it's important to view them cautiously in the context of federal AI R&D strategy, given the fact that many companies who may be eager to partner with the government on AI research are the same companies consolidating research power and resources in this field.

In order to better understand the tradeoffs of public-private partnerships in the AI research field, we advise OSTP to convene public workshops with civil society, academia and industry to develop a nuanced approach to public-private partnership that balances

[8]Stanford Institute for Human-Centered Artificial Intelligence. "National Research Cloud Call To Action" https://hai.stanford.edu/national-research-cloud-joint-letter; The National Security Commision on Artificial Intelligence. "NSCAI Submits First Quarter Recommendations to Congress." (Apr. 1, 2020). https://www.nscai.gov/2020/04/01/nscai-submits-first-quarter-recommendations-to-congress-2/

[9]Executive Office of the President. Executive Order on Promoting Competition in the American Economy. (Jul. 9, 2021). https://www.whitehouse.gov/briefing-room/presidential-actions/2021/07/09/executive-order-on-promoting-competition-in-the-american-economy/

[10]The National Science Foundation. "NSF partnerships expand National AI Research Institutes to 40 states." (Jul. 29, 2021). https://www.nsf.gov/news/news_summ.jsp?cntn_id=303176

the benefits of cross-sector collaboration with the risks of further consolidating power and resources within industry.

We also encourage additional historical analyses of analogous large-scale R&D investments from the federal government, in order to identify strategies for facilitating cross-sector collaboration, securing ethical and effective public-private partnership, and ensuring investment in technological development is rooted in maximizing public interest.[11]

3. **The Strategic Plan must begin to reckon with mounting evidence of the harmful impacts of large-scale AI systems, including discriminatory consequences for marginalized groups and long-term climate impact of this scale of computing.**

We encourage additions to Strategy 3 that engage more deeply with the body of critical research, press coverage, investigative reporting, and public discussion that has generated significant evidence of AI's harms[12], and raised fundamental research questions[13] about the ability of AI systems to operate safely and transparently in sensitive social domains.[14]

While proponents of rapidly expanding the use of AI often point to the potential for this technology to stimulate economic growth, many people—particularly marginalized communities—are already subject to the worst excesses, mistakes, and harms perpetuated by the oppressive and extractive use of powerful algorithmic technology.[15] These harms are not

---

[11] For example a historical analysis of U.S. media policy and radio technology. *See* Victor Pickard, (2022). "The Great Reckoning: Lessons from the 1940s media policy battles". *Knight Columbia*. https://knightcolumbia.org/content/the-great-reckoning, and *See* Pickard, Victor. (2021). A New Social Contract for Platforms: Historical Lessons for the Digital Age, in Dealing with Digital Dominance: Joining up the Policy Solutions (Damian Tambini & Martin Moore eds).

[12]Sisson, Patrick. Housing discrimination goes high tech: How algorithms, ad targeting, and other new technologies threaten fair housing laws." *Curbed*. (Dec. 17, 2019). https://archive.curbed.com/2019/12/17/21026311/mortgage-apartment-housing-algorithm-discrimination; Kirchner, Lauren. "Access Denied: Faulty Automated Background Checks Freeze Out Renters." *The Markup.* (May 28, 2020). https://themarkup.org/locked-out/2020/05/28/access-denied-faulty-automated-background-checks-freeze-out-renters ;Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30-56, https://doi.org/10.1016/j.jfineco.2021.05.047.

[13] Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2), 100205.

[14] Alkhatib, A. (2021, May). To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-9).

[15] Noble, Safiya Umoja. (2018). *Algorithms of oppression*. New York University Press. https://nyupress.org/9781479837243/algorithms-of-oppression/; Benjamin, Ruha. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*, Polity Press. https://www.wiley.com/en-us/Race+After+Technology:+Abolitionist+Tools+for+the+New+Jim+Code-p-9781509526437#:~:text=Presenting%20the%20concept%20of%20the,ultimately%20doing%20quite%20the%20opposite

abstract—these are tangible, often irreversible and irreparable harms that perpetuate inequality.[16] **Any effort that focuses on accelerating economic growth while these harms continue relatively unchecked, raises critical questions about whether certain communities are considered expendable and what kinds of harms are allowable in our society in the name of economic growth.**

Furthermore, many subcategories of AI, especially those that emphasize the research and development of large-scale data and computing (often called Large Language Models or LLMs), are uniquely prone to perpetuating social harms and entrenching biases.[17] These large-scale models, which are trained on troves of internet data from sources exhibit persistent discriminatory outputs.[18] Large-scale AI models are built on mass surveillance, which disproportionately impacts marginalized communities,[19] without implementing meaningful mechanisms for accountability or consent of the public.[20] **Their carbon cost is also substantial: the amount of processing required to train AI models is both financially and environmentally resource-intensive, and these costs are only likely to expand given industry standards that tie performance metrics to the size of the dataset used to train the model.[21] As a whole, the tech industry is responsible for a global carbon footprint comparable to the aviation industry, and data centers make up 45% of this footprint.**[22]

The Strategic Plan can begin to acknowledge these harms by being explicit about the wide range of people, organizations, and disciplines that also contribute to AI R&D. Major contributions to

---

[16] Hill, Kashmir. Wrongfully Accused by an Algorithm. *The New York Times*. (Aug. 3, 2020). https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

[17] Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, & Shmitchell, Shmargaret. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). https://dl.acm.org/doi/10.1145/3442188.3445922

[18] Abid, Abubakar, Farooqi, Maheen, and Zou, James. (2021). "Persistent anti-muslim bias in large language models." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298-306. https://arxiv.org/pdf/2101.05783v1.pdf

[19] Georgetown Law Center on Privacy and Technology. The Color of Surveillance." Conference. (Nov. 7, 2019). https://www.law.georgetown.edu/privacy-technology-center/events/color-of-surveillance-2019/

[20] Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, & Shmitchell, Shmargaret. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). https://dl.acm.org/doi/10.1145/3442188.3445922

[21] Schwartz, Roy, Dodge, Jesse, Smith, Noah A., and Etzioni, Oren. (2020). "Green ai." *Communications of the ACM 63*, No. 12 54-63. https://doi.org/10.1145/3381831.

[22] Dobbe, Roel, and Whittaker, Meredith. "AI and Climate Change: How they're connected, and what we can do about it." *AI Now Institute*. (Oct. 17, 2019). https://medium.com/@AINowInstitute/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c

this field are coming from scholars in disciplines like law,[23] anthropology,[24] and history,[25] and also from community organizations[26] gathering qualitative evidence and elevating the lived experience of the people who are surveilled, assessed, and otherwise subject to AI's determinations and predictions. These disciplines and approaches are most capable of analyzing and addressing the structural issues with AI and related technologies and should be included in a broad and flexible definition of AI R&D.

In a joint letter addressed to OSTP in July 2021, a coalition of civil rights and technology organizations called on the department to center civil rights concerns in AI and technology policy, emphasizing the need to fully incorporate the Biden Administration's Executive Order on Racial Equity into its AI policy priorities. This letter drew on years of evidence about the harms that AI is already causing, including perpetuating housing, financial services, and hiring discrimination.[27] This letter also calls on OSTP to "ensure that federal investment in research and development of AI technologies includes significant and immediate research on anti-discrimination measures and ways that AI systems can be used to advance equity, as well as investment in strategies to increase equity, diversity and inclusion in the tech industry."

While the Strategic Plan calls for R&D to develop AI architectures that incorporate societal concerns, we encourage reframing that centers equity and anti-discrimination as first principles and critical components of successful AI research. This is essential for facilitating rigorous, cross-disciplinary work, and avoiding the belated incorporation or retrofitting of ethics onto AI research or systems that were completed without the benefit or perspective of these disciplines.

---

[23]Richardson, Rashida, Schultz, Jason M. & Southerland, Vincent M. (2019). "Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems." *AI Now Institute*. https://ainowinstitute.org/litigatingalgorithms-2019-us.html

[24]Elish, Madeline Clare, and Watkins, Elizabeth Anne. "Repairing Innovation: A Study of Integrating AI in Clinical Care." *Data & Society Research Institute*. (Sept. 30, 2020). https://datasociety.net/library/repairing-innovation/; Brayne, Sarah. (2020). *Predict and Surveil: Data, Discretion, and the Future of Policing.* Oxford University Press. https://global.oup.com/academic/product/predict-and-surveil-9780190684099?cc=us&lang=en&

[25]Bouk, Dan. "House Arrest: How An Automated Algorithm Constrained Congress for a Century." *Data & Society Research Institute.* (Apr. 14, 2021). https://datasociety.net/library/house-arrest/

[26]For example, the Movement Alliance Project https://movementalliance.org/about/, Data for Black Lives https://d4bl.org/, Detroit Community Technology Project https://detroitcommunitytech.org/, Fight for the Future https://www.fightforthefuture.org/

[27]Akselrod, Olga. "How Artificial Intelligence Can Deepen Racial and Economic Inequities." *American Civil Liberties Union*. (Jul. 13, 2021). https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities

4. **The Strategic Plan should reflect the strategic importance of impact assessments as an accountability and governance mechanism for AI R&D.**

The Strategic Plan doesn't yet require the measurement of impacts of AI systems supported through these strategies. The harmful impacts of AI systems must be better understood in order to effectively and ethically commit public resources to those efforts.

Multiple jurisdictions are exploring algorithmic impact assessments as a means for regulating algorithmic systems and protecting the public interest. Data & Society's research on impact assessments maps the challenges of constructing algorithmic impact assessments by analyzing their use in other domains, including finance, environment, human rights, and privacy.[28] To ensure the effective development of algorithmic impact assessments as a governance mechanism, this report presents a framework for evaluating impact assessment processes. This framework deepens our understanding of the mutual shaping of accountability and practices of measuring harms through impacts.

Our research shows that impact assessments are a reliable way to ensure that companies study, explain, and report on how their proposals will affect society. Not only do impact assessments illuminate the harms of specific activities or products, they also establish a much needed standard for the disclosure of where and why algorithms are being deployed, who was consulted, and what steps were taken to mitigate or prevent risks.

A recent legislative proposal seeks to mandate impact assessments in instances where automated decision systems are used to make critical decisions about our lives, demonstrating the viability of impact assessments as an accountability mechanism.[29] However, more research is needed to understand how best to develop and use impact assessment methodologies. Additional research is also needed to understand how best to assess the impacts of basic research, along the lines of

[28]Moss, Emanuel, Watkins, Elizabeth Anne, Sing, Ranjit, Elish, Madeline Clare, & Metcalf, Jacob. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." *Data & Society Research Institute*. (June, 29, 2021). https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

[29] *See* "Wyden, Booker and Clarke Introduce Algorithmic Accountability Act of 2022 To Require New Transparency And Accountability For Automated Decision Systems." (Feb. 3, 2022). https://www.wyden.senate.gov/news/press-releases/wyden-booker-and-clarke-introduce-algorithmic-accountability-act-of-2022-to-require-new-transparency-and-accountability-for-automated-decision-systems#:~:text=Y.%2C%20today%20introduced%20the%20Algorithmic,every%20aspect%20of%20Americans'%20lives.

the technology impact work once conducted by the Office of Technology Assessment.[30] Without this research, the status quo in which companies conduct their own impact assessments (if at all) and don't publish the results, remains in place. This makes it difficult for researchers, policymakers, and the general public to understand and contest the impacts of AI systems.

We encourage OSTP to incorporate impact assessments into the Strategic Plan in two ways:

- **Impact assessments should be incorporated into Strategy 1 as a mechanism for determining which long-term investments in AI research maximize public interest while minimizing or preventing harm.** Our ability to control AI systems depends largely on what we know about how they work, where they are deployed, and in which ways they affect the general public. Impact assessments can make this information visible and can be useful for encouraging researchers to consider and mitigate negative downstream impacts of their research in advance of deployment. This is particularly important to do in advance of being granted access to public funding and public datasets.

- **Impact assessments should also be incorporated into Strategy 3 and Strategy 6 as areas where additional research is needed.** These assessments can expose preventable harms, encourage consultation with affected communities, and standardize the information we have available for further research about which AI systems are used in which contexts and for which purposes. The development of methodological standards for these assessments is especially critical for ensuring impact assessments are done in the public interest, and to prevent the proliferation of assessments that manipulate or obscure harmful impacts of AI systems.

Thank you for your openness to feedback on revisions to the Strategic Plan. We encourage further public engagement on areas the Strategic Plan can support that facilitate meaningful democratic control of AI and related technologies. We look forward to supporting OSTP in this effort.

Sincerely,

Brittany Smith, Policy Director
Melinda Sebastian PhD, ACLS Leading Edge Fellow

[30]Congressional Research Service Report. "The Office of Technology Assessment: History, Authorities, Issues, and Options." (Apr. 29, 2020). https://sgp.fas.org/crs/misc/R46327.pdf

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

DeepMind

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.

DeepMind Response: National Artificial Intelligence Research and Development Strategic Plan

At DeepMind we believe that AI's extraordinary potential will only be realized if its development and deployment uphold appropriate ethical standards and if it is purposefully directed towards benefitting society.

We welcome the opportunity to respond to the Office of Science and Technology Policy's Request for Information on the update of the National AI Research and Development Strategic Plan. The past years have seen extraordinary advances in the field, further evidencing the ways AI will impact all parts of society. **We believe the eight Strategies in the Plan reflect a comprehensive vision for AI R&D, and as such remain the right priorities to guide investment. At the same time, clarifying how the United States intends to weave them together into a whole-of-government approach, particularly given the emergence of many new AI-related government efforts in the past two years, remains crucial to continued leadership in this space.** We offer below some considerations to strengthen these strategies, and we look forward to future opportunities to input into OSTP's work.

About DeepMind

DeepMind[1] is a scientific discovery company, committed to 'solving intelligence' to advance science and benefit humanity. This requires a diverse and interdisciplinary team working closely together – from scientists and designers, to engineers and ethicists. AI has the potential to enrich the lives of billions and improve our understanding of the universe. Ultimately we hope that new scientific breakthroughs, driven by innovations in machine learning, can make the crucial difference in helping us prosper in an increasingly complex world, and respond to global challenges such as climate change and tackling diseases.

[1] We share these comments on behalf of DeepMind, and not on behalf of Google or any other entity in Alphabet, Inc.

For AI to benefit as many people as possible, it needs to be built and used responsibly. We view responsible AI as an ongoing process of ensuring our research and engineering are informed by the values, needs and expectations of society – with the goal to minimize risks, but also to accelerate and equitably distribute the benefits of AI. In practice, this means: (1) making sure our research addresses major scientific and social challenges; (2) anticipating and mitigating potential risks and harms; and (3) engaging with the wider world, and its complexities, challenges and possibilities.

The evolving AI research landscape

The 2019 update of the Strategy captured priorities in AI development that remain relevant today. We particularly welcomed its emphasis on safe, ethical and trustworthy AI; the recognition of the inherently interdisciplinary nature of AI R&D; and the adoption of a long-term perspective, building preparedness for outcomes that could take decades to manifest. Taking stock of the Strategic Plan is nonetheless pertinent, given developments in the field in recent years and the broader backdrop of tremendous public and private investment in the AI ecosystem.

The growing promise of using AI to advance scientific discovery

AI can deliver transformative benefits by accelerating scientific discovery, helping to address grand challenges like climate change and pandemic preparedness. These benefits are just beginning to be realized, as shown by the launch of a new National Science Foundation-funded institute to harness AI for accelerated discoveries in physics, astronomy and neuroscience last year.

We hope that DeepMind's deep learning system, AlphaFold, which can predict the 3D structure of a protein based solely on its genetic sequence, will herald a new age of AI-powered scientific breakthroughs. More than 350,000 users from more than 190 countries have already visited the AlphaFold Protein Structure Database, which we established in partnership with the European Molecular Biology Laboratory's (EMBL) European Bioinformatics Institute (EBI), and which now holds structures for almost one million proteins. We are seeing promising signs of scientists incorporating AlphaFold into their day-to-day work, and socially-beneficial research

emerging in areas like human biology, neglected diseases, and plastic-degrading enzymes.

We are also starting to see how AI can help to tackle climate change, by making existing infrastructure, such as data centers, more efficient, and by enabling better prediction models in areas like weather forecasting. AI's most transformative climate impacts will likely come from accelerating crucial longer-term breakthroughs in areas like fusion and battery design.

AI community efforts to pioneer responsibly

AI's potential transformative benefits to society are only possible if it is built and used responsibly, and if negative effects, which have to date disproportionately affected marginalized groups, are sufficiently mitigated. The AI community has been making progress in providing practitioners with resources and tools to identify, analyze and manage potential risks and benefits from their AI research and applications. For example:

- In 2020, the NeurIPS AI research conference introduced a new requirement for authors to produce an impact statement on the potential ethical aspects and societal risks of their work.
- Organizations deploying AI applications can increasingly draw on tools and best practice case studies shared by other organizations,[2] in areas like fairness, privacy, security and explainability[3].
- At DeepMind we invest heavily in safety, security, privacy, ethics and sociotechnical research both to inform our own approach to AI development and governance, and to help foster progress in the broader field. For example, we recently released research that assessed the near-term ethics risks, and longer-term safety risks, posed by large-scale language models - a priority area of focus for AI researchers.

[2] Google has made available a range of off-the-shelf and customisable risk analysis and mitigation tools, in areas like fairness, privacy and explainability

[3] For instance, DeepMind's recent work on glass dynamics demonstrates how a graph neural network can predict molecules' future movements based on their current structure. The DeepMind Safety Research Team published on its blog about 'what mechanisms drive agents behavior'. Another example is this paper which seeks to provide explanations for AlphaZero's chess moves.

Updating the National AI R&D Strategic Plan

Over the past three years, an increasing number of policy instruments have been established to support the AI ecosystem and demonstrate the US government's commitment to being a world leader in AI. We particularly welcome:

- The increased investment to accelerate AI innovation – including by significantly increasing non-defense AI funding and augmenting R&D in security and robustness – following various recommendations (including from the National Security Commission on AI).
- The establishment of the National AI Office within OSTP and the creation of platforms such as ai.gov to improve the visibility of ongoing AI-related work and to act as a connection point for stakeholders.
- The creation of a National AI Research Resource Task Force to democratize access to research tools and capabilities;
- The US's support and creation of international fora to improve convergence on AI governance questions, such as the EU-US Trade and Tech Council, and to facilitate the input of international experts, such as the Global Partnership for AI.
- The work initiated by the National Institute of Standards and Technology around trustworthy AI, including the development of a voluntary AI Risk Management Framework (RMF). We've welcomed the multistakeholder approach taken by NIST and the multiple opportunities to input as the RMF is being developed.

The National AI R&D Strategic Plan rightly highlights the importance of a whole-of-government approach; the recent multiplication of policy proposals around R&D investments shows coordination is all the more important today. In the next update to the Plan, we recommend prioritizing that a common vision be evident for how the government intends to drive forward its multiple strategies. We hope to see, for instance, how this Strategic Plan will link with OSTP's development of an "AI bill of rights," among other initiatives. Considering the complex nature of AI R&D and the overlaps that naturally exist between the eight strategies, the Strategic Plan could also benefit from a better acknowledgement of these links to prevent each strategy from becoming siloed, a challenge which figure 1 of the Strategic Plan tries to capture.

We share below research areas that could be emphasized and more general observations across the eight strategies.

1) *R&D areas*

Long-term investments in AI research (Strategy 1)

The Strategic Plan rightly highlights the long-term perspective needed when investing in AI, particularly in research aiming to achieve general-purpose artificial intelligence. We believe general-purpose learning systems are key to unlocking the long-term potential for AI to benefit society.[4] But as such AI capabilities become more advanced, they raise the possibility of novel safety risks that may manifest on different timelines. While this is also covered in Strategy 4, we recommend that this section place greater emphasis on AI safety research, and connect clearly to the subsection in Strategy 4 that focuses on long-term AI safety and value alignment.

Effective methods for human-AI collaboration (Strategy 2)

We strongly agree with the view outlined in the Strategy that "achieving effective interactions between humans and AI systems requires additional R&D to ensure that the system design does not lead to excessive complexity, undertrust, or overtrust." Enabling this careful calibration of the trust one places in an AI system is core to the notion of “trustworthy AI.” While we are seeing promising research in this space, it remains underserved, especially with regard to more advanced AI systems. For example, while researchers in the field of human computer interaction (HCI) have studied the trust placed by humans in narrow applications, such as doctors using ML-based tools for diagnostics, there is comparatively little research on how humans interact with broader or more advanced systems, such as large language models.

The Strategy also seems to focus quite heavily on human-computer-*collaboration* on a task and on human *augmentation*, in comparison to studying human perception and meaningful oversight of AI systems in human-AI collaboration settings. A

[4] In an example of how general learning systems are beginning to be applied for practical applications, DeepMind recently demonstrated how the application of our algorithm, *MuZero,* helped improve video compression, resulting in a 4% bitrate reduction across a large, diverse set of videos for YouTube.

relevant avenue of research in this direction would focus on how we can design meaningful human control, oversight, and accountability. The Strategy could also look at ways to create human-in-the-loop evaluation pipelines for AI, and how to provide humans with explanations for decisions and outcomes in a way that is judged by a human to be useful.

2) *Cross-cutting R&D Foundations*

Ethical, legal, and societal implications of AI (Strategy 3)

This Strategy rightly highlights the importance of understanding the ethical, legal, and social implications of AI, as well as developing methods for AI design that align with ethical, legal, and social principles. We need to continue encouraging investments in sociotechnical research, and institutionalizing best practices such as foresight and deliberation. The prompt that the NeurIPS conference introduced for researchers to produce an impact statement on the potential ethical aspects and societal risks of their work is something that could be further replicated in other conferences and processes to award grants.

We also agree that research benefits from multidisciplinary perspectives from computer science, the social and behavioral sciences, ethics, biomedical science, and other fields. Such an interdisciplinary approach will also be particularly important to driving forward much needed research on AI governance. At DeepMind, we have a multidisciplinary leadership group and dedicated internal teams that review research proposals and potential applications of our technology, consult external experts, and develop recommendations to maximize the likelihood and distribution of positive outcomes and minimize the potential for harm. In advance of our AlphaFold release, for instance, this group engaged with leading bioethicists and protein folding researchers to explore potential impacts on the research community, as well as any ways in which harmful actors might use our research. We also sought guidance from experts in areas with potential for beneficial impact, such as neglected diseases, to try and validate and accelerate these opportunities. We've also seen other organizations create fora for such discussions, in ways that match their particular research priorities.[5]

[5] Stanford University for instance created the Ethics and Society Review (ESR) to aid researchers in mitigating negative ethical and societal aspects of their research.

While this is already mentioned throughout the Strategy, we recommend creating a separate subsection on the importance of multi-stakeholder approaches, and on prioritizing diversity and inclusion in discussions on the ethical, legal and societal implications of AI. AI actors (researchers, developers, deployers, and more) need to develop normative thresholds to inform decision-making about what constitutes 'trustworthy enough' for an AI system to be deployed, and inclusiveness should be central to such decision-making. Developing normative thresholds requires identifying and engaging with groups that may be most at risk from AI systems. Sociotechnical research can help all sorts of stakeholders — including both parties involved in the design and development of AI systems and end users — better understand how society (including traditionally minority groups) may be affected by AI systems. [6] Participatory approaches with these groups can provide a source of expert insights and lived experience, and a way to empower those who may be most affected by AI systems.[7] There could also be value in mirroring ways in which some governments are creating fora for public engagement: the UK, for instance, has created an AI Council to engage with the broader AI community.

Safety and security of AI systems (Strategy 4)

Research into the safety and security of AI systems remains a crucial priority, and we consider the current description of the Strategy still relevant, since challenges such as improving trust and increasing the explainability and transparency of AI need continued research and funding.

The section on 'long term AI safety and AI alignment' would benefit from adopting a broader framing: ultimately, the focus of alignment research is to prevent powerful goal-directed systems from pursuing undesired goals. While recursive self-improvement is an important potential factor, it is only one (speculative) way

[6] By 'sociotechnical research', we mean research on the interaction and effects of AI when embedded in a specific social system. For example, DeepMind researchers have applied critical science and decolonial theory to AI to explore risks like algorithmic oppression, dispossession and exploitation, and analysed the potential positive and negative effects of artificial intelligence on queer communities.

[7] DeepMind researchers are exploring the potential role of participatory approaches in developing and/or evaluating AI systems. Such approaches are nascent in AI, but are more established in fields like human-computer-interaction (HCI), which we hope can serve as an important source of insights.

that such risks could emerge. We hence recommend this section take a wider lens, including specific longer-term AI safety problems that have been identified as priorities by the research community.[8] An example of such a problem is 'specification gaming', where an agent satisfies the literal specification of their objective, but does not achieve the desired outcome.[9] We also recommend considering how to attract more researchers into long-term AI safety. The field is home to a growing number of talented researchers, but their overall number remains limited. High-profile, coordinated activities – such as collective research agendas, dedicated major funds for such research, and/or competitions – could help to further develop the field.

The section on 'security against attacks' covers a robust set of priorities on security of AI systems, and the focus should remain on ensuring these systems are developed to detect when they are being attacked, and to withstand attacks (such as data poisoning to evade or manipulate models, model stealing, etc.). Two areas that could be explored further are the verifiability of AI systems and ensuring some level of provenance and traceability of the corresponding models and data of such systems. We recommend discussion on bringing more transparency and openness in deployed AI systems to address emerging risks and enable trust-building.

Shared public datasets and environments for AI training and testing (Strategy 5)

As the 2019 Strategic Plan outlines, it is critical for researchers to have easy access to reliable, clean, as well as findable, accessible, interoperable, and reusable (FAIR) data. We welcome the creation of the National AI Research Resource Task Force and the drafting of a roadmap to expand access to critical resources and educational tools relating to AI. The Strategic Plan could be updated to more directly support the National AI Research Resource Task Force's mandate.

We also encourage OSTP to consider updating the Strategic Plan to encourage opportunities to make data from countries around the world more available and

[8] DeepMind has a dedicated team of researchers that identify and work on such problems, in particular specification, robustness and assurance. A recent paper assesses the potential risks from misaligned Language Agents, such as producing language that is deceptive or manipulative.
[9] DeepMind collated 60 examples of specification gaming, such as an agent that was trained to carry a ball on its back but instead dropped the ball into a leg joint and wiggled across the floor without the ball dropping.

accessible and, where possible, to use harmonized methods to protect privacy and security.

Standards and benchmarks (Strategy 6)

The establishment of effective AI standards and benchmarks is key to trustworthy AI. We welcome the efforts NIST is leading in this space, along with many needed discussions in international fora. At the same time, it is helpful to acknowledge the need for more research and awareness of possible limitations of benchmarks, and the need for new accompanying tools to address this. For instance:

- Current benchmarks disproportionately evaluate accuracy on narrow tasks of interest, and neglect important areas like fairness and explainability, although useful benchmarks are emerging in these areas. Current benchmarks also don't necessarily test performance in a way that is relevant to real-world use. As highlighted in a paper by Deb Raji, Emily Bender and co-authors at Google, evaluating AI systems based on their aggregate performance on benchmark tasks may provide little indication of their real-world utility. For example, 80% accuracy on Iris classification might be sufficient for the botany world, but to classify a mushroom you need to ingest as poisonous or edible, you would need 99% (or higher) accuracy.
- Benchmarks that are analyzed in isolation may also give a wrong picture of a model, as pointed out by DeepMind researchers in this paper. They find that evaluating a model against toxicity can ignore or even introduce unfair bias, for example against texts about, and dialects of, marginalized groups, demonstrating that benchmarks must consider social harms in concert in order to not fix one problem by aggravating another.
- Most benchmarks are quantitative metrics on task performance, and there is a need for new types of qualitative analysis tools, including feedback from humans-in-the-loop and learnings from other domains, such as auditing practices that evaluate how something is being used in the real world.

National AI R&D workforce needs (Strategy 7)

The Strategic Plan mentions the importance of broadening participation among groups traditionally underrepresented in computing and related fields. Diverse

teams are indeed important - not only for the more innovative work that such teams produce, but also because of the diverse values, hopes, and concerns that diverse teams bring into AI design, risk/benefit assessment, mitigation development, and deployment. We believe this is essential and an area where the private sector also has a role to play: the Strategic plan could highlight opportunities to work with the private sector to tackle these challenges and create a strong diverse talent pipeline.

The Strategy could also include mention of the importance of supporting early-career researchers from under-represented groups to pursue postdocs and possibly transition to permanent positions in academia to become role models for the next generation of researchers. This is why we launched the DeepMind Academic Fellowships, in addition to the DeepMind Scholarship, which provides funding and mentoring to graduate students.

Public-private partnerships to accelerate advances in AI (Strategy 8)

The addition of this Strategy in 2019 reflected the growing importance of public-private partnerships in enabling AI R&D. The 2022 update might be the opportunity to create a separate section dedicated to the need for more international projects, across governments but also with the private sector and civil society, to accelerate advances in AI. We're particularly supportive of initiatives such as the recent US-UK Privacy Enhancing Technologies prize challenge, and encourage more international AI R&D collaborations.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Electronic Privacy Information Center (EPIC)

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.

epic.org

Electronic Privacy Information Center
1519 New Hampshire Avenue NW
Washington, DC 20036, USA

+1 202 483 1140
+1 202 483 1248
@EPICPrivacy
https://epic.org

COMMENTS OF THE ELECTRONIC PRIVACY INFORMATION CENTER

to the

Office of Science and Technology Policy

Regarding the

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

87 Fed. Reg. 5,876

March 3, 2022

The Electronic Privacy Information Center (EPIC) submits the following feedback to the request for information by the Office of Science and Technology Policy (OSTP) on behalf of the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence (Select Committee), the NSTC Machine Learning and AI Subcommittee (MLAI-SC), the National AI Initiative Office (NAIIO), and the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), hereinafter referred to as "agencies," concerning the Update of the National Artificial Intelligence Research and Development Strategic Plan.[1]

Interest of EPIC

EPIC is a public interest research center in Washington, D.C. that was established in 1994 to focus public attention on emerging privacy and related human rights issues and to protect privacy, the First Amendment, and constitutional values.[2] EPIC has a long history of promoting transparency and accountability for information technology.[3]

[1] Science and Technology Policy Office, Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan, 87 Fed. Reg. 5,876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and.

[2] EPIC, *About EPIC* (2021), https://epic.org/epic/about.html

[3] EPIC, *AI & Human Rights* (2021), https://epic.org/issues/ai/; EPIC, *Algorithms in the Criminal Justice System* (2021), https://epic.org/issues/ai/ai-in-the-criminal-justice-system/; EPIC, *AI Policy* (2021) https://epic.org/issues/ai/ai-policy/; EPIC, *Government AI Use* (2021 https://epic.org/issues/ai/government-use-of-ai/; EPIC, *Commercial AI Use* (2021) https://epic.org/issues/ai/commercial-ai-use/; EPIC, *Scoring and Screening (2021)* https://epic.org/issues/ai/screening-scoring/.

EPIC has a particular interest in promoting algorithmic transparency and has consistently advocated for the adoption of the Universal Guidelines for AI ("UGAI") to advance trustworthy use of algorithms and justice for individuals harmed by AI systems.[4] EPIC has advocated for transparency and accountability internationally in connection with the use of AI systems.[5] EPIC has litigated cases against the U.S. Department of Justice to compel production of documents regarding "evidence-based risk assessment tools"[6] and against the U.S. Department of Homeland Security to produce documents about a program purported to assess the probability of whether an individual committed a crime.[7] In 2018, EPIC and leading scientific societies petitioned the U.S. Office of Science and Technology Policy to solicit public input on U.S. artificial intelligence policy.[8] EPIC submitted comments urging the National Science Foundation to adopt the UGAI and to promote and enforce the UGAI across funding, research, and deployment of U.S. AI systems.[9] EPIC has also submitted comments to the National Security Commission on Artificial Intelligence, the U.S. Office of Science and Technology Policy, the European Commission, and the U.S. Office of Management and Budget urging the adoption of AI system regulation that meaningfully protects individuals.[10]

---

[4] *See, e.g.*, EPIC, *EPIC Comments to NIST in re Artificial Intelligence Risk Management Framework*, National Institute of Standards and Technology (Aug. 18, 2021), https://epic.org/documents/regarding-the-artificial-intelligence-risk-management-framework/; EPIC, *EPIC Comments to Comptroller of the Currency et al. in re Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning* (July 1, 2021), https://archive.epic.org/apa/comments/EPIC-Financial-Agencies-AI-July2021.pdf; EPIC, *EPIC Comments to the Federal Communications Commission Technolological Advisory Council*, Federal Communications Commission (Sept. 18, 2020) https://epic.org/documents/comments-to-the-federal-communications-commission-technological-advisory-council/; EPIC, *EPIC v. DOJ* (2020), https://epic.org/foia/doj/criminal-justice-algorithms/; EPIC, *EPIC Comments to the U.S. Patent and Trademark Office in re Intellectual Property Protection for Artificial Intelligence Innovation* (Jan. 10, 2020), https://epic.org/apa/comments/EPIC-USPTO-Jan2020.pdf; EPIC, *EPIC Comments to the Department of Housing and Urban Development in re Implementation of the Fair Housing Act's Disparate Impact Standard* (Oct. 18, 2019), https://epic.org/apa/comments/EPIC-HUD-Oct2019.pdf; Testimony of EPIC, Massachusetts Joint Committee on the Judiciary (Oct. 22, 2019), https://epic.org/testimony/congress/EPIC-FacialRecognitionMoratorium-MA-Oct2019.pdf; Statement of EPIC, *Industries of the Future*, U.S. Senate Committee on Commerce, Science & Transportation (Jan. 15, 2020), https://epic.org/testimony/congress/EPIC-SCOM-AI-Jan2020.pdf; EPIC, *EPIC Comments to the Office of Science and Technology Policy in re Request for Information: Big Data and the Future of Privacy* (Apr. 4, 2014), https://epic.org/privacy/big-data/EPIC-OSTP-Big-Data.pdf.

[5] EPIC, AI & Human Rights, *supra* note 3.

[6] EPIC, *EPIC v. DOJ*, *supra* note 4

[7] *Id.*; *see also* EPIC, *EPIC v. AI Commission* (2021), https://epic.org/documents/epic-v-ai-commission/; *EPIC v. DHS (FAST Program)* (2015), https://epic.org/documents/epic-v-dhs-fast-program/.

[8] EPIC, Petition to OSTP for Request for Information on Artificial Intelligence Policy (July 4, 2018), https://epic.org/privacy/ai/OSTP-AI-Petition.pdf.

[9] EPIC, Request for Information on Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan, 83 Fed. Reg. 48,655 (Oct. 26, 2018), https://epic.org/apa/comments/EPIC-Comments-NSF-AI-Strategic-Plan-2018.pdf.

[10] EPIC, *EPIC Comments to OSTP in re Public and Private Sector Uses of Biometric Technologies* (Jan. 15, 2022); EPIC, *EPIC Comments to OSTP in re Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource* (Oct. 1, 2021); EPIC, *EPIC Comments in re Solicitation of Written Comments by the National Security Commission on Artificial Intelligence* (Sept. 30, 2020), https://epic.org/apa/comments/EPIC-comments-to-NSCAI-093020.pdf; EPIC, *EPIC Comments to OMB in re Request for Comments on a Draft Memorandum to the Heads of Executive Departments and Agencies* (Mar.

The Agencies Should Rely on the Universal Guidelines for AI and the OECD AI Principles to Guide Updates to the Research and Development Plan

EPIC recommends that the agencies use the Universal Guidelines for Artificial Intelligence to guide updates to the Research and Development Plan. The UGAI, based on the protection of human rights, were set out at the 2018 Public Voice meeting in Brussels, Belgium.[11] The UGAI have been endorsed by more than 250 experts and 60 organizations in 40 countries.[12] The twelve guidelines are:

1. Right to Transparency
2. Right to Human Determination
3. Identification Obligation
4. Fairness Obligation
5. Assessment and Accountability Obligation
6. Accuracy, Reliability, and Validity Obligations
7. Data Quality Obligation
8. Public Safety Obligation
9. Cybersecurity Obligation
10. Prohibition on Secret Profiling
11. Prohibition on Unitary Scoring
12. Termination Obligation[13]

The agencies should also incorporate the AI principles adopted by the Organization of Economic Cooperation and Development ("OECD AI Principles"). [14] The OECD AI Principles were adopted in 2019 and endorsed by 42 countries—including several European Countries, the United States, and the G20 nations.[15] While largely aligning with the principles of the UGAI, the OECD AI Principles provide additional considerations that may be beneficial to the register The OECD AI Principles establish international standards for AI use:

1. Inclusive growth, sustainable development and well-being
2. Human-centered values and fairness
3. Transparency and explainability
4. Robustness, security, and safety

13, 2020), https://epic.org/apa/comments/EPIC-OMB-AI-MAR2020.pdf; EPIC, *EPIC Comments to the European Commission Fundamental Rights Policy Unit in re Request for Feedback in Parallel with the White Paper on Fundamental Rights* (May 29, 2020), https://epic.org/apa/comments/EPIC-EU-Commission-AI-Comments-May2020.pdf; EPIC, *EPIC Comments to the European Commission in re Proposal for a legal act of the European Parliament and the Council laying down requirements for Artificial Intelligence* (Sept. 10, 2020), https://epic.org/apa/comments/EPIC-EU-Commission-AI-Sep2020.pdf.

[11] *Universal Guidelines for Artificial Intelligence*, The Public Voice (Oct. 23, 2018) [hereinafter *Universal Guidelines*], https://thepublicvoice.org/ai-universal-guidelines/.

[12] *Id.*

[13] *Id.*

[14] *Recommendation of the Council on Artificial Intelligence*, OECD (May 21, 2019) [hereinafter *OECD AI Principles*], https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[15] *U.S. Joins with OECD in Adopting Global AI Principles*, NTIA (May 22, 2019), https://www.ntia.doc.gov/blog/2019/us-joins-oecd-adopting-global-ai-principles.

5. Accountability[16]

Agencies Should Expand and Substantiate Goals of Ethics and Accountability and Focus Development Strategies to an Appropriate Scope

The agencies should update the eight strategies currently comprising the National Artificial Intelligence Research and Development Strategic Plan with specific, actionable measures to "understand and address the ethical, legal, and societal implications of AI" and to "ensure the safety and security of AI systems."[17]

Five of the eight current strategies explicitly promote the development of AI. For example, making datasets publicly available and expanding public-private partnerships to accelerate advances in AI will support AI developers and increase the introduction and use of AI systems. Yet there are no specific parallel strategies to address "implications of AI" or "ensure . . . safety and security." If the Strategic Plan fails to establish concrete steps and benchmarks for safeguarding the public against AI, it is likely these goals will go unrealized or become watered down in the interests of rapid AI development.

EPIC provides feedback on the specific strategies below, recommending several new strategies and limiting ones that irresponsibly accelerate development and deployment of technologies without the requisite oversight and protections in place.

Strategies 3 and 4 Should Include Action Items to Protect the Public From Harmful AI Systems

Building AI oversight and regulatory capacity must be a top priority of federal agencies. To achieve this, EPIC recommends that Strategy 3 be expanded into specific action items, including:

- Prohibiting the use of AI systems that pose unjustifiable risks or which are otherwise ineffective, improper, inaccurate, or biased;
- Establishing prohibitions on AI systems that, alone or in combination with other technologies, are manipulative or facilitate mass profiling;
- Requiring agencies that use automated decision-making systems to publish vital information about those systems in a user-friendly inventory, expanding on requirements created by Executive Order 13,960;
- Developing impact assessment and reporting standards that users of a commercially developed AI tool must comply with when the tool is used in sensitive contexts such as hiring, credit determinations, and criminal justice;
- Imposing purpose specification and use limitation requirements on AI systems to mitigate mission creep;
- Providing an opportunity for individuals unfairly harmed by AI systems to obtain redress; and

[16] *Recommendation of the Council on Artificial Intelligence*, Organisation for Economic Cooperation and Development (May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
[17] National AI Research and Development Plan Strategies 3 & 4.

- Limiting the collection of personal information by AI systems without express, informed consent.

Similarly, the agencies should update Strategy 4 to include the following action items:

- Prohibiting the use of inherently unsafe AI applications;
- Establishing and enforcing minimum data governance and minimization requirements;
- Ensuring that agencies adequately evaluate and publish information about the AI systems they procure and develop; and
- Determining best practices for identifying which AI tools are discriminatory, inaccurate, or otherwise fundamentally incompatible with the protection of human and civil rights.

Strategy 8 Should be Narrowed to Prevent Surveillance, Bias, Corporate Capture, and Other Harms

Without adequate limits, the focus on AI public-private partnerships called for in the Strategic Plan will pose an unacceptable threat to civil and human rights. Under the current Plan, the agencies are complicating problems associated with AI that they are not yet putting adequate resources toward addressing. As EPIC warned the National Security Commission on Artificial Intelligence in September 2020, "incentivizing the adoption of commercial software tools and 'moderniz[ing]' solely to gain a competitive edge will undermine the U.S.'s principled leadership on AI."[16]

EPIC recommends that the agencies adjust Strategy 8 to limit public-private partnerships to circumstances where (1) there is a demonstrated need for a particular type of AI development or research, and such development or research can be accomplished consistent with the preservation of human and civil rights; or (2) where public-private partnerships are needed to improve AI oversight."

This shift in focus would allow the agencies to achieve their parallel goals of expanding public-private partnerships and protecting civil rights and civil liberties without hindering either. The agencies can identify appropriate and beneficial uses of AI while avoiding the facilitation of new AI systems purely for the sake of competition.

Conclusion

For the reasons above, EPIC recommends that the agencies prioritize the protection of the public in updating the National Artificial Intelligence Research and Development Strategic Plan.

Respectfully Submitted,

/s/ *Ben Winters*
Ben Winters
EPIC Counsel

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Freed/Choset/Mani, Carnegie Mellon University

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.

The Democratization of AI and Data: A Strategic Public-Private Partnership Policy for American AI in the Twenty-First Century

Ben Freed, Howie Choset, Ganesh Mani
School of Computer Science, Carnegie Mellon

Abstract

As our AI tools become more advanced, they are increasingly created and controlled by a select few organizations. By limiting the breadth of institutions, groups, and people who can create, use, and inspect AI tools, the AI oligarchy has negative impacts for individuals, society, as well as the progress of the field. We firmly believe that AI has the potential to bring a tremendous amount of good to the world, but only if developed and used responsibly, which is a conversation in which everyone should have a voice. We identify three key ingredients to advance AI, which we call *the three D's of AI:* data availability, developer accessibility, and democratization of AI tools. It is our view that to unlock the true potential of AI, and retain American competitiveness, we must conquer, and therefore invest in, the three Ds of deep learning: data, developer, and democratization. This investment should take the form of a public-private-academic partnership to promote education and development.

The Three-D's	Data availability	Developer	Democratization
Benefit	Benefits everyone Removes biases	More talent Uncork talent	Society engagement Equitable ownership
Problem	Cost to create Privacy Biases / silos	Cost to obtain Cost to maintain Lack of tools	Unchecked firms Monopoly / barriers No easy-to-use tools
Solution	Surrogate data Public funds reqs. Data efficient apps	Tool develop Shared resources K12 education Retain international	Easy-to-use tools User-owned data K12 competition

Introduction

Today, we are the indistinguishable leaders in AI and it is important, not only for our commercial success, but for National security, that we sustain this leadership role. We were the pioneers, innovators, and first-time-users in computer manufacturing, and today, we are not. It is even worse - the world is now suffering from lack of American leadership in computer manufacturing - just ask any potential customer in the automobile market about the microprocessors that are holding up production. We are the leaders in software development, as evidenced by the several companies formed

around software, and we run the risk of losing that edge as well. We cannot lose our leadership in AI.

Recent years have seen a boom of AI development, spurred by deep learning. Deep learning has revolutionized the way in which AI is applied to domains such as manufacturing, finance, medicine, energy, agriculture, security, retail, just to name a few. Deep learning can be viewed as a type of data science that can model and predict future outcomes from (an enormous amount of) data that is provided to it, during a training process, of a multi-layer neural network. Deep learning, along with many other AI technologies, are typically classified as *data driven*, because they primarily focus on extracting patterns from data, rather than relying on the knowledge of AI engineers or domain experts. This shift in perspective from the *good old-fashioned AI* (GOFAI) techniques of past decades has the benefit that it removes human bias from the system, allowing deep learning algorithms to discover their own data representations and decision-making procedures, yielding higher levels of performance compared to hand-engineered approaches.

The salient feature that delivers AI's greatest strength- its ability to process an enormous amount of data - is also a drawback: it requires an overwhelmingly large amount of data to be effective. Such data may not be available to the "common" developer (or user). In fact, lack of access to data is just one barrier of entry to enjoy the benefits of deep learning: an extensive and often time-consuming and expensive education is another requirement and therefore limits deep learning to highly educated and specialized PhDs with years of important education and training. These PhDs are great, but only represent the tip of the iceberg of potential developers that can contribute to and benefit from deep learning - not everyone can get into Carnegie Mellon. Finally, the computational resources to develop and use deep learning tend to be limited to the Google's, Facebook's, and perhaps some Universities of the world and yet many can contribute, if resources or low-overhead deep learning approaches were available.

An increasing portion of AI breakthroughs are being made using resources far outside the budget of the typical academic lab. Freelance and small companies also offer us opportunities that large companies and universities cannot, such as niche applications of AI to problems that might not be appealing to large companies; we do not want to lose them. Finally, improving access to data and AI tools also has the potential to reduce harmful bias in our AI technologies. If datasets are free and open, they can be inspected and are open to criticism by experts in fairness and ethics in AI. America may be at the lead of AI research, development and use, but frankly we are still doing it with one arm tied behind our back.

AI technology is at a level where the Internet was 40 years ago. Without DARPA's public investment, we would not have commercial success today. However, times are different. AI runs the risk of becoming an oligarchy run by a few companies, and perhaps even worse, by an even fewer number of foreign actors who may monopolize its true potential. We need investment and participation from all sectors - private, public, and academic - to develop, innovate, and utilize AI technology. Hence, we need to form a public-private partnership (PPP) to direct research for our National agendas; this partnership, however, must also include support and direction for both STEM K-12 education and workforce development. The AI PPP (AI Triple P) ensures that the US maintains its leadership role, and that we, as a society, will reap the full benefits of AI.

D1: Data availability

Deep learning-based approaches require a large amount of data to be effective. As can be expected, data availability plays a crucial role in the performance of our ML-based technology. High-quality datasets are necessary for the advances made in research settings to percolate into applied technologies, because availability of quality data plays a large role in determining the efficacy of a learned model in the real world. While it is a strength to process and ``learn from'' a large amount of data, often high quantities of data is required for development and training of AI approaches. Typical datasets used to train deep neural networks used for supervised vision models contain hundreds of thousands to millions of labeled datapoints (e.g., ImageNet, CIFAR10, CIFAR100). For example, state of the art natural language processing (NLP) models, such as GPT3, have been trained on hundreds of billions of words. State of the art reinforcement learning systems trained for two-player game-play, such as AlphaZero and AlphaStar, were trained on games. For typical academic labs, these large data requirements limit the applicability of deep learning to situations in which large datasets are freely available.

Challenges limiting data availability include D1P1) high cost or high required investment, D1 P2) privacy concerns, D1P3) data is siloed and D1P4) difficulty in obtaining high-quality labels.

D1P1: High cost / investment to acquire data: Gathering large datasets for custom applications is often a high-cost endeavor. For example, one recent publication that used reinforcement learning to learn robotic grasping required 14 expensive robotic arms and over 800,000 grasp attempts to achieve 80-90% grasp success rates using a 2-finger gripper [CITE Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection]. For many academic labs, this would be a

prohibitively expensive undertaking. Finally, in order to get the most out of our tax-payer dollars, we must share and document our publicly supported datasets.

D1P1.1: Data efficiency: To ameliorate the high costs associated with large dataset collection, we suggest that more funding be allocated to improving data efficiency in ML, through techniques such as transfer learning, semi-supervised learning, domain adaptation, and data augmentation. These approaches improve data efficiency on some target task by enabling data from another task, or unlabeled data, to contribute to the learning process.

D1P1.2 Leverage public funds: We can also leverage our existing investments by requiring that datasets generated by public funds should be available to the public: it was paid for by the public. Also, we believe that such a practice will inevitably promote the scientific process of validating results. In fact, we already see this practice taking place, as it is in the best interest of the scientist to promote their work and supporting data facilitates such promotion.

D1P1.3: Synthetic Data: Finally, we advocate for increased funding for research on synthetic data creation. Fake data is free but making it meaningful is hard; however, the ability to generate realistic synthetic data could at least allow AI researchers and practitioners to validate approaches and identify weaknesses before making costly investments in dataset collection. An additional potential benefit of synthetic data generation is that it avoids the privacy concerns typically associated with sensitive (e.g. medical) data.

D1P2: Privacy. Privacy permeates all issues that involve datasets. Obviously, privacy at the extreme inhibits the proliferation of datasets in order to protect the owners and subjects of the data. Failure to recognize this importance could be catastrophic. For one thing, we can compromise personal, organizational, and national security. Next, we could potentially lose the trust of those who contribute to the dataset. By no means do we, the authors of this document, claim to be privacy experts, nor understand the bounds of the implications of privacy. Therefore, we suggest that experts in privacy be included in the ideation of data availability and defer to them for specific suggestions. With such experts, we advise that approaches be developed to either develop surrogate data sets and other methods be created to strip private information from datasets and yet retain their salient properties.

D1P3: Data Silos. Improving data availability has the potential to both improve AI both as a research field and a technology. Data is the raw ore from which useful models are smelt, and machine learning is a fundamentally empirical science; hypotheses must be

tested on *real-world data* that are truly reflective of the situations in which they are intended to be used. For this reason, in many fields such as computer vision and natural language processing, large, representative, and high-quality datasets are absolutely crucial for fundamental advancement. It is also important to know when to take additional data to improve the predictive power of our AI system, as there are circumstances where the cost of acquiring data does not justify the improvement in predictive capabilities. We strongly advocate an increase in funding for transfer learning and imitation learning, but require disparate problem domains for which this research would be funded. We also advocate for the study of how data in one form correlates (perhaps unexpectedly) with data in another.

Additionally, we suggest that measures be taken to encourage inter-agency sharing of data, when possible. It stands to reason that different agencies may have some common denominators in the datasets they collect. It would be meaningful to understand the commonalities, to see what shared problems they're all solving, as well as the differences to see how we can round off each others' limited datasets. Moreover, agency x can stress-test its approaches using agency y's datasets. The problem is that, from the authors' distant perspective, agencies often have a hard time cooperating and sharing at deep levels. We suggest that the White House look at examples of where inter-agency cooperation has been successful, and one such example is the National Robotics Initiative, based out of the NSF.

D1P4: Labels

Labels typically refer to some form of identification or annotation placed on data by people. Obtaining high-quality labels can in many cases be the most expensive aspect of data collection. Often, gathering unlabeled data is cheap because it requires little human oversight (e.g., downloading text from wikipedia or images from Google images). In some applications, such as labeling of medical data or data from particle accelerator experiments, data must be labeled by domain experts, who’s time is very valuable.

D1P4.1: To ameliorate the difficulties associated with labeling large datasets, we advocate for increased funding in dataset generation. Generation of high-quality datasets with high-quality labels is not a flashy job, but it often spurs advances in the field (e.g., the ImageNet dataset, which was an expensive undertaking, but since its inception has served as an invaluable tool for the computer vision community). Of course, we must acknowledge that incorrectly labeled datasets can have a detrimental effect.

D1P4.2: We additionally suggest increased funding for machine learning approaches that make more efficient use of human experts, for example *active learning*. Active

learning is a form of machine learning that allows the ML system to query an expert or other knowledge source (e.g., a person, or a simulator) for labels during the learning process. Typically, active learning algorithms are designed so as to query the expert for the highly useful information, thereby reducing the number of labels that must be provided by the expert.

D2: Developer Access

Developer access relates to the resources and capabilities that people who develop AI technologies must possess in order to develop, and in many cases advance the state of the art, in AI and deep learning. One resource, as described above, is data. However, other resources are needed: computers, software tools, developer communities, etc. Just like data, a tremendous problem faced by deep learning developers is the quantity of computational resources and other developer access tools required. Most academic labs cannot compete with the massive GPU (and now TPU) clusters used by the likes of OpenAI and Google. As a result, high-powered private industrial companies such as Facebook and Google, would be the only ones who could enjoy the benefits of developer tools to advance the state of the art. This means that the most powerful AI algorithms are controlled by a few large companies. We should seek a policy of supporting research and education in empowering people outside these centers of machine learning excellence to create novel AI tools.

D2P1: Computing resources. The primary obstacle to developer accessibility is cost. The hardware cost for a single AlphaGo Zero system in 2017, including the four TPUs, has been quoted as around $25 million (according to wikipedia). Certainly, a tier 1 University lab, let alone a small company or citizen-scientist, cannot afford such computational resources. The trend towards ever larger models that yield better performance on popular benchmarks while requiring more computational power to train makes it increasingly difficult for labs with modest resources to compete with state-of-the-art (SOTA) performance on ML benchmarks.

D2P1.1: Shared Resources. Therefore, we suggest, just as the physicists can band together to raise funds for a common platform, such as a telescope, so should the academic AI researchers form a similar consortium for a shared resource. This could follow the already existing model of the Super Computer Centers, but some careful consideration must be given to the special needs of AI researchers and perhaps the more broad user community of such a resource.

D2P1.2: Efficient tool development. To better enable labs with smaller budgets and modest compute resources to compete with well-funded companies, we recommend that the NSF fund research in *computationally efficient* machine learning approaches, and *low-cost computing hardware (perhaps including robotics).* Investing more in computationally efficient ML approaches would have several benefits: firstly, it would provide more avenues of possible funding for labs that are capable of contributing, but cannot match the SOTA performance on benchmarks simply due to computational limitations. Secondly, the development of computationally efficient ML approaches would allow academic labs with modest budgets *to be competitive* with SOTA performance. Finally, efficient ML algorithms have the potential to lower the carbon footprint of ML research.

D2P2: Another challenge limiting developer accessibility is K-12 education. Opportunities for K-12 students to engage with computer science are not evenly distributed, putting segments of the population at a disadvantage when entering college. We therefore advocate for the expansion of computer science education in K-12. Computer science is unique in that compute resources, and even IT support for students can be easily shared by multiple schools.

The barrier to entry for becoming an AI developer for the community at large is unnecessarily high. Even state-of-the-art advances in machine learning can be broken down into a few basic steps: define the model, train the model, validate the model. While programming libraries exist (e.g., Keras) that massively streamline the machine learning development process, even just installing and using these tools requires a high degree of programming expertise and understanding of computer infrastructure, for example, proficiency in python and linux.

To lower the barrier to entry for potential AI developers, both in K-12 and in the community at large, we advocate for the creation and development of web-based tools, accessible to anyone with an internet connection, that allow machine learning workflows such as data set handling, model definition, and training, to be represented through a simple and easy-to-use graphical interface. Any program created in this interface could then be converted to (e.g. python) code for the purposes of further development or sharing with the AI community.

D3: Democratization of AI Tools and Data.

Beyond data and developers, AI tools are often out of reach of most people who want to use AI tools to solve problems for their own businesses, or just personal research and education. For the United States to reach its full potential in using advanced computing to compete and collaborate with our peers in Europe and Asia, we must get the AI solutions into the hands of everyone. We believe that in doing so, everyone has an opportunity to voice how AI tools are used - in other words, we must democratize the use of AI tools.

As stated already, having solutions in the hands of a few large companies will limit our ability to solve complex problems. We are quite fortunate to have Tensorflow and Pytorch, created (for public use) by Google and Facebook, but as AI tools increasingly shape our lives, it is increasingly important that the power of tech giants does not go unchecked. Having the citizen-AI-scientist using AI tools to solve similar problems may actually serve as a check and balance to large companies, whose initial goals were to generate profit, who may misuse or abuse their capabilities.

One major challenge toward democratizing AI tools is the fact that large tech companies (such as Apple, Google, and Facebook) control much of the data generating pipelines, because much of the data these companies run on is generated by users using their products. While these tech giants offer a tremendous benefit to our economy and society, we cannot allow them to monopolize the AI market in perpetuity. We are inhibiting our growth if we sustain long-term difficulties for small companies or non-profit open-source ventures to break into the market. This challenge overlaps heavily with the issues discussed in data availability; however, here we are mainly focused on the assumption that companies own the data generated by their platforms, and how this limits democratization of AI tools.

To encourage competition in the AI market, as well as decouple data from data-generation platforms, we advocate for measures to be taken that allow users of AI technology to *own their own data.* Users could then choose to sell their data at free market price on a data market, thus lowering the barrier to entry of smaller companies and research groups. Such a data ownership model would also give users the ability to *vote with their data:* if users do not like the way a particular company uses their data, they can choose to withhold their data from that company. Changing the data ownership model gives users of AI technology a seat at the table, instead of simply allowing big tech companies to be the sole arbitrator of how to use data and AI tools in whatever way makes them the most profit. Finally, having a free data market will have the positive side-effect of encouraging citizens to use and develop AI tools.

Public-Private Partnership

Sustaining American leadership in AI is an all-hands-on-deck endeavor. AI clearly has its birthplace in American academia and academia continues to innovate. Private industry also innovates but has resources that academia does not have. The government has a social responsibility to ensure our citizens benefit from the public investments in academia and industry, as well as an obligation to steward American innovation and education for the betterment of society. We therefore must have a National vision, and that vision must come from the community whether it is academics/private industry/public,or developer and user, or educator and innovator.

Therefore, a public-private partnership is necessary to ensure all perspectives are integrated and a plan is properly executed. This will of course mean significant investment in AI research and development, as well as the education of K-12 students and workforce training. At the risk of creating additional bureaucracy, funding for AI work for the Federal government can be funneled through a single task force, or we could create an advisory board that tracks and recommends all of the AI research agendas for different parts of the government.

However, such an entity has to be public/private to ensure engagement, buy-in and grounded research is produced. Co-creating AI solutions with the “producers” of AI involving the “consumers” of AI early on in the design and development process will increase both the usability and trust in AI, which naturally will result in wider use.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Gajos, Harvard University

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.

HARVARD School of Engineering and Applied Sciences

150 Western Ave,
Allston, MA 02134
USA

KRZYSZTOF Z. GAJOS
Gordon McKay Professor of Computer Science

March 4, 2022

To:
AI R&D RFI Response Team,

Re: RFI Response: National Artificial Intelligence Research and Development Strategic Plan

To the AI R&D RFI Response Team:

On behalf of the Intelligent Interactive Systems Group at Harvard School of Engineering and Applied Sciences, we thank the Office of Science and Technology Policy for the opportunity to provide input on the *National Artificial Intelligence Research and Development Strategic Plan*.

Since 2009, our group has been investigating the principles and applications of intelligent interactive systems, contributing computational and behavioral research addressing the following question: How do we design, build and evaluate AI-powered systems that support desired real-world outcomes and that produce predictable and reliable experiences for the users despite the fact that the underlying technology is inherently proactive, unpredictable, and occasionally wrong? We are primarily computer scientists, but we draw on insights from a range of other disciplines, including cognitive sciences.

Summary of concerns

In short, we observe that **many assumptions about human-AI interaction that inform the current Strategic Plan and that guide the contemporary computational research and development of AI-powered technologies are wrong or at least unverified**. For example, contrary to the common assumptions, evaluations of human-AI collaborations on actual decision-making tasks demonstrate that explanations do not substantially improve decision quality. People do not engage cognitively with the content of AI-generated explanations but, instead, process them as general signals of system competence. This, in turn, leads to overreliance on the AI systems. Similarly, there is growing evidence that requiring human oversight of the AI algorithms is probably insufficient to ensure the safety of decisions made by AI-powered systems. One consequence of building on unverified or wrong assumptions is that resources meant to support the development of novel computational solutions are misdirected toward solving wrong problems. Another is that ineffective or harmful solutions are being developed for practical deployment.

Second, we notice that **the current implementation of the Strategic Plan deprioritizes investments in research on human-AI interaction**. Specifically, we note that human-AI interaction is currently a part of the Computing-Enabled Human Interaction, Communications, and Augmentation (CHuman) Program Component Area (PCA). However, unlike other PCAs, the CHuman PCA does "not correspond directly to [any] individual coordinating [Interagency Working Group] IWG" (page 7, Networking and Information Technology Research and Development Program Review, January 2021). Consequently, the considerations related to human-AI interaction are tacked onto other programs without ever becoming a central priority. There is no entity championing or accountable for the goals of the CHuman PCA.

Recommendations

We recommend that the following minimum revisions be made to the current Strategic Plan and its implementation:

- Strategic Plan revisions:
    - Extend *Strategy 1: Make Long-Term Investments in AI Research* to include "Developing fundamental theory and applied principles of human-AI interaction"
    - Extend *Strategy 2: Develop Effective Methods for Human-AI Collaboration* to include fundamental research on Human-AI Collaboration. Specifically, add a subsection on "Seeking new paradigms for human-AI interaction".
    - In *Strategy 4: Ensure the Safety and Security of AI Systems*, in addition to "Improving explainability and transparency" and "Building trust", include fundamental research into developing novel human-AI interaction paradigms that would support safe and secure operation of AI-powered systems by people.
    - In *Strategy 4*, subsection "Enhancing verification and validation," recognize that AI-powered systems need to be validated not just on their own (i.e., not in terms of the outputs that they produce) but also in the sociotechnical contexts in which they are meant to operate (e.g., evaluate not just the accuracy of treatment selections predicted by an AI algorithm but also the quality of decisions made by doctors supported by the system).
    - Extend *Strategy 5: Develop Shared Public Datasets and Environments for AI Training and Testing* to also include standard procedures and outcome measures for testing the efficacy of human-AI collaborations.
- The implementation of the Strategic Plan:
    - Create an Interagency Working Group (IWG) to champion and be accountable for the CHuman PCA.

Elaboration

To further motivate our recommendations, we will focus on AI-supported decision-making (e.g., an AI-powered decision support system helping a clinician make a treatment selection decision), but our conclusions apply more broadly.

When the work on AI-supported decision making started, it was presumed that human-AI collaborations would produce better decisions than either people or AI systems alone [9]. This assumption was well grounded in earlier work on ensemble methods in the field of AI, recognizing that combining multiple complementary sources of expertise typically results in improved outcomes. However, recent studies that examined human-AI teams engaged in actual decision-making tasks have consistently demonstrated that such teams do not perform better than AI systems alone [3,8] – somehow the expected synergies are not occurring. Adding explanations to decision recommendations does not generally help and might even hurt the decision quality [1,8]. There is emerging evidence that people do not engage cognitively with AI-generated recommendations and explanations and, instead, process them as general indicators of system competence [1,3,4]. The result is human overreliance on the AI systems. Such overreliance can be reduced through timely cognitive interventions (e.g., asking a person to articulate their own decisions before viewing the AI recommendation and explanation) [3]. However, our recent work suggests that even such interventions do not result in deep processing of the AI-generated information [4]. Instead, more radical redesign of human-AI interaction may be needed, such as only providing people with supporting information but leaving them to engage in even minimal cognitive processing of that information before they arrive at a decision [4]. Of course, such redesigns of human-AI interaction would require more cognitive effort on the part of human operators and thus may be resisted. Thus extensive research and extraordinary care are needed to design novel human-AI interaction paradigms that result in high quality decisions and acceptance by people.

The above insights have not informed much of the computational work on AI-powered systems yet. Part of the reason for it is that empirical evaluations with people have been rare. Another reason is that most of the evaluations that have been conducted, used *proxy tasks* instead of actual decision-making tasks. When a proxy task is used, a person is presented with an AI-generated explanation or model description and is asked to predict what the model would recommend in a particular setting. Such tasks artificially

focus people's attention on the AI-generated information and produce unrealistically optimistic results compared to what would be observed if study participants were asked to perform actual decision-making tasks [2].

Next, AI-powered decision support systems frequently only address a part of the decision problem or a wrong framing of the problem. For example, a decision support system for medical treatment selection most likely will focus on the medical efficacy of the treatments leaving out factors such as cost or patient preference and tolerance of side effects. It is presumed that the clinician using the decision support system will integrate the system-contributed insights with other aspects of the problem to recommend a holistic solution. However, research has demonstrated that decision support systems that focus on a subset of the problem can cause the human operators to overemphasize the part of the problem targeted by the system in their decision making [6] potentially leading to worse decisions on the complete problem. Thus, we need to understand how to design human-AI interaction to support the whole problem decision-making even when the AI can only help with only a part of the problem.

Continuing with the medical treatment selection example, most systems in this space are designed to support the clinician. Meanwhile, clinicians point out that modern medical practice strives for shared decision making, where patient and clinician make decisions together (particularly with respect to side effects and other elements where patient preference is important). Clinicians in one of our studies pointed out that clinician-focused decision support systems detract from shared decision making and this, in turn, reduces the motivation of the clinicians to use such systems [7]. Thus, to improve health outcomes and adoption by clinicians, we should be designing human-AI interaction to support patient-clinician collaboration rather than to inform the clinician.

Regarding human oversight of AI-generated decision recommendations, recent work by Dr. Ben Green from University of Michigan synthesized evidence showing that, in general, people are not able to provide such oversight [5]. We add that the requirement of human oversight of AI algorithms forces human operators to solve a seemingly impossible cognitive task: They interact with a machine that will occasionally make decision recommendations different from what they, the human operators, would make. Sometimes, these decision recommendations will be brilliantly correct, going beyond the knowledge and experience of the operator and reflecting the machine's superior ability to find novel and nuanced patterns in the data. Sometimes, these machine-generated decision recommendations will be catastrophically bad, driven by complex but spurious patterns in the data. It is probably impossible (even with AI-generated explanations) to tell the difference between extreme brilliance and extreme stupidity. To enable safe and efficacious use of AI-powered decision support systems, these systems have to come with some guarantees regarding the frequency and extent of the mistakes that they make. Because rigid theoretical guarantees can rarely be provided for such systems, we need novel human-centered empirical approaches—perhaps based on randomized controlled trials conducted in real application settings—for demonstrating the safety and efficacy of such systems before they are deployed.

These are just a few specific research results and examples to illustrate the urgent need to prominently include human-AI interaction in the revised Strategic Plan and in its implementation.

Acknowledgments

Sohini Upadhyay and Carlos Borges Torrealba Carpi contributed to the preparation of these comments.

Respectfully yours,

Krzysztof Gajos
Gordon McKay Professor of Computer Science
Intelligent Interactive Systems Group
Harvard School of Engineering and Applied Sciences
Allston, MA, 02134

References

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 81, 1–16. https://doi.org/10.1145/3411764.3445717

[2] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20). Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. Proc. ACM Hum.-Comput. Interact., 5(CSCW1), April 2021. http://www.eecs.harvard.edu/~kgajos/papers/2021/bucinca2021trust.shtml

[4] Krzysztof Z. Gajos and Lena Mamykina. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22, 2022. http://www.eecs.harvard.edu/~kgajos/papers/2022/gajos2022people.shtml

[5] Green, Ben, The Flaws of Policies Requiring Human Oversight of Government Algorithms (September 10, 2021). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3921216

[6] Ben Green and Yiling Chen. 2021. Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 418 (October 2021), 33 pages. https://doi.org/10.1145/3479562

[7] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2021. Association for Computing Machinery. http://www.eecs.harvard.edu/~kgajos/papers/2021/jacobs2021designing.shtml

[8] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy Jr, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry, 11, 2021. https://doi.org/10.1038/s41398-021-01224-x

[9] Kamar, E. (2016, July). Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In IJCAI (pp. 4070-4073). https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/hi.pdf

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Global Catastrophic Risk Institute (GCRI)

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.

Global Catastrophic Risk INSTITUTE

March 4, 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

Dr. Alondra Nelson, Deputy Director of Science and Society of the Office of Science and Technology Policy (OSTP) and Performing the Duties of OSTP Director, the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence (Select Committee), the NSTC Machine Learning and AI Subcommittee (MLAI-SC), the National AI Initiative Office (NAIIO), and the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO):

Thank you for the invitation to submit comments in response to the Request For Information (RFI) to the National Artificial Intelligence Research and Development Strategic Plan. We, the Global Catastrophic Risk Institute (GCRI), are researchers with expertise on AI ethics and AI governance. We offer the following submission for your consideration.

We support the eight strategic aims described in the 2019 Update. As detailed below, we encourage specific changes to seven of the aims that support a more robust and inclusive approach to AI ethics and governance. These changes seek to ensure that the National AI R&D Strategic Plan supports broad and sustainable success and meets high social and ethical standards.

Our recommendations are as follows:

Strategy 1: Make long-term investments in AI research.

We recommend an emphasis on interdisciplinary research in which technological progress is oriented according to social and ethical values and in which technology governance is informed by a sound understanding of the nature of AI technology.

Investment in basic research is essential for improving the capability and reliability of AI systems. However, there has been a tendency for some research on AI systems to be focused on improving capabilities without substantial consideration of social and ethical

dimensions.[1] This risks the development of AI technology that can have inappropriate impacts on society. For example, machine learning research has often pursued larger neural network models to improve model accuracy without regard for the adverse energy resource and climate change consequences of larger models.[2] To align research on AI systems with social and ethical values, it is essential to have these values built into the core of the research, including the selection of which research directions to pursue.[3]

The ability of society to successfully govern AI technology additionally depends on basic research. AI technology has only recently risen to prominence as a societal issue, and the study of AI governance is likewise at an early stage. Research on AI governance has often focused on general concepts such as ethical principles, with less regard for how to operationalize them.[4] Given rapid ongoing changes in AI technology, an important challenge is to ensure that governance concepts are informed by a state-of-the-art understanding of the technology.[5] To formulate practical and technologically sound AI governance concepts, it is essential to invest in AI governance research in which the AI technology is not treated as a black box but instead is considered in detail.

Strategy 2: Develop effective methods for human-AI collaboration.

We recommend an emphasis on methods to ensure that human-AI collaboration is in the common public good and not just in the interests of the select few with control of advanced AI tools.

AI technology has enabled major advances in workplace productivity, bringing major economic benefits. However, this often comes at the expense of disadvantaged populations. For example, low-wage workers are forced to work unpredictable schedules that are optimized according to AI processing of last-minute data[6] and social

[1] Seth D. Baum, "On the promotion of safe and socially beneficial artificial intelligence", *AI & Society*, vol. 32, no. 4 (2017), pp. 543-551, https://doi.org/10.1007/s00146-016-0677-0.

[2] Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick, "Aligning artificial intelligence with climate change mitigation", 2021, https://hal.archives-ouvertes.fr/hal-03368037.

[3] Steven Umbrello and Ibo Van de Poel, "Mapping value sensitive design onto AI for social good principles", *AI and Ethics*, vol. 1, no. 3 (2021), pp. 283-296, https://doi.org/10.1007/s43681-021-00038-3.

[4] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi, "Operationalising AI ethics: barriers, enablers and next steps", *AI & Society* (2021), https://doi.org/10.1007/s00146-021-01308-8.

[5] Wendell Wallach and Gary Marchant, "Toward the Agile and Comprehensive International Governance of AI and Robotics", *Proceedings of the IEEE*, vol. 107, no. 3 (2019), pp. 505-508, https://doi.org/10.1109/JPROC.2019.2899422.

[6] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).

media websites use AI recommender algorithms that increase user engagement by promoting politically extremist content.[7]

The 2016 National AI R&D Strategic Plan recognizes that "the challenge of understanding and designing human-AI ethics and value alignment into systems remains an open research area." We agree that this is an important open research area. However, it is important to address: to whom specifically are AI systems aligned?[8] AI systems that are designed to support a single individual require fundamentally different designs than those that are designed to support society as a whole.[9] Only by including the common good of society in AI system design can human-AI collaboration advance the interests of all members of society instead of the select few.

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

We recommend a holistic and pluralistic approach to the ethical, legal, and societal (ELS) implications of AI, in particular, to address the full range of implications and to evaluate them in terms of a diversity of social and ethical perspectives.

It is vital that the ELS implications of AI are central to all work on AI and not tacked on as an afterthought. There are ELS implications in, among other things, (1) the selection of algorithms, determining which ethical concepts can be implemented in an AI system[10]; (2) the selection of the scale at which to implement the algorithm, determining the energy consumption of the AI system[11]; (3) the selection of training data, determining which problems and issues an AI system can be applied to[12] and the

[7] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta, "Recommender systems and the amplification of extremist content", *Internet Policy Review*, vol. 10, no. 2 (2021), https://doi.org/10.14763/2021.2.1565.

[8] Seth D. Baum, "Social choice ethics in artificial intelligence", *AI & Society*, vol. 35, no. 1 (2020), pp. 165-176, https://doi.org/10.1007/s00146-017-0760-1.

[9] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar, "Towards Pluralistic Value Alignment: Aggregating Value Systems through ℓ_p-Regression", Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022, May 9–13), https://filippobistaffa.github.io/papers/2022aamas.pdf.

[10] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein, "Implementations in machine ethics: A survey", *ACM Computing Surveys*, vol. 53, no. 6 (2020), article 132, https://doi.org/10.1145/3419633.

[11] Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick, "Aligning artificial intelligence with climate change mitigation", 2021, https://hal.archives-ouvertes.fr/hal-03368037.

[12] For example, an AI language processing system trained in a dominant language such as English will not function in other languages.

potential for biases in AI system outputs[13]; and (4) the deployment of AI systems, determining the specific societal and environmental impacts[14]. The breadth of ELS implications underscores the importance of embedding ELS at all points on the AI system lifecycle.

Work on ELS should additionally welcome a range of perspectives and support constructive and open debate. Recent work on AI ethics has emphasized the development of consensus-driven sets of ethics principles or guidelines. However, these guidelines come overwhelmingly from North American and European organizations, many of which had little or no public participation.[15] Marginalized populations, such as Indigenous peoples, may have differing perspectives on ELS issues.[16] Inclusion of diverse perspectives can help to overcome apparent gaps in existing ELS work, such as pertaining to the moral status of nonhumans.[17] Furthermore, active AI ELS debates remain unresolved, such as on the relative importance of near-term and long-term dimensions of AI.[18] Given the lack of universal consensus on AI ELS issues, it is important to support an inclusive and open-minded conversation about the issues.

Strategy 4: Ensure the safety and security of AI systems.

We recommend an emphasis on a dynamic research and development program to ensure that AI systems remain safe and secure as the technology and its usage evolve over time.

An essential aspect of AI safety and security is that as AI systems become more capable and become used more widely, the safety and security challenges become

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A survey on bias and fairness in machine learning", *ACM Computing Surveys*, vol. 54, no. 6 (2021), article 115, https://doi.org/10.1145/3457607.

[14] For example, decisions of when to deploy autonomous weapons; Hendrik Huelss, "Deciding on Appropriate Use of Force: Human-machine Interaction in Weapons Systems and Emerging Norms", Global Policy vol. 10, no. 3 (2019), pp. 354-358, https://doi.org/10.1111/1758-5899.12692.

[15] Daniel Schiff, Jason Borenstein, Justin Biddle, and Kelly Laas, "AI ethics in the public, private, and NGO sectors: a review of a global document collection", *IEEE Transactions on Technology and Society*, vol. 2, no. 1 (2021), pp. 31-42, https://doi.org/10.1109/TTS.2021.3052127.

[16] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung et al., "Indigenous protocol and artificial intelligence position paper", Indigenous Protocol and Artificial Intelligence Working Group and the Canadian Institute for Advanced Research, Honolulu, HI (2020), https://spectrum.library.concordia.ca/id/eprint/986506.

[17] Andrea Owe and Seth D. Baum, "Moral consideration of nonhumans in the ethics of artificial intelligence", *AI and Ethics*, vol. 1, no. 4 (2021), pp. 517-528, https://doi.org/10.1007/s43681-021-00065-0.

[18] Charlotte Stix and Matthijs M. Maas, "Bridging the gap: the case for an 'Incompletely Theorized Agreement on AI policy", *AI and Ethics*, vol. 1, no. 3 (2021), pp. 261-271, https://doi.org/10.1007/s43681-020-00037-w.

more important. Prior to the deep learning revolution, AI technology had limited usage[19] and likewise little need for safety and security. By now, AI technology is used widely across economic sectors and other areas of human society. Barring a new AI winter, i.e. another years-long drop in AI research if AI progress fails to live up to expectations, AI technology will only grow in its importance. As it does, the potential for AI systems to cause harm is likely to increase. In risk terms, harm from AI systems could become more frequent, due to their wider usage, and more severe, due to their usage in more high-stakes applications. Prospects for catastrophic harm may further increase via the use of AI in crucial sectors such as agriculture[20] and via increasingly general-purpose AI systems that are becoming widely used across sectors.[21]

Given the growing stakes, it is vital for AI safety and security to keep up with the changing technology. This is not a trivial challenge. Some aspects of AI safety and security may remain viable even as the technology becomes more advanced,[22] whereas other aspects may need to be customized for more advanced systems.[23] Research and development on AI safety and security must be forward-looking in order to keep society safe in the face of more capable and more widely deployed AI systems. Doing so will further be of economic and strategic benefit because safer and more secure AI technology would permit the technology to be deployed more widely, especially in more sensitive settings.

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

We have no comments on Strategy 5.

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

[19] Terrence J. Sejnowski, *The Deep Learning Revolution* (MIT Press, 2018).
[20] Victor, Galaz, Miguel A. Centeno, Peter W. Callahan, Amar Causevic, Thayer Patterson, Irina Brass, Seth Baum et al., "Artificial intelligence, systemic risks, and sustainability", *Technology in Society*, vol. 67 (2021), article 101741, https://doi.org/10.1016/j.techsoc.2021.101741.
[21] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al., "On the opportunities and risks of foundation models", Stanford Institute for Human-Centered Artificial Intelligence (2021), https://arxiv.org/abs/2108.07258.
[22] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete problems in AI safety", (2016), https://arxiv.org/abs/1606.06565.
[23] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019).

We recommend support for programs that can facilitate the ongoing development and adoption of standards across the AI industry, including via the development of regimes for the certification of AI systems.

A variety of AI standards and frameworks are in the process of being developed, including the National Institute of Standards and Technology (NIST) AI Risk Management Framework[24] and the Institute of Electrical and Electronics Engineers (IEEE) Standards Association P2863 Recommended Practice for Organizational Governance of Artificial Intelligence.[25] These standards and frameworks will provide voluntary guidance for developers and deployers of AI systems. As these and other frameworks are completed, two challenges will be faced. One is to ensure that the standards remain relevant and appropriate in the face of ongoing changes in AI technology and its various applications. The other is to facilitate the adoption of the standards by AI developers and deployers.

One valuable approach to facilitating adoption of standards is via certification regimes.[26] Certification regimes serve to address information asymmetries between insiders within an organization and outsiders who wish to know if the organization is complying with relevant standards. Certification is already widely used in many sectors, such as in the US EnergyStar program for consumer appliances, the ISO 9001 program for supply chains, and the LEED program for building design. AI certification regimes have been developed or proposed by, among others, the European Commission,[27] IEEE,[28] and the government of Malta.[29] When implemented effectively, certification regimes can incentivize adoption of standards. Certification regimes are also flexible in terms of being public or private (or both), voluntary or mandatory, and geographically local or global. These attributes make certification regimes an important element of AI standards adoption.

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

[24] https://www.nist.gov/itl/ai-risk-management-framework

[25] https://sagroups.ieee.org/2863

[26] Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum, "AI certification: Advancing ethical practice by reducing information asymmetries", *IEEE Transactions on Technology and Society*, vol. 2, no. 4 (2021), pp. 200-209, https://doi.org/10.1109/TTS.2021.3077595.

[27] European Commission, “White paper on artificial intelligence: A European approach to excellence and trust,” (2020), https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

[28] IEEE, “The ethics certification program for autonomous and intelligent systems (ECPAIS)”, https://standards.ieee.org/industry-connections/ecpais.html.

[29] Malta Digital Innovation Authority, “AI-ITA blueprint guidelines”, (2019), https://mdia.gov.mt/wp-content/uploads/2019/10/AI-ITA-Blueprint-Guidelines-03OCT19.pdf.

We recommend an emphasis on a multidisciplinary and multitalented workforce that is capable of developing and applying AI technology for the common good.

As AI technology grows in its ethical, legal, and societal (ELS) significance, it is vital for the computer scientists and engineers who design and deploy AI systems to be conversant in ELS topics. Historically, AI technology had little societal impact; AI computer scientists and engineers were likewise focused narrowly on how to increase the capabilities of AI systems with little regard for ELS implications.[30] The emergence of AI as a class of technology with significant ELS implications has created a need for cultivating an understanding of ELS among AI computer scientists and engineers. Some progress on this front has been made,[31] but this remains a major weakness of the AI workforce.

Concurrently, there is a need for ELS experts who are conversant in AI technology. Computer scientists and engineers play a vital role in AI technology design and in relating design details to AI governance. However, it is inappropriate to expect computer scientists and engineers to have a comparable depth of knowledge about ELS as people who are professionally trained in ELS fields such as moral philosophy, law, and the social sciences. In order for ELS experts to successfully contribute to AI governance, it is vital for them to have some knowledge (the more the better) of how AI technology works. This is not typically taught in ELS university programs. It is therefore important to support dedicated programs. The Technology, Management, and Policy Consortium[32] provides a valuable set of benchmarks for how such programs can be designed and executed.

Strategy 8: Expand public-private partnerships to accelerate advances in AI.

We recommend that public-private partnerships be pursued both domestically and internationally to ensure that the AI industry as a whole is oriented toward the common good.

In AI, as is the case in many industries, there is often tension between the private interest and the public good. For example, as noted above, companies running social media websites sometimes use AI recommender algorithms that increase user

[30] John Bohannon, "Fears of an AI pioneer", *Science*, vol. 349, no. 6245 (2015), p.252, https://doi.org/10.1126/science.349.6245.252.

[31] For example, the AAAI/ACM Conference on AI, Ethics, and Society, https://www.aies-conference.com.

[32] https://tmpconsortium.org

engagement by promoting politically extremist content[33] and AI developers sometimes build larger neural network models to improve model accuracy without regard for the adverse energy resource and climate change consequences of larger models.[34] Additionally, in the international context, competition between military adversaries can result in both sides pursuing unsafe AI technology.[35] These situations are known as collective action problems; they constitute an important class of challenges in AI governance.[36]

Public-private partnerships can play a valuable role in addressing AI collective action problems. Domestically, these partnerships can help support private firms as they orient their activities toward the common good. Internationally, the partnerships can facilitate the cooperation needed to solve AI collective action problems at the global scale. The US already participates in international public-private partnerships for mutual benefit with other countries.[37] Additionally, some international organizations already work to bring together public and private AI stakeholders; these include the Global Partnership on Artificial Intelligence (GPAI)[38] and the OECD Artificial Intelligence Policy Observatory.[39] The US should participate in these organizations as part of its program on public-private partnerships. The US should further seek to include diverse participants from the international community, including to support global justice in AI technology.[40] Additionally, where appropriate, including rival powers such as China in these forums would further facilitate the resolution of AI collective action problems. If successful, these partnerships could support a “race to the top” dynamic in which competition between companies and countries advances benefits for the national and global common good.[41]

---

[33] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta, "Recommender systems and the amplification of extremist content", *Internet Policy Review*, vol. 10, no. 2 (2021), https://doi.org/10.14763/2021.2.1565.

[34] Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick, "Aligning artificial intelligence with climate change mitigation", 2021, https://hal.archives-ouvertes.fr/hal-03368037.

[35] Richard Danzig, "Managing Loss of Control as Many Militaries Pursue Technological Superiority", Center for New American Security (2018), https://www.cnas.org/publications/reports/technology-roulette.

[36] Robert de Neufville and Seth D. Baum, "Collective action on artificial intelligence: A primer and review", *Technology in Society*, vol. 66 (2021), article 101649, https://doi.org/10.1016/j.techsoc.2021.101649.

[37] For example, the US-Israel Binational Industrial Research and Development Foundation (BIRD), established in 1977, aims to provide funding to joint projects of mutual benefit to both the US and Israel; see https://www.birdf.com.

[38] https://gpai.ai

[39] https://oecd.ai/en

[40] Eugenio V. Garcia, "The International Governance of AI: Where is the Global South?" (2021), https://www.researchgate.net/publication/348848134.

[41] Will Hunt, "The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry", Center for Long-Term Cybersecurity, University of California, Berkeley (2020), https://cltc.berkeley.edu/2020/08/11/new-report-the-flight-to-safety-critical-ai-lessons-in-ai-safety-from-the-aviation-industry.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Google

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.

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

March 4, 2022

As a leader in artificial intelligence research, Google welcomes the opportunity to provide comments in response to the Office of Science and Technology Policy (OSTP) Request for Information (RFI) on developing and updating the National Artificial Intelligence Research and Development Strategic Plan.[1] Google commends OSTP for its ongoing leadership in this area.

We encourage OSTP to build upon the foundation laid out in the 2019 AI Strategic Plan, which identifies eight strategies that remain the top priorities for guiding U.S. government suppot of AI research and development. As discussed in more detail below, we recommend that the updated Strategic Plan call on the U.S. government to take the following actions:

- Strategy 1. Strengthen partnership with the private sector on long-term investments in AI R&D and explore ways in which the updated Strategic Plan can further public-private research collaborations, as well as how such collaborations can contribute to subsequent strategic plans.
- Strategy 2. Continue to suppot research of AI's ethical, legal, and societal implications and call for research on those implications (e.g., civil rights and equity) in specific use cases (under Strategies 2 and 3).
- Strategy 3. Consider how governmental license agreements for open source models or datasets can be used to drive responsible AI adoption.
- Strategy 4. Continue investing in research to advance AI explainability and convene ML and AI researchers, practitioners, and other stakeholders to discuss these important questions.
- Strategy 5. Build on progress in making open government data available and more useful for AI (with appropriate privacy protections), including by making data from a range of government sources available in knowledge graphs, and reducing

[1] 87 FR 5876 (Feb. 2, 2022).

licensing restrictions on use of government data and data produced using government funding.

Strategy 6. Continue to support the development of AI standards and benchmarks, including those that incentivize use of privacy-enhancing technologies, along with testbeds to validate emerging technologies.

Strategy 7. Conduct and publish studies on the state of the national AI R&D workforce and the role of U.S. immigration policy on that workforce, and build on existing investments to broaden workforce participation by groups traditionally underrepresented in computing and related fields.

Strategy 8. Address the obstacles impeding research collaboration between the U.S. government and industry, and consider establishing a standalone strategy for increasing technical talent in government.

Strategy 1: Sustaining long-term investments in fundamental AI research

Google thanks the U.S. government for its continued prioritization of investment in long-term, fundamental research to advance AI and ML capabilities, including in the areas outlined in the 2019 Strategic Plan,[2] and we agree this investment is critical to the long-term success of the field. At the same time, companies including Google are also investing significantly in some of the fundamental ML and AI research areas that are highlighted in the 2019 strategy update, and we publish our work in these areas – for example, multimodal ML, perception, general AI, distributed and sparse models, robotics, AI hardware development, and application of ML to hardware development. In addition, in 2020, we provided more than $50M in external research funding to more than 350 academic researchers. We also support PhD Fellowships[3] and provide an AI Residency.[4]

We recommend that the Strategic Plan be updated to recognize that the private sector is an important partner when it comes to long-term commitments for investing in fundamental ML and AI research, and to consider the ways in which the U.S. government and industry can partner more closely to further joint objectives. It could be highly beneficial, for example, for the government to examine both public and private investment in key long-term research areas and to explore research topics where U.S. agencies should encourage more collaboration between academic,

[2] 2019 Strategic Plan, pp. 7-13.
[3] https://research.google/outreach/phd-fellowship/.
[4] https://research.google/careers/ai-residency/.

government, and industry researchers, as well as other stakeholders. This analysis could also identify the long-term research areas where the private sector is *not* publishing results so that government funding can be targeted to those areas.

The Strategic Plan could also specifically identify areas in which interdisciplinary collaboration on long-term research is essential. We have heard from our partners that government funding for interdisciplinary research can be difficult to obtain; federal funding review processes (and publishing systems) often focus on single disciplines. But as we move toward more implementation of AI in areas like health, agriculture, and economics, interdisciplinary collaboration will be vital. In addition to enhancing U.S. leadership in AI, these collaborations might also enhance our understanding in other domains: consider the possibility that neuroscience researchers collaborating with ML and AI expets might be able to identify fundamental principles of human understanding.

Strategy 2: Develop effective methods for human-AI collaboration

Ineffective methods for human-AI collaboration present practical and other barriers to AI adoption. But the need for effective human-AI collaboration extends far beyond the practical limitations and raises important questions about the transformational impact that AI might have on existing societal structures and policy frameworks.

OSTP should consider more directly connecting Strategy 2 and Strategy 3. As one approach, OSTP could specifically identify areas in which effective human-AI collaboration might raise ethical, legal, and societal implications, and it could support research to understand those implications. For example, purely automated tools that replace human labor entirely can raise significant concerns regarding civil rights and inequality. Human-AI collaboration also raises important questions about societal trust towards AI technology and decision making, as well as the safety of the tools as they are embedded within our environment.

Strategy 2 as described in the 2019 Strategic Plan highlights that there are certain areas for which human interaction will not be as relevant, such as deep sea exploration. But human involvement may be relevant (and useful) in even these cases, such as through remote interaction or management where potential physical dangers exist. The desirability of human intervention in even these scenarios is a potential area for further research, and we encourage OSTP to consider updating Strategy 2 accordingly.

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

Google agrees that long-term commitments from governments, industry, and academia are needed to understand the ethical, legal, and societal implications of AI and to develop AI systems that align with ethical, legal, and social principles. These effots must also remain flexible and adaptable as new innovations are discovered.

We have dedicated significant resources to understanding the implications of our technologies, and we have taken steps to enhance trust and mitigate AI risks. These actions are guided by our AI Principles, which include "Be socially beneficial," "Avoid creating or reinforcing unfair bias," "Be accountable to people," and "Be made available for uses that accord with these principles." Google assesses proposals for new AI research and applications for alignment with its AI Principles through a dedicated review process.

Ethical, legal, and societal implications of an AI technology cannot be completely understood in a vacuum: the impacts of a technology depend on the context. We therefore suggest that the updated Strategic Plan consider directing the research called for in Strategy 3 to specific applications of AI, such as the human-AI collaborations identified in Strategy 2. We also encourage funding of research to increase transparency at the AI model and system levels, so that we can better understand how systems work, their impacts, and appropriate mitigations.

OSTP may also wish to consider how governmental license agreements for open source models or datasets can be used to drive responsible adoption of AI tools across the U.S. ecosystem. These license agreements can enhance trust by, for example, providing recourse against deliberate misuse of open tools and data while still allowing for democratization. At the same time, as we explain below, license agreements should not become overly restrictive, or they otherwise risk jeopardizing the wide-ranging benefits of open tools and datasets.

Strategy 4: Ensure the safety and security of AI systems

Significant progress has been made in transparency and explainability research in recent years. Google Research has published many recent papers, including on Testing

with Concept Activation Vectors (TCAV) (2017), Estimating Training Data Influence (2020), Concept Bottleneck Models (2020), and Disentangled Simultaneous Explanations via Concept Traversals (DISSECT) (2021). While these advance our knowledge of how to explain an AI (from the model's algorithmic perspective), it is equally critical to identify what humans need to know from an AI (from the user's own perspective, sociocultural context, and worldview). For example, there are still a number of open questions surrounding topics such as:

- What kind of technical documentation is required for an AI system?
- What types of information about an AI system do users need to know, so that the AI is more likely to produce positive human outcomes (e.g. human productivity, effort, control, enjoyment)? How is that information best communicated, and what information is relevant for which types of users? (Note there has been recent progress on this topic of AI Onboarding: e.g. this 2019 paper on the onboarding needs of medical practitioners using AI, and this 2021 study on developing onboarding materials for AI users).
- How can the necessary information be provided while also protecting the security, privacy, and intellectual property of the system?
- How can the results of an AI system, and any related uncertainty or risks, be effectively communicated?

AI systems are rapidly increasing in size, acquiring new capabilities, and being deployed in high-stakes settings, which increases the importance of appropriately securing AI systems from malicious actors and unintended accidents. Further research is also needed to identify and reduce the risk of hazards from AI systems (including through making systems more robust). Google researchers worked with industry and academic partners on a roadmap for AI safety research in late 2021 in a publication titled Unsolved problems in ML safety. This research outlined four immediate problem areas—withstanding hazards (robustness), identifying hazards (monitoring), steering ML systems (alignment) and reducing deployment hazards (external safety)— and concrete research directions for each.

We recommend that the updated Strategic Plan call on the U.S. government to continue investing in research to explore what constitutes appropriate levels of explainability in specific contexts. In addition, we encourage OSTP to call on the U.S. government to convene ML and AI researchers, practitioners, and other stakeholders to share their progress on these questions to further accelerate the development of new knowledge.

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

This remains a critical component of the U.S. AI R&D strategy. As the 2019 Strategic Plan outlines, it is important to make data available, to clean and structure the data, and to ensure it adheres to FAIR principles.[5] The U.S. government has made great progress in both making data more open and making it usable for ML and AI research, and we welcome focused programs in this area such as NASA's Transform to Open Science initiative. We are also pleased that Congress established and OSTP and NSF are leading the National AI Research Resource Task Force, responsible for drafting a road map to democratize and expand access to critical resources and educational tools relating to AI, and hope that the NAIRR will be established and further accelerate access to data (from both public and private sources), as well as AI training and testing environments.[6] We strongly support establishing a NAIRR, and share the government's goal to make AI access more equitable through this resource. The NAIRR is a great opportunity to support increased access for a diverse range of researchers to critical AI and cloud resources such as storage, compute, databases, networking, data analytics, AI services and collaboration tools.

We encourage the updated strategy to build on this strong foundation. Below are a few examples of current challenges faced by ML and AI researchers and recommended actions that the U.S. government can take.

Researchers in AI and other domains often need data that exists from multiple sources, and sometimes in various formats. Compiling such data is time- and resource-intensive and prone to errors, which can limit the number of researchers capable of conducting certain analytics and might impact the resulting analyses. A way in which these challenges can be addressed is through publicly available knowledge graphs, which connect data from varied sources and make the resulting data and connections available in one location for modeling and analysis. The National Science Foundation has a Convergence Accelerator program focused on promoting knowledge in the U.S. government,[7] and the National Information Exchange Model is

[5] 2019 Strategic Plan p. 28; https://www.go-fair.org/fair-principles/.

[6] https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.

[7] https://www.nsf.gov/od/oia/convergence-accelerator/Award%20Listings/track-a.jsp.

also working on this approach.[8] Data Commons, a knowledge graph built by Google, is an example of a knowledge graph that makes public data available via one API, and this data is also available at datacommons.org.

Knowledge graphs are a key component of reducing the resources required to access datasets needed for research, which ultimately enhances the abilities of government, industry, and academic researchers to investigate important questions. We encourage the U.S. government to prioritize development of knowledge graphs that incorporate public data from all U.S. agencies, as well as publicly available private sector data. For example, Google makes over 100 datasets available for use by AI/ML researchers. We welcome opportunities to partner with the U.S. government in these efforts.

We also encourage the U.S. government to remove, to the greatest extent possible, restrictions on the use of public data, subject to reasonable and appropriate limitations necessary to safeguard the privacy and security of sensitive data. For example, the government could make its public data openly available under copyright, patent, trademark, and trade secret laws with clear references to public domain licenses like Creative Commons CC0. In addition, publicly funded researchers should be required to share data openly in a manner that is easy to access and preserves privacy and security. Where public data is already available online and is free from sensitive personal information, the government should also eliminate unnecessary hurdles to using that data such as by making data available in a machine-readable format, not requiring registration,setting up an account, or providing personal data, or otherwise requiring administrative steps. Provided it does not raise privacy concerns, public data should be made available as soon as possible from its source and with the highest possible level of granularity. Where privacy and security restrictions are needed, we encourage the U.S. government to adopt consistent limitations across agencies and datasets.

It is also critical that datasets used for AI (and other types of data science) are as diverse and representative as possible. OSTP should expressly call on the U.S. government to take action to make public datasets available for research and AI training as representative as possible of the challenge at hand.

Ideally, researchers should have easy access (e.g. through one knowledge graph) to reliable, clean, and FAIR data from around the world. We encourage OSTP to consider

[8] https://www.niem.gov/.

updating the Strategic Plan to call on the U.S. government to look for opportunities in international fora to make data from countries around the world more available and accessible.

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

Accepted, rigorous, and auditable standards and benchmarks are an important component of trustworthy and responsible AI. We commend the National Institute of Standards and Technology (NIST) for driving the development of AI standards and for participating in standards-making proceedings in international fora such as the International Organization for Standardization and the Institute of Electrical and Electronics Engineers. We recommend continued NIST participation and leadership to drive global AI standards harmonization, including as part of the EU-US Trade and Technology Council and the Global Partnership on AI.

We also applaud NIST's leadership in developing an AI risk management framework.[9] OSTP may wish to consider how standards and benchmarks may be used to support the framework, including how AI developers and deployers should identify and report performance against appropriate benchmarks for the application of AI in different domains or contexts. For example, it would be useful to have adversarial benchmarks, clear definitions and escalation plans for issue severity, and monitoring production systems to detect unreported concerns, among other items.

We also encourage OSTP in updating the Strategic Plan to support the development of standards that incentivize use of privacy-enhancing technologies such as differential privacy and federated learning and analytics to enable increased and shared value from AI while protecting individual privacy.

In addition, we support the call in the 2019 Strategic Plan for the government to establish testbeds to validate emerging technologies and commend NIST for its progress on this item. We also see a need for additional testbeds to help compare the effectiveness of new algorithms. In line with recommendations of the NAIRR Task Force, we suggest that the NAIRR could fill this gap. We recommend that the U.S. government consider supporting "living laboratory"-style testbeds to test deployment of AI models for specific applications, such as in medicine.

[9] https://www.nist.gov/itl/ai-risk-management-framework/.

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

The 2019 Strategic Plan highlighted the need for official AI workforce data and called for "additional studies on the current and future national workforce needs for AI R&D."[10] Google agrees with the call for these studies. It would be helpful for the government to regularly release information on its understanding of the state of the AI R&D workforce in the US and to share its upcoming research agenda on workforce needs.

We also recommend that the updated Strategic Plan expressly consider the effect of immigration policies on the current AI R&D workforce and on the U.S.'s long-term ability to attract more talent in the coming years. For example, the updated Strategic Plan could include plans to study these effects to inform potential changes to the U.S. immigration system, especially H-1B, H-4, and O-1 visas, to attract more talent or to clarify criteria and rules for those seeking to bring their skills to the U.S. Additionally, we note that the U.S. is currently experiencing a substantial backlog of highly skilled applicants for green cards and other work visas, and it would be helpful for the government to release more information about the number of people with skills in AI-relevant fields that are caught in the backlog.

Google also supports the 2019 Strategic Plan's call to broaden participation in the AI R&D workforce of groups traditionally underrepresented in computing and related fields.[11] The update should offer more clarity on specific investments that can help achieve that worthwhile goal. For example, the U.S. should invest in STEM education for K-12 students in low-income districts that typically have a low level of early exposure to STEM coursework. The updated Strategic Plan should also call for targeted research to understand the state of STEM education in community colleges and other underrepresented institutions and consider how they can expand access to STEM education. Through our Google Career Certificates initiative, Google has made access in STEM education a priority. The program offers certificates in IT Support and Data Analytics for free to all community colleges and career and technical education high schools in the U.S., with the goal of providing students with technical skills to go on to further education or directly into technical career pathways.

[10] 2019 Strategic Plan, p. 39.
[11] Id., p. 38.

The 2019 Strategic Plan correctly emphasizes the need for multidisciplinary expertise in AI research,[12] and we recommend that the updated Strategic Plan expressly call for greater investment in interdisciplinary higher education and professional programs that integrate AI approaches into their areas of study (e.g. social sciences, psychology, engineering). Additionally, the U.S. government can explore releasing greater educational resources on AI such as core curriculum for universities to adopt, advanced modules for those with relevant backgrounds, and more accessible resources for those with backgrounds in other fields.

Strategy 8: Expand Public-Private partnerships to accelerate advances in AI

The 2019 Strategic Plan correctly identifies public-private partnerships as a key to accelerating AI developments.[13] While progress has been made in this area, we also encourage the government to take steps to address remaining obstacles. For example, research collaborations can currently require months of work negotiating agreements and navigating procurement rules, which can impact the timeliness of the research and otherwise slow down the pace of innovation in public and private sectors. It would be useful for the government to explore mechanisms like agreement frameworks to enable more seamless research collaboration across agencies. Additionally, as mentioned under Strategy 1, the U.S. government may wish to consider establishing priority areas for public-private research collaboration.

The U.S. government would also benefit from additional AI talent in government to accelerate foundational and applied research within government. The updated Strategic Plan should call on the government to explore more investments in AI-oriented talent exchange programs with industry and academia, including fellowships and initiatives like the U.S. Digital Response. OSTP should also consider establishing "Increasing AI talent in government" as a standalone pillar under the updated Strategic Plan.

******

Google appreciates this opportunity to provide a response to the OSTP's request for information and looks forward to continued discussion to strengthen U.S. leadership on AI R&D.

[12] Id.
[13] Id., pp. 40-42.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Gursoy/Kakadiaris, Computational Biomedicine Lab, University of Houston

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.

RFI Response:
National Artificial Intelligence Research and Development Strategic Plan

Furkan Gursoy and Ioannis A. Kakadiaris*
Computational Biomedicine Lab, University of Houston, Houston, TX, USA

The increasingly decisive role of AI in people's lives necessitates a socio-technical viewpoint that encompasses everything from the conception of an AI system to the consequences of its use in the real world. Such a socio-technical viewpoint is concerned with people's interactions, complex relations, and broadly defined AI. The current version of the National Artificial Intelligence Research and Development Strategic Plan (the Plan) already addresses several socio-technical aspects. In this document, we present further recommendations to enhance the Plan towards achieving a trustworthy and safe AI that is welcome in society to progress the Nation towards a new techno-social paradigm.

I. Strategy 1 in the Plan describes fundamental AI research areas where further efforts are supported. Causality in AI is not included in the discussion. The topics around Causal AI are already receiving increasing attention from the machine learning community. However, it is still a domain with challenging questions and potentially significant benefits. Exploring causal relations in a system helps us understand the system and potentially improve the applications of AI and provides tools for Explainable AI and fairness, for instance, via counterfactual analysis.

Recommendation: The Plan should include Causal AI as an area that requires commitment for further long-term fundamental research. Future research potentially assists AI in advancing to the next stage in its capabilities, robustness, and explanatory power.
Relates to: Strategy 1.

II. Strategy 2 in the Plan addresses the human-AI collaboration. However, it primarily focuses on creating "AI systems that effectively complement and augment human capabilities." It acknowledges the challenges regarding human-aware AI, AI techniques for human augmentation, human-AI interfaces, and language processing systems. In general, these challenges are concerned with improving AI systems. However, improving human-AI collaboration depends on technical improvements regarding AI and its interfaces or mechanistic details of how humans collaborate with AI and requires an understanding and modification of how humans interact with the decisions or other outputs produced by AI systems. Human oversight of AI is an

*Corresponding Author
(Computational Biomedicine Lab, University of Houston, 4349 MLK Blvd, Rm 322, Houston, TX 77204-6022, USA)

area where further research is needed to understand how human decision-makers may influence or be influenced by AI decisions and to design appropriate and feasible monitoring and oversight mechanisms necessary to improve trust towards AI systems and minimize risks and harms.

Recommendation: The Plan should support research initiatives that tackle questions related to understanding and improving when and how humans can oversee and modify the decisions by AI systems such that the adoption of AI is relatively higher risk situations may be increased while avoiding unacceptable risks.
Relates to: Strategy 2.

III. Strategy 3 in the Plan lists and describes three key research challenges in AI's ethical, legal, and societal implications. These are (i) improving fairness, transparency, and accountability by design, (ii) building ethical AI, and (iii) designing architectures for ethical AI. However, as the way they are described in the Plan, these three challenges are largely overlapping without clear and intuitive distinctions. Also, the concept of explainability is discussed in Strategy 4 (which is concerned with the safety and security of AI systems). In contrast, we argue that it is more appropriate to discuss explainability within the scope of Strategy 3.

Recommendation: Strategy 3 should be rewritten to present notions and challenges concerning social implications and accountability of AI systems, which includes concepts such as responsibility, explainability, robustness, and fairness. It should also contain references to other related strategies such as Strategy 2 on human-AI collaboration, Strategy 4 on privacy and security of AI systems, and Strategy 6 on developing methods, metrics, benchmarks, and standards to evaluate AI systems.
Relates to: Strategy 3.

IV. Regardless of the efforts that are possibly spent to make AI systems safe, it is not typically attainable to ensure a given AI system is perfectly safe and free from risks. When due efforts are not provided, or unknown or undiscovered factors are in play, knowns risks increase, and unknown risks emerge. The trust to be placed in AI and its expanding role in society depends not only on the benefits of AI but also on its risks, potential harms, and remedies. To improve trust in future AI systems, on the one hand, the types and nature of unknown and typically undiscovered risks should be explored by future research. On the other hand, remedy mechanisms should be developed and put in place. Such efforts closely relate to risk ratings, certifications, and insurance for AI. Especially, given the unattainability of perfection for AI systems, insurance is a helpful and necessary development. However, for AI systems, evaluation of the probability and severity of risks and harms are not currently feasible, which provides an obstacle for AI insurance to emerge due to the uncertainties around pricing or settlements

Recommendation: The Plan should support research initiatives that tackle questions related to understanding and operationalizing the risks and harms of AI systems so that risk ratings, certifications, and insurance become feasible for AI systems.
Relates to: Strategies 3, 4, and 6.

V. The Plan addresses the increasing demand for AI researchers and practitioners. While it acknowledges that the AI workforce is not composed only of computer and information scientists and engineers but also includes multidisciplinary teams, it appears to present the other fields and domains as areas "in which AI may be applied." We suggest that multidisciplinary work where people from different disciplines work together is insufficient. Instead, an interdisciplinary and transdisciplinary approach that integrates knowledge from various disciplines and crosses disciplinary boundaries to employ a holistic perspective is needed. Accordingly, there is a growing need for social scientists with backgrounds in anthropology, economics, education, law, linguistics, political science, psychology, and sociology to conduct interdisciplinary and transdisciplinary research on the challenging problems at the crossroads of AI and social sciences.

Recommendation: Considering the emerging intertwined nature of AI and human lives, the importance of cultivating an interdisciplinary and transdisciplinary AI workforce should be emphasized.
Relates to: Strategy 7.

VI. Strategy 8 supports expanding public-private partnerships focusing on government-university-industry research and development partnerships. Given the social implications of AI, civil society organizations play a relevant and valuable role in voicing the expectations of the broader society.

Recommendation: Strategy 8 should be expanded to include collaboration with civil society organizations, particularly concerning future developments regarding the societal implications of AI.
Relates to: Strategy 8.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

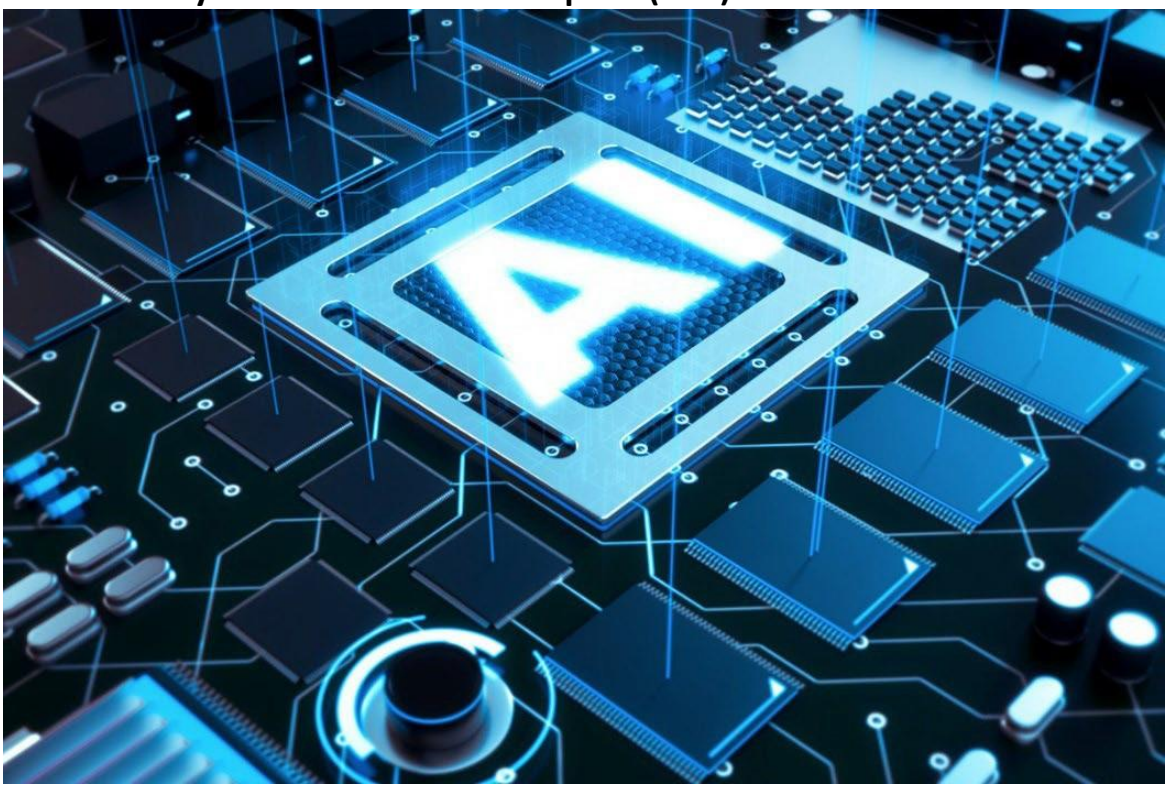
Hewlett Packard Enterprise (HPE)

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.

Hewlett Packard Enterprise

Response to the Office of Science and Technology Policy (OSTP) for Updating the National Artificial Intelligence Research and Development Strategic Plan Request for Information (RFI)

Submitted by Hewlett Packard Enterprise (HPE)



March 4, 2022

Request for Information

This page intentionally left blank

Table of Contents

This page intentionally left blank

Introduction

Hewlett Packard Enterprise (HPE) appreciates the opportunity to respond to the RFI providing input on updating the National Artificial Intelligence Research and Development Strategic Plan. We believe that artificial intelligence (AI) can amplify human capabilities and that investments into the cyberinfrastructure that fuels AI research and development is in our best national interest.

Driven by the promise of AI technologies, scientists across the country are eager to study AI and explore how it may benefit their work. We believe that to the extent the strategic plan can foster activities to capitalize on these scientists' enthusiasm and untapped potential by democratizing access to facilities, funds, data sets, knowledge, tools, and any other resources that would allow them to study and leverage AI. Beyond this recommendation, our response will offer commentary and suggested updates to 6 of the 8 strategies in the plan.

Comments and Suggested Updates

Strategy 1: Make long-term investment in AI research

- Advancing data-focused methodologies for knowledge discovery
- Enhancing the perceptual capabilities of AI systems
- Understanding theoretical capabilities and limitations of AI
- Pursuing research on general-purpose artificial intelligence
- Developing scalable AI systems
- Fostering research on human-like AI
- Developing more capable and reliable robots
- Advancing hardware for improved AI
- Creating AI for improved hardware

AI Computing Infrastructure

Creating and maintaining a shared computing infrastructure should leverage existing organizations that currently provide shared computing infrastructure to a wide community of students and researchers, including Department of Energy (DOE) computing user facilities and National Science Foundation (NSF) computing centers. Features to emphasize include an open and not proprietary technology in the resource; sufficient availability, scale, and capability of resources with a commitment to growth as needed; a strategic approach that encourages market competitiveness and ongoing technological innovation without sacrificing inter-operability; seamless interfaces and processes for inter-platform multi-modal data exploration and interrogation.

Democratized access to AI R&D infrastructure

Efficiently executing AI workloads is a very computing and data intensive endeavor. In the last decade, the AI R&D community has aggressively moved to a specialized hardware and software infrastructure, primarily using Graphics Processing Unit (GPU) acceleration, to achieve the desired cost/performance target. Democratizing access to AI R&D requires an open approach to accessing this specialized infrastructure, and the tools (such as runtime software, programming environment, and libraries) that are required to operate it.

There are a few obstacles to making this happen that an AI strategic plan has to consider. The AI accelerator market is experiencing a high degree of variability regarding openness in how accelerators are integrated into larger computing systems. Some AI systems are kept proprietary, not available in the open market, and only accessible through a specific interface, such as those offered by cloud service providers with dedicated proprietary accelerators. When the lack of openness raises to the software level, users are constrained in the way in which they can access and orchestrate the accelerators, and where the data they process can live.

Overcoming these limitations requires a "community call to action" that the AI strategic plan could sponsor, to preserve a healthy and open ecosystem that offers AI R&D users choice at all levels. While proprietary innovation is fundamental for the AI ecosystem to thrive as a business, there are important dimensions that can benefit from an open approach, such as storage, networking, virtualization, runtime, workflow, scaling and security. For example, the government could sponsor activities to produce open recommendations (possibly standards) and reference architectures for future accelerated systems. This will help create a blueprint for an open AI R&D infrastructure that can motivate all players to participate.

Building blocks

HPE believes that a few government agencies are already well positioned to provide the critical building blocks for an AI strategy. These include the DoE National Labs, the NSF advanced computing centers, National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA) and a few other agencies that have built a high end computing and data infrastructure as well as a surrounding ecosystem of system, applications, and other technical support. This existing infrastructure with some incremental funding could be leveraged to provide AI resources for a broader community. It is significantly more cost-effective to leverage an existing ecosystem, rather than starting up new centers from scratch.

There will be other considerations for the successful implementation that we'll identify for your consideration, as follows. Of critical importance are the data repositories. These will need to be federated, and to streamline access for users of the repository, may also need to be imported or securely linked into national AI resources. We believe some of the efforts by National Institute of Standards and Technology (NIST) to cultivate trustworthy AI systems, can also provide the complimentary foundational building blocks with regards to secure data repository access.

HPE recognizes that there will be specific use cases where Public cloud providers will provide additional resources to supplement those directly engaged by existing government agencies. These are often most effective for users who are just getting started and may not have computing resources beyond their personal client devices. It is also important to build on existing arrangements with a proven track record like the NSF-funded CloudBank model, which provides individual researchers access to commercial clouds for NSF-funded research.

Strategy 2: Develop effective methods for human-AI collaboration

- Seeking new algorithms for human-aware AI
- Developing AI techniques for human augmentation
- Developing techniques for visualization and human-AI interfaces
- Developing more effective language processing systems

HPE believes this strategy is fully developed and no additional comment is needed.

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

- Improving fairness, transparency, and accountability by design
- Building ethical AI
- Designing architectures for ethical AI

As AI Technologies become more ubiquitous and more essential to continued competitive leadership, it is essential that the National AI Strategic Plan enable increasingly diverse communities to be able to craft AI Ethics frameworks which are authentic to their experience. Our own recent HPE experience taking on this task is instructive. Under the joint leadership of Hewlett Packard Labs and the HPE Chief Compliance and Privacy Offices, a pan-HPE team has recently completed the drafting of the HPE global AI Ethics Principles. That team is now engaged in the even more challenging phase of operationalization of those principles into commitments and specifications to guide our team members, customer and partners in the utilization of AI across our products, our processes and our partnerships worldwide. While there is a large portfolio of AI Ethics principles and frameworks which have emerged from government, industry and academia, we felt that we needed to tailor a framework which was derived from our company purpose, to advance the way that people live and work, and was authentic to our roles in creating, supplying and consuming AI technologies.

As we gain experience applying these principles in practice, we're uncovering gaps where conventional AI applied to real world situations cannot be applied with confidence in meeting our AI ethics principles, revealing instead issues of Bias, Explainability, Trust and Robustness. It's not just a matter of being more careful or deliberate with today's state of the art, these are technology gaps and closing them will take engineering and ingenuity and it directly informs our Hewlett Packard Labs research agenda. That team is developing novel techniques and approaches to Model Synthesis and Analysis, the Data Foundation underpinning ethically robust AI, and Hardware Acceleration that will enable explainable, robust AI to be operated equitably and sustainably. By engaging in the hard work of crafting AI Ethics principles for our Enterprise, we not only enable the entire breadth of the enterprise to adopt AI technologies with confidence, we also directly shape the future of the technology. This should be directly mirrored in the National AI Strategic R&D Plan.

The National AI Strategic R&D Plan should embrace a dual mission of providing access to a robust, sustainable, holistic advanced computing ecosystem and data infrastructure and also the training and resources to enable producers and consumers of AI technologies to establish AI Ethics frameworks tailored to their communities and concerns. By enabling both producers and consumers of AI Technology with the means to create meaningful AI Ethics frameworks, the National AI Strategic R&D Plan could both enable the entire AI supply chain to both demonstrate with transparency how current technology can be applied with confidence and illuminate where current technologies fall short and innovation is required. Without mandating a particular ethical framework, the National AI Strategic R&D Plan focused on the dual mission of technology and ethical rigor could be establish benchmarks for due diligence in ethical application of AI technologies. With respect to traditionally underserved communities, this dual mission would provide not only access to technology but add the lived experience of those communities directly into the ethical discussions concerning their use.

Strategy 4: Ensure the safety and security of AI systems

- Improving explainability and transparency
- Building trust
- Enhancing verification and validation

- Securing against attacks
- Achieving long-term AI safety and value-alignment

Security

As you reevaluate the strategy, HPE believes these are key security principles to consider:

- Multifactor authentication to identify and categorize users and manage their appropriate access to the categories of assets
- Protection of the integrity of models and datasets and their inputs
- Encryption of data in motion and data at rest
- Recordation of state changes to assets with accountability to the level of a single individual or service

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

- Developing and making accessible a wide variety of datasets to meet the needs of a diverse spectrum of AI interests and applications
- Making training and testing resources responsive to commercial and public interests
- Developing open-source software libraries and toolkits

AI Datasets, FAIR access, Lineage and Provenance

Accessing high-quality government data sets remains challenging. One of the key reasons for this is the large number of government portals and initiatives that have been launched, often changing with each administration. This lack of a persistent, global data infrastructure often leads to confusion over where to find data sets, which data sets reflect the most current data and preparation process, and what data are best-suited to various AI applications.

A key need is to address questions of ensuring data lineage and provenance when making data sets available, but even more important is addressing the challenge of establishing and maintaining a persistent solution for making data sets accessible and available to researchers. This will include establishing capabilities within federal agencies to manage data availability and curation that persist across administrations and through budget cycles. Critically, however, this global solution must also be developed in close partnership with the research community, ideally leveraging non-federally owned or managed infrastructure and capabilities in non-profit entities such as universities or other research institutes. At the same time, the data infrastructure should not create a preference for one type of technology or limit competition from future technology and/or data solution providers. To ensure this, the government will need to reconsider its role in managing data sets, emphasizing the government's role in convening stakeholders and driving toward shared best-practices while federating data infrastructure management.

Some of the key challenges include:

- **Data set discovery** – Making it easy to find data sets and guidance on which data sets are considered the "gold standard" in various disciplines
- **Sharing** – Who can access data and how?
- **Trust** – How can users know they are accessing the data they think they are accessing, understand their lineage, assess their "quality" (e.g., reproducibility, etc.), and validate their provenance?
- **Security** – How can data sets be accessed and analyzed securely and in a way that maintains privacy?

- **Incentives** – Many critical data sets are not owned by the federal government. What incentives would help increase safe, trusted, and equitable data exchange?

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

- Developing a broad spectrum of AI standards
- Establishing AI technology benchmarks
- Increasing the availability of AI testbeds

Industry Consortia

HPE recommends engaging with Industry consortia like the ML Commons and other standards bodies such as Open Geospatial Consortium, as well as private sector solutions especially with highly data intensive applications like earth observations.

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

HPE does not believe we can provide significant input to understanding the AI R&D Workforce needs beyond what's already in the strategy.

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

Public-private partnerships

HPE believes public-private partnership is not just a mechanism to enable capacity or access to experts. Rather, HPE has a long history of developing new technologies, in partnership with our customers, to enable new capabilities. HPE believes that the AI researchers will identify gaps in capabilities that public-private partnership, through NRE funding, can fill. HPE and our government partners aspire to leverage current and future advances in artificial intelligence and high performance computing to dramatically accelerate scientific discovery and expand the frontiers of scientific knowledge to benefit US national security, scientific leadership, and industrial competitiveness.

Shared Principles should include area such as:

- Keep AI-related software open source as much as possible and prevent vendor lock-in
- Establish an inviting, standards-based open infrastructure environment to foster AI hardware and software innovations from the greater community. Standards proposed will take into account the need to ensure system-wide functionality with competitive performance relative to proprietary solutions
- Hide complexity from the user in order to make high-end AI accessible to a wider number and range of domain scientists
- Greatly improve the collection, curation, storage, movement and analysis of AI relevant data at scale, taking into account the expected greater volume, velocity and variety.
- Enable real-time and near real-time control/steering of experiments and scientific based operations through the use of AI connecting the system of systems
- Emphasize the development of high productivity software and tools that perform will at scale

Advisory Board

HPE recommends that an advisory board with membership in a diverse community of stakeholders be established to work with the office responsible for implementation, deployment, and administration.

Community Participation and Coordination

To ensure the National AI research resource is broadly available throughout the United States, in urban, suburban and rural communities, and especially within communities that are traditionally underrepresented in the development and use of technology, the federal government will need the participation of nearly every agency.

AI will heavily impact all Networking and Information Technology Research and Development (NITRD) participating agencies and programs. But there are many more agencies and programs beyond NITRD that need to participate. The fastest way to accomplish this is to keep most of the responsibility for implementation, deployment and administration within existing federal programs and agencies with input from the advisory board to assure the full diversity of the community is recognized. The big piece that is missing is the coordination piece. AI, including machine learning, deep learning, inferencing and other forms, is already computationally expensive and will get much more expensive in the coming years.

Furthermore, data is also growing exponentially, in part because of inexpensive sensors and the ability to collect a lot of data and can be exploited in the use of AI. AI requires storing and moving large amounts of data. When looking at AI users across communities, we see enormous opportunities to share AI artifacts, models, data, computational results and more. This would produce dramatic savings in resources, including power and reduce potentially massive redundancies. A key role of a research resource would be to bring communities together to improve coordination between organizations at the many levels they operate.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

IEEE-USA

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.

4 March 2022

AI R&D RFI Response Team

Re: ***RFI Response – National Artificial Intelligence Research and Development Strategic Plan***

IEEE-USA is pleased to submit these recommendations in response to OSTP's request for comments on the 2022 update to the National Artificial Intelligence Research and Development Strategic Plan.

We fully support the Administration's efforts to update and develop a comprehensive national AI strategy. This presents an opportunity to update the existing 2019 strategy in ways that reflect and address the actual, now more fully realized, impacts of artificial intelligence (AI) and automated decision systems (ADS) on our society. Advancements in AI/ADS and their proliferation in all sectors of life, work, and government directly impact citizens, domestic and national security, and geopolitical order. As a prerequisite to existing in today's society, we all participate in, interact with, and are subject to AI/ADS processes, data collection and analyses, and determinations by these systems that directly impact us via government and financial services, healthcare, and education, among others. Many of these processes and their impacts are largely not transparent on the human side and lack meaningful choice or control. Despite this reality, the U.S. lacks and would benefit from a comprehensive and cohesive federal regulatory framework for AI/ADS governance.

The very ubiquity of AI/ADS - deployed across all public and private sectors - presents an opportunity for the White House to provide much needed and necessary guidance to address the existing reality. AI/ADS methods and applications, plus a growing knowledge base about their development, use, and examples of potential and actual harms, have all expanded significantly since 2019. This knowledge can and should be utilized to update and generate an actionable governance framework that promotes efficiency and security while preserving civil liberties and individual choice, while minimizing harm. Thus, we laud this RFI's goal of updating the national AI strategy to further (1) reflect the existing realities and (2) anticipate future risks and opportunities from the existing and emerging systems that deploy AI/ADS.

A strong AI governance and federal regulatory strategy will guide AI/ADS development and use, *and* shape federal legislative efforts. Currently, the lack of clear, comprehensive federal AI/ADS regulation has resulted in states attempting to regulate these systems in ways that may have unintended consequences, conflict with constitutional protections and norms, create economic inefficiencies, and result in compliance uncertainties. For instance, Florida's SB 7072, signed into law in 2021 (although currently subject to partial injunction by court order), which regulates the use of AI/ADS by social media platforms and "journalistic enterprises" to "prioritize" content and posts in ways that confound the functional technology of such

IEEE-USA | 2001 L Street, N.W., Suite 700, Washington, D.C. 20036-4928 USA
Office: +1 202 785 0017 | Fax: +1 202 785 0835 | E-mail: ieeeusa@ieee.org | Web: http://www.ieeeusa.org

platforms, and directly impacts free speech, interstate commerce, federal communications regulation, and preemption under other federal laws.

To be effective, our National AI strategy should be broadly expanded to include strategies for understanding and addressing the ethical, legal, and societal implications of AI/ADS technologies, and to ensure their safety and security regardless of their complexity. Below you will find specific recommendations; we have also attached our recent position statements on the governance of AI/ADS.

IEEE-USA thanks the OSTP for considering these comments in the Office's revisions to the National AI Strategic Plan. We would welcome any further discussions with OSTP on these matters. If you have questions, please do not hesitate to contact Erica Wissolik at [redacted] or [redacted].

Sincerely,

Deborah Cooper

IEEE-USA President

IEEE-USA RECOMMENDATIONS: Updates to the National Artificial Intelligence Research and Development Strategic Plan (2022)

These recommendations are drawn directly from recently adopted IEEE-USA position statements directly applicable to, and useful in, updating and developing the National AI Research and Development Strategy. These recommendations provide building strategies that proactively address the ethical, legal, and societal implications of systems that deploy AI and automated decision systems (ADS) and identify more recent research, implementation, and standards frameworks.

1. **Create a clear legal and procedural framework for data ownership, data rights, and privacy**:
    - **a.** Update, and create clear, coherent, and comprehensive data protection law(s) at the federal level that;
        - **i.** Build legal standards on the limits of data use and privacy
        - **ii.** Require clear notice of data use practices that 'by-design' are explicit opportunities for proactive user consent
        - **iii.** Mandate transparency and user control for use of individual data

2. **Require redress mechanisms for systems that deploy AI/ADS**:
    - **a.** Create easily accessible pathways for stakeholders to review, verify, and contest personal data and decisions about that data
    - **b.** Require AI/ADS system developers to share well understood, explainable descriptions of systems that deploy AI/ADS at multiple levels of details including explanations that are transparent to the layperson
    - **c.** Enact and implement clear statutory culpability and means of civil redress for entities that deploy AI/ADS that are responsible for harm to individuals, groups, or environments.

3. **Address disparate impacts and harms of systems that deploy AI/ADS**:
    - **a.** Develop metrics for accountability fairness, privacy, safety, and security by engaging Ethics in Action and IEEE P7000™ working group experts from the series of standards already produced and those under development for support and guidance. The IEEE P7000 series addresses specific issues (such as accountability, fairness, privacy, safety, security) at the intersection of technological and ethical considerations. Working group members are experts on these standards.
    - **b.** Develop transparency mechanisms for the use and impacts of AI/ADS systems
    - **c.** Build financial support lines and grants to research how the use of AI/ADS may disparately impact or disadvantage vulnerable individuals or groups

4. **Promote transparency, human agency, and accountability in the design and use of systems that deploy AI/ADS**:
    - **a.** Create verification and validation procedures, transparency standards, and mechanisms for redress
    - **b.** Develop international agreements for AI/ADS systems' responsible use, governance, and impacts on human rights
    - **c.** Engage Ethics in Action and IEEE P7000™ working group experts from the series of standards already produced and those under development for support and guidance. The IEEE P7000 series addresses specific issues (such as accountability, fairness, privacy, safety, security) at the intersection of technological and ethical considerations. Working group members are experts on these standards.

5. **Provide for public input on the governance of systems that deploy AI/ADS,** by:
    - **a.** Increase investment in public education so laypersons (1) generally understand how AI/ADS function and are aware of the prevalence of their use across private and public sectors and communication platforms, and (2) develop awareness of the potential impacts of systems that deploy AI/ADS (promoting citizen resiliency)
    - **b.** Develop mechanisms for soliciting multi-stakeholder and diverse public input on the governance of systems that deploy AI/ADS, particularly from marginalized or vulnerable communities
    - **c.** Engage IEEE-SA Working Group members from the new P3119™ *Standard for the Procurement of Artificial Intelligence and Automated Decision Systems* which aims to address the needs of government workers, policymakers, and technologists. The IEEE P3119 Working Group intends to establish a uniform set of definitions and a process model for the procurement of AI and ADS by which government entities can address socio-technical and responsible innovation considerations to serve the public interest. The process requirements include a framing of procurement from an IEEE *Ethically Aligned Design* (EAD) foundation and a participatory approach that redefines traditional stages of procurement as: problem definition, planning, solicitation, critical evaluation of technology solutions (e.g., impact assessments), and contract execution. The purpose of IEEE P3119 is:
        - To establish a uniform set of definitions and process requirements that address the socio-technical and responsible innovation challenges in the procurement of AI/ADS,
        - To help support agencies adapt their processes for procuring AI/ADS systems responsible for the public interest, and
        - To promote ethically aligned values and robust public engagement in the process model.

6. **Develop an AI/ADS education pipeline**:
    - **a.** Develop resources for and investing in AI/ADS education at the elementary, secondary, and postsecondary levels that covers technical material and ethical considerations that arise when systems deploy AI/ADS

b. Develop and provide resources to assist affected or displaced workers (those impacted negatively from the systems that deploy AI/ADS)

c. Invest in explicit and purposeful recruitment of diverse human resources in AI/ADS-related fields.

IEEE-USA represents approximately 150,000 engineers, scientists, and allied professionals in the United States, many of whom are actively conducting research and development into artificial intelligence, software engineering, cybersecurity, and advanced computing, as well as other foundational and emerging technologies. We are the American component of the IEEE – the largest organization of technology professionals in the world, representing more than 400,000 engineers, scientists, and allied professionals worldwide.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Information Sciences Institute, University of Southern California

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.

Response: Update of the National Artificial Intelligence Research and Development Strategic Plan RFI

4 March 2022

We submit these comments on behalf of the University of Southern California's Information Sciences Institute.

We support the strategic aims of the National Artificial Intelligence Research and Development Strategic Plan. The goals and priorities itemized in A-G are well-thought-out and crucial for the U.S. to continue as a leader in Artificial Intelligence research and deployment. We were pleased that the National Defense Authorization Act for Fiscal Year 2021 authorized a significant increase of funding for Artificial Intelligence research and education through NSF, and we urge that the funds be appropriated to achieve this vision.

Our comments respond to the request that "...comments are invited as to existing strategic aims, along with their past or future implementation by the Federal government." We focus on three points:

- Support fundamental research on investigator-initiated topics, in addition to applied research.
- Create mechanisms for dedicated, long-term support for research projects.
- Reduce the administrative burden on researchers and their organizations.

Support long-term fundamental research on investigator-initiated topics in addition to applied research.

We focus our discussion on the National Science Foundation, which funds the majority of fundamental research in Artificial Intelligence under its mission ***"****to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..."*

The existing NSF and NIFA AI institutes represent a major investment for use-inspired research on topics such as AI & Agriculture, AI & Climate, AI & Chip Design, and AI & Physics; these institutes have started to achieve important goals in research, education, and outreach. However, the solicitation mechanism could be improved. Narrow topics are selected each year, with a short time to prepare a proposal involving multiple universities. For example, the topics "AI for Accelerating Molecular Synthesis and Manufacturing" and "AI for Discovery in Physics" were announced on October 8, 2019, with proposals due less than four months later, on January 28, 2020. These topics are no longer solicited; they were replaced with "AI to Advance Biology" in 2021 and "Intelligent Agents for Next-Generation Cybersecurity" in 2022. We believe a stronger approach would be to select criteria for AI Institutes and perhaps general areas, but not

to select narrow topics every year. This would allow the full creativity of the AI community to propose topics that meet the criteria of an institute but on topics that NSF did not preselect. We believe there is much to be gained by sustained research on fundamental AI research topics, such as commonsense reasoning, that do not lend themselves to immediate applications. Furthermore, a group of researchers may have a compelling vision for how AI can advance entomology, astronomy, or some other topic not preselected by NSF in the current year. In other cases, a team may have been almost funded one year, but their topic is no longer continued the following year, so their efforts and their collective innovations are largely wasted.

Furthermore, in our experience, putting a team together for a major institute takes much longer than 3 to 4 months, which is the time between the NSF announcement release and the due date. In universities we have been involved with, faculty aspire to create an Engineering Research Center or Science and Technology Center for a year or more, knowing that their topic will not be excluded from the call for proposals. The university may support this activity in numerous ways, such as rearranging the lead investigator's teaching schedule during the semester the proposal is due, or hosting multi-institution workshops on the topic ahead of the solicitation. Narrow topics and short deadlines make proposal development difficult and far from a well-planned systematic process.

Although NSF has a mission to promote the progress of science, its focus on these new large AI awards has primarily been in the context of use-inspired research. Of course, NIH, NIFA, and DoE labs focus AI research on specific topics. Even within DoD, the most significant growth has been in 6.2 and 6.3 funding (applied research), with little growth in fundamental 6.1 research. There is a need for fundamental research into the science of Artificial Intelligence, in addition to use-inspired research. Core problems, e.g., commonsense reasoning, can be finessed or simplified in focused problems, and this results in narrow systems without the breadth of intelligence exhibited even by small children. The U.S. needs long-term investments in these very hard fundamental research areas, but this cannot be done without appropriating funds for NSF to support this.

We contrast the funding for AI research to the funding for quantum information science. The hundreds of millions of dollars allocated to quantum initiatives are fully devoted to fundamental research since there are not yet quantum applications. (We know this because our institute is home to the first quantum computer in academia.) In contrast, very little of the hundreds of millions of dollars reported as federal AI funding goes to fundamental AI research. We often see reports of the very large amounts of funding being allocated for AI in different government agencies, but the vast majority is driven by in-use or applied research rather than by foundational AI questions.

Create mechanisms of dedicated, long-term support for research projects.

Although the NSF AI institutes represent a major investment in AI by NSF, usually they do not provide the majority, and in many cases even the plurality, of an individual

researcher's funding. Neither do other large funding opportunities, such as NSF ERCs, STCs, or Expeditions. Consequently, instead of being the sole focus of a sustained research effort, such funding is one of several grants for most researchers involved in an NSF AI Institute, with their attention divided between the AI Institute and other projects. $20M for an NSF AI Institute is significant funding, but when divided among 20-30 researchers at multiple institutions over five years, and with requirements for education, outreach, and the additional overhead of multi-institution research with travel for P.I. meetings, the amount available supports less than one-twelfth of an individual's research time (if that) and a single graduate student or postdoctoral researcher.

This hypothesis comes from our experience and from discussions with several NSF AI Institute awardees. A more systematic study would compare:

- The total funding per year of each researcher involved in an AI Institute
- The total number of active awards per year of each researcher involved in an AI Institute
- The total number of trainees supervised by each researcher involved in an AI Institute

to the similar data for the AI Institute.

We hypothesize that maintaining multiple awards and funding streams is not efficient for conducting long-term research on fundamental research problems. Having one-twelfth the attention of 30 people who are involved in multiple other projects may not be as productive as having one-half the attention of a smaller group of people. We believe such dedicated efforts are needed to solve fundamental problems. Without this, intelligent systems will plateau at high levels of competence on narrow problems but will have no capabilities beyond that narrow problem that we would associate with general intelligence.

Other federal agencies have recognized the problem and proposed solutions, for example, the NIH R35 program: *"The goal of the R35 is to help investigators make meaningful contributions … by providing greater funding stability, flexibility, and support for your overall research project. The R35 RPA is a funding mechanism that supports your research efforts by: Providing stable funding of up to $750,000 per year (direct cost funding) for up to eight years; Allowing you to focus on your work rather than spending valuable time continuously applying for funding; Allowing you to conduct long-term, rewarding research that is not tied to specific aims; and Providing flexibility to pivot as needed to emerging and timely topics."*

Although nothing prohibits a smaller group (e.g., 8 people) from proposing an institute, the practice and history at NSF would not make the concentrated effort of a mid-sized team competitive when compared to a larger team. One option would be to create a team mechanism similar to that of the R35 or to restrict the number of senior personnel on some grant programs, with the intent of an institute being the majority of funding for

an individual. We know that some agencies, e.g., DARPA, do support larger amounts of funds per person involved in a project and receive more dedicated attention from P.I.s and trainees.

A more stable source of long-term funding may also reduce the trend of researchers leaving academia for industrial research jobs where sustained funding is available. Such researchers, even if they maintain some part-time academic appointment, no longer participate in the training of the next generation of researchers or the fundamental AI research pursued in universities.

In other countries, there are long-term funding programs for foundational AI research. One example is the Hybrid Intelligence Center in the Netherlands, a consortium of universities awarded in 2019 with €20M for 10 years by the Netherlands Organization for Scientific Research's Gravitation program. Another example is the Pan-Canadian Artificial Intelligence Strategy, which funds Canada's three national AI Institutes: Amii in Edmonton, Mila in Montreal, and the Vector Institute in Toronto. China has several large institutes dedicated to AI and technology with substantial stable funding and little overhead. In contrast, in the U.S., we see low success rates of AI research proposals, even when rated highly, and short periods of performance for awards.

Reduce the administrative burden on researchers and their organizations.

Numerous reports have discussed the administrative burden placed on researchers, with more than half the research time spent writing proposals and agency reports. For example, see Reducing Federal Administrative and Regulatory Burdens on Research (archives.gov).

The burden of an NSF AI Institute seems exceptionally high, with strategic and implantation plans approaching 100 pages and annual reports far exceeding that. While accountability and a good return on investment from federal funding are desirable, we may have reached the point of spending more effort proposing and justifying research than conducting research. Academic AI researchers are already overburdened with the largest high-demand classes, a very rapidly advancing field that is hard to keep up with, and very low success rates of funding. Any government initiatives to reduce the administrative burden on grant proposals and reporting will contribute to creating a more effective academic research and education capability in the U.S.

Thank you for the opportunity to comment on the strategic aims of the National Artificial Intelligence Research and Development Strategic Plan. We hope our input on improving federal support for research in the field of artificial intelligence is useful.

Yolanda Gil, [redacted]

Michael Pazzani, [redacted]

Dr. Yolanda Gil is Principal Scientist and Senior Director of Strategic Initiatives in AI and Data Science at the USC Information Sciences Institute, and Research Professor in Computer Science and in Spatial Sciences. She received her M.S. and Ph. D. degrees in Computer Science from Carnegie Mellon University, with a focus on artificial intelligence and cognitive science. In 2019 she co-chaired the community report "A 20-Year Artificial Intelligence Research Roadmap for the U.S." She served in the Advisory Committee of the National Science Foundation's Directory of Computer and Information Science and Engineering. She initiated and led the W3C Provenance Group that led to a community standard that provides the foundations for trust on the Web. She is a Fellow of the Association for Computing Machinery (ACM), the Association for the Advancement of Science (AAAS), and the Institute of Electrical and Electronics Engineers (IEEE). She is also a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI) and served as its 24th President. Dr. Gil is an Advisory Board member for two of the NSF AI Institute awards.

Dr, Michael Pazzani is a Principal Scientist at the USC Information Sciences Institute. Dr. Pazzani was the Vice Chancellor for Research and Economic Development at the University of California, Riverside, where he was also a professor of computer science with additional appointments in statistics and psychology. From 2006-2012 he was the Vice President for Research and Economic Development at Rutgers, the State University of New Jersey, where he was also a Distinguished Professor of Computer Science. Prior to his appointment at Rutgers, Dr. Pazzani was the Director of the Information and Intelligent Systems Division at the National Science Foundation from 2002 to 2006. In addition, Dr. Pazzani coordinated NSF's homeland security research. He also served as a member of the Board of Regents of the National Library of Medicine at the National Institutes of Health from 2003 to 2005. In 2019, Dr. Pazzani was appointed to the Defense Science Board. Dr. Pazzani started his career as an assistant, associate, and full professor of Information and Computer Science at the University of California, Irvine. Dr. Pazzani is also managing director of an NSF-funded AI Institute led by the University of California, San Diego.

The University of Southern California's Information Sciences Institute (ISI) carries out basic and applied research in artificial intelligence, networks and cybersecurity, high-performance computing, microelectronics, and quantum information systems. Its $100M annual external funding comes from the NSF, DoD, IC, NIH, DoE, industry, foundations, and other sponsors. ISI is home to the first quantum computer in academia. Part of the USC Viterbi School of Engineering, ISI has more than 400 personnel that includes 28 faculty that advise 65 PhD students. ISI's Artificial Intelligence Division is one of the largest AI research groups in the U.S. ISI's AI systems for machine translation, online misinformation detection, and data-centric AI are first-rate and have been deployed to support many parts of DoD and hundreds of law enforcement agencies. Some of ISI's commercial spinoffs were acquired for tens of millions and contribute to a vibrant innovation ecosystem in Southern California.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Information Technology Industry Council (ITI)

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.

ITI Promoting Innovation Worldwide

NCO
AI R&D RFI Response Team
2415 Eisenhower Avenue
Alexandria, VA 22314

March 4, 2022

Re: ITI Comments in Response to Office of Science and Technology Policy Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan

Dear AI R&D RFI Response Team,

The Information Technology Industry Council (ITI) appreciates the opportunity to respond to the White House Office of Science and Technology Policy's Request for Information to Update the National Artificial Intelligence Research & Development Strategic Plan (the "Strategic Plan").

ITI represents the world's leading information and communications technology (ICT) companies. We promote innovation worldwide, serving as the ICT industry's premier advocate and thought leader in the United States and around the globe. ITI's membership comprises leading innovative companies from all corners of the technology sector, including hardware, software, digital services, semiconductor, network equipment, and other internet and technology-enabled companies that rely on ICT to evolve their businesses. Artificial Intelligence (AI) is a priority technology area for many of our members, who develop and use AI systems to improve technology, facilitate business, and solve problems big and small. ITI and its member companies believe that effective government approaches to AI clear barriers to innovation, provide predictable and sustainable environments for business, protect public safety, and build public trust in the technology.

We recognize that AI is an active area of research that is constantly evolving and improving. To harness this growth, we believe it is vital to both utilize AI's potential benefits while monitoring its impacts carefully, and research and development (R&D) is a critical enabler of those possibilities. Indeed, our *Global AI Policy Recommendations*, released in 2021, include an entire section devoted to facilitating innovation and investment in AI that emphasizes the critical role R&D must play in that effort. As such, we welcome the opportunity to provide input on how OSTP can best update the Strategic Plan to reflect recent evolution in the field of AI.

We previously responded to OSTP's RFI on establishing an Implementation Plan to guide the National Artificial Intelligence Research Resource, in which we outlined several themes that may also be relevant to consider in the update to the Strategic Plan.[1]

We think the overarching strategic aims remain an appropriate construct by which to pursue national AI R&D efforts. The strategic aims generally track with the structure of our *Global AI Policy Recommendations* and accurately capture the areas that we believe the USG should focus on.[2] We believe that many of the updates made to the Strategic Plan in 2019 remain relevant in 2022. Below, we offer commentary on areas of the Strategic Plan that we believe remain relevant, as well as those that may require updating. In commenting on the strategic aims, we also note government initiatives that may be worth referencing in the 2022 update. We especially highlight the Final Report released by the National Security Commission on Artificial Intelligence in 2021, which includes many recommendations that we support, and which is worth referencing in the update of the Strategic Plan.[3]

- ***Strategy 1:*** **Make long-term investments in AI research.**

This strategy remains relevant and is aligned with our *Global AI Policy Recommendations,* in which we stress the importance of continued government investment in AI R&D, including basic science. It also aligns with the 2019 update of Strategy 1, which emphasizes the importance of sustaining investment in fundamental AI research. This sort of continuity will be important to the continued evolution of AI and machine-learning. We continue to encourage the USG to invest in and support its investment in research fields specific or highly relevant to AI, including cyber-defense, data analytics, detection of fraudulent transactions or messages, adversarial machine learning/AI and how to secure ML/AI, privacy preserving machine learning (PPML), robotics, human augmentation, natural language processing, interfaces, and visualizations.

Even more importantly than advanced algorithms, specialized computing hardware, and high-quality data, skilled human expertise is essential to enabling machine learning and the success of AI. To remain the leader in AI R&D, the United States must continue to promote an entrepreneurial environment, research network, and openness to talent. One of the reasons the United States has succeeded in this space is that it has invested heavily in R&D. In 2019, for example, the United States led the world in total R&D expenditures, with combined public and private sector spending totaling $657 billion.[4] Maintaining American leadership in AI related R&D efforts will not only require continued government R&D investment, but in promoting

[1] ITI comments responding to the *RFI on an Implementation Plan for a National Artificial Intelligence Research Resource (NAIRR)* available here: https://www.itic.org/documents/artificial-intelligence/2021-9-30_ITICommentsNAIRRRFIFINAL(1).pdf

[2] ITI's *Global AI Policy Recommendations*, available here: https://www.itic.org/documents/artificial-intelligence/ITI_GlobalAIPrinciples_032321_v3.pdf

[3] Final Report from National Security Commission on AI, available here: https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

[4] Main Science and Technology Indicators in OECD.*Stat* database, available here: https://stats.oecd.org/Index.aspx?DataSetCode=MSTI_PUB

scientific collaboration among like-minded nations (see our recommendation on p. 8 to add a specific strategic aim focused on this). Additionally, maintaining leadership in AI R&D will require continued strong political support from Congress and the Executive Branch, as well as active participation from the private sector and society.

- ***Strategy 2:*** **Develop effective methods for human-AI collaboration.**

We generally remain supportive of the goals outlined under Strategy 2. However, in contemplating improvements to human-AI collaboration, we encourage OSTP to further emphasize the need for additional research related to the usability and confidence of explanations that may be provided to the human. Research demonstrates that the mere presence of explanations increases trust in a system and user deference to the system.[5] So, if a prediction is accompanied by an explanation, users assume that it is true even if they cannot understand the explanation or if the explanation is incorrect. It would be prudent to emphasize that research is needed not only to build better explanations, but also to determine how to make the user reckon with the explanation as well.

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

Since 2019, the USG, industry, and governments globally have taken steps to understand and address the ethical, legal, and societal implications of AI. However, additional R&D is necessary to further address potential challenges that may stem from the use of AI.

We appreciate that OSTP is considering how privacy and civil liberties questions intersect with AI R&D and encourage an update to the strategy that reflects the need for additional research in this area. Indeed, maintaining appropriate privacy protections is imperative to fostering trust in AI technology. As we noted in our 2018 submission on the AI R&D plan, as well as our response to the RFI on the NAIRR Implementation Plan, it is important to recognize that AI operates in an existing policy and regulatory environment, and accordingly, personal data and related privacy concerns should be accounted for.

We face important questions around striking the right balance between various objectives in the responsible development of AI, such as ensuring accountability, which requires some level of visibility into an AI system, while also protecting privacy. Adding to the complexity of a dynamic international environment of laws that are not always in alignment, the US currently has a patchwork of privacy policies and regulations that could become more complex and fragmented as additional states follow California's lead in establishing state-level comprehensive consumer privacy laws. Conflicts across these laws could have a chilling effect on AI advancement, as well as other data-driven technologies. To maximize the use of AI, we need strong, globally accepted privacy standards to enable trust and interoperability, and to incentivize investment in research to develop new techniques for even stronger privacy and

[5] See Bansal et. al., *Does the Whole Exceed its Parts? The Effect of AI Explanations on Team Performance,* available here: https://arxiv.org/abs/2006.14779

security guarantees. To achieve this, we continue to recommend the development of a national privacy law in the United States, consistent with *ITI's Framework to Advance Interoperable Rules on Privacy*.[6] We also recommend that OSTP reference NTIA's work on Privacy, Equity, and Civil Rights in its 2022 update, as we believe that activities undertaken, and information collected pursuant to this effort will intersect with and inform efforts pursued by OSTP.

Continued research on algorithmic transparency, and in particular explainability, will be important in understanding transparency as a tool in and of itself, as well instances in which explainability might be useful (and possible) and where it might not be. NIST has already started to undertake work on explainability with the publication of *NISTIR 8312: Four Principles of Explainability.* It may be worth referencing this work in the 2022 update of the Strategic Plan. Although NIST has undertaken this work, we encourage the federal government to continue to fund additional research in this space, ensuring that the appropriate stakeholders from both the public and private sectors are involved. This will help ensure that any future policy actions that contemplate requirements around transparency or testing will be well-informed and rest upon a solid foundation.

Beyond that, additional research is needed into how to properly evaluate and mitigate bias, as this is a major concern stemming from the use of AI and/or machine-learning in certain contexts. NIST has once again led the way in this space with the publication of *draft NIST SP 1270: A Framework for Identifying and Managing Bias in Artificial Intelligence.* Addressing bias will require collaboration across the public and private sectors in order to foster a practical understanding of how AI tools are designed, developed, and deployed and create state-of-the-art approaches to address identified challenges. It is also necessary to develop data-driven techniques, metrics, and tools that industry can operationalize to accurately measure and mitigate bias in concrete terms.

Since the 2019 update, several federal agencies have developed and promulgated ethics principles to guide the way in which AI is used and applied to support agency missions. For example, the Department of Defense adopted in 2020 and reaffirmed in 2021 *Ethical Principles for Artificial Intelligence*, while the intelligence community adopted *Principles of Artificial Intelligence Ethics for the Intelligence Community* as well as the *Intelligence Community Artificial Intelligence Ethics Framework.* We encourage the 2022 update of the Strategic Plan to reference these efforts.

- ***Strategy 4:*** **Ensure the safety and security of AI systems.**

The 2019 update rightly recognized that the security and safety of AI systems is something that needs to be addressed throughout the AI development lifecycle as it is foundational to trustworthy AI. The update additionally noted that adversarial AI is becoming a more prominent issue. We believe these tenets remain relevant in 2022. Indeed, we encourage public and

[6] *ITI Framework to Advance Interoperable Privacy Rules*, available here: https://www.itic.org/public-policy/FINALFrameworktoAdvanceInteroperableRules%28FAIR%29onPrivacyFinal_NoWatermark.pdf

private sector stakeholders to incorporate AI systems into threat modeling and security risk management, taking into account AI systems as a potential attack surface. We also encourage governments to invest in security innovation to counter adversarial AI and urge OSTP to elevate this as a priority for R&D funding moving forward, particularly as research undertaken by the Center for Security and Emerging Technology demonstrated that less than 1% of AI R&D funding is being allocated toward the security of AI systems.[7]

We additionally urge OSTP to consider that AI may also be used for cybersecurity defense purposes. Indeed, AI can be integrated into defensive cybersecurity technology to effectively respond to automated, complex, and constantly evolving cyberattacks. As such, it may be appropriate to also highlight this as an area for continued R&D.

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

We remain supportive of this strategic aim, as shared datasets are crucial to both training and testing AI and foundational to continued innovation. Indeed, a key recommendation in our *Global AI Policy Recommendations* is for policymakers to consider how to increase access to government sources of publicly available data in machine-readable formats and across borders to enable access to this foundational building block of AI. In updating the Strategic Plan, we encourage OSTP to reference the continued work of the National AI Research Resource Task Force as an important step toward creating an implementation plan and roadmap to govern the establishment of a shared data and compute resource.

In our response to the RFI posted by OSTP on the NAIRR, we referenced access to data sets as a potential limitation in democratizing access to AI R&D, particularly given the continuing lack of a comprehensive data-sharing strategy in the United States. Therefore, we urge the USG to develop a balanced framework for the responsible use of data. The more data are available, the more algorithms can learn, and the better AI offerings will be. Data must also be high-quality, credible, timely and available in machine-readable formats. Critically, to comply with privacy and other relevant laws data must also be anonymized, aggregated, or otherwise de-personalized such that the dataset excludes any personal information and cannot be re-identified. This ensures the beneficial use of the data in training intelligent systems while protecting consumer privacy and security. We recommend that the USG consider what technical and administrative hurdles exist to making data sets available (including any obstacles to international transfer and use of datasets and/or data residency requirements) and seek to overcome these hurdles so that developers can access this data. See our comments in response to the NAIRR RFI for additional comments regarding other ways the USG can encourage data-sharing.[8]

---

[7] *The Three Major Security Threats to AI,* available here: https://cset.georgetown.edu/article/the-three-major-security-threats-to-ai/

[8] ITI comments responding to the *RFI on an Implementation Plan for a National Artificial Intelligence Research Resource (NAIRR)* available here: https://www.itic.org/documents/artificial-intelligence/2021-9-30_ITICommentsNAIRRRFIFINAL(1).pdf

In the context of Strategy 5, it may also be useful to note that oftentimes when datasets are made open, provenance can be difficult to ascertain. The full context of data collection might be unclear, or it might be unclear how or whether consent to collect the data was obtained, or the conditions under which the labelling was conducted is not disclosed. There can be wide variation in how much testing for sampling bias is done, or there might be information missing that makes it difficult for those using that data to do their own bias testing. As it funds research in this field, the government might consider requiring and supporting provenance creation in order to improve this situation.

- ***Strategy 6:*** **Measure and evaluate AI technologies through standards and benchmarks.**

In the 2019 update, the Strategic Plan mentions that emphasis on standards and benchmarks has continued to rise in the U.S. and globally. In our view, this emphasis has increased even further, particularly as countries around the world led by the European Union are introducing legislation to regulate uses of AI technology. We have encouraged the European Union to rely upon international standards in implementing aspects of the AI Act, its proposal to regulate AI, including for the quality and risk management processes and data governance requirements it contains so as to avoid fragmentation in the global regulatory environment.

In the United States, NIST has undertaken significant work in the development of standards, guidance, and best practices, to include the work on explainability and managing AI bias as referenced earlier in our submission. In particular, Congress has directed NIST to develop an AI Risk Management Framework, a voluntary tool that organizations can use to evaluate, assess, and manage risks that may result from the use of AI. The Framework will leverage existing standards and best practices that organizations can use to achieve stated outcomes and is worth highlighting as an important and robust USG effort related to implementing this strategic aim in an update to the Strategic Plan.

However, the field of AI standards is somewhat nascent and continues to evolve, meaning that it is all the more important that the USG continues to support industry participation in international standards organizations, specifically ISO/IEC JTC 1 SC 42 (SC 42), where standards are being developed on many aspects of AI. We encourage NIST to continue its engagement in SC 42, where appropriate, as this will ensure that the AI RMF is aligned with international standards that are in the process of being developed and vice versa. We offer additional thoughts on international cooperation below, but international standards bodies are one forum that we believe the USG should prioritize in seeking to increase international coordination on AI-related R&D.

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

As with many of the other sections in the 2019 update, the points made remain relevant. Indeed, the lack of a skilled workforce remains a challenge for supporting and contributing to AI R&D. Therefore, the 2022 update of the strategy should continue to reflect the need to

prioritize and support policies that advance modernizing candidate recruitment, hiring, and training, and should establish and advance industry-led training and re-training programs to prepare individuals for the future of work, including an AI-enabled future.

That said -- and the 2019 update to the strategy rightly recognizes this -- AI is not just a function of STEM; true innovation requires contributions from workers in multiple disciplines, including the humanities and the social sciences. For example, most technology firms make ample use of anthropology and sociology in the initial design phases of new products to better understand where human need lies and the likely social dynamics of adoption. The humanities have been vital in the ongoing discussion and debates about AI ethics and can inform AI system design as much as policy. The arts are crucial to building technical systems that matter and will be used enthusiastically by people. Therefore, the best way to ensure access to an AI workforce is to invest broadly across all relevant university disciplines, including STEM, the social sciences, and the humanities, and to support data science programs that integrate humanistic approaches into the curriculum beyond a single, separate "AI ethics" unit.

- ***Strategy 8:*** **Expand Public-Private Partnerships to accelerate advances in AI.**

Public-private partnerships remain key. The National AI Advisory Committee and Subcommittee on AI and Law Enforcement, as well as the many public comment opportunities that have been provided to stakeholders as the US government seeks to implement the *National AI Initiative Act of 2020*, have been positive steps toward forging better collaboration between industry, government, and academia on AI. Given the rapid development and adoption of AI technologies in the commercial space, the need for consistent dialogue between the government and private sector to inform research priorities, from both technical and social impact perspectives, cannot be understated. We continue to recommend formally establishing a regular cadence of dialogues between the public and private sector on discrete AI-related issues and hope that the establishment of the National AI Advisory Committee will help to institutionalize public-private cooperation in this area. Incorporating regular private sector participation in all efforts will help to ensure that AI is advanced in a fashion that is broadly beneficial to all Americans.

Many emerging AI technologies are designed to perform a specific task, such as assisting human employees or making tasks easier. Our ability to adapt to rapid technological change is critical and we must continue to be prepared to address the implications of AI on the existing and future workforce. By leveraging public-private partnerships – especially between industry partners, academic institutions, and governments – we can expedite AI R&D, democratize access, prioritize diversity and inclusion, and prepare our workforce for the jobs of the future. We recommend that government tap into the commercial space and when appropriate, form more public-private partnerships to maximize the potential of AI.

We also recommend adding a new strategic aim, below:

- ***Strategy 9:*** **Enhance international cooperation with partners on AI R&D.**

We believe that adding a discrete strategic aim focused on international cooperation will elevate the importance of undertaking joint R&D work with trusted international partners. There are various national efforts taking place globally on AI that may impact how industry develops and deploys AI systems and how they mitigate potential harms, most notably the AI Act, which is under active consideration in the European Union. As such, it has never been more important for the United States to cooperate and coordinate with partners around the world to ensure that approaches are interoperable, or at least aligned, to the greatest extent possible. Driving such alignment will help to avoid regulatory divergence, while also facilitating the flow of information between the US and like-minded partners, which could in turn help to spur innovation and address any challenges that may emerge. In pursuing this aim, the USG should seek to support the development of voluntary, consensus-based standards, which we reference above in our response to Strategy 6. Cooperating to establish international standards will also help to serve several of the Plan's underlying goals, including ensuring the U.S. remains a leader in AI, helping to address the harms due to the disparate treatment of different demographic groups, and evaluating and managing bias, equity, and other concerns.

***

Once again, ITI appreciates the opportunity to provide feedback on the 2022 update to the Strategic Plan. As OSTP rightly recognizes, AI is a rapidly evolving technology. In order to ensure continued AI innovation, strategic guidance must also adapt. That being said, on the whole, we believe the strategic aims remain a relevant and appropriate construct by which to guide the United States' approach to AI R&D.

Sincerely,

John S. Miller
Senior Vice President of Policy
and General Counsel

Courtney Lang
Senior Director of Policy

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

International Business Machines Corporation (IBM)

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.

600 14th St. NW, Suite 300
Washington, D.C. 20005

March 4, 2022

AI R&D RFI Response Team
Office of Science and Technology Policy
Attn: NCO
2415 Eisenhower Avenue
Alexandria, VA 22314

Subject: "Update of the National Artificial Intelligence Research and Development Strategic Plan" [FR Doc 2022-02161]

To Whom It May Concern:

On behalf of International Business Machines Corporation (IBM), we welcome the opportunity to respond to the Office of Science and Technology Policy's (OSTP) request for information (RFI) regarding the "Update of the National Artificial Intelligence Research and Development Strategic Plan" of 2019.

IBM is an artificial intelligence (AI) and hybrid cloud technology leader that is engaged in research and development across a broad set of scientific and industry domains. IBM has long been committed to the research, development, and deployment of new technologies – including AI. IBM believes that carefully constructed research and development (R&D) frameworks are critical to realizing AI's vast potential, and are vital to mitigating any potential impacts. In addition, IBM recognizes that AI is a rapidly evolving technology that requires a carefully calibrated AI R&D strategic plan to ensure that federal investment priorities keep pace with the rapidly evolving technical environment to drive AI innovation.

Overall, IBM believes the overarching aims in the Strategic Plan of 2019 remain an appropriate construct by which to advance national AI R&D efforts. Previously, IBM responded to OSTP's RFI on establishing an Implementation Plan to guide the National Artificial Intelligence Research Resource (NAIRR), in which we outlined several themes that may be relevant to consider in the update of the Strategic Plan for 2022. Below, we offer commentary on areas of the 2019 Strategic Plan that we believe remain relevant, and suggest updates. For any questions, please contact Mr. Jeffrey Brown, Science & Technology Policy Executive at jeffrey.brown@ibm.com.

Respectfully,

Jeffrey J. Welser
COO IBM Research
VP Exploratory Science & University Collaborations

IBM Response to OSTP RFI – "Update of the National Artificial Intelligence Research and Development Strategic Plan" [FR Doc 2022-02161]

COMMENTS ON EXISTING (2019) STRATEGIC AIMS:

Strategy 1) Make long-term investments in AI research

As part of the effort to sustain long-term investments in fundamental AI research, the 2022 R&D Strategic Plan should support the continued expansion of the National AI Research Institutes Program. The Program is establishing a nationwide network of AI research clusters that can support sustained, large-scale, and multidisciplinary research into pressing challenges. To date, the Program has established eighteen Research Institutes that each operate as hubs for research into how AI can be used to address a broad range of societal and technological challenges, including climate change, the agricultural supply chain, and cybersecurity. By bringing together researchers from multiple academic institutions, as well as experts from industry, government, and NGOs, each of the Program's Research Institutes are helping to break down siloes and foster coordination across the United States' dispersed AI R&D ecosystem. The 2022 R&D Strategic Plan should highlight the critical importance of the Program, and signal support for its continued expansion.

Strategy 2) Develop effective methods for human-AI collaboration

IBM is supportive of the 2019 goals outlined under Strategy 2. In contemplating improvements to human-AI collaboration as part of the 2022 strategy, we recommend an increase in support for AI R&D related to the usability and explainability of decisions provided to the human. Such research is needed not only to build better explanations, but to determine how to help the user reason with the explanation as well. To address challenges related to explainability, common sense, robustness, and human-AI collaboration, IBM supports additional R&D focus on neuro-symbolic AI, which combines statistical learning with logic-based AI techniques and allows data-driven machine learning to be combined with symbolic reasoning, decision making, and human-machine interaction.

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

Since the 2019 Strategy was released, government, industry, and academia have made strides in understanding and addressing the ethical, legal, and societal implications of AI. However, many challenges related to transparency, fairness, explainability, robustness, and privacy remain unresolved and require continuous research and monitoring if they are to be properly addressed. AI is a rapidly evolving technology and its applications are multiplying, so it is important to also provide all AI stakeholders – including researchers and developers – with multistakeholder environments where it is easy to clearly identify new challenges (such as those related to deep fakes, tech addiction, and large language models).

The 2022 R&D Strategic Plan should incorporate principles for trustworthy AI, including fairness, explainability, and transparency, including the work being undertaken by the National Institute for Standards and Technology (NIST) to develop a cross-sectoral AI Risk Management Framework. To support bias mitigation strategies, updates to the 2022 Strategy should proactively implement AI ethics principles and practices and ensure appropriate governance is in place to provide ongoing review and oversight. IBM has defined a risk-based AI governance policy framework, Precision Regulation, that calls for fairness and security validated by testing for bias before AI is deployed and re-tested as appropriate throughout its use, especially in automated determinations and high-risk applications. IBM also encourages OSTP to use the NAIRR Task Force as a forum to expand efforts to develop and deploy safe and ethical AI to researchers across the U.S.

Strategy 4) Ensure the safety and security of AI systems

If researchers and developers are not trained to recognize possible sources and types of bias in their decisions and in the AI models they build, AI techniques based on data-driven machine learning could perpetuate and amplify bias and discrimination. However, if carefully educated and trained, they can use AI to identify and mitigate discrimination, and build a transparency layer on top of such models, so others can easily understand outstanding risks. Technical tools need to be combined with human multistakeholder consultations and training, so AI R&D can become increasingly more socio-technical rather than just a technical effort.

The 2019 update rightly recognized that the security and safety of AI systems is something that needs to be addressed throughout the AI development lifecycle as it is foundational to trustworthy AI. The update additionally noted that adversarial AI is becoming a larger issue. We believe these tenets remain relevant in 2022. Indeed, we encourage public and private sector stakeholders to incorporate AI systems into threat modeling and security risk management, taking into account AI systems as a potential attack surface. We also encourage governments to invest in security innovation to counter adversarial AI and urge OSTP to elevate this as a priority for R&D funding moving forward. As part of its 2022 update, we urge OSTP to consider how AI can be integrated into defensive cybersecurity technology to effectively respond to automated, complex, and constantly evolving cyberattacks. As such, it may be appropriate to also highlight this as an area for additional R&D focus and funding.

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

In IBM's response to the NAIRR RFI, we noted that limited access to data is a potentially serious roadblock to democratizing access to AI R&D, particularly given the lack of a comprehensive data-sharing strategy in the U.S. The 2022 Strategic Plan should prioritize the development of shared public datasets and environments for AI training and testing that include data that is high-quality, credible, timely, and machine-readable. Data should adhere to FAIR guiding principles to improve the findability, accessibility, interoperability,

and reuse of data, thereby helping to democratize AI R&D to a wider cross-section of Americans.

To fuel AI R&D efforts, the U.S. should develop national AI testbeds and define application targets for U.S. industry that allow it to create communities of discovery that accelerate efforts around, for example, the development of drugs and vaccines. For example, the COVID-19 High-Performance Computing Consortium illustrates the enormous value of creating platforms for data sharing and computational resources to accelerate scientific discovery. In line with our recommendations on public-private partnerships outlined in Strategy 8, creating communities of discovery furthers AI R&D by pairing the development of physical infrastructure with communities focused on application.

Additionally, as a federated and heterogeneous system-of-systems, the NAIRR should include a number of testbeds for evaluating and researching ethical AI across myriad disciplines and implementations – including evaluating and testing for bias. In this way the NAIRR could potentially be a useful mechanism to aid in operationalizing the aims associated with OSTP's AI Bill of Rights.

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

The 2019 Strategic Plan emphasizes the need for AI standards and benchmarks in the U.S. and around the world. NIST has developed standards, guidance, and best practices, to include the work on explainability and managing AI bias as referenced in Strategy 3. Congress has directed NIST to develop an AI Risk Management Framework, a tool that organizations can use to evaluate, assess, and manage risks that may result from the development and use of AI. The Framework will leverage existing standards and best practices that organizations can use to achieve stated outcomes and is worth highlighting as an important and robust U.S. effort related to implementing this strategic aim in an update to the 2022 Strategic Plan.

Despite these recent developments, the field of AI standards remains nascent, and, in order to ensure the ecosystem continues to evolve, it is important that the United States government continues to support industry participation in international standards organizations. Specifically, we suggest a focus on ensuring participation in ISO/IEC JTC 1 SC 42, where standards are being developed on many aspects of AI. This direction of development and adoption of AI standards and benchmarks is even more critical for emerging applications, such as AI for climate change and AI for drug discovery. Incentivizing worldwide efforts in this direction will help AI transform the solution space for these critical challenges, in terms not only of costs and speed of discovery, but also of expanding the search space and discovering new modes of actions. We offer additional thoughts on international cooperation in Strategy 9, where we make the point that the U.S. Government should increase international coordination on AI-related R&D.

Strategy 7) Better understand national AI R&D workforce needs

Attaining the AI R&D goals outlined in the 2019 Strategy will require a well-trained AI R&D workforce capable of fueling innovation over the long run. In line with our 2021 NAIRR RFI response, IBM believes that democratizing AI R&D is critical to broadening and diversifying the AI workforce. We recommend updates to the Strategy in two areas:

A) Build a Pipeline of Human Capital**:** IBM strongly supports the National Artificial Intelligence Initiative, which created grants at the National Science Foundation for AI research, including workforce training and career and technical education programs and activities, undergraduate, graduate, and postdoctoral education, and informal education opportunities.

Updates to the Strategy should ensure that such grants and programs are aligned with the NAIRR – which will be a lynchpin in the 'democratization' of AI R&D to underserved regions and user groups. Once the physical infrastructure of the NAIRR is in place, researchers will face the challenge of developing the skills and competencies needed to adopt and use the resource to accelerate their AI research.

Further, updates to the 2022 Strategic Plan should incorporate the collection and analysis of data to forecast the skills and competencies that will be required to operate and use the NAIRR, and, more broadly, fuel a broader range of translational AI R&D. Government should work to better incentivize employers to improve their data collection methods, consolidate existing public workforce datasets, and support the creation of more modernized labor market databases that track in-demand AI skills. This analysis could then be matched with educational pathways and potential retraining opportunities to ameliorate the long-term growth of the AI R&D workforce.

B) Promote AI R&D at Regional Hubs: IBM strongly supports the establishment and strengthening of regional hubs to promulgate the development and deployment of emerging technologies such as AI, which would in turn advance workforce training, economic opportunity, and representation. As outlined in Strategy 1, IBM supports the extension of the National AI Research Institutes Program alongside regional hubs to ensure that more Americans participate in AI R&D. Regional hubs should also be a locus for the development and deployment of AI testbeds outlined in Strategy 5.

Strategy 8) Expand Public-Private Partnerships to accelerate advances in AI

Public-private partnerships can advance AI R&D, democratize access, and equip an AI-ready workforce. At a technical level, public-private partnerships should focus on the pooling of resources, including facilities, datasets, and expertise, to advance science and engineering innovations. For example, efforts to pool and share resources through the NAIRR offer a path for leveraging the unique capabilities of the public, private, and academic sectors to advance AI R&D.

As part of the effort to sustain long-term investments in fundamental AI research (Strategy 1), the 2022 R&D Strategic Plan should support the continued expansion of the National AI Research Institutes Program. The 2022 R&D Strategic Plan should acknowledge the critical importance of the Program, and support its continued expansion. To date, the Program has established eighteen Research Institutes that operate as hubs that promote translational research into how AI can be used to address a broad range of societal and technological challenges, including climate change, agricultural supply chain challenges, and cybersecurity.

IBM believes that public-private collaboration is key to generating translational applications that can address societal issues such as climate change, the future of work, and healthcare, especially in communities that have transitionally been underserved. AI R&D investments should be targeted at growing translational R&D that puts emerging AI technologies into immediate practice to address societal impacts. One specific example of such an existing collaboration is the IBM-Mila partnership, which is focusing on developing AI solutions for tackling overlooked global threats such as antimicrobial resistance (AMR). AMR causes 4.95 million deaths per year globally, with an economic impact of $55 billion in the U.S. alone.

Finally, for the 2022 update, IBM recommends adding a new strategic aim:

Strategy 9) Increase international collaboration to amplify the benefits of AI R&D

Given the rapid pace of AI innovation and the patchwork of R&D investment worldwide, updates to the Strategy should focus on boosting international collaboration amongst academic institutions, industry, and governments in like-minded partner countries. Aligning U.S. R&D priorities with investments made by partner countries will act as a force multiplier that complements domestic capabilities. In particular, international collaboration will amplify efforts to boost the AI workforce, help to resolve societal challenges, and ensure that more Americans benefit from opportunities afforded by AI R&D.

IBM supports increased U.S. participation in the Global Partnership on AI (GPAI), which assembles like-minded countries to engage in joint research and the tackling shared global challenges. Current IBM-GPAI initiatives include privacy-preserving technology, AI for climate change, responsible AI, and AI for drug discovery. This type of collaboration helps promote regulatory convergence, while also facilitating the flow of information between the U.S. and like-minded partners, which could in turn spur domestic innovation and job creation. Funding a coordinated international effort to tackle emerging global challenges like antimicrobial resistance or climate change will also reduce costs for each nation and accelerate the discovery cycle by encouraging development and adoption of open data and knowledge sharing platforms and AI benchmarks and standards.

Finally, updates to the Strategy should emphasize the creation of joint international research programs in AI across agencies. For example, by teaming NSF with corresponding agencies in the European Union to define research programs of joint interest in core and

applied AI, as well as in the context of large centers such as NSF AI Institutes. Establishing collaborative and reciprocal AI R&D centers should include scientists and industry participants working jointly to advance critical topics in AI related to contextual AI, trustworthy AI, foundation AI, AI engineering, and AI hardware. The creation of such institutes should be linked to regional innovation hubs outlined in Strategy 7.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

John Wright

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.

Subject: response to request for input on Machine learning strategic plan
Date: Saturday, March 5, 2022
From: John Wright
To: AI-RFI

Response to RFI in document reference 87 FR 5877 on an Update of the NaTonal ArTficial Intelligence Research and Development Strategic Plan.

From Dr. John C. Wright < [redacted] > on behalf of the Machine Learning SciDAC project: "Accelerating Radio Frequency Modeling Using Machine Learning".

Our comment applies to Strategies 5 and 7 in the current Strategic Plan.

Strategy 5: Develop shared public datasets and environments for AI training and testing.
We suggest that these public datasets are created for different disciplines. Further, that these datasets should be accompanied by models in a portable format such as ONNX. These datasets should also be curated especially those created by simulations for surrogate models.

Strategy 7: Better understand the national AI R&D workforce needs.
In our experience a combination of machine learning experts and domain experts is necessary to achieve good results in machine learning models, especially surrogate models intended to reproduce simulation models.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Kaiser Permanente (KP)

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.

KAISER PERMANENTE®

Kaiser Foundation Health Plan
Program Offices

March 4, 2022

ATTN: RFI Response: National Artificial Intelligence Research and Development Strategic Plan

The White House
Office of Science and Technology Policy
AI R&D RFI Response Team

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

Kaiser Permanente (KP) appreciates the opportunity to offer comments on the above-captioned request for information (RFI).[1] We applaud efforts by the White House Office of Science and Technology Policy (OSTP) to update the current *U.S. National AI Research and Development Strategic Plan* and consider incorporating the initiatives identified in the *National AI Initiative Act of 2020 (NAIIA).*

The Kaiser Permanente Medical Care Program is the largest private integrated health care delivery system in the U.S., delivering health care to over 12 million members in eight states and the District of Columbia[2] and is committed to providing the highest quality health care.

As a health care provider, we recognize the potential of artificial intelligence (AI) to help drive value in healthcare through targeted enhancements to patient care, better patient-provider encounters, and greater efficiency. AI can achieve these aims across seven key levers in most healthcare organizations: personalizing healthcare, reducing complexity, driving quality

[1] 87 Fed.Reg.5876

[2] Kaiser Permanente comprises Kaiser Foundation Health Plan, Inc. and its health plan subsidiaries outside California and Hawaii; the not-for-profit Kaiser Foundation Hospitals, which operates 39 hospitals and over 720 other clinical facilities; and the Permanente Medical Groups, self-governed physician group practices that exclusively contract with Kaiser Foundation Health Plan and its health plan subsidiaries to meet the health needs of Kaiser Permanente's members.

outcomes, reducing costs, speeding execution, fueling innovation, and fortifying trust with trustworthy AI.

Potential revisions to the strategic plan to reflect updated priorities related to AI R&D

We recommend that OSTP explicitly incorporate the seven goals and priorities defined in the current National AI Initiative (from the NAIIA) into the specific strategies of the new strategic plan and to define the baseline, processes, resources, metrics, and evaluation strategies for each of these new goals. This will ensure progress in achieving these goals can be accurately measured and documented.

We also recommend highlighting sector-specific priorities for investment, adoption, and use of AI. Examples of the priorities for incorporating the use of AI in health care are further described below.

The Strategic Plan should include a core set of AI policy principles including:

- Public Trust and Acceptance: Safety, accountability, accuracy, reliability, and security are key to building acceptance and maintaining trust.
- Data Quality and Governance: Data completeness, reliability, and governance practices are essential.
- Accountability: AI systems should enable accountability by involving human input and review across the entire design, implementation, and monitoring process.
- Safety, Security and Effective Performance: Predictable and replicable performance of AI systems is critical.
- Equity: AI systems should reflect policies and frameworks that minimize bias by ensuring equity, fairness, and inclusivity.
- Transparency and Explainability: AI systems must include evidence about the reasoning for all outputs. At the same time, transparency and explainability should not impinge on the economic value of intellectual property.
- Functional Frameworks: AI policy frameworks should focus on the functional aspects of AI that define nationally recognized, technology neutral standards and measures of AI outcomes.
- Responsible design: AI policy frameworks should promote the responsible design, development, deployment, use and evaluation of AI.
- National oversight frameworks: Oversight frameworks should be flexible, adaptable, and evidence-based.
- Appropriate incentives: The Federal government should expand funding and incentives to support development of new and innovative AI technologies.
- Workforce development: Federal initiatives should be aimed at creating a diverse, well-trained workforce to develop, implement and manage AI technologies.
- Additional strategic aim: Fundamental cognitive-oriented research is necessary to address many of the cross-cutting R&D foundations. Foundations include research in Knowledge Representation, continuation of the long-standing debate between symbolic and connectionist computational theories, then ultimately, the evolution from Narrow AI to Generalized AI.

Applicability of the eight original strategic objectives

The original eight strategic aims from the 2019 plan are still applicable. Specific to the healthcare sector, AI investments should be directed to improving affordability, equity, effectiveness, efficiency, transparency, and trust. We also propose modifications and additions to the current eight strategic aims, and suggest mapping of the NAIIA goals:

Strategy 1: Make short-term and long-term investments in AI research and development.

Align the following specific NAIIA sections to OSTP 2019 (as originally enumerated):

(A) Determine and prioritize areas of AI research, development, and demonstration requiring federal leadership and investment.

(B) Support long-term funding for interdisciplinary AI research, development, demonstration, and education.

(E) Provide or facilitate the necessary computing, networking, and data facilities for AI research and development.

We recommend additional federal investment to explore research avenues in *Federated Learning* and incorporate advanced cryptographic protections for data and privacy. It will also be important to research *Knowledge Brittleness*, inference on small data sets, and the concept of models that know what they do and do not know (resiliency).

We further suggest expanding R&D in *Intelligent Agents*, where agents can explore and discover data and knowledge for human partners, not just serve as assistants.

Lastly, we recommend researching the area of *Model Integration*, where multiple models are connected for higher level decisions (including the provenance of data and decisions by agents, objects, successful and failed executions).

Strategy 2: Develop effective methods for human-AI collaboration

No comments or recommended additions.

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

Align the following specific NAIIA section to OSTP 2019 (as originally enumerated):

(C) Support research and other activities on ethical, legal, environmental, safety, security, bias, and other appropriate societal issues related to AI.

We recommend promoting additional research on symbolic mechanisms that support machine-interpretable representation of legal knowledge and ethical principles, suitable for inference.

Strategy 4: Ensure the safety and security of AI systems.

This strategy should leverage and apply appropriate design principles (e.g., fault-tolerance) to AI systems from aviation, nuclear power plant, space and other operationally critical environments

and applications. In addition, we recommend continuing research into threats against AI systems from nefarious actors, especially nation states. Lastly, this strategy should address research advanced verification and validation methods for AI systems, testing for high availability and safety.

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

Align the following specific NAIIA section to OSTP 2019 (as originally enumerated):
(D) Provide or facilitate the availability of curated, standardized, secure, representative, aggregate, and privacy-protected data sets for AI research and development.

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

No comments or recommended additions.

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

Align the following specific NAIIA section to OSTP 2019 (as originally enumerated):
(F) support and coordinate federal education and workforce training activities related to AI.

Strategy 8: Expand Public-Private Partnership.

Align the following specific NAIIA section to OSTP 2019 (as originally enumerated):
(G) support and coordinate the network of AI research institutes.

Incorporating the seven new goals from the 2020 National AI Initiative Act

See our response on pages 1 and 2, above.

AI R&D focus areas to create solutions to address societal issues

We suggest AI R&D focus on the following areas and activities:

- Develop reliable metrics to assess the degree to which AI manages different types and forms of bias; AI transparency, reliability and reproducibility of its methods and results; and equity in it use.
- Use AI systems to predict and prevent inequities in the delivery of healthcare services.
- Use AI systems to improve affordability and access to healthcare services.
- Impact of privacy in the development and use of AI systems.
- Advance AI systems designs ranging from deployment on neuromorphic chips to quantum-based platforms, with emphasis on algorithms, architecture, end-to-end cybersecurity and privacy. In addition, focus on AI systems to support home-based healthcare for our aging, veteran, and handicapped populations (especially in mental health areas).
- Develop tools and resources to identify these issues and to modify AI systems as appropriate.

Addressing issues related to equity and disparate treatment

AI R&D should strive for appropriate and balanced recruitment practices to include underrepresented populations (race, ethnicity, SES, other). We also recommend monitoring AI risks and issues across different categories to mitigate harms due to disparate treatments of different demographic groups. These categories include underrepresentation of certain groups, general bias in existing data, lack of experienced AI talent, inadequate governance over AI applications, and non-transparent or "black box" algorithms that are harder to explain and harder to justify why a model is performing a specific way, possibly impacting protected classes differently.

Trustworthy AI is an example of an industry leading methodology to prevent unfair or bias algorithms. Trustworthy AI usually includes the following areas: *Fair AI, Explainable AI, Accountable AI, Robust and reliable AI, Privacy, and Security*.

R&D in AI model governance along the *Cross-Industry Standard for Data Mining* (CRISP-DM) AI model development lifecycle is crucial to address bias, equity, or concerns related to the development, use, and impact of AI. Federal standards regarding best practices at each of the stages should align to CRISP-DM.

* * *

Kaiser Permanente appreciates OSTP's consideration of our comments and recommendations for updating the *National Artificial Intelligence Research and Development Strategic Plan*. We would be pleased to provide additional information or answer any questions.

Sincerely,

Jamie Ferguson
Vice President, Health IT Strategy and Policy
Kaiser Foundation Health Plan, Inc.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Kapoor/Kshirsagar/Barocas/Arvind,
Princeton University

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.

PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY
PRINCETON UNIVERSITY

March 4, 2022

White House Office of Science and Technology Policy,
NCO,
2415 Eisenhower Avenue,
Alexandria, VA 22314

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

Thank you for the opportunity to respond to the Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan ("Strategic Plan"). We are academic researchers associated with the Center for Information Technology Policy (CITP) at Princeton University,[1] Microsoft Research, and Cornell University, and write to provide suggestions for how the Strategic Plan can focus resources to address societal issues such as equity, especially in communities that have been traditionally underserved. We also discuss how AI R&D can support research that informs the intersection of AI R&D and its application with privacy and civil liberties.

1. Strategy 1: Sustaining long-term investments in fundamental AI research requires supporting research on its impact on equity.

The 2019 Update and the original Strategic Plan rightly emphasize the importance of sustaining long-term investments in fundamental AI research. One core area for support that the Strategic Plan highlights is investments to advance trust in AI systems, which includes requirements for robustness, fairness, explainability, and security. This area of research has only become more important to sustain as AI systems have become embedded in public life. But, we suggest, the Strategic Plan should also explicitly include a commitment to making investments in research that examines how AI systems can affect the equitable distribution of

[1] In keeping with Princeton's tradition of service, CITP's Technology Policy Clinic provides nonpartisan research, analysis, and commentary to policy makers, industry participants, journalists, and the public. This response is a product of that Clinic and reflects the independent views of the undersigned scholars.

resources. Specifically, there is a risk that without such a commitment, we make investments in AI research that can marginalize communities that are disadvantaged. Or, even in cases where there is no direct harm to a community, the research support focuses on classes of problems that benefit the already advantaged communities, rather than problems facing disadvantaged communities.

We recommend that the Strategic Plan outline a mechanism for a broader impact review when funding AI research. As we argue below (Section 2), existing institutional mechanisms for ethics review of research projects do not adequately identify downstream harms stemming from AI applications. When deciding where to invest resources, the government and its funding bodies should take into account not only the potential positive impacts of research, but the potential negative impacts as well. The Strategic Plan should include mechanisms that take advantage of the government's unique position to steer the research community away from research questions that pose obvious risks of downstream harm without any clear benefits, such as the many phrenology-like studies in computer vision that have generated recent controversy.[2]

Because AI research can sometimes result in rather general knowledge or techniques with a broad range of potential applications, it may be challenging to determine what kind of impact it might have. In fact, many AI research findings will have dual use: some applications of these findings may promise exciting benefits, while others would seem likely to cause harm. While it is worthwhile to weigh these costs and benefits, decisions about where to invest resources should also depend on distributional considerations: who are the people likely to suffer these costs and who are those who will enjoy the benefits? Research should not only have a positive broader impact; its benefits should be distributed equitably. In fact, even research that only seems to have a positive upside should be assessed with distributional concerns in mind to ensure that the benefits don't accrue primarily to those who are already advantaged in society. While there have been recent efforts to incorporate ethics review into the publishing processes of the AI research community,[3] [4] adding similar considerations to the Strategic Plan would help to highlight these concerns much earlier in the research process. Evaluating research proposals according to these broader impacts would help to ensure that

[2] Luke Start and Jevan Hutson. "Physiognomic Artificial Intelligence." Fordham Intellectual Property, Media & Entertainment Law Journal, 2021.
[3] Brent Hecht et al. "It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process". ACM Future of Computing Blog, 2018.
[4] Priyanka Nanayakkara, Jessica Hullman, Nicholas Diakopoulos. "Unpacking the Expressed Consequences of AI Research in Broader Impact Statements." AIES, 2021.

ethical and societal considerations are incorporated from the beginning of a research project, instead of remaining an afterthought.

2. Prioritize research on the downstream implications of AI research and applications under Strategy 3 of the Strategic Plan.

The Strategic Plan correctly focuses on supporting research that designs architectures for ethical AI. But, on privacy issues, ethical AI has sometimes been framed incorrectly as merely concerning the data collection and management process.[5] We suggest that a larger threat comes from the downstream impacts of AI applications such as face recognition,[6] workplace surveillance,[7] and behavioral advertising.[8]

The current Strategic Plan focuses on two notions of privacy: (i) ensuring the privacy of data collected for creating models via strict access controls, and (ii) ensuring the privacy of the data and information used to create models via differential privacy when the models are shared publicly. Both of these approaches are focused on the privacy of the people whose data has been collected to facilitate the research process, not the people to whom research findings might be applied. Take, for example, the potential impact of face recognition for detecting ethnic minorities.[9] Even if the researchers who developed such techniques had obtained approval from the IRB for their research plan, secured the informed consent of participants, applied strict access control to the data, and ensured that the model was differentially private, the resulting model could still be used without restriction for surveillance of entire populations,[10] especially as institutional mechanisms for ethics review such as IRBs do not consider downstream harms during their appraisal of research projects.[11]

While it is critically important to protect the privacy of the people whose data are being used in the research process, such protections do nothing to ensure

---

[5] Vinay Uday Prabhu and Abeba Birhane. "Large image datasets: A pyrrhic win for computer vision?" arXiv preprint arXiv:2006.16923, 2020.
[6] Antoaneta Roussi. "Resisting the rise of facial recognition." Nature news feature, 2020.
[7] Kyle Wiggers. "Workplace surveillance algorithms need to be regulated before it's too late." VentureBeat, 2021.
[8] Charles Duhigg. "How Companies Learn Your Secrets." New York Times, 2012.
[9] Richard Van Noorden. "The ethical questions that haunt facial-recognition research." Nature News Feature, 2020.
[10] Solon Barocas and Helen Nissenbaum. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, Eds. Cambridge University Press, NY, 2014.
[11] Jacob Metcalf. 2017. "The study has been approved by the IRB': Gayface AI, research hype and the pervasive data ethics gap." Pervade Team.

that the resulting discoveries do not threaten other people's privacy. Even if the data used for creating the models is stored privately, the models created using this data can still be used for privacy-breaching inferences. In fact, even if the data that was collected for training the AI model is later deleted, the models trained using this data can still be used for such inferences. And models that are differentially private are just as good at privacy-breaching inferences as those that are not differentially private.

The Strategic Plan must therefore grapple with the fact that AI applications are a powerful tool for privacy-breaching inferences—even when the underlying research has taken the privacy interests of research subjects into account. We recommend that the Strategic Plan include as a research priority supporting the development of alternative institutional mechanisms to detect and mitigate the potentially negative downstream effects of AI systems. In addition, we recommend that the Strategic Plan include provisions for funding research that would help us understand the impact of AI systems on communities, and how AI systems are used in practice. Such research can also provide a framework for informing decisions on which research questions and AI applications are too harmful to pursue and fund.

3. Prioritize systematic studies of reproducibility under Strategies 5 and 6.

Many studies that purport to rely on AI have results that are overly optimistic and lack reproducibility.[12] Indeed, we found 18 reviews across 15 scientific fields that find errors in a total of 304 papers that use ML-based science (*see* Figure 1 below). Given the adoption of ML methods across scientific fields, there is an urgent need to address reproducibility issues in ML-based science. But there are challenges in creating the incentives for researchers to independently and rigorously examine scientific claims that the Strategic Plan can help overcome.

Evaluating academic claims about machine learning is challenging. First, the code tends to be complex and lacks standardization, which makes it difficult to understand and replicate models. Second, there are subtle pitfalls for researchers who fail to differentiate between explanatory and predictive modeling. Third, the hype and overoptimism about commercial AI often spills over into machine learning research and obscures the findings.[13] All these, of course, are in addition

[12] Sayash Kapoor and Arvind Narayanan. 2021. "(Ir)reproducible Machine Learning: A Case Study." Preprint available at reproducible.cs.princeton.edu.
[13] Joelle Pineau et al. 2020. "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)." arXiv preprint arXiv:2003.12206.

to the pressures and publication biases present in all disciplines that have led to reproducibility crises.

Systematic reviews have started to identify reproducibility issues and overoptimistic results in many academic fields that are adopting machine learning methods. But this is complex and expensive work. One estimate suggests that we spend over $28 billion a year on preclinical research in the United States that is not reproducible.[14] As machine learning methods spread across academic fields, focusing on the reproducibility of that research is critical to ensure its validity.

One of the major roadblocks to reproducibility research is that appropriate computing resources are difficult to secure. While researchers can rely on cloud services such as Amazon AWS, Google Cloud and Microsoft Azure for compute-intensive AI research, there are fewer resources available for those seeking to vet claims of performance. This problem has intensified with the shift of private firms undertaking research into new AI models. For example, natural language processing models routinely require large amounts of computational resources. But the cost of computational resources to replicate performance claims are often beyond the reach of independent researchers at research universities. This further makes the reproducibility of research output by private companies inaccessible due to issues with data sharing and lack of access to computational infrastructure.

We recommend that the Strategic Plan prioritizes the support of systematic reviews of published research across fields adopting machine learning methods to address the reproducibility crisis in ML-based science. The Strategic Plan could also incentivize work on the creation of computational reproducibility infrastructure and a reproducibility clearinghouse that sets up benchmark datasets for measuring progress in scientific research that uses AI and ML.[15] Finally, the Strategic Plan could make government funding conditional on disclosing research materials, such as the code and data, that would be necessary to replicate a study. A similar step is already underway for NIH funded studies.[16] Taken together, these steps would lead to significant strides towards the aim of promoting transparent, effective, and responsible research.

---

[14] Leonard P. Freedman , Iain M. Cockburn, Timothy S. Simcoe. 2015. "The Economics of Reproducibility in Preclinical Research." PLoS Biology 13(6).

[15] Benjamin Haibe-Kains et al. 2020. "Transparency and reproducibility in artificial intelligence." Nature 586, E14–E16.

[16] "Final NIH Policy for Data Management and Sharing". Notice Number: NOT-OD-21-013. https: /grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

Field	Paper	Year	Num. papers reviewed	Num. papers w/pitfalls	Pitfalls
Medicine	Bouwmeester et al.	2012	71	27	No train-test split
Neuroimaging	Whelan et al.	2014	—	14	No train-test split; Feature selection on train and test set
Autism Diagnostics	Bone et al.	2015	—	3	Duplicates across train-test split; Sampling bias
Bioinformatics	Blagus et al.	2015	—	6	Pre-processing on train and test sets together
Nutrition research	Ivanescu et al.	2016	—	4	No train-test split
Software engineering	Tu et al.	2018	58	11	Temporal leakage
Toxicology	Alves et al.	2019	—	1	Duplicates across train-test split
Satelitte imaging	Nalepa et al.	2019	17	17	Non-independence between train and test sets
Clinical epidemiology	Christodoulou et al.	2019	71	48	Feature selection on train and test set
Brain-computer interfaces	Nakanishi et al.	2020	—	1	No train-test split
Histopathology	Oner et al.	2020	—	1	Non independence between train and test sets
Computer security	Arp et al.	2020	30	30	No train-test split; Pre-processing on train and test sets together; Illegitimate features; others
Neuropsychiatry	Poldrack et al.	2020	100	53	No train-test split; pre-processing on train and test sets together
Medicine	Vandewiele et al.	2021	24	21	Feature selection on train-test sets; Non-independence between train and test sets; Sampling bias
Radiology	Roberts et al.	2021	62	62	No train-test split; duplicates in train and test sets; sampling bias
IT Operations	Lyu et al.	2021	9	3	Temporal leakage
Medicine	Filho et al.	2021	—	1	Illegitimate features
Neuropsychiatry	Shim et al.	2021	—	1	Feature selection on training and test sets

Figure 1 [from Kapoor and Narayanan]: a list of systematic reviews that highlight overoptimism and irreproducibility in applied machine learning research across academic fields.

4. Build and maintain infrastructure designed to independently test the validity of the claims of AI performance across applications under Strategy 6.

Recently, the industry has converged on a troubling and widespread practice that applies the label of AI to applications that do not and cannot work. We dub this phenomenon of using a veneer of AI to lend credibility to pseudoscience as *AI snake oil*. The proliferation of AI snake oil in such applications is a distinct issue from concerns around bias, but is a major contributor to the negative consequences that result.

AI-based research has led to genuine and rapid progress in many domains, but it is important to distinguish between the classes of problems where AI tools have been shown to be effective. For example, AI has made significant progress in aiding with perception tasks, but it has struggled to predict outcomes involving complex social phenomena. Applications that claim to predict social outcomes but in fact do not have any predictive power are unfair even if they are technically unbiased, since they mask the fact that they do not work as promised and end up perpetuating outcomes that differ from their stated purpose. This is especially true when such applications dictate important life outcomes.

As an example, consider the AI tools that are purportedly designed to automate hiring decisions. The main claim made by many companies producing these tools is that AI can analyze body language and speech patterns to determine candidates' personality traits or competencies from short video interviews and function as "algorithmic pre-employment assessments" to make hiring decisions easier. But it is generally understood by experts that these tools have significant shortcomings when it comes to predicting actual job performance. Nevertheless, Raghavan et al. describe how 18 companies working on algorithmic hiring systems have collectively raised over $200 million in funding over the last few years, though not all of these companies offer AI assessments of job candidates.[17]

Similar claims prevail in a large number of applications where AI systems are claimed to predict social outcomes such as the likelihood of recidivism or identifying at-risk kids. But recent research shows that AI systems today are no better than simple rules at predicting social outcomes.[18] [19] However, this does not

[17] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." ACM Conference on Fairness, Accountability, and Transparency.
[18] Matthew J. Salganik et al. 2020. "Measuring the predictability of life outcomes with a scientific mass collaboration." Proceedings of the National Academy of Sciences 117 (15).
[19] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism." Science advances, 2018.

stop companies from marketing AI-based systems that claim to solve these problems, and as a result industrial applications of AI that purportedly predict social outcomes are proliferating. This phenomenon has a further pernicious effect of fueling the hunger for personal data for these fundamentally dubious applications of AI and giving rise to "black box" algorithms that cannot be explained. Furthermore, these applications tend to distract attention from designing more effective interventions to address these important social issues.

As a result, we see evaluating validity as a core component of ethical and responsible AI research and development. The strategic plan could support such efforts by prioritizing funding for setting standards for and making tools available to independent researchers to validate claims of effectiveness of AI applications.

5. Incentivize and promote effective data stewardship under Strategy 5.

The creation of datasets has been pivotal in the development of AI applications. But there is an underexplored dark side to supporting the broad release of datasets without mechanisms of oversight or accountability for how that information can be used. Such datasets raise serious privacy concerns and they may be used to support research that is counter to the intent of the people who have contributed to them. The Strategic Plan can play a pivotal role in mitigating these harms by establishing and supporting appropriate data stewardship models.

Consider the challenge of "runaway datasets" as an example of a problem that the Strategic Plan might address. In the last few years, many datasets have been retracted due to ethical concerns. But our research has documented how, even after retraction, these datasets can remain widely available and are used across the industry and in research labs.[20] This phenomenon has been dubbed the problem of "runaway datasets." Of course, the ethical issues that caused the researchers to retract the original dataset persists in AI applications that continue to use these datasets after retraction. This highlights the necessity of dealing with ethical issues throughout the lifecycle of the dataset instead of addressing ethical issues only when the dataset is released.

In the same vein as our point about downstream impacts (Section 2), existing ethical oversight mechanisms within academia such as IRBs are poorly suited to deal with runaway datasets. "Human subjects research" has a narrow definition in the context of IRBs and thus many of the datasets and associated research that have caused ethical concern in machine learning would not fall

[20] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. "Mitigating dataset harms requires stewardship: Lessons from 1000 papers." NeurIPS 2021 (Datasets and Benchmarks track).

under the purview of IRBs. This compounds issues with runaway datasets and exacerbates ethical concerns with the creation and use of datasets.

The Strategic Plan can address this gap by supporting the development of centralized data clearinghouses to regulate access to datasets. Such clearinghouses could include safeguards for monitoring ethical concerns through the lifecycle of the use of the datasets. Finally, the Strategic Plan could establish mechanisms for exercising responsible data stewardship that can make decisions about the ethical uses of datasets at the time they are being created and while they are in use. While some research projects already follow such a procedure when releasing datasets, institutional support including providing funding towards data stewardship committees would help reduce the ethical risks of AI applications due to runaway datasets.[21]

* * *

We appreciate the opportunity to provide these comments and welcome the opportunity to discuss any questions.

Respectfully submitted,

Sayash Kapoor
Graduate Student, Department of Computer Science, Princeton University

Mihir Kshirsagar
Technology Policy Clinic Lead, Center for Information Technology Policy, Princeton University

Solon Barocas
Principal Researcher, Microsoft Research and Adjunct Assistant Professor, Department of Information Science, Cornell University

Arvind Narayanan
Associate Professor of Computer Science, Princeton University

[21] Ian Lundberg, Arvind Narayanan, Karen Levy, and Matthew J. Salganik. 2018. "Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge." Socius, 5.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Li/McGovern/Diochnos/Ebert,
University of Oklahoma

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.

Feedback on “THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE”

Yifu Li, Amy McGovern, Dimitris Diochnos, David Ebert
University of Oklahoma

Introduction

At the University of Oklahoma, we have been performing fundamental research on AI, grand challenge problem-driven AI research, and deploying AI-enabled solutions to address societal challenges. Below we suggest improvements and updates to the National Artificial Intelligence Research and Development Strategic Plan to incorporate new innovations, challenges, and needs that have emerged over the past four years for the successful research, development, and deployment of impactful AI.

Strategy 2: Develop Effective Methods for Human-AI Collaboration

One of the key, emerging areas of research that needs further advancement in creating, robust, adaptable, responsive, trustable AI solutions are **advances in interactive, human-guided machine learning and AI systems.** The coupling of human expertise, contextual information, and additional relevant nondigital has the promise of creating more accurate, robust, context-relevant AI solutions to solve problems in ever changing environments with contexts that have never been previously experienced (e.g., unprecedented weather, environmental, and growing conditions).

Expanding this strategy or creating a new strategy specifically focusing on **trustable, transparent/understandable AI** will create AI advances that are readily adopted, deployable to more situations, and enable AI to have a broader impact on our sociotechnical society.

Strategy 4: Ensure the Safety and Security of AI Systems

It is crucial to look at the interaction between adversarial settings and emerging aspects of machine learning, such as transparency and fairness. Transparency, trust, and explainability are mentioned in isolation in Section 4, but in Section 3, we see that they are blended together under "Improving fairness, transparency, and accountability by design." Research is needed to investigate tradeoffs not only among these notions but also about these notions in an adversarial setting and discuss what can be achieved by design.

Additionally, **in terms of verification and validation, we must include not only malfunction, but also fairness, transparency, and their impact on the precision, recall, and deployability of these systems.** Often, as we make a system more transparent or fair (by design), we inherently impose certain limitations on achieving certain goals at the end of the learning process. For example, how would/could

transparency-design and fairness-design decisions impact complex performance measures (e.g., precision, recall, upon deployment of these mechanisms?

Strategy 5: Develop Shared Public Datasets and Environments for AI Training and Testing - Advanced-Data Ecosystem for Improved Adaptability and Availability under the Community Contribution

Thanks to the rapid development of data collection, there has been a dramatic increase in data sources and information availability. With the increasing amount and the variety of available datasets, expert-led data analysis faces challenges in promptly evaluating the quality of raw information sources. Specifically, we need sufficient time and domain expertise to perform data capture, curation, and analysis for potentially valuable datasets under an enormously data-rich environment.

We have seen the great potential of creating dynamic and agile data repositories through the help of the data science community and standardized data storage language/format. Under the government-led efforts, we encourage **building an open-access repository paradigm** facilitated by a collection of moderators for each area to review the datasets submissions and oversee the sharing policy. The moderators can help recategorize the datasets with different values, assign categories, or reject their upload for inappropriate information. A comprehensive list of moderators should consist of government officials, the AI community, and the corresponding experts from a variety of domains.

To further improve the exposure and the accessibility of the open-access repository. We recommend **improving the availability of open-source AI software libraries** that can easily access such a repository (both for uploading and downloading). For example, we have seen AI software, including but not limited to OpenNLP, Sckit-Learn, Weka have been gaining a significantly larger AI user community over the past years. As a result, we can provide high-level programming language-based development toolkits to bridge the gap between the above-mentioned open-source software and the open-access repository to encourage the repository's further development.

Finally, we need to ensure that these resources are available to all and **democratize access to AI data, software, and hardware.**

Strategy 6: Measure and Evaluate AI Technologies through Standards and Benchmarks - Integrating the Domain Expertise and Expectations for AI standards and Benchmarks

It is important to promote the participation of both the industry and academia in measuring and evaluating AI technologies as they are the primary sources that generate and benefit from AI technologies. On examining the effectiveness of AI solutions, **there is a lack of a balanced participant pool from both the AI technology developers and users, especially in ensuring diversity, equity, inclusivity, and justice.** Although researchers from academia and industry R&D organizations have the motivations and incentives from directly engaging in AI solution development, other sides of the AI community, who can benefit from AI, but lack the background knowledge in this domain, are easily ignored. The practitioners carry critical domain expertise on the testbed of AI,

and their expectations can play a major role in developing AI technologies. As a result, it is crucial to broaden the education of AI to a variety of industries and encourage the AI community to **further engage in standards development for evaluating AI technologies**. Furthermore, it is even more crucial to bridge the gap between the expectation from practitioners and AI researchers to achieve a harmonious development cycle between AI technology developers and users.

Strategy 7: Better Understand the National AI R&D Workforce Needs

- Strengthen the Involvement of the Research Community Throughout the Studies

This report has identified the urgent need for additional studies on the need for AI R&D from the current and future US workforce to prepare the AI community's future, such as the educational pathways and retraining opportunities. It is important to emphasize the accuracy of such studies by strengthening the involvement of the research community during the study and **creating national accreditation and educational standards programs.**

There is a diverse industry workforce that can become the cornerstone of AI technology development. **We need to ensure tight partnerships and industry engagement in our educational program development, review, and improvement to ensure that academia is producing AI graduates that meet industry needs.** We also need to ensure that we create AI training across all fields of study since data-driven and AI-driven technology and solutions will impact all aspects of the workplace. However, it is difficult to create a continual retraining and upskilling environment to keep the workforce current with the research frontier of AI technologies and ensure they are fully aware of the capability of the rapidly advancing AI technologies. Enhancing the research community's involvement and strengthening the industry workforces' understanding of AI technology will ensure that the study result accurately reflects and understands the national AI R&D workforce needs. Similar problems exist in cybersecurity and medical fields, and adapting and modifying continual training programs based on their best practices will ensure the best AI workforce. With a better understanding of the current and future AI R&D workforce needs, we can develop more tangible plans and follow-up actions to help alleviate the challenges of the US workforce.

Conclusion

The report provided by the select committee on the artificial intelligence of the national science and technology council offered clear and tractable expectations of the federal AI research and development. Building upon the original report, we have provided several directions to help the United States continuously serve as the world leader on the AI knowledge frontier. We believe that such efforts will further strengthen our nation's security, economic growth, and our citizens' life quality.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Massachusetts Institute of Technology (MIT)

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.

Massachusetts Institute of Technology

Response to RFI on

National Artificial Intelligence Research and Development Strategic Plan

March 3, 2022

For Further Information Contact: Prof. Daniela Rus, Director, MIT Computer Science and Artificial Intelligence Laboratory; or David Goldston, Director, MIT Washington Office

MIT appreciates the opportunity to provide input on the update of the National Artificial Intelligence Research and Development Strategic Plan. If properly pursued, AI has the potential to improve the economy, education, and our daily lives. It could help address fundamental problems from reversing climate change to curing disease. At the same time, AI's deployment raises critical questions of ethics and equity. It is the essential that the national AI strategy recognize the need for additional research funding across a wide range of AI topics to further enhance AI while simultaneously addressing its societal implications.

While this submission will focus on three specific issues, we want to make clear at the outset the wide range of work on AI that is needed.

We mention just a few of the key challenges here. One involves data. Today's AI methods require massive data sets that must be manually labeled and are not easily obtained in every field. The quality of that data needs to be very high, and it needs to include critical corner cases for the application at hand. As we gather more data to feed into these AI systems, the risks to privacy will grow. Also, today's AI systems are black boxes -- there is no way for users of the systems to truly learn anything based on the AI system's workings or to fully understand how the systems have reached their conclusions. This makes it difficult to detect behavior that could undermine security or safety, and it means we cannot anticipate failure modes tied to rare inputs that could lead to potentially catastrophic consequences.

It is critically important to advance AI and computer science technologies more broadly with a focus on reliability and robustness, privacy and security, transparency, explainability and accountability.

We highlight three important areas of investment for AI.

<u>Algorithms for Causal and Compact AI models</u>

While neural networks can learn representations very well, it is notoriously difficult to understand how they come up with decisions. This impedes their deployment in high-stakes applications such as embodied autonomy. A large body of research focuses on designing

methodologies to explain the "black box," hoping to alleviate this limitation to at least some extent.

However, no existing approaches are scalable and trustworthy. An important area for research investment is the development of novel machine learning algorithms that would possess interpretable skills inherently. This requires fresh algorithmic approaches to machine learning that could enable sustainable and effective closed-loop deployments for real-world, safety-critical applications.

Today's deep learning technologies – whatever their beneficial capabilities – are growing uncontrollably in size while leaving us with fundamental sociotechnical challenges such as causality, interpretability, fairness, accountability, out-of-distribution generalizability, and carbon footprint. It is important to rethink the algorithmic design choices for AI learning systems and to develop new AI systems through interact-to-understand mechanisms, interpretability through causal manipulation, and closed-form decision-making.

Auditing and monitoring of AI systems

There is a growing consensus that methods must be developed to "audit" and monitor AI systems. A cornerstone of the European Union's AI legislation is the ex-ante auditing and the ex-post monitoring of "high-risk" AI systems. Moreover, in the joint statement released last September at the U.S.-EU Trade and Technology Council summit, the parties agreed to discuss "measurement and evaluation tools and activities to assess the technical requirements for trustworthy AI, concerning, for example, accuracy and bias mitigation." In short, AI systems will need to be properly audited and regularly monitored to identify and mitigate risks, both technical (e.g.., accuracy, reliability, robustness) and socio-technical (e.g., bias, privacy).

Yet there are many unresolved technical questions about how to effectively audit and monitor AI systems, and the scalability of auditing is also emerging as a significant practical challenge. As AI systems proliferate and find their way into more realms of human activity, how will it be feasible to conduct thousands of audits in multiple domains? The AI strategy plan needs to recognize and address these challenges head-on.

Creating effective auditing and monitoring capability at the needed scale will require developing scalable auditing techniques, training enough people to carry them out, and building up institutional capacity in government and industry to undertake, oversee and respond to audits.

Some current trends complicate the situation further, particularly the increasing outsourcing of AI deployment and the emergence of a two-stage architecture in which very large "core" or "foundational" models are pre-trained and then simpler models are fine-tuned for specific tasks. These trends raise questions about how to appropriately distribute auditing responsibilities and capabilities between the outsourced development of core models and the deployment (fine-tuning) of those models.

Policymakers will need to define "control points" (the point at which auditing occurs in an AI system) and will have to identify systemic risks. Outsourcing and the use of two-stage architecture can add to those risks by increasing reliance on a small number of AI architectures.

Applications of AI for drug discovery

AI has the potential to considerably improve the drug discovery process, speeding the discovery of new modes of action, and reducing costs. The pharmaceutical industry is beginning to recognize this, and has invested $5 billion in AI-related funding deals in the last several years (https://www.rootsanalysis.com/reports/ai-based-drug-discovery-market.html).

However, industry lacks the financial incentives to adequately address many key threats to public health, including the development of future pandemics, and antimicrobial resistance. (Antimicrobial resistance is already costing the U.S. economy billions of dollars, and we could see a 10-fold increase in problem in the next two decades. See https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(22)00087-3/fulltext.)

The AI strategic plan should recognize the need for major government investment in developing new ways to apply AI to drug discovery. It should also recognize the need to ensure that the public and private sectors are deploying that capability in a way that focuses sufficiently on key public health challenges. That will entail changing incentives so that the private sector will focus more on societal needs.

The government might explore approaches like those being tried in the United Kingdom to increase the development and production of antibiotics. The government could also experiment with funding mechanisms for private sector research that would include open-science and open-data contractual clauses. That would enable more sharing of data with others in academia and industry. More international efforts are needed to address public health threats, as well as purely domestic ones.

Work on these problems will require more cross-disciplinary work because we need to learn more about both technology *and* about ourselves. Such work needs to be supported across many federal agencies, but could be a particular focus of the Advanced Research Project Agency-Health, if it is established.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Matias/Wright, Cornell University

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.

Innovation & Governance of Adaptive AI with Community Science

J. Nathan Matias & Lucas Wright, Citizens and Technology Lab, Cornell University

Actor Jennifer Lawrence was caught in an escalating cycle of human-AI behavior in 2014 when her intimate photos were stolen and circulated on the social platform Reddit. As people clicked and up-voted the pictures, Reddit's algorithms showed them to even more people, encouraging the algorithm further. When asked about the experience, Lawrence told Vanity Fair, "It just makes me feel like a piece of meat that's being passed around for a profit."[1]

Because the 2019 U.S. National R&D Strategy on AI doesn't include the kind of AI systems that caused Lawrence and others so much harm, the U.S. risks being unprepared to lead safe innovation or reliably govern escalating catastrophes of human and AI behavior. Feedback happens when humans and adaptive algorithms react to each other in ways that change algorithm behavior without further involvement from engineers. And this feedback is everywhere—directing law enforcement, managing financial systems, shaping our cultures, coordinating remarkable generosity, and shaping democratic participation. Yet according to Meta's President of Global Affairs, even this leader on artificial intelligence is unable to reliably predict or prevent catastrophes of human and algorithm behavior (Clegg 2021), a view that many scientists agree with (Bak-Coleman et al 2021).

This comment has bearing on Strategy 3 on the ethical, legal, and social implications of AI, Strategy 4 on the safety and security of AI, and Strategy 8 on public-private partnerships.

We are researchers at the Citizens and Technology Lab at Cornell University (CAT Lab), who conduct citizen / community science on our digital environments. CAT Lab works alongside the public for a world where digital power is guided by evidence and accountable to the public, inspired by the history of community science on food safety, consumer protection, and the environment. Throughout the history of the U.S., industry-independent testing and accountability from community groups, academic researchers, and government agencies has contributed significantly to the parallel growth of industry and public safety.[2] CAT Lab conducts research and development to create the science, policy, and community capacities for similar progress in artificial intelligence and digital technology.

This week, we published a new article with the Social Science Research Council[3] about impact assessment of human-algorithm feedback, with recommendations for scientists, policymakers, and communities. In this note, we summarize key recommendations for the U.S. National AI R&D Strategic Plan. We also attach the longer article for your reference. Thank you for this opportunity to provide input.

---

[1] Kashner, Sam. 2014. "Exclusive: Jennifer Lawrence Speaks About Her Stolen Photos." *Vanity Fair*, October 20, 2014.

[2] Carpenter, D. (2014). *Reputation and power*. Princeton University Press.

[3] Matias, Nathan and Lucas Wright. "Impact Assessment of Human-Algorithm Feedback Loops." Just Tech. Social Science Research Council. March 1, 2022. DOI: https://doi.org/10.35650/JT.3028.d.2022

Develop a strategy for adaptive AI

Proposals for regulating AI often focus on the bias and accuracy of decisions made by algorithms. Bias is a useful concept for evaluating judgments that need to be impartial and independent (Barocas et al 2019). Unlike decision making systems, adaptive algorithms are not intended to make impartial, consistent decisions every time, regardless of context. Instead, adaptive systems designed for social media content recommendations, predictive policing, financial trading, and route-mapping are continuously changing their behavior. Consequently, static data evaluations cannot protect people from runaway feedback loops between humans and these algorithms (Ekstrand et al 2022; Lucherini et al, 2021).

Support involvement from affected communities at all levels of AI strategy and policy

Much of the most influential research and policy work on AI policy has come directly from affected communities, despite substantial attempts from technology operators to hinder independent investigation and oversight (Charles 2020; Cox 2017). One promising model for improving AI safety and equity is to conduct impact assessments, a model from environment management that includes affected communities in risk and benefit assessment in advance of system development and introduction (Moss et al 2021; Reisman et al 2018).

The people affected by an issue have the most at stake and the greatest understanding of the context, making them essential conversation partners at all levels of AI development and use. Algorithm designers sometimes involve communities in the design and training of AI systems systems (Halfaker & Geiger, 2019). In Chicago, researchers and community organizations coordinated with formerly gang-involved youth to develop an alternative to the city's Strategic Subjects List (Frey et al., 2020). Affected communities have also pioneered algorithm monitoring and accountability, often out of necessity (Matias, 2015; Matias & Mou, 2018). By acknowledging and resourcing community contributions as part of a national strategy, the U.S. can strengthen the excellence, safety, and equity of AI systems.

Support community-engaged basic research

A national R&D strategy for Artificial Intelligence can draw from the enthusiasm of our nation's citizens to make discoveries to ensure that AI serves the common good. Community-led, use-inspired basic research has been a staple of effective research, development, and governance of other complex industries, including food safety, water safety, the environment, and automotive testing (Stokes 2011; Blum 2019; Dietz & Ostrom 2003; Merrell et al 1999). For example, the E.P.A recently dedicated $20m in grants to support communities and tribal groups to install air quality sensors that could advance basic science while also providing early warnings of harmful pollution.[4] The E.P.A. also funds water quality monitoring, bacteria monitoring, and crowdsourced environmental violations reporting. With this country's AI strategy, we can integrate community involvement from the start rather than remediate disasters decades later.

[4] EPA-OAR-OAQPS-22-01 Enhanced Air Quality Monitoring for Communities. U.S. Environmental Protection Agency

Impact Assessment of Human-Algorithm Feedback Loops

J. Nathan Matias and Lucas Wright

When Wilmer Catalan-Ramirez was tackled in his Chicago home by immigration agents, threatened with deportation, and detained in March 2017, he couldn't understand why. The agents had no warrant. He had no criminal record, and his only encounters with police had been random stops in his neighborhood. Unknown to Catalan, his name was added to a Chicago predictive policing system after he was injured in a neighborhood drive-by-shooting (Moreno, 2017). This Strategic Subjects list, according to police, was designed to improve community support. In practice, police targeted people and communities on the list, trapping them in a cycle of escalating suspicion (Asher & Arthur, 2017).

The actress Jennifer Lawrence was caught in another escalating cycle when her intimate photos were stolen in 2014 and circulated on the social platform Reddit. As people clicked and up-voted the pictures, Reddit's algorithms showed them to even more people, encouraging the algorithm further. When asked about the experience, Lawrence told Vanity Fair "it just makes me feel like a piece of meat that's being passed around for a profit" (Kashner, 2014). Reddit made enough money from the incident to fund its computer systems for a month (Greenberg, 2014).

Could the harms experienced by Catalan-Ramirez and Lawrence have been predicted or prevented? Advocates have called for algorithm impact assessments to help manage the risks from algorithms. But unlike detection, risk assessment, and decision-making systems that are designed for accuracy and consistency, these algorithms were designed to change over time. When algorithms adapt, their behavior is hard to predict, fairness is usually unattainable, and effective remedies are even harder to identify—for now.

What makes these stories different? Catalan-Ramirez and Lawrence were both victims of feedback between human and algorithm behavior. Feedback happens when humans and algorithms react to each other in ways that change themselves without further involvement from engineers. And it's everywhere—directing law enforcement, managing financial systems, shaping our cultures, and flipping a coin on the success or failure of movements for change. Since human algorithm feedback is already a basic pattern in society, we urgently need ways to assess the impact of ideas for steering those patterns toward justice.

A Social Justice View of Human-Algorithm Behavior

Conversations about social justice and algorithms often focus on the bias and accuracy of decisions by algorithms. Bias happens when an algorithm for facial recognition, court sentencing, or credit scoring tends to mistake or wrongly penalize a group of people on average. Bias is a useful concept for evaluating judgments that need to be impartial and independent (Barocas et al 2019).

Evaluations of bias could not have protected Catalan-Ramirez or Lawrence because adaptive algorithms cannot be tested for bias—they work like a mirror rather than an independent, impartial judge. The Reddit ranking system was designed to reflect a list of most popular content rather than provide an unbiased judgment of each picture's social value. Chicago's Strategic Subjects List was a popularity ranking for policing. The system was deliberately designed to direct police toward people who already had more contact with public services and the police. Since bias applies to impartial judges but not to mirrors, we need different language to recognize the problems of adaptive algorithms. We also need different forms of power to solve those problems.

Scholars of social justice also argue that bias reduction is too small a vision in a world where injustice is the norm. Society would not be better off if Reddit equally amplified stolen intimate pictures of famous men or Chicago also over-policed a few White neighborhoods. Instead, we need to imagine a positive vision of a just society and work toward that goal (Costanza-Chock 2020).

What do we mean by social justice? In a persistently-unequal society built with stolen land, labor, and opportunity, social justice involves economic justice (McGhee 2021). Criminal and border justice would require reforms to a system that has made African Americans second-class citizens in their own country (Alexander 2000) and has created an underclass from the undocumented Americans who power our economy (Chavez 2012). Health equity would address a history of mistreatment and under-provision of health resources to marginalized communities (Reverby 2012; Donohoe 2012).

A positive vision of social justice also includes the institutions of democracy and collective social power. The right to vote is a cornerstone of justice in a country where power-holders have continually suppressed voting rights since the beginning (Anderson 2018). Beyond voting, collective behavior arises from how people think and act. That's why epistemic justice—what people know and believe about others through communication and media—also forms a basic building block of social justice (Fricker 2007).

As adaptive algorithms orchestrate more of our economy, criminal justice systems, healthcare, and democratic institutions, the future of social justice depend on our ability to understand and change human-algorithm feedback.

Understanding Human-Algorithm Feedback

Both Chicago's predictive policing system and the Reddit algorithm were designed to adapt to the world around them. These systems observe changing situations and adjust their behavior in turn. Feedback happens when these adaptive algorithms react to the behavior of humans, who are also reacting to the algorithms.

Investors learned about feedback in 2010 when they lost billions of dollars in just a few minutes. In high-frequency stock market trading, investors use algorithms to buy and sell thousands of investments per day. When humans start selling investments rapidly, algorithms can react by selling stocks too, spiraling rapidly into a stock market "flash crash" (Virgilio, 2019).

Chicago's predictive policing system created a similar spiral of injustice. With each police patrol, the software was fed new information about who officers met. The software then updated the list to prioritize people who had more contact with the police. When police followed those recommendations, the feedback increased policing for communities that were already over-policed (Saunders et al 2016). In the stock market and the streets of Chicago, harmful outcomes were caused by a cycle of influence between humans and algorithms.

Feedback can happen with one person or a million. When you use a predictive keyboard on your phone, your choice of words is shaped by your keyboard's suggestions. In turn, this personalized software also adapts its suggestions to the messages you write, making your language less creative (Arnold 2020). Other systems like Reddit's rankings aggregate behavior across many people (Ekstrand et al., 2011). Some systems do both (Li et al., 2010).

Adaptive algorithms do not create injustice on their own, but they do amplify it when they respond to unjust laws, institutions, and human behavior. Reddit did not invent a sexist culture of voyeuristic abuse when its aggregator encouraged people to view Lawrence's pictures without permission. Celebrity gossip websites also published the images, which were also distributed on file-sharing servers (Sparkes, 2014). But in a world where sexism is

already powerful and profitable, Reddit's algorithm poured gasoline onto the open flame of misogyny.

Because feedback amplifies collective behavior, it can also grow collective power for social change (Yasseri et al., 2016). When Black Lives Matter activists organized social media hashtags and in-person protests (Jackson et al., 2020), aggregators promoted the movement, and journalists responded to the attention by changing how they wrote about police violence (Freelon et al., 2016; Zuckerman et al., 2019). When news media gave more attention to those stories, people adapted too, showing up at protests in greater numbers and posting about it on social media, influencing the algorithms further (De Choudhury et al., 2016). These spirals of attention are so powerful that activists have created special advocacy software to coordinate cycles of feedback during campaigns (Wardle 2014).

Patterns of Impact from Human-Algorithm Feedback

How can we prevent feedback that amplifies injustice while also using it to create power for change toward a just society? Experts have proposed impact assessments as a process for governing algorithms (Reisman et al., 2018). These impact assessments would involve identifying the impacts of feedback, deciding whether those impacts are good or not, assigning responsibility, and deciding what actions to take (Moss et al., 2021).

Successful assessments start with the ability to identify and name impacts. To identify discrimination by decision-makers, researchers created a measurable statistical concept of "bias" (Becker, 1957; Narayanan 2018). Although bias is sometimes used to deflect support for accountability (Daumeyer et al 2019), it has also become a basic tool for regulating discrimination (Barocas et al 2019).

Can we create new ways to assess the impacts of human-algorithm feedback that are at least as useful and powerful as the idea of bias? Assigning responsibility and providing guidance on change can be difficult when humans and algorithms are continuously changing (Kitchin, 2017). But some patterns of impact are now common enough that people have started to name them, even if we don't fully understand yet how they happen.

Reinforcing: When people worry about feedback trapping people in racist, sexist, or extremist views of the world, they are concerned about personalized feedback that reinforces a person's beliefs and behaviors. For example, when Black teens select media that feature people who look like them, algorithms may show even more content produced by industries that dehumanize people of color (Epps-Darling, 2020). Because personalized algorithms make each person's behavior more consistent with their past (Arnold et al 2020; Negroponte, 1996), this reinforcement can entrench epistemic injustice and make personal change more difficult.

Herding

When Black Lives Matter activists built a movement through local organizing and hashtags, their message was partly amplified by aggregator algorithms (Jackson et al 2020). These aggregators amplify herding behavior by informing and encouraging people to do something already popular (Broussard, 2019; Salganik et al., 2006). Even before software recommendations, adaptive systems like the Billboard Charts encouraged people to flock to popular songs. But herding can also be a risk for marginalized groups. When enough people engage in collective discrimination, adaptive systems such as Chicago's policing software can further entrench injustice (Brayne, 2020). When discrimination goes viral in the media, the consequences include racial trauma and further marginalization (Bravo et al 2019). And in societies that normalize violence toward women, aggregator algorithms have further inflamed harassment mobs (Yasseri et al., 2016; Massanari 2017).

Suppressing:

Advertising markets sometimes learn to violate employment law by showing fewer job and housing opportunities to women and people of color (Sweeney, 2013). If a history of exclusion drives people away from certain opportunities, an algorithm might learn to hide the opportunity altogether. Aggregators can also suppress the interests of a minority group in favor of a dominant group. On predominantly White online platforms like Reddit, algorithms have adapted to suppress the views of people of color even in spaces created by and for communities of color (Harmon, 2019; Matias et al., 2017). Yet algorithmic suppression is also a widely-used tool for reducing the spread of misinformation designed to dissuade people of color from voting in elections (Funke 2018).

Clustering

People who have never met each other can be treated by algorithms as groups when they view similar websites or act in similar ways. These clusters already connect people with vital advice, and opportunities such as people with common medical conditions who come together to advocate for health equity (Wicks et al., 2010). But clustering can sometimes be dangerous. When people's interests are hateful and violent, they can be grouped in ways that grow hate groups (Tufekci, 2018; Paul, 2021).

Dividing

Can human-algorithm feedback divide society into opposing groups? In this pattern of polarization, groups become more and more socially separated, more opposed to each other, and less understanding of each other as humans (Finkel et al., 2020). This power to influence group dynamics is a basic tool for political organizing, whether organizers are building power within marginalized groups (Squires, 2002) or developing broad coalitions for change (Mcghee 2021). The role of algorithms in social division has been hard to study because polarization and racial resentment in the US are maintained by powerful politicians and media corporations that benefit from conflict and hatred, with or without algorithms (Benkler et al., 2018; Nyhan, 2021).

Optimizing

Pricing algorithms can reinforce economic injustice when they steer people toward behavior that benefits others against their own interests. Algorithms designed by the logistics company Uber prioritize corporate profits and rider convenience by nudging drivers to take fewer breaks and accept less well-paid work (Scheiber, 2017). When these market-management algorithms adapt to the competing preferences of customers, workers, and platform owners, they can further entrench discrimination, charging higher fares to people who live in non-White neighborhoods (Pandey & Caliskan, 2021). Some researchers have proposed that algorithms could optimize for algorithmic reparations, steering prejudiced societies toward economic justice despite themselves (Abebe & Goldner, 2018; Davis et al., 2021).

Knowing How to Create Effective Change

Impact assessments also need to include recommendations. To make these suggestions, we need usable knowledge about what kinds of power will lead to meaningful change. That's a problem because scientists can't yet reliably provide that knowledge (Bak-coleman et al 2021), although we can describe what it would look like.

When Instagram responded to advocacy groups in 2012 on mental health, self-harm, and eating disorders, they faced a herding problem similar to Reddit's disaster with Jennifer Lawrence's stolen pictures. Instagram tried to make people safer by changing its algorithm to restrict harmful searches. Instead, promoters of self-harm and eating disorders adapted to the

changes and gained popularity (Chancellor et al., 2016). Instagram could change the source code of its software, but it couldn't change the code of human culture. In the absence of reliable evidence on effective interventions, activists successfully and unknowingly forced companies to make the platform more harmful to young people.

Feedback is hard to change because change might need to come from the algorithm, from humans, or both. It's an algorithm-specific version of a common problem faced by policymakers. In the mid-20th century when civil rights activists ended school segregation through laws and court cases, they hoped to mitigate inequality by changing racist education policies. But policy alone couldn't change the underlying racism of American society. Over sixty years after Brown v. Board of Education, students of color in the US predominantly continue to study in racially segregated schools that receive far fewer resources than primarily-White schools (McGhee 2021). If we change algorithms without changing underlying behavior, algorithm policies could fail in similar ways.

Feedback is also hard to change because adaptive algorithms are less predictable than laws or institutions. Since the algorithms respond quickly to changing surroundings, attempts to change feedback patterns can have unpredictable consequences. One group of public health experts recently wrote that "we lack the scientific framework we would need to answer even the most basic questions that technology companies and their regulators face." They argued that it's currently impossible to tell whether a given algorithm will "promote or hinder the spread of misinformation" (Bak-Coleman et al 2021). Some scholars even claim that general knowledge about how to change algorithm behavior might be impossible (Kitchin, 2017). Other scientists see this as an advantage, arguing that the consistency of human behavior is the greater barrier to equality. They argue that if humans will never end patterns of discrimination, it might be better to end injustice by giving more power to algorithms (Mullainathan, 2019).

Companies point to these disagreements when trying to avoid regulation (Orben, 2020). According to Facebook's Nick Clegg, since predictable algorithms respond to unpredictable humans, the company bears less responsibility for how its algorithms behave. To improve safety, Clegg argues, Facebook should monitor people more closely and enforce policies on humans rather than regulate algorithms (Clegg, 2021).

So how can we develop the knowledge needed to advance justice and reduce the harms of human-algorithm feedback? When creating new knowledge, researchers differentiate between general and context-specific knowledge.

Context-specific knowledge can help advocates and policymakers monitor and respond to problems as they happen. Tech companies like Facebook are using context-specific knowledge when they monitor the actions of billions of people in real-time to decide what content to remove and which accounts to ban (Matias et al 2020; Gillespie 2018). But companies have also strongly resisted attempts at similar transparency for their algorithms. Even if society required algorithm inspections similar to car inspections, we do not yet have the testing technology or the staff to evaluate the many adaptive algorithms in use today.

General knowledge can help advocates and governments develop policy solutions that work in more than one situation and for more than one algorithm. Social science theories of behavior, simulations and mathematical proofs might potentially provide this knowledge if it is possible to obtain. This general knowledge could help society prevent injustice and advance a positive vision rather than simply monitor and respond to problems as they happen.

Creating reliable knowledge about feedback should be an urgent priority for everyone who cares about social justice. Without this knowledge, technology makers are introducing

algorithms into the world at a massive scale without the ability to predict the social impact. Without new scientific breakthroughs on human-algorithm feedback, we also face the risk that interventions for social justice could be ineffective or cause even more harm.

What Can We Do About Human-Algorithm Feedback?

What kinds of actions could impact assessments recommend? The search for effective interventions is just getting started. Here are some of the most commonly-discussed ideas.

The simplest option is to ban the algorithm. Bans are appropriate when the harms strongly outweigh the benefits of a system and when no one has (yet) invented a way to manage those harms effectively. In 2020 for example, Santa Clara became the first U.S. city to ban the use of predictive policing algorithms, citing how they amplify injustice (Asher-Schapiro, 2020). But banning algorithms can also introduce new harms. For example, banning automated content moderation systems with known discrimination problems (Davidson et al., 2019; Dias Oliva et al., 2021) would also expose millions of people, including marginalized communities, to violence and racist threats.

Governments sometimes require algorithm operators to remove records on harmful human behavior before algorithms can learn from it. This poorly compensated content moderation work, done by hundreds of thousands of people together with further algorithms, creates a heavy cost on mental health (Roberts, 2019).

Another way to govern feedback is to exclude people from environments where they might influence an algorithm in harmful ways. Between 2019 and 2021, Reddit "quarantined" and then banned several communities with a history of manipulating algorithms to amplify hatred (Isaac, 2020; Menegus, 2016). For example, one community of color on Reddit excludes White people from some conversations to protect the aggregation algorithm from non-Black voices and votes (Harmon, 2019). Predictive policing and decision-making algorithms often inform these bans and exclusions (Geiger, 2016; yellowmix, 2015; Zhang et al., 2018).

If an algorithm routinely participates in harmful cycles, designers could change the algorithm (Stray 2021). Companies frequently publish claims that they have adjusted their algorithms in response to public concerns (Hansell, 2007; Bossetta, 2020). Yet without systematic, public evidence on the effects of those changes, it is impossible to know what those changes have accomplished.

What if researchers could reliably determine the risk posed by an algorithm before putting it into the world? Algorithms with the potential to cause harm could be tested in the lab before being allowed to enter the market (Ohm & Reid, 2016; Tutt, 2017). For example, recommendation algorithms can be tested with simulation software that mimics people reading a news feed (Ie et al., 2019; Wainwright and Eckersley, 2021; Lucherini et al, 2021). As with all lab research, the effectiveness of these tests will depend on how realistic they are.

When feedback can't be tested in the lab, one option is to respond to problems after they happen. In a recent blog post, the Federal Trade Commission (FTC) announced that it will consider enforcing laws prohibiting unfair or deceptive practices against companies that produce or use "racially biased algorithms" (Jillson, 2021). Yet the government may struggle to prove deception if an algorithm's creators themselves can't reliably predict in advance how an algorithm would behave. In the stock market, authorities have introduced "circuit breaker" regulations that monitor markets and temporarily limit or even halt the trading of stocks when signs of volatile feedback emerge (*Staff Report on Algorithmic Trading in U.S. Capital Markets*, 2020). Similarly, some governments shut down the internet entirely during elections

when they expect violence or opposition—policies that are widely criticized by human rights advocates (Freyburg & Garbe, 2018; Howard et al., 2011; Kathuria et al., 2018; West, 2016).

How can we learn the safety of algorithms or the effectiveness of policies in real-world situations—while still limiting the risks? Field experiments, when conducted with public consent and careful oversight, enable researchers to audit systems and test policies in controlled circumstances with as few people as possible (Matias et al., 2016). Since the 1970s, researchers and regulators have used audit studies to study discrimination by humans and algorithms (Pager, 2007; Barocas et al., 2019). Field experiments also provide essential evidence on which policies are effective and which ones backfire. In one study, for example, a community learned effective ways to prevent harmful content from being promoted by Reddit's algorithm—without needing to alter the underlying software (Matias, 2020b).

Might restricting data collection make people safer (Zuboff, 2019)? While data governance is an important policy area, no evidence reducing the information available to algorithms will steer feedback in beneficial ways. Restrictions in data collection can also lead to color-blind policies that can hinder social justice (Bonilla-Silva 2006; Powell 2008).

Advocates sometimes argue that algorithm operators could prevent problems by hiring diverse teams. They argue that the technology industry's history of discrimination (Hicks, 2017; Kreiss et al., 2020) has hindered companies' ability to protect marginalized groups (Daniels et al., 2019). Yet new employees responsible for ethics and safety can struggle to create changes best led by more senior leaders, especially if new hires are considered outsiders (Dunbar-Hester, 2019; Kreiss et al., 2020; Metcalf & Moss, 2019; Silbey, 2009).

Who Does Governance?

Through a consultation process, impact assessments provide an opportunity for people and organizations across society to contribute to the governance of algorithms (Moss et al., 2021). Why does it matter who gets a voice in the process? Debates about governance always involve struggles over who is responsible and who should be trusted with the power to intervene (Dietz et al., 2003; Matias, 2015).

The city of Chicago faced one of these struggles over the Strategic Subjects list—a conflict that involved governments, companies, universities, and citizen groups. The software was initially commissioned by the city government, developed by the Illinois Institute of Technology, and funded by the U.S. Department of Justice (Asher & Arthur, 2017). But affected communities had to force designers and governments to listen to them. The city halted the program in 2020 after years of lawsuits, research studies documenting its problems, and public pressure (Charles, 2020).

Policy debates rarely include the people affected by an issue, although they have the most at stake and the greatest understanding of the context. Yet communities do sometimes have significant governance power, especially when designers give communities tools to adjust and manage algorithms in context (Matias, 2019; Matias & Mou, 2018). Algorithm designers also sometimes involve communities in the design and training of systems (Halfaker & Geiger, 2019). In Chicago, researchers and community organizations coordinated with formerly gang-involved youth to develop an alternative to the city's Strategic Subjects List (Frey et al., 2020). Affected communities have also pioneered algorithm monitoring and accountability, often out of necessity (Matias, 2015; Matias & Mou, 2018).

Civil society groups also contribute to the governance of algorithmic feedback. Activists and non-profits organize to influence governance indirectly through lawsuits,

research, lobbying campaigns, and many other tactics. By telling people's stories and conducting investigations (Diakopoulos, 2015), journalists create scandals that alert policymakers about problems and pressures companies to change (Bossetta, 2020).

When algorithm operators like Instagram try to manage human-algorithm feedback by changing code or increasing human surveillance, they are doing governance. Public pressure has also created a market for risk-management businesses that manage other companies' algorithm problems. For example, the UK government is supporting the development of UK-based, for-profit companies to create and rapidly-scale safety-focused technologies (*Online Harms White Paper*, 2020). Unfortunately, algorithm operators and risk management companies rarely evaluate their policies or publish the results, making it difficult for anyone to know whether their actions are beneficial or not (Matias et al., 2016).

Because research can alert people to problems, test products, and evaluate policies, researchers play a critical role in governance. But not all research advances the public interest. Governments and corporations often announce high-profile hires and donate money to universities to deflect criticism while avoiding change (Silbey, 2009; Weiss, 1979). While industry research partnerships can guide wise decisions, industry-funded research is often distrusted by the public (Johnson, 2019). Just as in food safety and the environment, independent research is essential for understanding and governing human-algorithm feedback (Matias, 2020a).

Effective governance often involves conflict between different actors (Dietz et al., 2003). But social movements can also develop collective power across groups. Whether they use the language of reform, de-colonization, or abolition, these movements succeed by organizing many different kinds of activities for justice (Benjamin, 2019; Costanza-Chock, 2020; Couldry & Mejias, 2020). Creating and sustaining social justice will require powerful coalitions of advocates, technology employees, researchers, journalists, and policymakers who combine efforts to understand, re-imagine, test, and change complex systems.

The Future of Impact Assessments

When algorithms and humans adapt to each other at scale, the resulting patterns have powerful consequences for every part of people's lives. For Wilmer Catalan-Ramirez, who was wrongfully detained, it led to ten months in prison and injuries that could leave him paralyzed for life. For millions of people on Reddit, it spread a culture of sexism and disrespect for women that is already endemic in society. Yet feedback is also a basic building block of movements for social change.

Anyone trying to advance social justice should consider these three basic points about feedback and how to assess its impact:

1. Impact assessments will fail if they focus exclusively on algorithms without considering the underlying human causes of problems
2. Impact assessments can do more harm than good without new kinds of knowledge: both general science to guide policy-making and monitoring systems to spot problems as they occur
3. Impact assessments need to include affected communities as equal partners in understanding and solving problems

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Medical Imaging & Technology Alliance (MITA)

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.

MITA®
MEDICAL IMAGING
& TECHNOLOGY ALLIANCE
A DIVISION OF NEMA®

1300 North 17th Street ▪ Suite 900
Arlington, Virginia 22209
Tel: 703.841.3200
Fax: 703.841.3392
www.medicalimaging.org

March 1, 2022

AI R&D RFI Response Team
Office of Science and Technology Policy

Re: 87 FR 5876, "Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan"

Dear AI R&D RFI Response Team:

As the leading trade association representing the manufacturers of medical imaging equipment, radiopharmaceuticals, contrast media, and focused ultrasound therapeutic devices, the Medical Imaging & Technology Alliance (MITA) applauds the Office of Science and Technology Policy (OSTP) and its continued dedication to AI excellence. The following comments reflect our dedication to improving American leadership in the development of Artificial Intelligence (AI) for both medical imaging and the nation.

Today, the strategic plan focuses on the development process for AI algorithms. Strategies 2, 3, 4, and 6 are concerned with algorithm creation, technical mechanisms, design, and technical standards, respectively. These are important areas to research and will yield important insights. However, concerns about the data generated for and used by systems are not highlighted. This discrepancy is a missed opportunity for OSTP and the strategic plan.

Healthcare data today is fragmented, often incomplete, and difficult to assess. This limits AI capabilities. The most significant improvement to AI algorithms—for safety, reliability, and trust—can be obtained through focus on improvements to data access, standards for metadata that capture important social characteristics, and a balance that achieves privacy for the individual and enables ethical, legal, and societal validation. Research on policies which seek to improve systemic issues which are reflected in available data would also be a welcome addition to the strategic plan.

In medical imaging, such a change would improve access to high-level care for underserved communities. Use of AI by radiologists increases the standard of care by enabling faster, more accurate decisions, which leads to better patient outcomes and lower costs. This supports the Quadruple Aim by improving the work-life balance of practitioners and leading to efficiencies in the delivery of care.

We also encourage OSTP to recognize the work done by healthcare and medical imaging on providing safe, trusted products, with meaningful benefits to patients through existing quality management systems and regulatory processes. These mechanisms respond quickly and effectively to potential issues. As the adoption of AI in healthcare encounters new challenges, the existing mechanisms can be readily adapted to address those challenges.

We look forward to continued engagement with OSTP in pursuit of AI excellence. If you have any questions, please contact Zack Hornberger, Director of Cybersecurity & Informatics, at [redacted] or by phone at [redacted].

Sincerely,

Patrick Hope
Executive Director, MITA

MITA is the collective voice of manufacturers of medical imaging equipment, radiopharmaceuticals, contrast media, and focused ultrasound therapeutic devices. It represents companies whose sales comprise more than 90 percent of the global market for medical imaging innovations. These products include: magnetic resonance imaging (MRI), medical X-Ray equipment, computed tomography (CT) scanners, ultrasound, nuclear imaging, radiopharmaceuticals, and imaging information systems. MITA Member company technologies are an important part of our nation's healthcare infrastructure and are essential for the screening, diagnosis, staging, managing and effectively treating patients with cancer, heart disease, neurological degeneration, and numerous other medical conditions.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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.

[1] See AI Risk Management Framework | NIST.

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).

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.[3] 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)[4], 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.[5] 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.[6] 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[9], 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[10] and the National Security Commission on AI[11], 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/advmlthreatmatrix: Adversarial Threat Matrix.
[10] See Implications of Artificial Intelligence for Cybersecurity A Workshop | National Academies.
[11] See, for example, chapter seven of the 2021 Final Report - NSCAI

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

National Oceanic and Atmospheric Administration (NOAA) AI Executive Council

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.

Subject: RFI Response: National Artificial Intelligence Research and Development Strategic Plan

From: V Ramaswamy - NOAA Federal

To: AI-RFI

Comments from NOAA AI Executive Council

Name: Tyler Christensen

Comments:
A key investment that the Federal government could make would be to improve the AI-readiness of our open datasets. OSTP could play a couple of very important roles in that process. This would be a task under Strategy 5, “Develop Shared Public Datasets and Environments for AI Training and Testing”

1. Establish a US government standard to define AI-readiness (via NIST?).
2. Understand the specific, detailed community requirements that make a dataset “ready” for use in AI and machine learning applications
3. Encourage Federal agencies to openly release training datasets by providing guidelines for agencies, priority research areas, quality guidelines, etc.
4. Establish a data search portal on ai.gov that allows users to narrow down their search to find data that exactly meets their particular needs.
5. Promote AI-ready data collections as a way to incentivize Federal agencies to modernize their data collections and dissemination platforms, to the benefit of all data users.
6. Collaborate with non-governmental and international organizations to promote consistent standards for AI-ready open data.

Tyler and Ramaswamy

--
V. "Ram" Ramaswamy

Director
NOAA/ GFDL
Princeton, NJ 08540

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Nicole Renae Marcy

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.

Subject: RFI Response: National Artificial Intelligence Research and Development Strategic Plan
Date: Wednesday, February 23, 2022 at 7:57:49 PM Eastern Standard Time
From: Nicole Renae Marcy
To: AI-RFI

RFI Response: National Artificial Intelligence Research and Development Strategic Plan

Prioritize, in the top three priories, the creation of specifically AI researchers pipeline for succession planning in AI workforce in USA.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)

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.

RFI Response: National Artificial Intelligence Research and Development Strategic Plan

Submitted by: Amy McGovern on behalf of the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)

Strategy 2: Develop Effective Methods for Human-AI Collaboration

AI2ES is actively focused on creating trustworthy AI for a specific user population of scientific end-users, composed primarily of forecasters, emergency managers, and researchers. As a comment on the current strategy and investments at the national level, we need additional investment into a wide variety of end-user scenarios because trust in a human working alongside an AI will vary with the needs of the end-user. To fully achieve a synergistic system for many scenarios, we need to invest in studying the needs of the end-users and what their needs for AI actually are.
Fully exploring and developing effective methods for Human-AI collaboration across all domains where it could have potentially high public value will require strategic investments in partnerships that include both AI and domain scientists as well as decision makers and end users in those domains. Further (bringing in Strategy 3 on ethics), we need to ensure we are involving the end-users in the development of the products from the beginning to ensure that AI products actually meet their needs and do not create injustices.

As an integral part of developing effective human-AI collaborations, there should be additional investment into developing explainable AI methods (also relevant for Strategy 4). The current set of methods focuses on the scientific end-users and not on the general public. While this is an important audience, additional explanation methods that explain AI effectively for diverse end-users are required for widespread adoption of AI.

Searches for *new algorithms for human-aware AI,* and *general ways of allowing machine learning algorithms to incorporate domain knowledge* (p 17) might explicitly address the multiple ways in which humans, decision contexts, and ways of knowing are diverse.

Strategy 3: Understand and Address the Ethical, Legal, and Societal Implications of AI

The goal of developing ethical and responsible AI is critical especially as trustworthy AI has become a large focus within the AI community. We encourage additional national R&D investment into the following.

1. Development and testing of principles for ethical and responsible Al: while there is unlikely to be a single guiding set of principles for all Al methods to be used in an ethical and responsible manner across all domains, it is critical that such principles be addressed and used. Funding agencies should encourage their development and use for each use-case that practitioners are examining.

2. Developing approaches to identifying bias across datasets: Although many high-profile cases have come to light where Al unintentionally recreated biases from its training data, it is not always clear what biases exist in new training data and how to identify them. For example, weather and climate are often perceived to be "objective" as much of the data comes from sensors such as radar and satellite but the coverage of these sensors is not uniform and can provide geographic biases. Identifying new biases in data is critical to creating Al that will be used responsibly and ethically.

3. Identifying legal implications of both trust and distrust of Al systems: If an Al system is trusted and used in appropriate situations yet a failure occurs, such as missed cancer diagnosis or a missed tornado warning, and people die, who is responsible? How does this affect the development, testing, and use of Al? Likewise, if an Al system outperforms humans at a task such as cancer diagnosis, what are the implications of not using the Al system? All of these legal implications must be studied and addressed through R&D and policy.

4. Prioritizing investments in fundamental research that also advance Strategy 3.

5. In addition to considerations of transparency, explainability, accountability, and fairness, sustainability and climate impact should be included among the ethical, legal and societal considerations in Al.

Strategy 5: Develop Shared Public Datasets and Environments for Al Training and Testing

Al-ready data is key to making reproducible Al models, which facilitates the development of trustworthy Al. However, the current state of the art for creating and sharing data is still haphazard at best. A national effort to provide online curated repositories of data across a variety of application domains would significantly increase the development of Al methods and provide equitable access to less well funded institutions.

A key part of this strategy and strategy 7 will be partnerships with the private sector. At the national level, we need to invest in additional mechanisms to facilitate collaboration with the private sector, through funding and through novel mechanisms that allow academia, NGOs, and other researchers to use private sector resources.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Pangiam

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.

PANGIAM

7950 Jones Branch Drive
McLean, VA 22102

March 4, 2022

Office of Science and Technology Policy
The White House
1600 Pennsylvania Ave NW
Washington, DC 20500

RE: RFI Response: Update of the National Artificial Intelligence Research and Development Strategic Plan

Mapping to the strategic plan point, "Develop Shared Public Datasets and Environments for AI Training and Testing," Pangiam offers the following response to the Office of Science and Technology Policy's (OSTP) request for information, "Update of the National Artificial Intelligence Research and Development Strategic.[1] Pangiam is a national security software and technology company applying computer vision to define the future of trusted movement and security. We are revolutionizing the future of operations, security, and safety at airports, seaports, and land crossings as well as on air force bases using emerging technologies. Our technology and expertise in computer vision and artificial intelligence (AI) is recognized by the National Institute of Standards and Technology (NIST). Our facial recognition algorithm recently achieved a top three ranking in the NIST Face Recognition Vendor Test 1:N Identification and is ranked the fastest in the world by NIST's most recent 1:1 test.

We are also experts at deploying biometric technology for identity verification in a variety of trade and travel use cases, working in partnership with private sector entities and

[1] OSTP, Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan, 87 FR 5876 (February 2, 2022).

government. In the private sector, Pangiam's technology is trusted and used by the departure control system of over thirty-five airlines. For government, Pangiam developed the U.S. Customs and Border Protection's cloud-based facial biometric matching service, known as the Traveler Verification Service. In a collaborative effort, Pangiam is the technology provider that enables a partnership between Delta Air Lines and the U.S. Transportation Security Administration for their digital ID program.

Pangiam employs an industry-leading rigorous governance program to ensure the accuracy, efficiency, and security of our technology and build trust with clients and the users of our technology. Our company has expended private resources to develop artificial intelligence algorithms with the highest ethical standards. As federal entities like OSTP and the National Artificial Intelligence Initiative Office consider policies to support U.S. AI research and development (R&D), we propose two key recommendations to ensure the ethical advancement and use of biometric and AI technology and to keep the United States competitive edge with international competitors. Our first recommendation is to provide private companies access to government datasets to train their algorithms and, second, for the U.S. government to leverage technical safeguards and policy incentives to ensure ethical AI development by the private sector.

Beyond developing and deploying AI for identity verification, Pangiam has partnered with Google Cloud to further automate threat detection at passenger screening checkpoints using AI. This solution, named Project DARTMOUTH, allows for better detection of prohibited items, and introduces a new capability of aggregated threat detection for complex, coordinated threats across screening lanes, checkpoints, and airports. This technologic development,

however, has outpaced the United States' capability to acquire and deploy it. While already in pilot in the UK, there are legal, policy, and operational barriers regarding the use of AI by federal entities that need to be addressed to be able to leverage this capability. To overcome these barriers, we recommend that the forthcoming update to the National AI Research and Development Strategic Plan include research into the legal, policy, and operational impacts of AI on federal agencies.

Artificial Intelligence in Identity Verification

The use of AI for identity verification in the trade and travel domain offers several key benefits to create operational efficiencies, increase the security and safety of air travel, and improves the traveler experience. Biometric identity verification automates manual human processes which reduces labor costs and errors caused by fatigue. This automation verifies passenger identity more quickly than manual processes, cutting time for processes like boarding almost in half, reducing potentially costly late departures, and allowing for growth of aviation operations within existing infrastructure constraints.[2]

Identity verification in travel is the cornerstone of aviation security. Using biometric technology to verify a passenger's identity increases the security of travel as algorithms outperform humans in identifying imposters.[3] From a safety perspective, at several points in a passenger journey the passenger physically exchanges documents to verify their identity.

[2] https://www.flydulles.com/news/emirates-deploys-one-step-biometric-boarding-dulles-international-airport-veriscan-1

[3]https://www.nist.gov/system/files/documents/2021/05/12/frgc_face_recognition_algorithms_surpasshumans.pdf

Biometric identity verification is contactless, reducing the potential points of transmission for viruses and enabling physical distancing at otherwise crowded chokepoints.

Our use case is narrow, explicitly used to facilitate the movement of people and goods. Despite this narrow application, there are still instances in which the companies developing this technology could cause harm. First, is in the development of AI algorithms. The size and quality of datasets used to train algorithms could cause disparity in performance across different demographics. Without high performance across all demographics, the facilitation and efficiency benefits for large, diverse populations such as the traveling public are lost.

Second, it is costly and time consuming to acquire large and diverse datasets, which has led some organizations to blur and cross ethical and legal lines in acquiring them. Additionally, poor data security and data protection practices can leave biometric information vulnerable to nation-states, hackers, or even the highest bidder through third party sales. In a recent example, one company improperly harvested customer data through deceptive methods, training its algorithm on a consumer-facing application designed to acquire dataset images without informing the consumer. Further, consumer data was retained indefinitely even after accounts were deactivated. This is just one example, but many more exist and will continue to so long as these datasets remain a barrier to development.

Access to and Ethical Safeguards for Datasets

Pangiam has proactively adopted its own Biometrics Principles to fully realize the benefits and neutralize the potential harms of this rapidly advancing technology. From data sourcing and acquisition to deployment, Pangiam's industry-leading principles protect the

integrity of the technology and process against malfeasance and abuse. While our company has adopted this policy voluntarily, if not properly regulated, biometric technology has, like all technologies, the potential to be misused. If the United States is to maintain its lead in trustworthy AI R&D, it must promote the adoption of rigorous ethical behavior for AI research in the private sector.

U.S. federal initiatives and policymaking to promote this ethical behavior must start with how algorithms are developed. In this first stage, private companies must acquire or create datasets for algorithms to be trained on. This is a costly, time-consuming step that has already led to many examples of unethical shortcuts in the industry, breaching legal agreements and eroding the confidence of the American public in AI. Access to government biometric datasets would speed the development of trustworthy U.S. algorithms, leapfrogging U.S. companies past this initial barrier to development.

Internationally, governments are already sharing their datasets with their private sector to help advance AI performance. In our closest competitors, this sharing is a key reason for the rapid advancement of their capabilities, but few, if any, ethical guardrails are required for companies to access to billions of images. In the United States, access to government datasets can be done in an ethical fashion with technical safeguards and policy incentives. Technical safeguards, such as training algorithms behind a firewall, can ensure companies only have access to results rather than the underlying data. Policy incentives, such as requiring ethical corporate behavior either by audit, pledge, or other disclosure and no past abuse of consumer data, can ensure that only responsible companies that abide by an ethical code of conduct can take advantage of this resource.

The U.S. government is already sharing data with the private sector for AI development. The National Geospatial-Intelligence Agency (NGA) shares geospatial data with the private sector to help the agency solve some of its current AI challenges and employing advanced algorithms. NGA takes its engagement with the private sector a step further and runs an accelerator to grow the number of advanced solutions available to NGA. The Pentagon's Joint Artificial Intelligence Center uses the Joint Common Foundation, a cloud-based platform that enables users to access Defense Department data and develop AI solutions in a secure environment. These are just two examples, but there are more that could be leveraged as a model for sharing datasets with the private sector for training facial recognition algorithms.

The U.S. government has a responsibility to its citizens to ensure the ethical development and use of AI. Federal agencies have their own requirements for data access, but facial recognition requires the highest ethical corporate behavior. Pangiam offers the following principles for consideration as ethical requirements for U.S. companies to access federal datasets for AI R&D for facial recognition.

- **Data Transparency**. Regularly and publicly communicate how information is stored, transmitted, and accessed, and the privacy policies governing biometrics use simplified, easy to understand language.
- **Opt-In Databases**. Travelers to affirmatively opt-in to biometric collection.
- **Opt-Out Operations**. Areas where biometrics are captured are clear and obvious and an operational policy for those who opt-out is deployed.
- **Privacy-by-Design**. Systems are designed and implemented that protect the privacy of the traveler.
- **Security Safeguards**. Use encryption and biometric algorithms whose face templates cannot be reverse engineered to identify a traveler.

- **Performance and Accuracy**. Use only biometric algorithms that have the highest rates of accuracy and precision as determined by The National Institute of Standards and Technology.
- **Domestic Development**. Algorithms are developed in the United States.
- **Ethically sourced Datasets**. Datasets are ethically and legally sourced and contain a wide range of demographics to reduce bias.
- **No Third-Party Sales**. Prohibit the sale of biometric and biographic information to third parties; and
- **A Track Record of Trust**. Organizations that are known or found to have abused customer privacy, misled consumers, or insufficiently protected data are excluded from accessing government datasets.

While ethical behavior by the private sector is paramount to developing trustworthy AI in the United States, any federal effort to verify the desired corporate behavior must keep in mind the speed at which biometric technology is being developed worldwide. In order to maintain a competitive edge against global competitors, this ethical verification effort cannot be so cumbersome for the private sector as to inhibit innovation and therefore counterproductive to the United States' overall aim at leading AI R&D. In the same vein, any technical safeguards must not be so complex as to prohibit the ability of an ethically verified company to use the dataset.

Beyond incentivizing ethical behavior, access to government datasets would address a key cause of inconsistent performance across different demographic groups or "algorithmic bias." There is an overrepresentation of Caucasian images in many public and private datasets and, when algorithms are trained on datasets that lack diversity, it leads to bias in facial matching of underrepresented groups. Pangiam is committed to eliminating bias in its facial recognition models and we already strive to train our models on datasets that have equal

representation across gender and ethnicity. We do this because of our strong ethical principles, and this diligence results in better parity of performance across demographic groups. The U.S. government can help the private sector writ large produce higher performing models with access to diverse government datasets.

AI Enablement Beyond Research and Development

In mid-2020, airport operators, regulators, control authorities and industry bodies from around the globe, including the United States, endorsed an Open Architecture policy for airport security systems.[4] This policy was a paradigm shift in transportation security, decoupling original equipment manufacturers from the software used on transportation security equipment. Removing vendor lock-in on detection algorithms opened the door for non-traditional software developers to enter the transportation security market. Third-party threat detection AI algorithm developers quickly demonstrated interest and capability in this opportunity.

These new entrants provide a full generational leap forward in AI use for threat detection technology, allowing for better detection of prohibited items, more advanced detection capabilities, and reducing false alarms. More advanced threat detection capabilities, known as aggregated threat detection, can detect complex, coordinated threats across lanes, checkpoints, and even airports, all while maintaining consistently high checkpoint throughput numbers. While U.S. support was a key factor in the Open Architecture policy successfully encouraging AI R&D in transportation security, there are non-technology barriers to the United States benefitting from this achievement.

[4] Open Architecture for Airport Security Systems (aci-europe.org).

There is a misconception that AI R&D is a purely technologic concern. The development of AI for threat detection has revealed, however, that many barriers to its deployment are not in R&D, but in support and policy functions. Therefore, any strategic plan to advance AI R&D include understanding the impact of AI on enterprise and operational support functions. AI R&D will result in legal, policy, and operational issues in almost all federal agencies. The federal government will need to review its own internal policy and processes to ensure readiness to enable the acquisition, deployment, and regulation of this technology. AI R&D is moving quickly, and these ancillary functions must not be overlooked in the forthcoming strategy update if the United States is to maintain leadership in AI R&D.

Pangiam welcomes the opportunity to participate in further discussions to ensure that AI R&D is conducted ethically and enables the United States to maintain its global competitive advantage. We are also willing to share insights on non-R&D opportunities for the federal government to enable development of this technology.

Respectfully submitted,

Shaun Moore
Chief Artificial Intelligence Officer
Pangiam
7950 Jones Branch Drive
McLean, VA 22102

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Q Bio

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.

Stacy Murphy
Operations Manager,
White House Office of Science and Technology Policy

Via Email

Re: RFI Response: National Artificial Intelligence Research and Development Strategic Plan

Dear Ms. Murphy,

Q Bio is responding to the White House Office of Science and Technology Policy (OSTP) *Request for Information (RFI) to the Update of the National Artificial Intelligence Research and Development Strategic Plan*.

Q Bio recognizes that Artificial Intelligence (AI) and Machine Learning (ML) technologies have the potential to transform health care by deriving new and important insights from the vast amount of data generated in the industry. Medical devices are using these technologies to innovate their products to better assist health care providers and improve patient care. There has been early adoption of AI / ML in the medical imaging industry and with the early adoption, there have been questions raised around guidance of use. As a leader in health technology and in the space of imaging and medical bioinformatics, Q Bio believes the following principles around AI adoption and use should be considered.

Most techniques labeled as AI / ML at this time are **in essence statistical algorithms** that learn from a large set of consistent training data and can make inferences from new data that has very similar properties as the training data. The **benefit of this approach is identifying correlations in the data that are not easily described by humans** and lead to very powerful applications with promising performance in many applications. **However, most of these techniques do not "understand" the data. They operate on correlation but not causation.** There should be caution around application as they can create data by interpolating from past training data. In the academic world, there has been significant progress in recent years on techniques allowing causal inference and associated network techniques, which if supported for further development could mean a leap forward in AI, i.e. networks would not only be able to detect correlations between smoking and lung cancer, but actually be able to infer that smoking causes cancer. We believe further research in this domain will have an impact on public health management and disease discovery.

We are concerned with the **overuse and perhaps misappropriation of neural networks in medical imaging applications**. In many cases, networks are being used to provide de-noised, resolution-increased, or even image contrast transferred data (i.e. a CT image generated from MRI data by a neural network) for diagnosis. While there is a certain "wow factor" with these applications, there is a lack of awareness that these networks are simply interpolating data, or in some cases restoring data using correlations from the training dataset. In other words, while these applications may help display images in a form more familiar to a physician, they don't actually enhance data, or show smaller structures in super-resolution images, as the information is not contained in the original data. This can be very powerful, but the **output is non-deterministic and not easy to audit**.

We also observe in our industry that networks have been used to supplant algorithms that are analytical and verifiable, often by laws of physics or mathematics, resulting in a new application that is no longer as easy to verify as the analytical solution. **Clear guidelines should be in place around the communication of the limitations and risks associated with the use of neural networks**, especially outside single-purpose applications. While the FDA is implementing stricter regulations around the verification of such applications, we believe that even the R&D world should be focused on the appropriate use of neural networks, rather than a wide range of transitions to neural networks, where they are not necessary or even more harmful.

We believe **a good use of AI and neural networks and other machine learning techniques is in the acceleration of conventional algorithms**. For example, in the area of optimization, techniques such as auto-differentiation have made an impact by enabling minimization of cost functions that only few trained specialists were able to address in the past. By embedding ML into conventional algorithms, verification of the solution and detection of failure is much easier to ensure.

Neural networks are safest to use when we may not be able to compute an answer we want directly from raw inputs, but **can verify correctness of the neural network output** efficiently. An example of this is protein folding. It is computationally intractable to directly compute how a protein will fold. If we train a neural network to predict how it will fold, we can use the laws of physics to then verify that the output of the neural network is at least a valid solution because it does not break those laws. On the other hand, neural networks are most dangerous to use where it is not possible to directly compute answers over inputs, and it is not tractable to verify correctness of the answer given by algorithm.

At Q Bio we believe neural networks are a very useful tool, but for the foreseeable future they will be tools that help clinicians make better and more efficient decisions and will not be a replacement for human decision making. Neural networks will just supply another input that clinicians can use in their practice. We will continue to focus on using neural networks on problems where we can verify the accuracy of the output of the neural network easily, or where we are confident we can train neural networks on data that uniformly samples entire input spaces, not subspaces.

Overall, we continue to believe in the power of AI/ML and with responsible application of this technology, we believe the U.S. will continue to be a leader in innovation. We strongly support ongoing investment in Artificial Intelligence and Machine Learning Research and public private partnerships. We at Q Bio will continue to invest in our applications of AI/ML in bringing cheaper, faster, better and quantified imaging to all.

Thank you for the opportunity to provide Q Bio's perspective regarding the update of the National Artificial Intelligence Research and Development Strategic Plan. We welcome the opportunity to serve as a resource to OSTP, the Select Committee, NAIIO, the NSTC Subcommittee on Machine Learning and AI, and the NITRD AI R&D Interagency Working Group.

Sincerely,

Clarissa Shen

Chief Operating Officer

[redacted]

About Q Bio

- Q Bio is a health information technology company that provides the first clinical, whole-body Digital Twin platform that measures a large set of health variables and biomarkers in individuals to provide clinically useful information to empower physicians and patients in their health care decisions. We combine MRI data with genomics, biochemical markers, and patient history to track health changes over time.
- Physical 2.0 as a Service — we are not a Care Provider, but provide a clinical information platform to providers and their patients to empower health insights and longitudinal tracking of changes across anatomical measurements, biochemistry, and other key vitals.
- Q Bio offers a new, highly efficient, whole body-scanning technology that is faster and cheaper than existing options. Our Q Bio Gemini platform supports existing MR scanners as well in collecting whole-body anatomical measurements efficiently and reproducibly.
- Q Bio has an interdisciplinary team of pioneers in artificial intelligence, applied math, computational biology, computational physics, computer science, software engineering, genetics, medicine and radiology from

Our Vision: *"A world where treatable diseases no longer take lives, and each generation is healthier than the last."*

- Reduce Biases/Improve Outcomes: Q Bio seeks to provide personalized care, reduce existing bias in medicine and improve patient outcomes by building the first "Clinical Digital Twin Platform" to track a well-defined set of biomarkers over time for each individual, and to make such a program accessible for all individuals and their health care providers. This will allow for the detection of diseases based on an individual's data, rather than population-level data, at the earliest stages where an intervention or treatment may be more likely to succeed.
- Personalized Medicine: Q Bio's platform will allow for the detection of diseases based on an individual's data rather than only population level data and at the earliest stages where an intervention or treatment may be more likely to succeed. We are building the missing tool to make Precision Medicine possible in proactive primary care.
- Population Health Management: Q Bio aims to build a population scale triage network for primary care physicians that is fully automated–allowing physicians to quickly access critical data for their patients. With such a network, physicians could save highly valuable time by interacting directly with their patient's health data instead of depending solely on a manual physical examination, or on previously existing broad sets of population data.
- Reduce Disparities: Q Bio strives to close the significant equity gap in medicine, by helping providers focus on access to care based on need rather than offered on a first-come, first-served basis, or ability to pay.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Shah, University of Washington

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.

Subject: RFI Response: National Artificial Intelligence Research and Development Strategic Plan
Date: Saturday, February 5, 2022 at 8:07:32 PM Eastern Standard Time
From: Chirag Shah
To: AI-RFI

Hello,

I would really like to emphasize the importance of developing shared public datasets and tools for AI R&D. Currently, almost all the large-scale pre-trained models (also referred to as foundation models) are developed and controlled by private industry, which not only makes it difficult (or impossible) for educational institutions, non-profits, and government agencies to do impactful original work outside of these realms, but this over reliance on a few large models controlled by even fewer commercial entities can be a security risk. Developing comparable datasets and tools for AI development outside of these profit-based organizations must be given high priority and necessary support for more sustainable and democratic advancement of AI.

Thank you.

Chirag Shah, PhD
Associate Professor, Information School (iSchool)
Adjunct Associate Professor, Paul G. Allen School of Computer Science & Engineering (CSE)
Adjunct Associate Professor, Human Centered Design & Engineering (HCDE)
University of Washington
Director, InfoSeeking Lab
Co-Director, Center for Responsible AI Systems & Experiences (RAISE)
Editor-in-Chief, Information Matters
[redacted]

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Society for Industrial and Organizational Psychology (SIOP)

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.

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

The Society for Industrial and Organizational Psychology (SIOP) is submitting these comments in response to a request for information released on February 2, 2022, by the White House Office of Science and Technology Policy (OSTP) seeking input on a proposed revision of the National Artificial Intelligence Research and Development Strategic Plan.

Industrial and organizational (I-O) psychology is a dynamic and growing field that addresses workplace issues at the individual and organizational level. I-O psychologists apply research that improves the well-being and performance of people and the organizations that employ them. Collectively, I-O psychologists bring over a century of expertise in understanding and predicting workplace behavior. SIOP is the professional organization representing a community of over 10,000 scientists, academics, consultants, practitioners, and students of I-O psychology and working to promote evidence-based policy and practice of the science of the workplace. Many I-O psychologists specialize in topics related to education, development, and workforce training, as well as the emerging technology-enabled workforce. SIOP believes this expertise is well suited to address the issues at hand in the National Artificial Intelligence Research and Development Strategic Plan.

Bias, Fairness, and Standards in AI-Based Hiring Assessments

Decisions about whom to hire are made many thousands of times each day and many organizations, including federal agencies, are looking for ways to make these decisions more accurately and efficiently to remain competitive in a demanding marketplace. To this end, there has been a growing interest in the use of artificial intelligence (AI) for pre-employment screening of job candidates. AI in this context refers to a broad range of technologies and statistical techniques that have the potential to identify patterns in candidate information that are predictive of future job performance. At the same time, there have been increasing calls for scrutiny of AI-based assessments, reflecting concerns over privacy, fairness, lack of transparency, and the accuracy of their predictions.

To address these concerns, AI-based decision tools require the same level of scrutiny that traditional employment tests have been subjected to for decades. In fact, state and federal regulatory control specific to the use of AI in organizational decision making has already occurred in some places and seems imminent in others. To aid policy makers and employers looking for guidance on AI-based hiring tools, SIOP has published the *Principles for the Validation and Use of Personnel Selection Procedures*, which is updated regularly to reflect current scientific research and best practices in hiring and promotion. This document

summarizes the fundamental requirements for selection procedures that should guide the evaluation of assessments. Importantly, these professional guidelines are applicable to *all selection procedures*, including technology-based hiring and promotion procedures that incorporate AI, machine learning, and other novel assessment techniques (e.g., game-based assessments, evaluation of voice and facial characteristics).

Building on the guidelines published by SIOP, there are five key criteria for evaluating AI-based assessments:

1. AI-based assessments should produce scores that are considered fair and unbiased.
2. The content and scoring of AI-based assessments should be clearly related to the job.
3. AI-based assessments should produce scores that predict future job performance (or other relevant outcomes) accurately.
4. AI-based assessments should produce consistent scores that measure job-related characteristics (e.g., upon re-assessment).
5. All steps and decisions relating to the development and scoring of AI-based assessments should be documented for verification and auditing.

These five key criteria are intended to represent the ***minimal*** requirements necessary to justify the use of AI-based assessments for hiring and promotion decisions.

SIOP encourages OSTP to include a focus on research exploring bias and fairness in AI-based hiring and selection assessments in the next revision of the National Artificial Intelligence Research and Development Strategic Plan under "*Strategy 3:* Understand and address the ethical legal and societal implication of AI" to ensure this rapidly expanding topic does not exacerbate biases and can contribute to a strong workforce across the federal and private sectors. Additionally, SIOP encourages OSTP to include a focus on the development and implementation of metrics for fair and unbiased AI-based selection and hiring systems under "*Strategy 6:* Measure and evaluate AI technologies through standards and benchmarks."

AI and the Technology-Enabled Workforce

As the rate of technological change continues to accelerate, understanding how these changes affect American workers has never been more critical. These changes impact not only the products that our workforce creates and sells but also the work environment itself, such as increased coworking with robots and AI. Federal agencies have done well to address this change as the current strategic plan notes in its 2019 update, highlighting opportunities from NSF, NOAA, NIH, and DOE.

AI and automation are transforming the production of goods and services. Workforce trends require employees to develop new routines, skills, and competencies to better work alongside automated systems. I-O psychologists have deep expertise in both preparing for this shift and understanding how current workers will react to it. Assessing job demands and developing responsive employee training programs using new technologies are necessary to inform future efforts to reskill employees. As jobs are transformed or replaced, it is also critical to

understand how American workers react and respond. Assessing, interpreting, and anticipating human reactions to automation and other new technologies is a core I-O competency. I-O psychologists also have expertise in effective strategies and processes to retrain workers, including identifying current skill needs and projecting the skills of importance for the future. Efficient job retraining can address the need for lifelong skills education, keep older workers in the workforce, and combat talent shortages in areas critical for societal well-being.

SIOP applauds current efforts focused on bolstering the AI-enabled workforce, including understanding the changing skills needs. We encourage OSTP to continue this focus under "*Strategy 7:* Better understand the national AI R&D workforce needs." SIOP additionally encourages OSTP to expand the focus on workforce needs to include research on improving and disseminating strategies for effective upskilling and job training, particularly workplace-based training that allows employees to continue working while expanding job capabilities and skills. Finally, SIOP recommends that OSTP include a focus on anticipating and responding to worker reactions to the increased use of AI and other automation in the workplace to ensure that workers are willing to accept these new technologies and remain productive in the face of the challenges associated with their implementation.

Expert Contacts

SIOP welcomes the opportunity to submit these comments and further provide expertise and insight as OSTP seeks to update the National AI R&D Strategic Plan. Please reach out to the following SIOP issue experts with additional comments or questions:

Christopher Nye, Ph.D.

Associate Professor of I-O Psychology, Michigan State University
Expertise: Employee assessment and hiring; technology-based assessments; quantitative modeling and biases in the hiring process

Richard N. Landers, Ph.D.

John P. Campbell Distinguished Professor of I-O Psychology, University of Minnesota
Expertise: Application of technology in the workplace; job candidate assessment; employee learning and behavior; research methods

Joseph A. Allen, Ph.D.

Professor of I-O Psychology, University of Utah
Expertise: Education and training guidelines; workplace meetings; organizational community engagement; occupational safety and health

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Software & Information Industry Association (SIIA)

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.

Comments of the Software & Information Industry Association on the Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan

Submitted to the Office of Science and Technology Policy

March 4, 2022

On behalf of the Software & Information Industry Association (SIIA), we appreciate the opportunity to provide input on the Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan.

SIIA, a non-profit organization, is the principal trade association for the software and digital information industries worldwide. Our members include over 450 companies reflecting the diversity of the information landscape, from creation to dissemination to productive and responsible use. They include digital content providers and users in academic publishing, education technology, and financial information, along with creators of software and platforms used by millions worldwide, and companies specializing in data analytics and information services. Our members support policies that foster innovation and a healthy digital ecosystem, including consumer privacy protections, responsible and ethical AI, and DEI initiatives.

Input on Strategy 3: Understand and address the ethical, legal, and societal implications of AI; and Strategy 4: Ensure the safety and security of AI systems.

The OECD Principles on Artificial Intelligence were only a month old when the U.S. government issued the 2019 Strategic Plan. Since then, public and private research into ethical and responsible AI and methods to operationalize these objectives has grown exponentially. Governments around the world have launched councils to study ethical AI, international and multi-stakeholder efforts at the OECD and through GPAI have brought needed high-level engagement, and academic research centers have pioneered new ways to hone AI. This still-burgeoning field has led to numerous advances in the quality of AI datasets and algorithms that are more accurate, transparent, and fair. Moreover, private sector companies have taken steps to build and enhance internal governance frameworks to operationalize principles.

We support continued federal funding of AI research and development into the ethical, legal, and societal implications of AI along with initiatives to ensure the safety and security of AI systems. We support efforts by the National Science Foundation, the Department of Energy and other agencies, and the National Institute of Standards and Technology (NIST) to support work in this area. Likewise, we support the work of the Office of Science and Technology Policy (OSTP) to craft an “AI Bill of Rights” that would reflect principles for responsible and ethical development and use of AI that support diversity, equity, inclusion, accessibility, and fairness in how AI systems are applied.

Need for an AI Governance Framework

Despite these efforts, alongside world-leading innovation, research, and investment, the United States has fallen behind in formal AI governance. The European Union (EU) is advancing an AI Act that has many admirable qualities but will also impose significant costs on U.S. companies and barriers to innovation. These costs and barriers are likely to make it harder to maximize the benefits of AI while achieving appropriate protections individual rights, privacy, and security. Other jurisdictions around the world, such as the UK, Brazil, and Singapore, have made significant strides in formulating AI governance frameworks, and some, like China, have already implemented algorithm regulations. Municipalities in the United States are beginning to do the same.

Because these laws will have effects on U.S. companies and their innovation, we have concern that the United States is missing out on an opportunity to shape AI governance framework as has happened in the privacy context following implementation of the EU's General Data Protection Regulation.

We recommend that the Strategic Plan include as a strategic aim a dedicated effort to develop frameworks of rules and regulations to guide AI R&D and use by both the private and public sectors. To achieve this, we recommend that OSTP convene a council of legal, policy, and technical experts with a concrete timeline to generate specific proposals.

We have suggested elements and the structure of such a framework at the back of this submission. (See "*Annex – Advancing a Democratic Vision of Data Governance*".)

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

The availability of robust, reliable, and trustworthy data sets is a key impediment to AI innovation. While data is an essential component of the AI stack, developing robust data sets that meet the standards for responsible AI and minimize privacy concerns is extremely costly for most companies and entrepreneurs. That cost both limits the potential of AI and allows AI tools to be built on unreliable, untrustworthy, or potentially biased information. Data sets that do not comport with standards of accuracy, reliability, trustworthiness, and bias present significant societal risk.[1]

Shared public datasets are critical to fostering new and better uses of AI technologies and ensuring that the data relied on by AI algorithms meets quality standards. We support continued work by the U.S. government in this space. Below we suggest specific additional measures that the U.S. government can undertake to create more robust shared datasets.

[1] Joshua New, AI Needs Better Data, Not Just More Data, Center for Data Innovation (Mar. 20 2019); Tasha Austin, et al., Trustworthy Open Data for Trustworthy AI, Deloitte Insights (Dec. 10, 2021).

Public-Private Dataset Initiatives

We recommend that the Strategic Plan advance efforts to build public-private partnerships to develop large, high-quality, and privacy-protective data sets that are accessible and usable by a wide range of actors would promote innovation and entrepreneurship while greatly reducing these risks. We offer two proposals for consideration.

First, we recommend a public-private effort to create **large synthetic data pools** that would be accessible by researchers, government, and across industry. Synthetic datasets can enable algorithms to run on data that reflect, rather than rely on, real-world data. This approach would allow for the creation of a robust data lake that can be vetted to ensure accuracy, reliability, fairness, and so on. Moreover, it would not present privacy and individual rights concerns that may arise from the collection, retention, sharing, and use of datasets that are built directly from personal information. We understand there is interest in the private sector to work with the government on this sort of initiative.

Second, we recommend a public-private effort to create **large open data sets** of personal information collected through enhanced notice and consent procedures. This could be modeled on the Casual Conversations dataset developed by Meta.[2] That dataset consists of over 45,000 videos of conversations with paid actors who consented to their information being used openly to help industry to test bias in AI systems.

Under either proposal, we would recommend that NIST lead the effort to ensure that the large data set is appropriately screened before it is put into wide use. The data pools should be subject to intensive test, evaluation, verification, and validation procedures in accordance with NIST standards and with the involvement of government and private sector experts.

Improving International Collaboration through Shared Datasets

We also encourage the Strategic Plan to advance recommendations for international, multilateral, and bilateral research efforts that can use shared datasets to support innovation. Two such models are as follows:

For example, an international or multilateral research effort built on shared datasets could be modeled on the Multilateral AI Research Institute (MAIRI), recommended by the National Security Commission on Artificial Intelligence.[3] As proposed, MAIRI would "facilitate joint efforts to develop technologies that advance responsible, human-centric, and privacy-preserving AI/ machine learning (ML) that better societies and allow allies to pool their talents and resources. It will provide a model for equitable, multilateral research, facilitate AI R&D that builds on like-minded countries' strengths, and foster a global AI workforce for the next generation." It would be a U.S.-led effort that would benefit from a federated network of global research institutes and leverage shared datasets to foster innovation in line with democratic technology values.

[2] Meta AI, Casual Conversations Dataset (April 2021).
[3] National Security Commission on Artificial Intelligence, Final Report (March 2021) at 192, 244, 249, 535-40.

Another approach would be to encourage more targeted, bilateral efforts, with much the same objectives as MAIRI.[4] A bill introduced in the current Congress would create a joint U.S.-Israel center "to leverage the experience, knowledge, and expertise of institutions of higher education and private sector entities in the United States and Israel to develop more robust research and development cooperation" in several critical AI areas.

Input on Strategy 8: Expand Public-Private Partnerships to accelerate advances in AI.

We encourage the Strategic Plan to call for further public-private efforts to advance AI, especially in areas of democratic affirming technology and technology that provides other important societal benefits.[5] Below, we highlight two areas where we believe government action can help to address societal challenges and foster greater innovation.

Promoting Privacy Enhancing Technologies (PETs)

Privacy enhancing technologies (PETs) refer to a group of technologies that protect the privacy and security of sensitive information. NIST, a long-time champion of PETs, will recognize that a host of what were once "emerging" PETs—such as homomorphic encryption, differential privacy, federated learning, and synthetic data—now have established uses in a wide range of contexts, including research, health care, financial crime detection, human trafficking mitigation, intelligence sharing, criminal justice, and more.[6]

PETs can be an essential part of a democratic model of emerging technology in practice, as a counter to a model that sacrifices privacy, trust, safety, and transparency.[7] PETs can enable the secure sharing of data between entities and across jurisdictional boundaries, expanding data access and utility and enabling organizations to reduce risk while making faster, better-informed decisions.[8] PETs are one way to solve (as a technical but not legal matter) privacy-based restrictions on EU-US data flows.[9]

---

[4] S.2120, The United States-Israel Artificial Intelligence Center Act (June 17, 2021).

[5] SIIA recently published a series of case studies highlighting uses of our members' technology to assist in a range of socially beneficial efforts, including tracking financial crime and human trafficking, finding missing children, supporting and advancing resources for the disabilities community, and more.

[6] The Center for Data Ethics and Innovation, PETs Adoption Guide, Repository of Use Cases. See also, e.g., Kaitlin Asrow and Spiro Samonos, Federal Reserve Bank of San Francisco, Privacy Enhancing Technologies: Categories, Use Cases, and Considerations (June 1, 2021); Luis T.A.N. Brandao and Rene Peralta, NIST Differential Privacy Blog Series, Privacy-Enhancing Cryptography to Complement Differential Privacy (Nov. 3, 2021).

[7] Andrew Imbrie, et al., Privacy Is Power: How Tech Policy Can Bolster Democracy, *Foreign Affairs* (Jan. 19, 2022).

[8] Two use cases involving SIIA members will help to illustrate this point. First is a partnership between Enveil (an SIIA member) and DeliverFund, the leading counter-human trafficking intelligence organization, which leveraged Enveil's PETs-powered solutions to accelerate reach and efficiency by allowing users to securely and privately screen existing assets at scale by cross-matching and searching across DeliverFund's extensive data. Second is Meta's use of secure multi-party computation, on-device learning, and differential privacy tools to minimize the amount of data collected in the advertising space while ensuring that personalized content reaches end users.

[9] See Asrow and Samonos, supra, at 4.

This is not just an industry view. As the White House stated in announcing the new US-UK challenge, PETs "present an important opportunity to harness the power of data in a manner that protects privacy and intellectual property, enabling cross-border and cross-sector collaboration to solve shared challenges."[10] In addition, the U.S. Census Bureau plans to launch a series of pilot projects to deploy PETs to "to build a platform that will enable secure multi-party computation, encryption technologies, and differential privacy to promote better data sharing both domestically and abroad."[11] This energy complements growing global interest. For example, the UK Information Commissioner's Office is exploring guidance on PETs and ways to incorporate PETs into data regulations. Recently, according to reports, the United Nations launched a "PETs Lab" to test PETs against data sets from the United States, the UK, Canada, Italy, and the Netherlands, and work with researchers and the private sector to develop use cases and create guidance.[12]

We therefore recommend that Congress and the executive branch **incentivize PET adoption by public and private entities**. The GDPR and other new privacy regimes have helped to foster increased attention in PET capabilities abroad. Official action by the U.S. government can have a similar effect and lead to the development and use of PETs designed to address critical needs around information privacy and security – enhancing innovation in the United States and helping to drive behavior globally. PETs can also help to drive up compliance with a range of laws and regulations in ways not possible when those laws and regulations were drafted.

We similarly encourage additional **public-private partnerships designed to establish use cases**. A bill under consideration in Congress, the Promoting Privacy Technologies Act (H.R. 847),[13] would promote fundamental research into PETs. The government has a critical role to play in fostering fundamental research – especially in areas, such as PETs, where robust markets have yet to develop. In addition to fundamental research, the government should look to establish partnerships focused on the application of already-mature PETs to new areas.

Promoting Content Provenance to Combat Disinformation

Maintaining a trustworthy digital ecosystem, one that addresses growing and malign influence efforts, is important for the health of the internet and entire digital ecosystem. Disinformation can erode social cohesion and human rights,[14] with a disproportionate effect on marginalized communities.[15] AI supercharges the ability of state and non-state actors to spread disinformation

[10] White House Office of Science and Technology Policy, US and UK to Partner on Prize Challenges to Advance Privacy-Enhancing Technologies (Dec. 2021); White House, Remarks of Jake Sullivan (July 13, 2021).

[11] White House, Fact Sheet: The Biden-Harris Administration is Taking Action to Restore and Strengthen American Democracy (Dec. 8, 2021).

[12] United Nations. Global Platform: Data for the World; The Economist, The UN is testing technology that processes data confidentially (Jan. 29, 2022).

[13] H.R.847 - 117th Congress (2021-2022): Promoting Digital Privacy Technologies Act (Jan. 19, 2022).

[14] Carme Colomina, et al., The impact of disinformation on democratic processes and human rights in the world, European Parliament (April 2021).

[15] Center for Democracy and Technology, Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender (2021).

creating a systemic risk for the entire information environment.[16] Synthetic media, including deepfakes, provide a special challenge because of how they deliberately distort existing images, video, and audio.[17]

We are encouraged by work underway around **content provenance and authenticity** as one way to combat the scourge of disinformation and deepfakes. The Content Authenticity Initiative[18] is a cross-industry coalition of content creators, technology companies, and others dedicated to using technology to fight the scourge of disinformation through content authenticity. A related project, the Coalition for Content Provenance and Authenticity[19] recently issued a series of technical specifications designed to certify the provenance of media content.[20] This builds on work of several private firms, including one of our member companies, Adobe.[21]

This is a core area where further efforts within the U.S. government and between the government and private firms would be extremely beneficial and we recommend including this as a strategic aim in the Strategic Plan. We support the Deepfake Task Force Act[22] as one important way to achieve this. That Act would create a task force within the Department of Homeland Security and coordinate efforts with the private sector to fight deepfakes.[23] This is exactly the sort of coordination that we believe is critical in addressing one of the most challenging digital threats facing society today.

* * *

Thank you for the opportunity to provide input on the Strategic Framework. We would be pleased to discuss any of these issues in further detail. Please direct any inquiries to **Paul Lekas, SIIA Senior Vice President for Global Public Policy** [redacted]).

[16] Katerina Sedova, et al., AI and the Future of Disinformation Campaigns, Georgetown Center for Security and Emerging Technology (Dec. 2021).

[17] Kartik Hosanagar, Deepfake Technology Is Now a Threat to Everyone. What Do We Do?, Wall Street Journal (Dec. 7, 2021); Tim Hwang, Deepfakes: A Grounded Threat Assessment, Georgetown Center for Security and Emerging Technology (July 2020).

[18] Content Authenticity Initiative, https://contentauthenticity.org/.

[19] Coalition for Content Provenance and Authenticity, https://c2pa.org/.

[20] Coalition for Content Provenance and Authenticity, C2PA Specifications.

[21] Eric Abent, Adobe Expands Content Authenticity Initiative Tools to Fight Misinformation, SlashGear.com (Oct. 26, 2021).

[22] Deepfakes Task Force Act, https://www.congress.gov/bill/117th-congress/senate-bill/2559/text.

[23] U.S. Senate Comm. on Homeland Security & Govt. Affairs, Tech Leaders Support Portman's Bipartisan Deepfake Task Force Act to Create Task Force at DHS to Combat Deepfakes (July 30, 2021).

Annex - Advancing a Democratic Vision of Data Governance

Advancing a model for the responsible development and use of emerging technologies is among the most important components of a U.S. approach to fostering economic and competitiveness. The global nature of data and information means that many U.S.-based companies and the strength of the U.S. innovation climate are directly affected by laws and regulations implemented in foreign jurisdictions. It also means that what the United States does in terms of establishing rules of the road can have a noticeable effect on how other nations develop their own technology policies.

While the U.S. government has made this a priority in its foreign policy,[24] the nation still lacks a fundamental data governance framework, with no general application privacy law and no clear vision for advancing a regulatory framework for AI. Other jurisdictions have stepped up to fill this gap. The European Union's (EU) GDPR has become the benchmark privacy framework for jurisdictions worldwide, and the EU is looking to do the same with AI. The United Kingdom (UK) is among many nations that have advanced a concrete vision of data governance. The UK Data Protection Act 2018 is now entering its fourth year and the UK Information Commissioner's Office is actively exploring measures to update the Act and provide guidance to the public and business around emerging technologies.[25] China, too, has passed a consumer data privacy law along with regulatory frameworks for AI.

As described below, we recommend that the Strategic Plan convey the importance of (1) passing a comprehensive federal privacy law and (2) developing a formal framework to guide AI development and use.

Federal Privacy Legislation as a Necessary First Step to Providing a Baseline for Emerging Tech

SIIA and its members have advocated strongly for federal privacy legislation for years and are active in engaging with members of Congress and administrations of both parties. A federal privacy bill is the number one solution to closing the gaps on the use of personal data and data-driven technologies and driving innovation in the U.S. economy. Currently, the patchwork of state laws across the nation create uncertainty for consumers and businesses, burden companies with duplicative compliance costs (estimated at $1 trillion over 10 years) and have a disproportionate impact on growth and innovation for

[24] As the White House states in its Indo-Pacific Strategy: "We will also work with partners to advance common approaches to critical and emerging technologies, the internet, and cyberspace. We will build support for an open, interoperable, reliable, and secure internet; coordinate with partners to maintain the integrity of international standard bodies and promote consensus based, values-aligned technology standards; facilitate the movement of researchers and open access to scientific data for cutting-edge collaboration; and work to implement the framework of responsible behavior in cyberspace and its associated norms." White House, Indo-Pacific Strategy of the United States (Feb. 2022).

[25] The U.K. government launched its AI strategy later in 2021 and is piloting an AI technical standards hub through the Alan Turing Institute. See, e.g., UK Office for Artificial Intelligence, Department for Digital, Culture, Media & Sport, and Department for Business, Energy & Industrial Strategy, UK National AI Strategy (September 2021).

small- and medium-sized businesses.[26] This will grow as additional states pass privacy legislation – unless Congress acts first.[27]

The benefits of a federal privacy law for fostering a stronger innovation climate for emerging technologies and U.S. competitiveness are many. Such a law should both protect consumers from harm caused by the unreasonable collection and misuse of their personal data and prevent and remedy data practices that stifle innovation and stagnate data flows and routine business processes. Enactment of that statute would create both international and domestic benefits.

A federal privacy law is essential to digital trade and transatlantic data flows. As you are no doubt aware, those flows are in a state of flux due to the EU's invalidation of the Privacy Shield, and recent developments have threatened the use of foundational technologies by multinational companies. The United States' failure to offer a competing vision of privacy has allowed the GDPR to become the dominant approach to regulating personal data in the world. As the EU acts to restrict the activities of U.S. firms, the United States will lose competitiveness in the production, sale and distribution of data-driven technologies and services. Maintaining U.S. competitiveness requires the development of a U.S.-based common set of definitions and policies that can help build alignment on data flows with the rest of the world.

Domestically, the benefits of a federal privacy regime include creating baseline harmonization of consumer and business expectations surrounding personal information; supporting and fueling further competitive innovation in emerging technologies; and more deeply embedding diversity, equity and inclusion into privacy, emerging tech, and AI policies and practices.

We therefore encourage the Strategic Plan to emphasize the need a comprehensive federal privacy law that will provide strong and meaningful consumer protections (such as individual rights to notice, access, control, correction, deletion, and portability), permit socially beneficial uses of consumer data - particularly publicly available information - and promote innovation and competition in the American economy.

Developing a U.S. Legal Framework for AI

The United States risks missing an opportunity to shape AI regulation in a manner that will promote U.S. innovation and reflect core U.S. values. The EU is open about its goal of establishing the ground rules and guardrails for AI as it has done with privacy. The EU is not alone; indeed, many nations as well as U.S. states and localities are exploring and enacting rules that will have a direct impact on U.S. companies and U.S. innovation.

We recommend that the U.S. government take a more proactive approach to building an AI governance framework to avoid a repeat of what happened with privacy. We offer a few overarching

[26] See Information Technology & Industry Foundation, The Looming Cost of a Patchwork of State Privacy Laws (Jan. 24, 2022).

[27] Three states have comprehensive privacy laws (California, Colorado, and Virginia), and dozens of bills have been introduced in most of the remaining states. See IAPP, State Privacy Legislation Tracker (Feb. 2022).

principles for building a regulatory (or legislative) framework that would advance responsible AI and fairness, promote U.S. innovation, and provide necessary guardrails around both public and private use.[28]

First, we encourage efforts currently underway in the U.S. government to develop **guidelines for responsible and ethical use** of AI technologies – which should include those technologies that collect and use data embedded in emerging technologies. NIST's work in developing a risk management framework and establishing guidelines to address algorithmic bias is especially encouraging.[29] Alignment of key stakeholders, including industry groups, government stakeholders, private companies, and consumer advocacy groups to existing principles and standards like those adopted by NIST will lead to further harmonization of policy and technical foundations for how information is collected and used.

We also recommend that the government endorse a **technology-neutral approach** to regulating AI. We believe the focus should be on the type of data being collected, building transparency around notice and consent of data collection, and clear basis for the use and processing of information, rather than the technological tools— and the underlying algorithms—that facilitate that collection and use. The distinction may be subtle, but it is an important one. As noted, there are a wide variety of emerging technologies (and nuances within each); overarching rules must be sufficiently flexible to provide a groundwork as advances occur, with detail to be worked out through standards and regulations.

This approach also recognizes that technologies are tools that in most cases are not inherently good or bad. Because of this, we recommend that the focus be on how the models are used in practice. We believe the **risk-based approach**, now widely endorsed, provides the right lens. With respect to emerging technologies, there is a range of risks associated with the type of information collected, how that information is stored and used, and what consumers' expectations are. Responsibility, security, and accountability for data-driven technologies should be commensurate with risk.

One way to think about how to build out a risk-based approach is to **identify the potential harms**—such as discrimination and privacy violations—that could arise out of use, misuse, or abuse of the data-driven technologies.[30] This should be done through a sector-by-sector assessment and ead to **targeted, contextual restrictions on collection, storage, and use** to prevent those harms. We believe such restrictions or guardrails are critical, but they should be as narrowly tailored as possible to support innovation and positive use cases that benefit society.

In addition, we recommend that there be **no broad copyright exception** for the use of data in artificial intelligence. Facts, of course, are free for the taking.[31] And fair use permits the analysis of a copying and analysis of protected works for purposes that do not displace the markets for the original,

---

[28] Among the many existing frameworks, we commend those published by Google and, in January 2022, by the Business Roundtable as especially worthy of attention.

[29] See NIST, AI Risk Management Framework Concept Paper (Dec. 13, 2021); NIST, A Proposal for Managing and Identifying Bias in Artificial Intelligence (June 2021).

[30] See, e.g., FTC Staff, Comment to NTIA (Nov. 8, 2018) (identifying four categories of informational injuries: financial, physical, reputational, and unwanted intrusion).

[31] *Feist Pubs. v. Rural Telephone, Inc.*, 499 U.S. 341 (1990).

for example by creating an electronic card catalog that allows the user to search both the title and text of particular works.[32] Fair use and other doctrines (absent from the laws of different countries) similarly permit ordinary internet-based activities.[33] At the same time, however, the fact-based nature of fair use inquiry allows courts to examine and proscribe uses that would result in substitutes for the author's original creation. There is nothing about artificial intelligence that requires recalibration of the copyright law's current balance.

Lastly, we recommend separate frameworks that **distinguish between public and private** collection, storage, and use of information. In the private context, we note that many companies have built robust frameworks for assessing how they collect and use personal data. Many already adhere to industry standards and have established self-regulatory frameworks and principles[34] and conduct pre- and post- deployment impact assessments when collecting and using personal information. .

The public sector or government experience in collecting and using personal information for AI models presents a different situation. Governments can and do make use of personal information to verify access to benefits, for national security, public safety, and law enforcement purposes, to counter fraud, to assist in providing public health services, and so on. Beyond the restrictions of the Privacy Act of 1974, questions abound about what information the government may appropriately collect, how that information is obtained, and how the information is used consistent with the rights guaranteed by the Constitution.[35] Concerns have been raised about whether the Fourth Amendment provides sufficient protection to individuals—and clarity to government actors—about what information may be collected and how it can be used.[36] Congress has introduced several bills that would direct the U.S. government on use (or non-use) of facial recognition technologies, for example,[37] yet much can be accomplished through Executive Branch action alone. Many states and local governments are grappling with this issue. It is important that the federal government confront these issues and provide a model for the nation.

---

[32] *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).

[33] E.g., *Field v. Google*, 412 F.Supp.2d 1106 (D. Nev. 2006) (copying of copyright-protected work allowed via implied license and other doctrines).

[34] See, e.g., Adobe, Ethical Approach to AI; Google, AI Principles.

[35] See, e.g., Matthew Doktor, Facial Recognition and the Fourth Amendment in the Wake of *Carpenter v. United States*, *Univ. of Cincinnati L. Rev.*, v.89, issue 2, at 552-74 (2021); Glenn Gerstell, "Failing to Keep Pace: The Cyber Threat and Its Implications for Our Privacy Laws," remarks to Georgetown Cybersecurity Law Institute (May 23, 2018) (reprinted at *Lawfare*).

[36] Ibid.

[37] See, e.g., H.R.3907 (117th Cong.), Facial Recognition and Biometric Technology Moratorium Act of 2021 (June 2021) (proposing to "prohibit federal government use of facial recognition technologies, provide a private right of action, and remove federal aid from agencies using the technology"); S.1265 (117th Cong.), The Fourth Amendment Is Not For Sale Act (Apr. 2021) (proposing to "prohibit law enforcement from purchasing personal data without a warrant and prevent law enforcement and intelligence agencies from buying illegitimately obtained data"); see also S.3035 (117th Cong.), Government Ownership and Oversight of Data in Artificial Intelligence Act of 2021 (Nov. 2021) (proposing interagency working group to develop guidance for the federal government and contractors around the development and use of AI).

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Stanford Institute for Human-Centered Artificial Intelligence (HAI)

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.

Response to Notice of Request for Information (RFI) to the Update of the National Artificial Intelligence Research and Development Strategic Plan
Stanford Institute for Human-Centered Artificial Intelligence

The Stanford Institute for Human-Centered Artificial Intelligence (HAI) offers the following submission for consideration in response to the Request for Information (RFI) by the White House Office of Science and Technology to the Update of the National Artificial Intelligence Research and Development Strategic Plan. Our submission recommends:

- For Strategy 1: Boost non-defense AI R&D budgets, particularly on AI-related infrastructure, to support long-term investments.

- For Strategy 2: Increase support for interdisciplinary and multidisciplinary AI research on human-AI collaboration that expands beyond exclusively technical research.

- For Strategy 3: Please refer to the Stanford HAI letter submitted in January 2022 in response to the White House Office of Science and Technology proposal for an AI Bill of Rights that safeguards the American public against powerful technologies.

- For Strategy 4: Develop appropriate acquisition strategies and update existing procurement regulations to respond to AI procurement and acquisition challenges in the federal government.

- For Strategy 5: Expand government data access to academic researchers to train AI models and develop frameworks for government agencies to evaluate such datasets and their applications in tandem.

- For Strategy 6: Establish a mechanism to evaluate AI models within the exact context of their intended use to ensure safe deployment as well as designate NIST in collaboration with other federal agencies to benchmark AI models in institutional contexts.

- For Strategy 7: Update immigration policies to attract talent in AI and other technical fields as well as develop federal programs to hire AI talent and re-skill civil servants with both technical capacity and institutional knowledge.

- For Strategy 8: Expand public-private partnerships to accelerate advances in AI.

Strategy 1: Make long-term investments in AI research.

Recommendation: Boost non-defense AI R&D budgets, particularly on AI-related infrastructure, to support long-term investments.

A long-term commitment to sustained federal research and development (R&D) funding in AI is critical to advance the United States' leadership in global innovation. The federal government should increase non-defense investment in AI and basic research to strengthen research in critical fields of AI R&D, including healthcare, education, finance, and more, that underpin economic stability and robust growth. Such investment should reflect a multidisciplinary approach, focused on advancing basic and applied R&D, research on AI governance and norm-setting, and supporting research infrastructure with multi-agency collaboration.

Current federal funding for non-defense AI R&D, however, does not meet the needs of the fast-growing AI field. The public non-defense AI R&D budget requested by 25 federal agencies participating in the Networking and Information Technology Research and Development (NITRD) program and the National Artificial Intelligence Initiative in FY 2022 represents an increase of just an 8.8 percent increase over what was spent in FY 2021.[1] In contrast, the National Security Commission on Artificial Intelligence (NSCAI) recommended in its final report to increase public funding for AI R&D at compounding levels, doubling annually to reach $32 billion per year by FY 2026.[2]

Federal long-term non-defense investments-and high return-on-investment basic research funding[3]-can address the challenges the AI innovation ecosystem is currently facing in the United States. For example, the high cost of compute and the lack of access to critical data are hindering efforts by academic researchers to engage in cutting-edge AI R&D.[4] The federal government is lagging behind the private sector on AI development, and federal standards for technical and ethical AI are sorely needed.[5] Long-term public investment in AI-related infrastructure can strengthen AI R&D by supporting a variety of federal initiatives, including the National Artificial Intelligence Research Resource (NAIRR) that aims to expand access to "critical resources and educational tools that will spur AI innovation and economic prosperity nationwide."[6] Another example of such an initiative is the Multilateral AI Research Institute (MAIRI), recommended by the NSCAI report, that would "facilitate joint efforts to develop technologies that advance responsible, human-centric, and privacy-preserving AI/machine learning (ML) that better societies and allow allies to pool their talents and resources."[7]

[1] The Networking & Information Technology R&D Program and the National Artificial Intelligence Initiative Office, "Supplement to the President's FY 2022 Budget," December 2021, https://www.nitrd.gov/pubs/FY2022-NITRD-NAIIO-Supplement.pdf.
[2] Eric Schmidt et al., "Final Report," National Security Commission on Artificial Intelligence, March 2021, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital.pdf.
[3] Benjamin F. Jones and Lawrence H. Summers, A Calculation of the Social Returns to Innovation (Nat'l Bureau of Econ. Research, Working Paper No. 27863, 2020).
[4] Daniel E. Ho, Jennifer King, Russell C. Wald, and Christopher Wan, "Building a National AI Research Resource: A Blueprint for the National Research Cloud," Stanford Institute for Human-Centered Artificial Intelligence, October 2021, https://hai.stanford.edu/sites/default/files/2022-01/HAI_NRCR_v17.pdf.
[5] David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano Florentino-Cuellar, "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies," Stanford Law School, February 2020, https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.
[6] "The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force," The White House (The United States Government, June 10, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.
[7] Schmidt et al., "Final Report."

Strategy 2: Develop effective methods for human-AI collaboration.

Recommendation: Increase support for interdisciplinary and multidisciplinary AI research on human-AI collaboration that expands beyond exclusively technical research.

Intentionally building trustworthy AI that is unbiased and supportive of human flourishing is crucial to ensuring the successful development and deployment of human-centered AI. A key part of that effort requires an interdisciplinary and multidisciplinary approach, involving collaboration from a variety of fields to develop the hardware and software, to understand and design for people's behaviors and expectations when interacting with AI in different institutional contexts, and to establish policies and regulations to determine human responsibilities as well as the required domain knowledge for various applications. A human-centric approach to AI calls for wide-ranging collaboration among multiple disciplines. Harnessing the potential of AI with rapidly growing capabilities while addressing its impact on existing structural inequalities and biases cannot rely on the voices of computer scientists and engineers alone.

Yet current policies do not necessarily match this need. The National Science Foundation program on algorithmic fairness, for instance, calls for interdisciplinary perspectives while stating that "this program supports the conduct of fundamental computer science research" and requiring the PI to "bring computer science expertise to the research."[8] But producing fairness-aware algorithms or just understanding the concept of fairness requires knowledge and expertise outside the computer science field to incorporate the social and legal contexts in which AI systems will be deployed.[9] The federal government should expand the support of multidisciplinary AI research for human-AI collaboration to include critical academic fields such as the social sciences, law, ethics, and the humanities-all of which should have a prominent voice in providing the necessary frameworks for understanding AI today and in the future.

Strategy 3: Understand and Address the Ethical, Legal, and Societal Implications of AI

With respect to recommendations for Strategy 3, please refer to the Stanford HAI letter submitted in January 2022 in response to the White House Office of Science and Technology proposal for an AI Bill of Rights that safeguards the American public against powerful technologies.[10]

Strategy 4: Ensure the Safety and Security of AI Systems

Recommendation: Develop appropriate acquisition strategies and update existing procurement regulations to respond to AI procurement and acquisition challenges in the federal government.

Public sector AI can rely heavily on contracting and procurement with external vendors to build up technical capacity. Research shows that almost half of identified use cases of federal agencies' use of AI came from external sources, with one-third coming from private commercial

[8] "NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon (FAI)," The National Science Foundation, 2021, https://www.nsf.gov/pubs/2021/nsf21585/nsf21585.htm.
[9] Andrew D. Selbst et al., "Fairness and Abstraction in Sociotechnical Systems," *Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 59-68, https://doi.org/10.1145/3287560.3287598.
[10] Michele Elam and Rob Reich, "Stanford HAI Artificial Intelligence Bill of Rights," Stanford Institute for Human-Centered Artificial Intelligence, January 2022, https://hai.stanford.edu/white-paper-stanford-hai-artificial-intelligence-bill-rights.

sources via the procurement process.[11] Compared to the internal sourcing of AI systems that may be more policy-compliant and more accountable, the uses of procured AI in government raise several concerns in terms of AI trustworthiness, transparency, and safety. The federal acquisition regulation (FAR), for example, provides strong IP protection for vendors.[12] But such protections can obscure certain information about the inputs their tools use and how the tools operate behind trade secrecy claims, which in turn prevents appropriate analysis, audit, and testing to ensure the fairness of their use.[13] Moreover, those protections may also create uncertainty for acquisition aimed at the black-box nature of AI systems. FAR makes a clear distinction between rights to "software" and rights to "data," but AI systems, particularly machine learning (ML), integrate customer software with the new data generated in the process of training, and current procurement policies do not sufficiently address how rights to that data and the resulting AI systems are to be distributed, and under what constraints and conditions.[14]

The federal government should develop appropriate acquisition strategies and update existing procurement regulations to help address some of the public sector's challenges in evaluating, monitoring, and using AI systems.[15] Specific examples include developing clear standards that call for the disclosure of data and information on the design and operation of contractors' algorithms, requirements that ensure contractors adhere to ethical AI standards, and testing infrastructures that allow for iterative testing and evaluation.[16]

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

Recommendation: Expand government data access to academic researchers to train AI models and develop frameworks for government agencies to evaluate such datasets and their applications in tandem.

While there are publicly available resources for AI development, there is still more work to be done to promote the open and collaborative non-commercial use of training and testing data and environments. For example, access to data resources sufficient for training AI systems is increasingly limited to large private companies, which in turn direct resources toward developing applications with a focus on private profit instead of public interest.[17] Because large platforms have unequaled access to data for AI development, smaller actors, some of which may legitimately lack the financial resources to invest in building training data from scratch, are incentivized to mine the public sphere for data, violating individual privacy expectations and creating privacy risks both for individuals and society at large. At the same time, there are

[11] Engstrom et al., "Government by Algorithm."
[12] "Federal Acquisition Regulation Part 27 - Patents, Data, and Copyrights," U.S. General Services Administration, accessed March 2022, https://www.acquisition.gov/far/part-27.
[13] Deirdre K. Mulligan and Kenneth A. Bamberger, "Procurement As Policy: Administrative Process for Machine Learning," *Berkeley Technology Law Journal* 34 (2019), https://dx.doi.org/10.2139/ssrn.3464203.
[14] Ken Farber, Kristine Lam, and Ellery Taylor, "From Ethics to Operations: Current Federal AI Policy," Advanced Technology Academic Research Center, October 4, 2021, https://atarc.org/wp-content/uploads/2021/10/Current-Federal-AI-Policy-Assessment-20211004-for-public-comment.pdf.
[15] Laura Gerhardt and Mark Headd, "Digital Service Delivery: Why We Love Modular Contracting," 18F, April 9, 2019, https://18f.gsa.gov/2019/04/09/why-we-love-modular-contracting/.
[16] Engstrom et al., "Government by Algorithm."; Lavi M. Dor and Cary Coglianese, "Procurement as AI Governance," *lEEE Transactions on Technology and Society* 2, no. 4 (2021): pp. 192-199, https://doi.org/10.1109/tts.2021.3111764.
[17] Jathan Sadowski, "When Data Is Capital: Datafication, Accumulation, and Extraction," *Big Data & Society* (January 2019), https://doi.org/10.1177/2053951718820549; Ho et al., "Building a National AI Research Resource."

significant barriers for interagency and external researchers to access a rich portfolio of public sector data (e.g., employment, healthcare, education).[18]

As a starting point, the executive branch can use the NAIRR as an opportunity to make more and better quality government data available to the research community at no cost.[19] In doing so, the federal government should weigh considerations such as privacy, security, and fairness, and it should begin to develop its own frameworks for evaluating such datasets and their applications in tandem, informed by important developments under the Foundations for Evidence-Based Policymaking Act of 2018 and the National Secure Data Service. Considering the lack of a standardized framework to test and evaluate AI models for safety and security, the government should pursue creating a testing environment for this purpose, especially for government-procured AI systems.[20]

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

Recommendations:

- Establish a mechanism to evaluate AI models within the exact context of their intended use to ensure safe deployment.
- Designate NIST in collaboration with other federal agencies to benchmark AI models in institutional contexts.

Understanding the true in-domain accuracy, or the accuracy of AI systems' deployment in specific contexts (e.g., in different industries, with different subpopulation groups), is crucial for the federal government to capture the capabilities of the technology and ensure safe deployment. Many current performance evaluations do not comprehensively assess how AI systems would perform in a real-world context.[21] Object recognition systems, for example, are often evaluated against large-scale benchmark datasets to validate their performance, but such datasets are limited in their coverage of non-Western contexts and temporally bounded, capturing only parts of the real world.[22] The same gap is observed in language models where these state-of-the-art systems amplify human bias and discriminate against minority users, and their performance degrades when given out-of-domain text.[23] All told, the accuracy of AI systems in one domain does not automatically translate to its uses in other domains, and changing context can significantly impact performance.

---

[18] Amy O'Hara and Carla Medalia, "Data Sharing in the Federal Statistical System: Impediments and Possibilities," *The Annals of the American Academy of Political and Social Science* 675 (2018): 138-50, https://www.jstor.org/stable/26582286.

[19] Ho et al., "Building a National AI Research Resource," 35.

[20] Note that we recommend government agencies withhold certain data to allow sufficient testing. For example, NIST's Face Recognition Vendor Test (FRVT) challenge does not provide training images because "the tests seek to mimic operational reality and, there, algorithms are almost always shipped and used 'as is' without any training or adaptation to customer data." Read "Face Recognition Vendor Test Ongoing-Frequently Asked Questions (FAQs)," National Institute of Standards and Technology, November 6, 2019, https://www.nist.gov/system/files/documents/2019/04/22/frvt_frequently_asked_questions.pdf; Avrim Blum and Moritz Hardt, "The Ladder: A Reliable Leaderboard for Machine Learning Competitions," arXiv.org, February 16, 2015, https://arxiv.org/abs/1502.04585. For creating a testbed for the federal government, see Tina Huang, "Creating an AI Testbed for Government," Institute for Progress, January 2022, https://progress.institute/wp-content/uploads/2022/01/Creating-an-AI-Testbed-for-Government-_final.pdf.

[21] Inioluwa Deborah Raji et al., "AI and the Everything in the Whole Wide World Benchmark," arXiv.org, November 26, 2021, https://arxiv.org/abs/2111.15366.

[22] Raji et al., "AI and the Everything in the Whole Wide World Benchmark."

[23] Kawin Ethayarajh and Dan Jurafsky, "Utility Is in the Eye of the User: A Critique of NLP Leaderboards," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2020, https://doi.org/10.18653/v1/2020.emnlp-main.393.; Thomas Manzini et al., "Black Is to Criminal as Caucasian Is to Police: Detecting and Removing Multiclass Bias in Word Embeddings," *Proceedings of the 2019 Conference of the North*, 2019, https://doi.org/10.18653/v1/n19-1062.

The White House should consider a proposal to charge the National Institute of Standards and Technology (NIST), in collaboration with federal agencies that have regulatory oversight for AI-powered products, such as the U.S. Food and Drug Administration (FDA), Consumer Financial Protection Bureau (CFPB), and the National Highway Traffic Safety Administration (NHTSA), to develop improved AI benchmarking protocols.[24] Such benchmarks should explicitly address and incorporate the institutional contexts in which the AI systems are developed (e.g., commercial settings) and deployed (e.g., border control). NIST should also consider how to measure the effectiveness of deploying AI in real-world settings where enabling technologies (e.g., cameras, microphones, computing hardware) and other factors may vary. For instance, one field experiment of an earlier generation of predictive policing algorithms found that models that worked well in the lab did not perform well in the field, failing to reduce crime when used in context.[25] However, it should do so from a human-centered perspective grounded in ethical frameworks, such as privacy by design, and not focus exclusively on identifying technical benchmarks. This requirement suggests a potentially different set of roles and expertise than NIST has had in the past.

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

Recommendations:

- Update immigration policies to attract talent in AI and other technical fields.
- Develop federal programs to hire AI talent and re-skill civil servants with both technical capacity and institutional knowledge.

Future U.S. leadership in AI hinges on the country having the necessary talent-generation process and hiring pipeline-as well as the ability to attract and retain talent that already exists. The country needs individuals who are not only equipped with the skills to build AI systems, but who also know when, where, and how to ask the right questions when such systems pose risks to individuals and society and/or break the law. While not every AI expert has or should have a technical background, understanding the technology is essential to designing and implementing accountable AI initiatives. Such talent could also be a useful tool of accountability, helping design and maintain transparent, auditable, and responsible systems as well as engaging with stakeholders to ensure trustworthiness in AI systems.[26] This could include expediting the hiring process for certain kinds of AI talent,[27] updating immigration policies to attract and retain

[24] For FDA, see "Artificial Intelligence and Machine Learning (AI/ML) Medical Devices," U.S. Food and Drug Administration (FDA, September 22, 2021), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices; for CFPB, see Patrice Alexander Ficklin, Tom Pahl, and Paul Watkins, "Innovation Spotlight: Providing Adverse Action Notices When Using AI/ML Models," Consumer Financial Protection Bureau, July 7, 2020, https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models/; for NHTSA, see "Automated Vehicles for Safety," National Highway Traffic Safety Administration, accessed March 2022, https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.
[25] Daniel E. Ho, Emily Black, Maneesh Agrawala, and Li Fei-Fei, "Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains," Stanford Institute for Human-Centered Artificial Intelligence, November 2020, https://hai.stanford.edu/sites/default/files/2020-11/HAI_FacialRecognitionWhitePaper_Nov20.pdf; Priscillia Hunt, Jessica Saunders, and John S. Hollywood, "Evaluation of the Shreveport Predictive Policing Experiment," National Institute of Justice, 2014, https://nij.ojp.gov/topics/articles/evaluation-shreveport-predictive-policing-experiment.
[26] Engstrom et al., "Government by Algorithm."
[27] See, e.g., problems hiring young tech talent: Jack Corrigan, "The Government's Struggle to Hire Young Tech Talent Is Worse Than You Thought," *Nextgov*, December 1, 2017, https://www.nextgov.com/cio-briefing/2017/12/governments-struggle-hire-young-tech-talent-worse-you-thought/144225/.

technical talent,[28] and expanding opportunities for permanent residency to those with AI and other technical degrees, as well as opportunities for entrepreneurs.[29]

Improving public sector capacity can be essential to accountability and oversight to ensure AI systems are (or are not) built and deployed in ways that promote, rather than degrade, the public interest. Currently, the federal government faces numerous challenges in hiring AI talent, including competing with private-sector salaries and benefits, competing with shorter private-sector hiring timelines, and enticing applicants who face an onerous and often confusing federal hiring process versus an often faster and easier one in industry or civil society.[30] The White House should work on its own, as well as with Congress and with industry partners, to identify the biggest challenges in attracting AI talent to the government and ways to potentially resolve them. Several agencies have initiated such efforts. In 2019, the Office of Personnel Management (OPM) established a classification of information technology (IT) positions to ease the hiring burden of the federal government's competitive service and the "direct hire" appointing authority for several STEM and IT positions for agencies with critical hiring needs.[31] In 2021, the U.S. General Services Administration (GSA) launched a two-year fellowship aimed at placing early-career software engineers, data scientists, and others with technical skills in federal agencies.[32]

Finally, it bears mentioning that the government should take a long-term view to developing future AI talent and expertise by investing in technical skills development in primary and secondary education through STEM-focused educational initiatives.

Strategy 8: Expand public-private partnerships to accelerate advances in AI.

Recommendation: Strengthen partnerships with academic institutions and build a framework for a public-university-industry AI R&D ecosystem to drive AI development forward.

The federal workforce to date can still lack appropriate training in AI to use it effectively in public operations and to discharge regulatory responsibilities. A Government Business Council and Accenture survey found that more than 60 percent of federal employee respondents worry about the "lack of technical support and user training" for public AI deployment.[33] An evaluation of the U.S. Customs and Border Protection (CBP)'s facial recognition program used at air exit by the Government Accountability Office (GAO) found that agents on the ground received little

[28] Tina Huang and Zachary Arnold, "Immigration Policy and the Global Competition for AI Talent," Center for Security and Emerging Technology, June 2020, https://cset.georgetown.edu/publication/immigration-policy-and-the-global-competition-for-ai-talent/.
[29] The House version of the America COMPETES Act of 2022 is one possible starting point. See "H.R. 4521 - Bioeconomy Research and Development Act of 2021 [America COMPETES Act of 2022]," House of Representatives Committee on Rules, February 2022, https://rules.house.gov/bill/117/hr-4521.
[30] Joan Timoney, "Building a Federal Civil Service for the 21st Century: The Challenge of Attracting Great Talent to Government Service," Pan-Organizational Summit on the US Science and Engineering Workforce: Meeting Summary (U.S. National Library of Medicine, 2003), https://www.ncbi.nlm.nih.gov/books/NBK36382/.
[31] Margaret M. Weichert, "Delegation of Direct-Hire Appointing Authority for IT Positions," Office of Personnel Management, April 5, 2019, https://chcoc.gov/content/delegation-direct-hire-appointing-authority-it-positions.
[32] Chris Kuang, "Introducing the U.S. Digital Corps: A New Path to Public Service for Early-Career Technologists," U.S. General Services Administration, August 30, 2021, https://www.gsa.gov/blog/2021/08/30/introducing-the-us-digital-corps-a-new-path-to-public-service-for-early-career-technologists.
[33] Kristen Vaughan, Britaini Carroll, and Michael R. Gavin, "Federal Workers Ready to Thrive in the Age of AI," Accenture, February 15, 2019, https://www.accenture.com/us-en/insights/us-federal-government/ready-thrive-ai.

training to use a feature of the system.[34] Not only does this undermine agencies trying to carry out their missions, but individuals using AI systems without appropriate training can also create or exacerbate threats to privacy, civil liberties, and even safety. Supporting AI education and research in the university environment can help address some of the government's talent problems and help fill the talent pipeline for the public sector.

This kind of government-academic collaboration in scientific and technological areas can also fuel innovation. For example, after World War II, the U.S. Department of Veteran Affairs (VA) collaborated with academic institutions (specifically medical centers) to meet the increasing medical needs of returning veterans.[35] The collaboration between academic medicine and the VA helped to revolutionize VA healthcare and spurred innovation in healthcare at many levels, including, for instance, the invention of pacemakers and CAT scan prototypes.[36]

* * * *

As lead authors, we proudly submit this response on behalf of our colleagues and the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Daniel E. Ho, J.D., Ph.D.
William Benjamin Scott and Luna M. Scott Professor of Law, Stanford University; Faculty Associate Director, Stanford Institute for Human-Centered Artificial Intelligence (HAI)

Jennifer King, Ph.D.
Privacy and Data Policy Fellow, Stanford Institute for Human-Centered Artificial Intelligence (HAI)

Russell C. Wald
Director of Policy, Stanford Institute for Human-Centered Artificial Intelligence (HAI)

Daniel Zhang
Policy Research Manager, Stanford Institute for Human-Centered Artificial Intelligence (HAI)

[34] Adam Hoffman et al., "CBP and TSA Are Taking Steps to Implement Programs, But CBP Should Address Privacy and System Performance Issues," U.S. Government Accountability Office, September 2020, https://www.gao.gov/assets/gao-20-568.pdf.
[35] "75th Anniversary of VA's Academic Mission," U.S. Department of Veterans Affairs, January 12, 2021, https://www.va.gov/OAA/75th_anniversary.asp.
[36] Rob Marek et al., "Actions Needed to Help Better Identify Agency Inventions," U.S. Government Accountability Office, April 2018, https://www.gao.gov/pdf/product/691501.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

State University of New York Canton

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.

SUNY CANTON
State University of New York • 34 Cornell Dr. • Canton, NY 13617-1096 • www.canton.edu

March 3, 2022

RE: RFI Response: National Artificial Intelligence Research and Development Strategic Plan" in the subject line of the message.

Dear Office of Science and Technology Policy,

The Advanced Information Security and Privacy (AISP) Research Lab is housed at The State University of New York in Canton and is the hub for national and international collaboration on research and education efforts on Information Technology and Artificial Intelligence. The AISP Lab has more than 16 Research Fellows, Graduate and Undergraduate students, visiting scholars and has helped over 30 members with their Doctorate dissertations, Masters' theses and Internship projects over the years. The AISP Lab is the focal point for not only the State University of New York's CyberSecurity program, but also a place for our community to educate and enhance their knowledge of AI and CyberSecurity. More information can be found at: www.ghazinour.com/AISP

The AISP Lab respectfully submits its comments on THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN as follows below.

Upon review of the 2019 plan in light of recent events, we request the Government prioritize the following three areas for immediate, robust and sustainable focus and funding:

A. Strategy 4: Ensure the Safety and Security of AI Systems

1) Securing against attacks. With conflicts rising globally and the high-stakes of AI and cyberspace infrastructure, protecting and defending our AI systems is not a luxury option but an immediate need.
2) Autonomous attack analysis and countermeasures. Expand on DARPA's efforts that "involved AI agents autonomously analyzing and countering cyber-attacks" (p. 26) Although major targets such as government agencies, and sensitive major businesses already benefit from sophisticated and costly existing autonomous systems, we would like to emphasize that we strongly believe such autonomous systems should be widely available and affordable for others to ensure that universal, accessible and responsive AI agents protect a wide array of traditionally underprepared potential targets, including:
    - Health care providers
    - Small to mid-size financial institutions
    - First responders

- Small businesses
- Industries involved with cyber-infrastructure
- K-12 schools
- Local governments
- Faith communities
- "Last mile" grid/infrastructure providers
- "First mile" food providers – especially farmers

B. Strategy 5: Develop Shared Public Datasets and Environments for AI Training and Testing

1) Making training and testing resources responsive to commercial and public interests (p. 34)
   - Expand efforts similar to DHS IMPACT program which "supports empirical data sharing between the international cybersecurity R&D community, critical infrastructure providers, and their government supporters" ideally with "dynamic, agile repositories" with the broadest view of infrastructure to ensure agility in data collection, analysis and responsiveness to diverse interests of industry and citizenry adrift and drowning in a sea of unstructured data.

C. Strategy 8: Expand Public–Private Partnerships to Accelerate Advances in AI

1) Advancing our Nation's AI innovation ecosystem. The vast subject area of AI and its diverse active players within it is too complex to be only run by the public or private entities separately. The collaboration and partnership between them needs to be expedited and facilitated if we are hopeful to benefit from advancements made by both parties.
2) Expand Government-Industry-University research partnerships that engage efforts to communicate clearly and simply best practices in daily life for cybersecurity to motivate students, employees, warfighters and academics to exchange applied experiences in a rapid, agile and iterative process with researchers to quickly disseminate life-saving AI tools, develop novel, "US-first or best" applications and disarm domestic, foreign and rogue AI threats.

I am happy to report that The AISP Research Lab, The Cybersecurity Program and the Center for Criminal Justice, Intelligence and CyberSecurity with the collaboration of our partners offer:

A. Suggestions for priorities

a. Investing on research regarding Privacy and Security aware AI systems
b. AI systems that interact with non-traditional legacy systems such as IoT, pervasive networks and social media platforms

B. Key concerns to address

a. Ethics in AI is placed as one of our highest key concerns. With the use of advanced AI algorithms, discriminatory and biased algorithms are increasingly available and

it is essential to educate the workforce on basic ethics of using such powerful AI systems.

b. AI-inspired Social Engineering attacks. With the advancement in collecting and analyzing digital footprints from cyberspace users, the adversary can build a system that expertly impersonates individuals and can create a more believable platform to use for social engineering attacks.

C. Aspirations to sustain success

a. The AISP Research Lab with its successful track record of receiving funds from government agencies ($2.5M since 2019), and a large, collaborative group of research fellows, scholars and students is capable of assiting government agencies and private organizations in proper use of AI-inspired systems in a safe and secure manner.

b. The AISP lab with over 60 peer-reviewed publications in national and international journals and conferences plays an important role in dissemination of knowledge to empower our community to utilize AI systems with security and privacy in mind.

D. Lessons learned from our work

We have learned that having the most secure and trusted AI systems are only one piece of the puzzle. The most important piece is to build secure, affordable and user-friendly AI systems that provide utility, security and other benefits to individual users. It has been proven over and over that systems that are secure but introduce a lot of complexity are not effective and they can be counterproductive as they discourage users to engage with the AI systems. An additional risk is that overly complex AI systems can become vulunerable to adversaries using ever-increasing malicious techniques to misuse their data.

E. An invitation to share

a. The AISP Lab invites the Government to a no-cost platform to promote the AI R&D Plan as a key part of our annual National Cyber Security Month Virtual Symposium – we would be honored to host sessions with a dedicated conference track on the subject so your experts can describe the plan and invite a diverse group of AI experts to join the effort to implement the plan.

b. The AISP Lab commits to promote the AI R&D plan through alignment with our curriculum to train the next generation of AI researchers, developers, policy makers and entrepreneurs.

c. The AISP Lab plans to establish a multidisciplinary, applied internship program with SUNY cybersecurity and other students and our Innovation Fellows to help develop and implement the AI R&D plan on a local level with small businesses, K-12 schools, first responders, city government and other traditionally

underprepared sectors as a model to be replicated with universities across the country.

We hope the above points would respectfully assist your office and we are looking forward to our collaboration in building safe and secure AI enabled communities.

Sincerely,

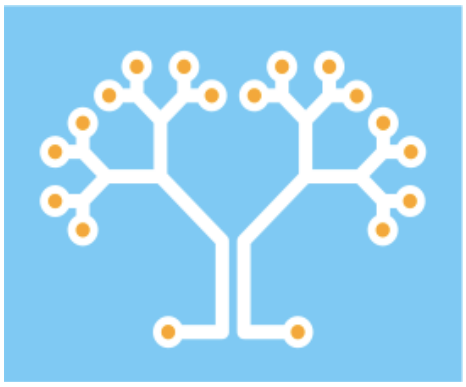
Kambiz Ghazinour, PhD.
Director of the Advanced Information Security and Privacy (AISP) Research Lab
Associate Professor, State University of New York at Canton.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

The Enterprise Neurosystem

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.



The Enterprise Neurosystem

RFI Response:

Update of the National Artificial Intelligence Research and Development Strategic Plan

Introduction

The Enterprise Neurosystem is an open source community of Fortune 500 companies, enterprise technology vendors, and academia. It is ultimately designed to address the fundamental challenges of large-scale AI infrastructure to protect our global ecosphere. The founding academic institutions include Stanford SLAC, Harvard Analytics and UC Berkeley Data-X. Participating firms include America Movil, Equinix, Fiducia | AI, IBM Research, Intel, Kove, PerceptiLabs, Verizon Media/Yahoo!, Red Hat, Reliance Jio and others.

The Challenge

Climate change, energy security, agricultural challenges, political unrest, and mass migration pose immediate threats to global stability and the planet's health as a whole. The exponential growth of technology provides us with the means to address these challenges, but only if we, as a nation, can build an enabling infrastructure to leverage that technology effectively. The National Artificial Intelligence Research and Development Strategic Plan guides our opportunities to unlock innovations that the culmination of emerging heterogeneous systems will merge to reveal cross-domain patterns in data, algorithms, technology, and discoveries that lead to robust solutions for the longevity of our public, private, and commercial way of life.

The Enterprise Neurosystem community sees the tailoring of the OSTP Strategic Plan as a means to more optimally guide the construction of a wide-ranging AI infrastructure, one with broad public and private buy in, that will enable rapid application of US creativity to solve the challenges that lie ahead – identifying and circumventing environmental threats, providing advanced insight into human migration, integrating climate-adaptive agriculture from local to global scales, dynamically accommodating supply chain and energy grid disruptions, and ultimately advancing secure fusion and nuclear energy in a heterogeneous supply producer ecosystem. The mission here is to promote sustainably accelerated advancement that is consistent with our hypothesis that diversity and heterogeneity in technology as in thought will yield resilient solutions in a changing world, a hypothesis support by US history and by Nature's own demonstration of using diversity as a tool for bending but never breaking under stress.

Planetary scale objectives begin with smaller developmental steps at the intersection of the scientific community and the Fortune 500 enterprise market. For lasting and meaningful change, government sponsored scientific advancement must quickly integrate into the commercial market and there grow legs of its own to live and breathe beyond the initial research seed funding. By nurturing a mid-tier environment for the national AI technology resource, it extends far beyond academic research and yields a continuous stream of commercially viable architectures that can self-coordinate as an autonomous and self-aware AI ecosystem. This cannot be overstated, the rapid spin-off productization enhances the next technological development, and this compounded multiplicative effect is exactly the root of the Kurzweilian exponential technological growth law.

A mere handful of major companies cannot harvest the rich intellectual capital in the US research and technology sectors. This Intellectual capital is best served by a quintessentially US culture that encourages and rewards ambitious entrepreneurialism and creative expression. In this way, our national culture provides ideal raw creativity to incubate the solutions needed for our–and the world's–challenges. Honing the forward-minded National AI Research Initiative will provide the tools and the connectivity to exponentially accelerate US innovation in AI to solve the globe's largest problems.

In the field of AI, innovation always begins from a very data and compute intensive position. New ideas require enormous high-performance computing (HPC) and data resources unavailable to garage-level startups. The strategy of the National AI Initiative ideally creates a national infrastructure that more than enables, it encourages exploratory development and provides the tools and platforms that foster discoveries and faster time to market. Having the US government serve this role facilitates a fair entry to the field for newcomers, accelerating novel hardware and algorithms, all while centralizing our ability to forecast the importance of emerging solutions and use that value forecasting to unlock computing resources that otherwise are not available to rising researchers and entrepreneurs.

The Solution

There is a need not only for the curation of datasets and published "data challenges" but also the access to the appropriate computing infrastructure that only the US Federal Government can support. There is an emerging ecosystem of heterogeneous AI accelerating hardware; the hobbyist scale of compute is no longer a viable option to make full use of the current state of the art in e.g., transformer models like BERT and GPT. The core of such models are trained using a vast set of exemplar training samples without the need of computationally expensive round truth labels.

This modality is well positioned for so-called federated machine learning, whereby participants can effectively share their data without exposing that data explicitly to others.

The core transformer benefits tremendously from generalization across data sources, while the final training for niche tasks relies on a vastly smaller sub-set of well labeled data with only the peripheral neurons as “last mile” trainable parameters. This tailoring phase is therefore compatible with the smaller scale computing resources available to the individual niche stakeholders. Since the High Performance Computing resources like clusters of GPUs, TPUs, or IPUs and such provide the rapid training of the core transformers, allowing access to AI testbeds that are managed by the federal government is in perfect alignment with using public funding for private and public explorative computing that leverages the latest in emerging AI acceleration hardware.

With truly world leading AI infrastructure resources as federal government managed HPC infrastructure, a quantifiable metric system with a standardized use-intention classification system can be developed that allows broad public exploration during low-load periods. As is common in Department of Energy cluster management, exploratory jobs would be preempted when high peak load needs arise. The social/scientific impact metrics and data-producer intention identifiers can be used to encourage participants to follow best practices to promote high value and ethical outcomes in the form of carrots rather than sticks; high value and high ethics models and data will be the last to be preempted from compute resources. Being managed by the federal government and luring participation by the broader AI development community, the US could model a system of autonomous bias monitoring that could effectively warn users when their models might grow corrupted due to either accidental or even nefarious dataset poisoning. Such an agent would be a digital analog of biological T-cells, autonomously scanning the central models and distributed datasets for indications of mis-use or unintentional outcomes. Altogether, the federal government solves multiple challenges simultaneously; it exposes the next generation of AI developers to the latest high performance technology, it attracts commercial financial and hardware support for community shared resources, and it holds the world’s most advanced AI infrastructure at the ready so that when catastrophe does strike, all resources can be immediately turned to the singular emergency response.

In the Enterprise Neurosystem community, there is a study group dedicated to potential ethical and safety issues in using AI technologies that influence or govern humanity. Our community also recognizes the underlying often accidental societal and cultural bias in curated data sets, affecting even the most granular AI determinations. Advanced feature engineering techniques can be broadly encouraged to help mitigate latent human bias, manage preference and develop ancillary AI models that assist the core intelligence and overall system toward driving greater accuracy and equity. Such an ethics system would reside at the kernel level within the primary guidance model to encourage beneficial outcomes for humanity.

Our community has noted the preponderance of AI models being deployed in the enterprise. They currently exist as bespoke solutions that lack deep integration with other models or domains, thus precluding the innovation that only arises from the cross-fertilization of collective findings. A connective data fabric and a centralized cross-correlation model are required for a national AI infrastructure that aligns with the long-term objective to monitor the planet and take real-time actions as needed. This connectivity will involve the emerging swarm of Edge IoT devices, data sources, and AI models, with a core interpretive model and recommendation engine.

A distributed infrastructure paradigm shift needs to take place. Standardized and composable feature extraction methods should be applied *in situ* to improve efficiencies in distributed model training with enforced confidentiality. Data should only be moved on demand or by way of interpretive metadata or interface layer, one that both increases security and reduces expense and extraneous network traffic. Metadata can be delivered in a tiered fashion to reduce latency and shorten the time to action. Data generalities will only give way to finer granularity based on user requirements, permissions, and authentication. For instance, data sampling techniques that include anonymization through hashing can lead to smaller data sources maintained in their respective silos in a federated model. Data and related features are shared only on an as-needed and as-granted basis.

The national data fabric infrastructure with a curated and multi-tiered security system to authenticate users and enable targeted data sharing would be a requisite primary focus of this new architecture. Independent AI-powered security instances will travel the network to scan and authenticate new users and data sources. In a non-intrusive manner, round-the-clock penetration testing will be enabled, and remote decryption key monitoring and related pattern analysis will be implemented. Instead of granting access to the entire network, focusing on Layer 7 application connectivity based on mTLS and Zero Trust Network Architecture (ZTNA) will isolate and reduce the impact of intrusions to a single application.

Although Public Cloud Offerings have demonstrated industries' acceptance and hunger for cloud-based HPC services, the newly emerging Edge computational and AI paradigms create novel security challenges as these resources are decentralized. Emerging advanced hardware is not currently available in Public Cloud Infrastructure. At the same time, users can already see tremendous benefit from a Federated model that exposes innovations in a secure sandbox. In such a sandbox, both scientific researchers and industry innovators could explore tailored hardware, even work together to co-design new technology for their specific research or market needs. A federated model would also allow data to remain resident in individual silos, fostering a safe but rich collaborative outcome via its tiered metadata capabilities.

Production Architecture

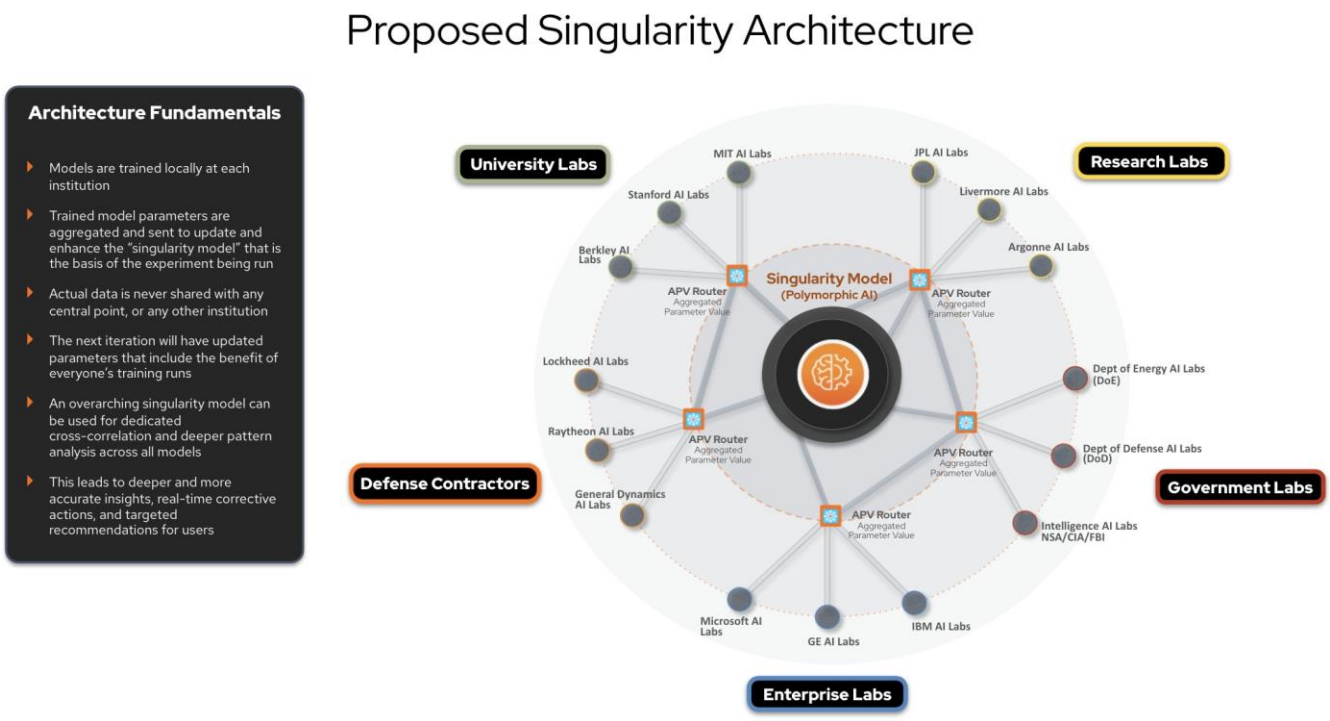


Illustration 4 - Tiered Cross-correlation Framework

Singularity refers to a pervasive synergy between human and artificial intelligence: an analytical engine that conducts ongoing pattern analysis, security reviews, and health checks across all the activities under its purview. As an opt-in model, not every function will require or be granted access to this full suite of resources, but the participating entities, be they private, public, or federal, would leverage a real-time analytic engine to advance science and directly inject technological discoveries into industry development. The complete body of federal research then becomes an immediate national AI resource. Enterprise and AI startups can as well propel intellectual property forward with new feature discovery, digital twin testing, and rapid time to market in an environment with maximal national exposure.

A series of intelligence engines, working together to unify streaming real-time and historical data to develop a more profound and instructive understanding of our environment and all its possibilities with the forecast possibility to enable dynamic response to challenges as they happen.

An Open Source Community Approach to National AI Infrastructure and Research

In terms of all RFI elements, it seems clear that an open community approach to the ownership and maintenance of this resource would be optimal. Essentially, a federated compute, storage, and AI development environment composed of multiple labs and data centers, connected on a highly secure and lossless framework, with a small group of dedicated paid resources to manage the common infrastructure instance. This funded management team would coordinate architectural upgrades and responses to technical issues.

The related hardware and software resources would be donated and shared by the various organizations and participants in a federated network architecture. A governance model for this resource could be based on open community principles via the Linux Foundation or similar community frameworks. Furthermore, given element D, it requires a multi-tiered approach to data access and provenance tracking through Distributed ledger-based technology. All domains will likely benefit significantly from HPC-enabled digital twin simulation and creation. This motivates an infrastructure that enables a unified framework among a diversity of components, from the IoT Edge to the HPC Core. Such heterogeneity will be critical to the success of this endeavor, and several platforms and emerging techniques can be combined to help address this requirement.

Commonly known required components:

- Open and closed data sets that help participants enable model training capability to be identified, built, and curated. Metadata frameworks would be created to increase efficiency and help navigate issues of privacy and bias.
- A generalized digital twin environment that supports discovery and data set generation.
- A user-permission and authentication mechanism.
- A platform to create, distribute and manage AI models, with capabilities including ground-up development, pre-built models, pipeline workflow, lifecycle management, and drift correction.
- A hardware environment including all necessary resources, including storage, processors (GPU/TPU/IPU/FPGA, x86, ARM), networking components, and software infrastructure platforms (Kubernetes, container management, databases, memory optimization, etc.).

Unique elements would include:

- A flexible and dynamic resource federation model that is based on an automated data fabric that extends across all elements. This helps private sector firms gain access to and directly build better hardware and software via a shared space that is co-designed with public sector emerging needs.
- A marketplace that enables both open and private science and industry, with models available to any organization. In essence, a library of templated architectures and base pre-trained AI models for related research purposes that ultimately enable cross-correlation of models, incorporating relevant heterogeneous data sources and sensors that can generate pattern analysis in real-time to enable deeper insights and rapid course correction.
- A library of composable transformation and featurization layers that aid user adoption of the templated architectures and enable security and provenance tracking from the beginning of the development cycle.
- A software-defined memory allocation and virtualization framework (Kove, etc.) that eliminates bottlenecks for large-scale AI workloads and optimally capitalizes on both on-premise and distributed data stores.
- A tiered security architecture that includes:
    - Ongoing 24/7 penetration testing with non-intrusive security scans.
    - Distributed ledger-based user/resource authentication.
    - Layer 7 integration strategy (intrusions relegated to a single application).
    - Biometric authentication and resource isolation within the federated network for sensitive workstreams.
    - Monitor decryption keys for data provenance, audit, and automated discovery of abuse patterns.
    - Independent software instances that autonomously assign themselves to assess and scan newly federated hardware.
- Open source fairness and bias management tools, provided by a dedicated community ethics group. This team will apply operational parameters to data sources, metadata layers, and cross-correlation engines to maintain systemic neutrality. This includes a tight focus on privacy and the protection of individual rights.

Capabilities and services for prioritization

- Curated and openly available data sets would be the top priority in our estimation.
- Various users and institutions can access openly available models in a registry.
- Readily available hardware infrastructure for model training, openly available to all member institutions. Beyond training requirements, the use of more expensive chipsets (GPUs, TPUs, IPUs, FPGAs, etc.) should be evaluated from a price/performance

perspective against standard chip architectures (x86, etc.) in terms of production lab use. This will lead to a balanced workload approach, cost savings, and co-designed hardware and algorithm compositions as well as encourage community exploration of emerging new technology.

Current building blocks and resources

The various national lab environments (Argonne, Oak Ridge, Stanford SLAC, etc.) can be unified as a federated infrastructure, using containerized technologies and integration via container platform and data fabric/layer 7 application networks. With a metadata engine and digital twin layer, the proposed federated approach would enable secured data to remain on-premises but would allow each of the partner labs to leverage the computing resources across the DOE complex.

Public-private partnerships

Community-based research organizations like the Enterprise Neurosystem serve as examples of grass-roots efforts that act as a crossroads between academia, government, and the private sector, with a common goal of creating global scale infrastructure for humanity and the environment. The high level of private interest in such an organization shows the appetite for participation in a fair and federally regulated AI infrastructure. By encouraging partnerships between industries, academia, and government, the resulting ecosystem provides an open arena for encouraging equitable access, early experience with developing technologies, and diverse insight into the strengths and weaknesses of participant approaches that ultimately yields a more robust AI infrastructure and industry.

With common overarching objectives, a multi-tiered project approach has been robust and proven in practice. Members and the private sector both support and contribute to the Enterprise Neurosystem proof of concept. The membership of the community provides the related infrastructure as an open-source and shared approach to hardware infrastructure, with set permissions for users mapped to projects. We have security development tracks, application layer connectivity, and a distributed ledger authentication end state. And the ethical guideline development track is a fascinating exercise in canvassing operational philosophies and multicultural guidelines to create an intelligence that will non-intrusively contribute to society in a positive and unbiased manner. As we have a lab and private sector participation with shared hardware and open and available data sets, this navigation has been pain-free to date.

There is strong support for academic, government, and industry projects within this community, and these activities allow the private sector to find benefits through federal research and vice versa. For example, creating a similar neurosystem for the Fortune 500 would support a Business Singularity for each corporation. A real-time intelligence that helps each corporation would be an evolutionary step in AI development that spans both AIOps and Business Intelligence functionality; creating mutual benefit for all parties is a primary objective.

Democratic access to AI R&D

Data sensitivity to sharing across organizations is a significant challenge for a national AI infrastructure. Federated Machine Learning and confidential computing best practices will certainly help address the challenge by maintaining the integrity of individual data silos and adding the abstraction layers of Federation and metadata generation to provide accurate results without compromising privacy or security.

Contributing Authors and Review Committee:

Ryan Coffee
Senior Staff Scientist
SLAC National Accelerator Laboratory

Ben Cushing
Federal Health and Science Lead
Red Hat Inc.

Erik Erlandson
Senior Principal Software Engineer
Red Hat Inc.

Ganesh Harinath
CEO
Fiducia | AI

Vishnu Hari Kumar
Product Manager
Meta AI

Sanjay Aiyagari
Principal Architect
Red Hat, Inc.

Pierre Mathys
Global Senior Director, Telco Edge Solutions
Red Hat, Inc.

John Overton
CEO
Kove

Audrey Reznik
Senior Software Engineer
Red Hat Inc.

Dinesh Verma
CTO, Edge Computing and IBM Fellow
IBM Research

Bill Wright
Head of AI/ML and Intelligent Edge
Industries and Global Accounts
Red Hat, Inc.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

The MITRE Corporation

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.

©2022 The MITRE Corporation. ALL RIGHTS RESERVED. Approved for public release. Distribution unlimited. Case Number 21-01760-16.

MITRE Response to OSTP's RFI Supporting the National Artificial Intelligence Research and Development Strategic Plan

March 4, 2022

For additional information about this response, please contact:
Duane Blackburn
Center for Data-Driven Policy
The MITRE Corporation
7596 Colshire Drive
McLean, VA 22102-7539

policy@mitre.org
(434) 964-5023

<<This page is intentionally blank.>>

About MITRE

The MITRE Corporation is a not-for-profit company that works in the public interest to tackle difficult problems that challenge the safety, stability, security, and well-being of our nation. We operate multiple federally funded research and development centers (FFRDCs), participate in public-private partnerships across national security and civilian agency missions, and maintain an independent technology research program. Working across federal, state, and local governments—as well as industry and academia—gives MITRE a unique vantage point. MITRE works in the public interest to discover new possibilities, create unexpected opportunities, and lead by pioneering together to bring innovative ideas into existence in areas such as artificial intelligence (AI), intuitive data science, quantum information science, health informatics, policy and economic expertise, trustworthy autonomy, cyber threat sharing, and cyber resilience.

MITRE's 50-year history of AI research and application, in partnership with federal agencies, has led to developing and supporting ethical guardrails to protect people and their personal data. Our team's experience with the entirety of the AI and machine learning (ML) adoption cycle has strengthened our ability to anticipate and solve future needs that are vital to the safety, well-being, and success of the public and the country.

Overarching Recommendations

From healthcare to national security, recent advances in AI can improve how we live our lives, modernize government operations, and increase national security. These same technologies can also create unintended consequences for democratic processes, mission-critical systems, and citizen privacy. Until recently, with rare exceptions, the idea of safeguards for AI systems was an afterthought. More needs to be done to mitigate bias, increase transparency, defend against attacks, secure the AI supply chain, and ensure the overall trustworthiness of AI systems so they perform as intended and are successfully applied in mission-critical environments. AI's potential will only be realized through collaborations that produce reliable, responsible, fair, explainable, transparent, traceable, privacy-preserving, and secure technologies.

AI is a complex technology that is unfamiliar to most of our citizens, despite the growing impact it has on their lives—and their personal use of it. Public perceptions of AI can range from hype that borders on fantastical Hollywood depictions to glowing press releases to dire predictions of dystopian futures. The same holds true for legislators and policymakers, except they are also bombarded by policy advocate messaging for and against the technology—with wildly varying degrees of accuracy. AI policies need to have scientific integrity[1]; they need to be based on rigorous evidence and methods rather than political objectives, fantasy, or fear.

The White House recently called for "widespread training for agency scientists so they can communicate scientific findings effectively to nonscientists in their agencies and to lay audiences, with the idea of helping ensure that policies and actions are based on an accurate

[1] Protecting the Integrity of Government Science. 2022. National Science and Technology Council, https://www.whitehouse.gov/wp-content/uploads/2022/01/01-22-Protecting_the_Integrity_of_Government_Science.pdf.

understanding of the science."[2] A focus on such accurate and effective communications is needed in each of the Strategies within this Strategic Plan. While outside the norm of National Science and Technology Council (NSTC) activities, there is precedent, as the prior NSTC Subcommittee on Biometrics and Identity Management focused significantly on communications matters so that policymakers, federal agencies, and the public better understood the then-nascent technology.[3] MITRE also recommends that sociological-based and stakeholder communication research be included in this R&D Strategic Plan to help understand the most effective way for scientists to explain AI, its applications, and issues to non-experts. Doing so will help advance the overall Strategic Plan while also ensuring that future policy deliberations will be based on evidence rather than hyperbole.

Questions Posed in the RFI

Input on potential revisions to the strategic plan to reflect updated priorities related to AI R&D. Responses could include suggestions as to the addition, removal, or modification of strategic aims, including suggestions to address OSTP's priorities of ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security, and to maintaining the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.

Overall, the 2019 Strategic Plan remains valid today, which underscores that we have been focusing on the proper areas of research, which are both high-priority and difficult to solve. Wholesale changes are not required, though some elements could be refined based on advancements and discoveries that have been made over the ensuing years.

<u>New Discussion (or Organization) on Trustworthy AI</u>
The AI community is now more regularly including ethical, legal, and social implications of AI within concepts of "trustworthy" AI, which in the past had predominantly focused on safety and security concerns. Doing so elevates human-centered concerns into the mainstream consciousness of technologists building systems. MITRE therefore recommends a similar linkage within the 2022 strategy—either combining existing strategy element 3 (Understand and Address the Ethical, Legal, and Societal Implications of AI) and element 4 (Ensure the Safety and Security of AI Systems) into a new overarching "Ensure Trustworthy AI" strategy, or otherwise crafting a stronger linkage between the two existing strategies via text discussion.

We also note that, throughout the 2019 Strategic Plan, there is discussion on topics such as AI explainability, transparency, traceability, trust, fairness, ethics, bias, equity,

[2] White House Office of Science & Technology Policy Releases Scientific Integrity Task Force Report. 2022. The White House, https://www.whitehouse.gov/ostp/news-updates/2022/01/11/white-house-office-of-science-technology-policy-releases-scientific-integrity-task-force-report/. Last accessed February 13, 2022.

[3] A National Science and Technology Council for the 21st Century. 2021. MITRE, https://www.mitre.org/sites/default/files/publications/pr-21-2388-national-science-technology-council.pdf.

responsibility/accountability, reliability/robustness, safety, and security—all of which are aspects of trustworthy AI. We maintain that it is important for these to remain important considerations within each Strategy in the document, but this content would be enhanced by a coordinated, collective discussion of trustworthy AI, with clear definition of the elements that comprise trustworthy AI.

MITRE previously crafted the following graphic, which links seven important elements into an umbrella concept of trustworthy AI. Adopting this, or a similar unifying concept, and bringing the Plan's content on trustworthy AI elements together will increase the importance, clarity, organization, and understanding of this topic, as well as provide a foundational reference to ensure that all elements of trustworthy AI are considered within each of the Plan's other Strategies.

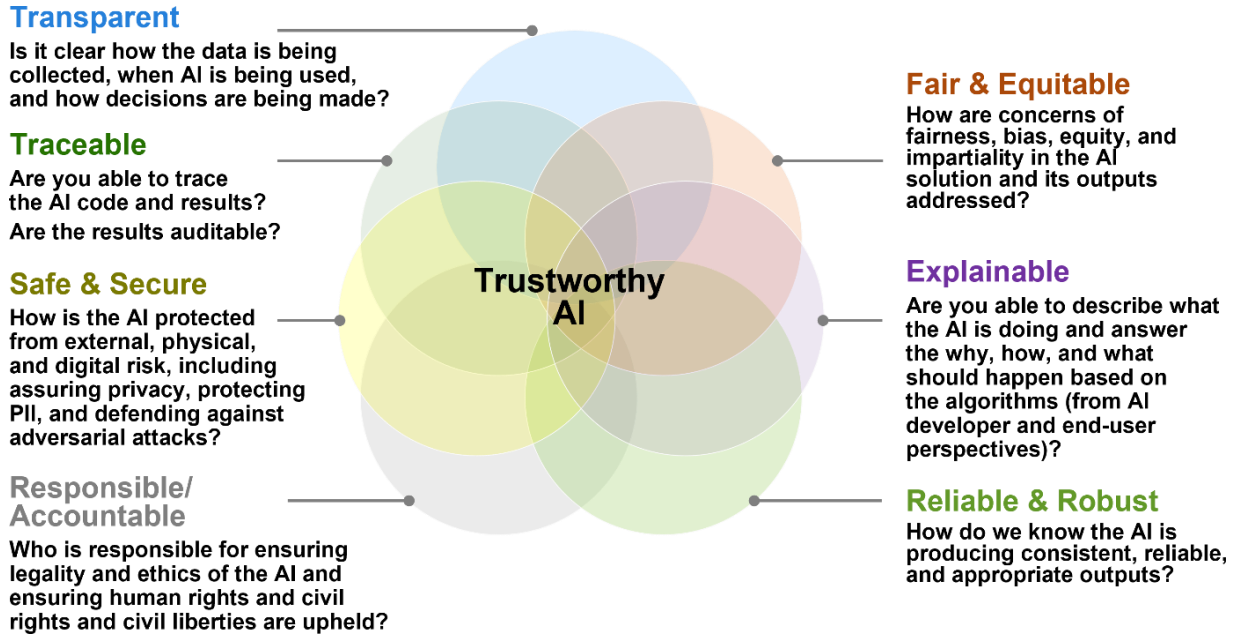


Figure 1 - Seven Connected Elements of "Trustworthy AI"

Existing Strategy 3

The legal implications of poorly performing AI on various parties (including leaders, business owners, project and acquisition managers, system designers, developers, testers, evaluators, operators, and maintainers) need to be researched. The results of such research will better ensure each of these professionals handles their responsibilities to advance and build AI capabilities. Acceptable use considerations, ethical principles, and a desire for equitable outcomes must also be incorporated into each lifecycle phase (e.g., needs validation, requirements definition, capability design, capability development, testing, system and process integration, operations, maintenance, system termination, and disposal).

Existing Strategy 5

Strategy 5 could be enhanced to better show how providing environments and datasets for researchers can also help overcome researcher inequities. ML, particularly deep learning, at scale

for large or complex problems requires massive computing resources. In 2018, estimates were that the amount of computing required for the largest AI training runs had increased by 300,000 times in just 6 years[4]. This yields a systemic imbalance between the "haves" and "have-nots" in terms of ability to acquire computational resources to use ML to solve their problems.

In addition to the computational resource problems, data availability and quality present a modeling challenge. For many modeling challenges, such as machine translation for less commonly spoken languages, the data simply does not exist[5], and gathering that data represents a prohibitive cost to organizations wishing to build AI systems. Just as with computational resources, this data challenge can leave groups behind and perpetuate systemic inequities.

Strategy 5 could be further expanded to call for additional shared resources that will enhance the R&D community's ability to "meet the needs of a diverse spectrum of AI interests and applications."[6]

1. Expand on the short mention at the end of the gray box at the top of page 29 about studying and investing in shared computational resources to promote AI R&D.[7]
2. Add a new subsection calling for research to study the needs for shared computational infrastructure to support AI R&D. Such shared computational resources will be needed to support access to and use of shared datasets and AI training and testing environments.
3. After the last paragraph on page 30 (the 2nd paragraph under "Developing open-source software libraries and toolkits"), add the following paragraph:
    a. Innovation requires adoption of AI technologies, meaning that people are using the AI and the AI is delivering mission value. Use of open-source software libraries and toolkits accelerates the transfer to AI technologies from R&D to missions in implementing agencies. Accelerated technology transfer, in turn, facilitates and accelerates AI adoption and innovation in government.

While computational resource and data democratization research are important elements, it is essential to research trustworthy AI approaches that do more with less. Approaches that improve and consider algorithmic efficiency in research address some computational resource concerns[8]. R&D into quantifying data needs for algorithmic performance prior to data acquisition and using this information to more effectively map problems to the best algorithmic solutions will reduce potential resource burdens as barriers to success. Finally, research into hybrid approaches that combine knowledge-based AI with ML to reduce the training burden should be pursued.

<u>Existing Strategy 7</u>

To incorporate AI into K-12 and undergraduate classrooms, educators must be empowered to quickly research, obtain, modify, create, and share lessons. Initial investment is needed to create

[4] D. Amodei and D. Hernandez. AI and compute. 2018. Open AI, https://openai.com/blog/ai-and-compute/. Last accessed February 25, 2022.

[5] C. Cieri and M. Liberman. TIDES Language Resources: A Resource Map for Translingual Information Access. 2002. Language Resources and Evaluation Conference (LREC) 2002.

[6] Quote is from the existing Strategic Plan, page 29.

[7] This research is currently being carried out by the National AI Research Resource Task Force. MITRE recommends the groundbreaking work of the Task Force be highlighted in the Plan's update, along with encouragement for sustained support as determined appropriate. See https://www.ai.gov/nairrtf/.

[8] R. Schwartz, et al. Green AI. 2019. arXiv, https://arxiv.org/pdf/1907.10597.pdf.

high-quality, domain-specific, and appropriately challenging lessons. To increase the likelihood of use in the classroom, educators must be trained on these materials through incentivized training targeted toward foundational and domain-specific AI competencies. This training needs to be supplemented with access to a platform that encourages inter-institutional data and code sharing and accessible development tools, including free code repository systems and development languages (e.g., R, Python). Given the current state of educator resources, equitable content adoption is dependent on additional policy-driven incentivization and support.

MITRE's existing Generation AI Nexus (GenAI) program[9] has provided a platform and incentives that make these sharing opportunities possible. MITRE is working with universities and private industry to develop students across the United States into thought leaders who can leverage the power of AI and data science and extend opportunities into areas outside of classic computer science programs. GenAI aims to help this generation become as comfortable with and proficient in working with AI as a prior generation was with the internet. GenAI makes available highly curated datasets and curriculum (lecture notes, in-class notebooks, and homework assignments) that any member can use, with the requirement that each educational entity also create and make available additional course material for others to use. By design, these lessons are built to leverage free, open-source, and commonly available private sector tools and platforms that reduce barriers to access, sharing, and usage.

The existing discussion to broaden participation among traditionally underrepresented groups should also be enhanced. Wholesale, national advancements may also require Department of Education attention. Educators have uneven levels of systematic incentives and flexibility that allow them to put in the necessary time and energy to learn, use, and build AI educational lessons for students. Thus, to broaden participation among groups traditionally underrepresented and enable equitable access, rural and minority-serving institutions need access to supplementary foundational content, incentivized support, and a free analytics infrastructure that is web-based, integrates with open-source tools and training data, and is accessible on mobile phones. While there may be an abundance of interest (and capabilities, such as in GenAI) in widening opportunities for students in data science and AI capabilities, educators need the necessary top cover, flexibility in curriculum development, easing of existing time demands, and establishment of personal growth opportunities and recognition to encourage participation in programs to introduce AI into their learning environments.

Suggestions of AI R&D focus areas that could create solutions to address societal issues such as equity, climate change, healthcare, and job opportunities, especially in communities that have been traditionally underserved

To address societal issues via AI R&D, the R&D strategy should prioritize interdisciplinary research efforts that involve social scientists and community engagement with those who are experiencing the problem being addressed, starting at the design of the project. An analysis of AI research publications has found that the distance between social science fields and AI research has grown over the past decades, likely driven by the more technical focus of industry-funded

[9] Generation AI Nexus. 2019. MITRE, https://ainexus.org/home. Last accessed February 23, 2022.

R&D.[10] AI development efforts lacking social analysis are likely to oversimplify the intended task, encounter challenges in social adoption (Did the intended users want the capability? Does it meet their needs?), and neglect to anticipate secondary social consequences[11]. The U.S. government can help reverse this trend.

The strategy should also further address application-driven research by promoting research projects that have explicitly predicted impact on a set of real-world benchmarks of national objectives, with special attention to social equity. The U.S. government can complement the AI for Social Good[12] movement, which promotes the United Nation's Sustainable Development Goals (SDGs) as objectives for AI projects, with projects aligned with the SDGs.

Just because a research area could relate to an SDG does not mean it will advance the SDG and may even work against other SDGs. Therefore, the connection to all relevant SDGs should be expressed in a logic model[13], specifying how the research project might effect change, building beyond the traditional broader impact statement in research proposals[14]. Such an impact prediction should also make the case for differentiated impact—why an AI solution is better than a non-AI based one, including the current state. If a major AI R&D investment requires a cost-benefit analysis, the analysis should consider the distributional effects of the benefit[15], with special consideration to the effects on disadvantaged populations, akin to an equity assessment for federal programs and policies[16].

Community engagement, participatory design, social science co-authorship, and social impact/equity impact logic models support social impact and harm reduction on a project-by-project basis. There are also some technical research fields of AI/ML that, when integrated into R&D projects with participatory design and social analysis, may advance social impact and equity goals more readily than others. For example:

- On-device AI, in conjunction with participatory design and socio-technical research into point-of-care use and expectations, to aid healthcare and other service delivery and various jobs in low-resource communities
- Language processing for underserved languages and dialects to increase equity of access to services, education, and job markets

[10] M. Frank, et al. The evolution of citation graphs in artificial intelligence research. 2019. Nature Machine Intelligence, https://doi.org/10.1038/s42256-019-0024-5. Last accessed February 28, 2022.

[11] E. Dahlin. Mind the gap! On the future of AI research. 2021. Humanities & Social Sciences Communications, https://doi.org/10.1057/s41599-021-00750-9. Last accessed February 28, 2022.

[12] N. Tomašev, et al. AI for social good: unlocking the opportunity for positive impact. 2020. Nature Communications, https://doi.org/10.1038/s41467-020-15871-z. Last accessed February 28, 2022

[13] Systems Engineering Guide. P. 76-81, MITRE. 2014. https://www.mitre.org/sites/default/files/publications/se-guide-book-interactive.pdf.

[14] Broader Impacts Improving Society. 2022. National Science Foundation, https://www.nsf.gov/od/oia/special/broaderimpacts/. Last accessed February 28, 2022.

[15] N. Nelson and A. Bohmoldt. 2021. MITRE, Benefit-Cost Analysis and Consideration of Distributional Effects and Social Equity.

[16] A Framework for Assessing Equity in Federal Programs and Policies. 2021. MITRE, https://www.mitre.org/sites/default/files/publications/pr-21-1292-a-framework-for-assessing-equity-in-federal-programs-and-policy.pdf.

- Graph ML and optimization algorithms on temporal social networks so that AI-aided policy evaluation and resource allocation decisions can better account for different needs, relationships, and resources between communities
- Approaches to decision making with inconsistent, indirect, qualitative, and limited data, such as transfer learning or learning with Partially Observable Markov Decision Processes or Bayesian Machine Learning, since critical social impact applications may not have large datasets or, if they do, may not include sufficient data for historically under-represented populations.
- Causal ML and other approaches that can evaluate the impact of interventions, instead of relying on correlative models that may pick up on the fact that socio-economic and health disadvantages are correlated
- Multi-agent reinforcement learning, distributed control, and other areas that support dynamic resource allocation in complex systems to consider equity and relative experiences in a changing environment
- AI R&D focus areas that could create solutions to address climate change include computer vision and time series analysis to support spatiotemporal causal modeling of the complex relationships between climate and human health. Better understanding these relationships can help local communities make more informed decisions about preventative programs and build resilience plans.
- AI R&D in the health domain should include securing diverse and accessible datasets. Special attention will need to be provided to the underprivileged and to rural areas that may have less capacity (such as qualified staff, staff time, or financial resources) to dedicate to data collection and reporting.

Comments for the strategic plan are welcomed regarding how AI R&D can help address harms due to disparate treatment of different demographic groups; research that informs the intersection of AI R&D and application with privacy and civil liberties; AI R&D to help address the underrepresentation of certain demographic groups in the AI workforce; and AI R&D to evaluate and address bias, equity, or other concerns related to the development, use, and impact of AI.

Bias and Equity

Bias is and will remain a persistent issue for AI, as it is both inherently probabilistic and a creature of the data used for training during its development. "Bias" is also a word that has different meanings to different people[17], leading to inaccurate conflations that have negatively impacted policy deliberations, research, and usage of AI. While this has been a longstanding issue, it has significantly increased since the 2019 Strategic Plan and needs to be directly addressed in the 2022 update. Technical biases, operational biases, and prejudicial biases are not equivalent, though they are often discussed as such in policy advocacy materials, press articles, and the public's deliberations on social media. Similar to MITRE's recommendations in our

[17] D. Blackburn. When and How Should We "Trust the Science"? 2021. MITRE, https://www.mitre.org/sites/default/files/publications/pr-21-1187-when-and-how-should-we-trust-the-science_0.pdf

response[18] to OSTP's prior *RFI on Public and Private Sector Uses of Biometric Technologies*, we recommend that this Strategic Plan properly describe and focus on each type of bias distinctly and accurately. Neglecting to do so will likely mean that these conflations continue to occur, to the detriment of the Plan's ability to advance national capabilities. Bias is obviously a key component within the "fair and equitable" element of trustworthy AI, as described above. Research to identify and minimize technical, operational, and prejudicial biases of AI should be a part of this Strategic Plan so that we can achieve equity by overcoming harms due to disparate impacts on various demographic groups.

Diversifying project teams also ensures a variety of experiences and perspectives (including gender, ethnic, financial, geographic, educational, and use case experiences) during the development of AI capabilities and systems that incorporate them. Leveraging these varied insights at each lifecycle stage will likely reduce differential performance issues.

Research Approaches

Federated learning is an approach that can be used to protect privacy in AI research as it involves a decentralized training method of algorithms. In healthcare-focused research, for example, several participating institutions could train ML algorithms locally, without sharing patients' data outside of the hospital. Subsequently, they share only model characteristics with external partners to improve decision making. Studies showed that such an approach performs comparably to other ML models. But the advantage of this collaborative technique is that sensitive data does not leave the hospital.

The U.S government can prioritize research on applying equity-centered design principles to AI design, as well as research projects that adopt participatory approaches in the research process. AI design that involves the communities that will be interacting with the AI system yields better solutions than AI systems only assessed for bias at the conclusion of the design phase. Fully participatory methods are not always feasible, but advancements in human-in-the-loop simulation and computational social models can support more socially responsible ML model design and training. The federal government can invest in networks of public-private partnerships to build social models and collect public interest datasets (such as those supporting national social priorities) that commercial R&D does not have the immediate incentive or scale to collect. Federal investment in such data can model fairness best practices, such as requiring researchers to create transparent datasheets for datasets[19] and perform bias audits.

Ensuring equitable impact also requires research on the social and behavioral interaction between the AI systems and populations served, as well as environments—in vivo, qualitative human feedback, simulated, hybrid, and computational social models[20]—that explore the downstream distributional effects of AI systems on individuals, communities, and institutions. This impact may extend beyond those directly at the receiving end of the AI system; the U.S. government

[18] MITRE Response to OSTP RFI on Public and Private Sector Uses of Biometric Technologies. 2022. MITRE, https://www.mitre.org/sites/default/files/publications/pr-21-01760-11-mitre-response-information-on-public-and-private-sector-uses-of-biometric-technologies.pdf.

[19] T. Gebru, et al. Datasheets for Datasets. 2021. Communications of the ACM, https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/fulltext. Last accessed February 28, 2022

[20] J. Egeth, et al. Sociocultural Behavior Sensemaking: State of the Art in Understanding the Operational Environment. 2015. MITRE, https://www.mitre.org/sites/default/files/publications/SocioculturalSensemaking.pdf.

should support research into effects on those left behind by the adopted AI solution, and the impact of the *use* of the AI solution, not just the AI algorithm itself.

Several ML algorithm research areas are likely to support more equitable AI, such as causal ML, multi-model techniques that can incorporate human feedback into the system's response, and federated learning, as discussed above. The federal government should prioritize fundamental technical research that makes the case for its utility to fair and equitable AI development.

Comments on strategic directions related to international cooperation on AI R&D and on providing inclusive pathways for more Americans to participate in AI R&D

Recommendations from (or based on) the National Security Commission on Artificial Intelligence Final Report[21] apply here and provide guidance on directions related to international cooperation on AI R&D, such as shaping global norms and standards and expanding cooperation with allies.

Comments are invited as to existing strategic aims, along with their past or future implementation by the Federal government

Comparing Performance and Use

The use of AI systems is often denigrated or abandoned because these systems cannot meet desired (sometimes unrealistic) performance and use expectations, even though their usage would be significantly better than the status quo. AI-enabled identity solutions, for example, are often referred to as "nascent" or "too immature" if they are not perfect or completely absent of performance differentials, even though neither is theoretically possible to obtain and these AI solutions' current capabilities are already significantly better than what can be provided by trained humans alone. Similar issues occur in health-based clinical settings because of the need to justify or defend the decision made based on the prediction. For instance, if a predictive model detects cancer in a screen, a process of verification must then occur. In some situations, such as when the detection is obvious, this unnecessarily delays patient treatment. It is essential for government R&D and implementing agencies to properly deliberate appropriate use in the future.

The updated Strategic Plan should address this issue, enabling the community to address unrealistic lobbyist messages about research advancements and helping to open up avenues for appropriate usage. Research could also be performed to analyze prior applications and begin developing models to help match appropriate AI to different use cases, thus increasing overall trustworthiness of AI. This process of modeling AI-enhanced decision making has the added advantage of showing areas where additional research into human-machine teaming are needed to increase trustworthiness.

[21] Final Report – National Security Commission on Artificial Intelligence. 2021. National Security Commission on Artificial Intelligence, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

Technology Transfer and Feedback

Existing strategic aims and their future implementation by the federal government can be achieved with greater success if the R&D community improves how technologies are transferred from R&D to missions (in implementing agencies) and how mission user needs inform R&D efforts. One way R&D can improve this two-way exchange is to establish "hubs" with collaborative spaces and shared, reusable resources including datasets, software libraries and toolkits, previous models, lessons learned, subject matter expert (SME) points of contact, training environments, test and evaluation environments and testbeds, compute infrastructure, storage, and standards. Such reusable resources and collaborative spaces promote cross-pollination and information sharing and can advance the existing Strategic Plan's aims of government AI innovation and "fostering AI R&D in the open world to provide design of AI systems that incorporate and accommodate the situations and goals of users" (see page 15 of the current Plan). Hubs with collaborative workspaces provide a mechanism for team members from R&D and implementing agencies to work together on R&D projects with increased potential for transfer, implementation, and adoption in agencies.[22]

Cross-Functional Teams

In any AI R&D effort and in any government AI adoption effort, a cross-functional team should be involved from the very beginning and throughout the effort to ensure that all perspectives and aspects of the mission are included. This cross-functional team should include ethicists, privacy and personally identifiable information SMEs, community juries (at the right time), legal and civil rights/civil liberties SMEs, as well as team members in AI governance, project management, acquisition management, organizational change management, business process design, data management, AI model development and operations, and IT systems and infrastructure (to address enterprise architecture, interfaces, and integration). Bringing all of these skills and expertise together from the beginning and throughout the duration of R&D efforts will help the federal government achieve its aims set forth in this Strategic Plan.

[22] The writers of the updated Strategic Plan should consider how the government's strategic investment in NSF AI Research Institutes might be a vehicle for implementing R&D hubs as described. See https://beta.nsf.gov/funding/opportunities/national-artificial-intelligence-research-institutes.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

Twilio

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.

Subject: RFI Response: Na-onal Ar-ficial Intelligence Research and Development Strategic Plan
Date: Friday, March 4, 2022
From: Michael Seeds
To: AI-RFI

March 4, 2022

White House Office of Science and Technology Policy
Executive Office of the President
Eisenhower Executive Office Building
1650 Pennsylvania Avenue NW
Washington, D.C. 20504

Re: Comments of Twilio on the Update of the National Artificial Intelligence Research and Development Strategic Plan

Twilio welcomes the opportunity to comment on the White House Office of Science and Technology Policy (OSTP) request for information regarding updates to the *National Artificial Intelligence Research and Development Strategic Plan*.

Twilio supports the overall goals of the Strategic Plan and welcomes OSTP's contemplated revisions. We support the approach of the U.S. government in facilitating impactful research and development (R&D) in the AI space, especially through the support of a diverse AI workforce. We believe that facilitating industry-led and internationally-coordinated research, development, and standardization efforts are crucial to maximizing the benefits of AI while minimizing potential harms.

I. About Twilio

Twilio is a leading b2b global cloud operator that enables other businesses, governments and nonprofits to embed communications, such as voice, text messaging, email, chat and video, into their existing web and mobile applications to enhance their engagement with their customers and constituents. Organizations have used Twilio to allow their end-users to contact their teacher or students, alert the public about an emergency, video chat with their doctor, speak with their rideshare driver, make a bank transaction, shop online, authenticate an account, and interact with elected officials, among many other activities.

Founded in San Francisco in 2008, Twilio now has 26 offices in 16 countries and the infrastructure to support communications worldwide. Twilio provides services to more than 256,000 enterprises globally and powers more than 1 trillion interactions between them and their customers every year. Twilio customers range from small and medium-sized enterprises (SMEs) to the world's largest corporations and come from a broad range of industries including financial services, health care, manufacturing, retail, education, and logistics. Twilio's non-profit arm, Twilio.org, supports charitable organizations to deliver their communications needs, such as the American Red Cross. Twilio is also a technology partner and supporter of the United Nation's Vaccine Alliance GAVI.

II. How Twilio uses Artificial Intelligence

Twilio leverages AI in order to support the company's mission of enhancing communications and customer

engagement. Twilio uses AI and high-quality data training sets to create products that allow innovative companies to use AI tools that drive efficiency, responsiveness, and customer satisfaction. In its internal processes, Twilio works to ensure responsible use of AI, with particular focus on the protection of accuracy, privacy, security, and transparency.

Twilio currently offers one AI-powered service, *Twilio Autopilot*. Autopilot is an AI interface that bridges the gap between human agents and self-service bots. It allows developers to build intelligent Interactive Voice Response (IVR) systems, bots, and applications that are powered by Twilio-built Natural Language Understanding and Machine Learning frameworks. Through these technologies, Autopilot is able to turn nested phone trees into simple "what can I help you with" voice prompts and allow customers to use voice search to access a knowledge base. In addition, Twilio is actively looking at ways to incorporate AI into future products that will benefit consumers.

III. Twilio Supports OSTP and US Government Leadership at Home and Abroad

Twilio supports the Administration's efforts to facilitate the development of AI and the Plan's focus on researching the ethical, legal, and societal implications of its uses. Security, privacy, and the integrity of data sets are critical for effective AI, and encouraging R&D in these areas is important to developing effective solutions, mitigating potential harms, and fostering public trust.

It is also critical to coordinate with industry and strategically aligned international partners to enhance the Administration's efforts in this realm. As jurisdictions around the world develop their approach to AI, global alignment on regulatory approaches and standards is critical to fully realizing the benefits of the technology. To this end, Twilio supports international efforts to coordinate AI R&D and stresses the need to ensure that any policies seek to be future-proof and forward-looking. In this respect, U.S. leadership is important to ensure that global AI policy efforts are informed by consultation with leading private sector companies and to ensure that innovation, commerce, and critical user services are not impeded by well-meaning legislative or regulatory actions.

IV. Pursue Consensus-Driven International Standards in Support of OSTP's Policy Priorities.

Standards and benchmarks are a foundational component of the *National AI R&D Strategic Plan*, recognized both in their own right and as a cross cutting element across strategic goals. This is rightly embodied in *Strategy 6: Measure and Evaluate AI Technologies through Standards and Benchmarks.* Standards and benchmarks are also a crucial element that can enable progress towards other goals embodied in the strategy, including addressing the ethical and society implications of the technology, ensuring security, and fostering the long-term scaling of the AI ecosystem.

As AI technology continues to change and evolve, industry is continually developing and refining approaches to mitigate potential risks. These efforts bring critical nuance and expertise that can inform both those deploying AI and policymakers seeking to regulate it. Some of the most important challenges in AI today, e.g. developing and scaling effective and commonly accepted means to detect and mitigate bias in AI systems, depend on the formalization of technologies, processes, common metrics, and techniques that can be trusted and verified. We encourage the US government to work with private industry and coordinate with its trusted international partners to support consensus-driven international standards that can inform responsible AI development and thoughtful policy both at home and abroad.

V. Support Inclusive Pathways for Americans to Participate in AI R&D

Twilio also supports OSTP's and the National Artificial Intelligence Initiative Office (NAAIO)'s efforts to ensure AI technology R&D can benefit everyone in the United States. As a company that develops AI tools

for clients around the world, we acutely understand the importance of considering fairness and inclusivity and ensuring it is reflected throughout each stage of the AI lifecycle.

One key component of this is the AI workforce. In order to build a more diverse workforce capable of more holistically addressing societal impacts, Twilio supports federal efforts to coordinate education and training activities; funding for AI research, development, and demonstration; and the network of artificial intelligence research institutes.

Additionally, ongoing efforts to “maximize the benefits of science and technology to advance health, prosperity, security, environmental quality, and justice for all Americans” are an especially critical goal of the strategic initiative. As such, Twilio supports contemplated revisions to the National AI R&D Strategic Plan to include more inclusive pathways for more Americans to participate in AI R&D. By increasing the number and diversity of students, researchers, and professionals active in AI, America’s national AI base will be better prepared for the challenges of tomorrow.

VI. Conclusion

Twilio supports the Administration’s efforts to revise the National AI R&D Strategic Plan and OSTP’s efforts to focus revisions on strategic international cooperation and inclusive pathways for more Americans to participate in AI R&D efforts. Twilio welcomes the opportunity for future engagement with the Biden Administration on this important policy matter.

Best,

Michael Seeds
Director, Government Affairs

What’s Twilio? It’s your new superpower to build conversations anywhere. See how it works.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

U.S. Chamber of Commerce Technology Engagement Center

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.

U.S. Chamber of Commerce

1615 H Street, NW
Washington, DC 20062-2000
uschamber.com

March 4, 2022

Office of Science and Technology Policy
NCO
2415 Eisenhower Avenue
Alexandria, VA 22314

Re: Update of the National Artificial Intelligence Research and Development Strategic Plan

To Whom It May Concern:

The U.S. Chamber of Commerce's Technology Engagement Center ("C_TEC") appreciates the opportunity to submit feedback to the Office of Science and Technology Policy ("OSTP") in response to its request for information on how to best update "the National Artificial Intelligence Research and Development Strategic Plan.[1]" C_TEC appreciates OSTP's effort to determine "ways in which the strategic plan should be revised and improved.[2]"

1.) Suggestions as to the addition, removal, or modifications of strategic aims. Address OSTP priorities of ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security. Maintaining the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.

C_TEC strongly believes that Artificial Intelligence will significantly change our world for the better. It is paramount for the United States to lead the development of Artificial Intelligence for economic and security purposes. The United States' underlying democratic values provide us with a foundational set of rights such as liberty, free speech, freedom, due process of law, and freedom of assembly, all essential factors in facilitating the development of fair and ethical uses of Artificial Intelligence. These values are coupled with Americans'[3]belief that Artificial Intelligence can help society at large provide a significant advantage to the United States.

However, while these values are essential and can serve as a catalyst, the United States' values and optimism will not be enough for the U.S. to lead globally. Global

[1] https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and
[2] https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and
[3] https://americaninnovators.com/wp-content/uploads/2022/01/CTEC-US-Outlook-on-AI-Detailed-Analysis.pdf

leadership in Artificial Intelligence is heavily predicated on providing the correct regulatory climate that does not stifle innovation and provides the necessary investments into areas such as: education, hardware, and overall innovation. This is why C_TEC strongly supports the "National Artificial Intelligence Initiative Act[4]" (NAIIA). The NAIIA provides significant federal funding for the development of a "diverse workforce pipeline for the science and technology for Artificial Intelligence systems[5]" as well as grants for the development of "hardware for accelerating Artificial Intelligence systems[6]" and the development of the "National AI Research Resource Task Force."

Furthermore, the ANAIIA also requires the development of a voluntary risk management framework, which is currently being developed through the National Institutes of Standards and Technology ("NIST"). We strongly support the development of a Risk Management Framework around the development and life cycle of Artificial Intelligence systems and appreciate NIST's work to bring together "stakeholders to collaborate and address how to mitigate risks stemming from AI."[7] We believe that the development of "trustworthy AI is a partnership[8]" and that "governments alone cannot promote trustworthy AI. The Chamber believes that governments must partner with the private sector, academia, and civil society when addressing issues of public concern associated with AI."[9]

OSTP has a significant opportunity to be the driving force within the federal government in executing the National Artificial Intelligence Initiative. This effort will help ensure the U.S. government can meet the rising challenges through its continued support for increasing investments in research and development and supporting the development of AI-related voluntary consensus standards.

2.) Responses could include suggestions on AI R&D focus areas that could create solutions to address societal issues such as equity, climate change, healthcare, and job opportunities, especially in communities that have been traditionally underserved.

C_TEC strongly supports research and development on Artificial Intelligence that addresses societal issues. Continued research and development can provide significant

[4] https://www.uschamber.com/technology/coalition-letter-the-national-artificial-intelligence-initiative-act-naiia-of-2020
[5] https://science.house.gov/imo/media/doc/AI_initiative_SST.pdf
[6] https://science.house.gov/imo/media/doc/AI_initiative_SST.pdf
[7] https://www.uschamber.com/technology/coalition-letter-the-national-artificial-intelligence-initiative-act-naiia-of-2020
[8] https://www.uschamber.com/technology/us-chamber-releases-artificial-intelligence-principles#:~:text=Fostering%20public%20trust%20and%20trustworthiness,explainability%2C%20fairness%2C%20and%20accountability.
[9] https://www.uschamber.com/technology/us-chamber-releases-artificial-intelligence-principles#:~:text=Fostering%20public%20trust%20and%20trustworthiness,explainability%2C%20fairness%2C%20and%20accountability.

benefits to develop data-driven solutions in equity, climate change, healthcare, and employment, especially for those who have been traditionally underserved.

For example, the federal government is increasingly relying on AI R&D to help mitigate and "reduce the impacts of extreme weather events[10]" through efforts such as the development of the National Oceanic and Atmospheric Administration Center for Artificial Intelligence (NCAI) which looks to utilize Artificial Intelligence to "accelerate advances and serve as force multipliers to solve tough problems."[11] This is why Artificial Intelligence is increasingly being utilized in disaster risk management applications. From predicting the effect of an upcoming disaster to determining what resources and aid are needed post-disaster, AI is becoming an important tool to help Americans in times of need. For this reason, we encourage further investment into AI R&D for risk mitigation, which has great promise in helping to save lives and protect communities.

Artificial Intelligence has an excellent opportunity to assist in healthcare. From using AI and Machine learning to help diagnose cancer to using AI as a testbed for the development of future drugs, the future of our healthcare sector has never been brighter. Furthermore, AI can advance precision medicine and unearth insights that may not have been previously identified for the development of innovations for specific groups or individuals. However, we believe it is important to highlight that AI for healthcare is differentiated and can involve research that is predicated on looking for differences in these groups and should not be viewed as creating adverse bias. In terms of healthcare, AI can serve as a tremendous tool, and the different applications and contexts, as well as existing regulatory frameworks for healthcare, should be considered.

How AI R&D can help address harms due to disparate treatment of different demographic groups:

a. Research that informs the intersection of AI R&D and application with privacy and civil liberties.

C_TEC strongly supports the continued research and development to address harm to different groups. This includes the current guidance being developed at the National Institute of Science and Technology (NIST) on a "Proposal for Identifying and Managing Bias in Artificial Intelligence.[12]" We look forward to continuing to engage with NIST in their work to develop "advance methods to understand and reduce harmful forms of AI Bias.[13]" Furthermore, we support NIST's views and knowledge that "not all types of bias are negative.[14]" We encourage NIST in the future to facilitate larger conversations around data

10 https://sciencecouncil.noaa.gov/Portals/0/2020%20Cloud%20Strategy.pdf?ver=2020-09-17-150020-887
11 https://www.noaa.gov/noaa-center-for-artificial-intelligence/about-ncai
12 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf
13 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf
14 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf

representation and accuracy while representing laws and regulations protecting consumer privacy and rights, as this is necessary to improve model and performance.

Finally, because of NIST's ongoing work and stakeholder engagement, we highly recommend that any other federal agencies' work align with NIST's developing guidance (SP 1270) on identifying and managing bias in AI.

b. AI R&D to help address the underrepresentation of certain demographic group groups in the AI workforce.

We strongly believe that a diverse workforce is critical. This is why we strongly support the National Artificial Intelligence Initiative, which put forth grants through the National Science Foundation for "Artificial Intelligence research, and postsecondary education program and activities, including workforce training and career and technical education program and activities, undergraduate, graduate, and postdoctoral education, and informal education opportunities.[15]" These grants are meant to support a "diverse workforce pipeline for science and technology with respect to Artificial Intelligence systems." Additionally, they are intended to "support efforts to achieve equitable access to K-12 Artificial Intelligence in diverse geographic areas and for populations historically underrepresented in science, engineering, and Artificial Intelligence fields.[16]" This is an important effort to ensure underrepresented communities are provided the necessary training and skillsets to enter into the AI workforce.

3.) Strategic directions related to international cooperation on AI R&D and on providing inclusive pathways for more Americans to participate in AI R&D.

C_TEC strongly supports establishing and strengthening regional hubs throughout the United States to advance workforce, training, representation, and overall digital equity. Regional innovation centers and involvement help develop and meet the needs of those regions. They can also help foster an environment that allows academic scientists and industry peers to regularly work together to drive sustained innovation and solution developments to those specific regions.

Conclusion:

We firmly believe that U.S. leadership in the development of Artificial Intelligence is pivotal for both the economy and security of the United States, and appreciate OSTP's work to update the National Artificial Intelligence Research and Development Strategic Plan. Thank you for considering these comments. We are happy to further discuss any of these topics.

[15] https://www.congress.gov/congressional-report/116th-congress/house-report/617
[16] https://www.congress.gov/congressional-report/116th-congress/house-report/617

Sincerely,

Michael Richards
Policy Director
Chamber Technology Engagement Center
U.S. Chamber of Commerce

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

University of Alabama - Tuscaloosa

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.

RFI Response: National Artificial Intelligence Research and Development Strategic Plan
Organization: The University of Alabama, Tuscaloosa, AL 35487

This document summarizes the opinions, suggestions, and comments of the faculty interest group of Artificial Intelligence and Machine Learning at our institution, responding to the RFI: National Artificial Intelligence Research and Development Strategic Plan. The contributing authors are listed below.

Names: Jiaqi Gong, Associate Professor of Computer Science
Jeff Carver, Professor of Computer Science
Brendan Ames, Assistant Professor of Mathematics
Erik Johnson, Assistant Professor of Economics
Eyun-Jung Ki, Professor of Advertising and Public Relations
Jiyoung Lee, Assistant Professor of Journalism and Creative Media
Mark Cheng, Professor of Electrical and Computer Engineering
Yuanyuan Chen, Assistant Professor of Information Systems
Joshua Eyer, Associate Research Professional, Alabama Life Science Institute
Jiaze He, Assistant Professor of Aerospace Engineering and Mechanics
Andrew Raffo Dewar, Professor of Interdisciplinary Arts
Kasra Momeni, Associate Professor of Mechanical Engineering

Comments and suggestions:

Prioritizing AI research on social science and humanities

Existing AI research and development have been motivated by business interests seeking to profit through the efficiency gains made possible by autonomous services. These AI-led technologies are now integrated into the public sector, law enforcement, banking, manufacturing, and medical services. However, the continued integration of AI-enabled services is dependent upon public support. The link between public support and continuous integration has been revealed in the waves of discussion and protests cascading across social media. Therefore, to continue the current pathway of deep integration is by leveraging the social sciences and humanities to ensure that all stakeholders in society are served by and aware of the benefits of these technological advances. Of particular concern is the unequal distribution of the benefits of technological progress, as seen in the urban-rural divide in the United States. To this end, anticipated research topics include 1) Foundational frameworks of social and behavioral AI; 2) Mechanism-driven AI in human physical and psychological functioning across different scales; 3) Expanding AI‘s capabilities to extract current constructs or understanding of human behavior to include processing associated with emotional states and traits; 4) Better reinforcement strategies to train AI approaches for conceptual frameworks in social science and humanities, such as enhancing detection of emotion processing. Success on all of these research topics will help to ensure that the public will support continued investments and progress in AI facilitated efficiency gains.

Developing AI techniques for creativity and expressivity.

Previous strategic themes, especially theme 2, have included human-AI collaboration as one of the topics. However, the term “collaboration” might need more research and investigation to define or expand. Significantly, the typical applications of human-AI collaboration were described in industrial scenarios, such as manufacturing, driving, and other purposeful activities. Another significant aspect of the humanities is creativity and expressivity. The question of how AI could be developed to create collaborations that gesture toward research into expressivity, creativity, interaction – characteristics.

Developing AI techniques to promote every voice being heard.

Technologies seem to be becoming bottlenecks and barriers to communications in public and societies now, especially in the era of social media. This has caused a critical divide of communication issues among stakeholders, such as police and citizens, politicians and public, and physicians and patients. These conflicts are derived from the natural limitations of technology development and becoming more and more severe in lower socioeconomic regions. Although previous efforts have been invested into the explainability and transparency of AI techniques, long-term research in developing communication strategies of AI to promote every voice being heard in the era of social media is needed. Particularly exploring specific features of AI-synthetic media can enhance individuals’ informed decision-making. Considering that many societal issues are related to lack of information resources, a question that asks whether AI can serve as an effective tool to deliver information should be addressed.

Creating AI governance frameworks to unify data cooperations

Existing data science efforts, including NIH data science strategies and NSF data sharing and management efforts, are inefficient to promote data cooperation. Barriers and bottlenecks to defining the ownership, data governance, data usage, data transparency, access openness, and benefit-sharing of the value-generated data are ubiquitous and very vague for researchers and the public to trust these data-driven technologies. Therefore, developing an AI governance framework facilitates public and private data consortium and data sharing.

Pursuing research in AI for societies, especially the lower socioeconomic regions.

Although previous strategic aims have included scalable AI approaches and systems, we felt that it is inadequate to meet the needs of societies, especially the lower socioeconomic regions. Most crises, including climate change, drug abuse, and infectious

diseases, struck these regions first and lasted longer than other areas. There are plenty of opportunities for AI techniques to help and improve the regional economy, such as optimally guiding transportation infrastructure planning to facilitate workforce development and access to opportunities for traditionally underserved populations, reducing health disparities for the underprivileged populations or for the areas that do not provide access to healthcare, and investing environmental sensing (pollution/ invasive species) for preventing climate change influences. However, the investment of AI techniques needs economic booster and policy stimulation that apparently cannot be afforded by the lower socioeconomic regions. The challenges of creating appropriate societal strategies for developing and adopting AI techniques for these regions need long-term research in these topics.

Creating AI education programs and institutions promote public engagement and prevent disparities and inequality.

The digital divide between urban-rural America has been defined as “the gap between individuals, households, businesses, and geographic areas at different socioeconomic levels with regard both to their opportunities to access AI techniques and their use of the AI for a wide variety of activities.” The existing digital divide is exacerbated by the inherent bias in the previous data-driven AI models, resulting in an emerging AI divide. The access divide (the first level AI divide) is the inequality of access to AI techniques and knowledge in homes and schools. The capability divide (the second-level AI divide) is the inequality of the capability to exploit AI arising from the first-level AI divide and other contextual factors. The outcome divide (the third-level AI divide) is the inequality of outcomes (e.g., bias, imbalance, underrepresented) of exploiting AI arising from the second-level AI divide and other contextual factors. Training and education can lower the barriers to entering the new AI technology-related industry. Community-based educational/training programs would be an effective method. Perhaps, scholars in education might have a better idea of developing such programs and institutions.

Promoting societal and community-based approaches of evaluating AI capacities and limitations and establishing standards and benchmarks.

The previous strategic plan listed the research aims for this topic, such as exploring the capacities and limitations of AI establishing standards and benchmarks. However, we argue that it would be beneficial for the lower socioeconomic regions to emphasize and prioritize the societal and community-based approaches for evaluating AI capacities and limitations, thus establishing standards and benchmarks. Mainly, AI is more or less driven by intensive data and workforce, but the data and workforce are not sufficiently represented by the underserved populations, particularly in rural America. The evaluation,

standards, and benchmarks of AI technologies need to acknowledge this underrepresentation, and approaches are required to solve this issue theoretically and practically.

Existing strategic aims lack adequate support of AI strategies for material discovery.

Responding to existing strategic aims and implementation, the priorities should be emphasized for pursuing shared datasets and environments for AI training, testing, and knowledge discovery. For instance, specific AI strategies for material discovery beyond the Materials Genome Initiative are missing. Specifically, only a handful of databases can be used for training ML models, which generally ignore the kinetics in the material discovery process. Also, there are few databases of materials beyond atomic scale, determining the critical material properties. There is also an unmet need to design benchmarks for AI models utilized.

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

University of California - Berkley

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

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

Dear Dr. Alondra Nelson, Deputy Director of Science and Society of the Office of Science and Technology Policy (OSTP) and Performing the Duties of OSTP Director,

We thank the OSTP, the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence, the NSTC Machine Learning and AI Subcommittee, the National AI Initiative Office, and the Networking and Information Technology Research and Development National Coordination Office for the opportunity to submit comments in response to the update of the National Artificial Intelligence Research and Development Strategic Plan. We are professors and researchers with expertise in AI research and development, policy, and ethics, affiliated with centers at the University of California, Berkeley, including the Berkeley AI Research Lab; the Division of Computing, Data Science, and Society; the AI Policy Hub; the Center for Long-Term Cybersecurity and its AI Security Initiative; the CITRIS Policy Lab; the CITRIS and the Banatao Institute; the Center for Human-Compatible AI; the Human Rights Center at the UC Berkeley School of Law; as well as external technology and governance non-profit research organizations including the Future of Life Institute, the Digital Life Initiative at Cornell Tech, and The Future Society.

In this document, we affirm the continued importance of the eight strategic aims described in the 2019 Update. However, we advocate for modest changes to each aim that take into account the continued learning across the AI R&D landscape. Lastly, we advocate for the inclusion of a ninth strategy—one that draws attention to the need for research on transparency and documentation of AI systems and applications. We believe this strategy is a necessary addition to support responsible and sustainable advances in this technology. Our recommendations are intended to help ensure the National AI R&D Strategic Plan enables sustained technological innovation, supports broad inclusion, economic prosperity, and national security, and upholds essential democratic values.

We have included a one-sentence summary of our main recommendation for each strategy below.

Strategy 1: Make long-term investments in AI research.

We encourage a strengthened focus on multidisciplinary research that supports AI robustness, ethics, transparency, and security integrated with long-term investments in fundamental research.

We agree that long-term investments in fundamental research are needed to continue building on

previous discoveries in AI. Specifically, we advocate that the sustained funding of R&D is an essential element that advances the trust in AI systems necessary to ensure they meet society's needs and adequately address requirements for robustness, ethics, transparency, and security. However, these research threads should not be seen as disparate, but as mutually reinforcing and essential to the development of AI.[1] AI advances are always socially driven. We should not seek advances at all costs, but in such a way that is safe, secure, responsible, and ethical. We note that research on general-purpose and scalable, multi-AI systems should be pursued cautiously and with these properties at the forefront, given the extreme potential risks from such systems.

For example, one of the subsections of research encouraged in the 2019 Update is the development of more capable and reliable robots. Indeed, it is not helpful or desirable simply to have more capable robots if they are not also reliable. In fact, the more capable the robot, the greater people may come to depend on or interact with it, implying the need for higher reliability and trustworthiness. This is particularly apparent in the context of domain-specific applications– a "reliable and safe" autonomous drone is unlikely to interact physically with humans as frequently as a self-driving car or Amazon warehouse robot.

Our point is consistent with the legal guidance of the National AI Initiative Act, in which Congress specifically states that the "United States government should use this Initiative to enable the benefits of trustworthy artificial intelligence while preventing the creation and use of artificial intelligence systems that behave in ways that cause harm."[2] The National Science Foundation (NSF) is additionally called upon to work on "research areas that will contribute to the development and deployment of trustworthy artificial intelligence systems."

Strategy 2: Develop effective methods for human-AI collaboration.

We encourage greater focus on assessing the appropriateness of varying human-machine teaming arrangements and on understanding the associated human labor implications.

We emphasize the importance of trust and alignment in enabling human-AI collaboration. As described in the 2016 Plan and mentioned in the 2019 Update: "Appropriate trust of AI systems requires explainability, especially as the AI grows in scale and complexity. … This research area reflects the intersection of Strategies 2 and 3, as explainability, fairness, and transparency are key principles for AI systems to effectively collaborate with humans. Likewise, the challenge of understanding and designing human-AI ethics and value alignment into systems remains an open research area."

[1] Sheila Jasanoff, "Ordering Knowledge, Ordering Society," in Jasanoff, ed., *States of Knowledge*, pp. 13-45. http://sheilajasanoff.stsprogram.org/wp-content/uploads/Jasanoff_Ordering-KnowledgeOrdering-Society.pdf.

[2] National Defense Authorization Act for Fiscal year 2021. HR 6395. Division E - National Artificial Intelligence Initiative Act of 2020. https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210.

We encourage additional investment in research on human-AI collaboration, including technology and policy strategies that may be pursued to support greater efficiency, effectiveness, and equity. We appreciate that the National AI R&D Strategic Plan outlines key areas of research, including where AI performs functions alongside humans, in instances where humans experience high cognitive load, and in lieu of humans where they have limited capabilities. Additional research is needed on human-machine teaming arrangements and the safeguards that must be in place to ensure that they function safely and without undue risk. This is an example of an area where closer coordination between DARPA, USD (R&E), and the National AI Initiative Office could support research advances, promote shared learning, and ensure maximum benefit from taxpayer dollars to support AI R&D, in both defense and civilian contexts. Furthermore, there should be support for efforts geared toward understanding the human labor impacts, including the toll on workers asked to interact with and rely on AI systems as well as workers involved in the development of AI systems such as data annotators[3] or UX and UI professionals.[4]

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

We encourage strengthened research and transparency in the integration of ethical, legal, and societal concerns throughout all stages of the AI lifecycle, as well as on the detection of malicious uses of AI including potential human rights abuses.

This strategy remains critical, and we underscore the importance of enabling more R&D resources that target the integration of ethical, legal, and societal concerns throughout all stages of the AI lifecycle, rather than simply after development or deployment. We also highlight the importance of research on varying interpretations of relevant, but contested terms such as "fairness" and "explainability" and their application in practice.[5,6,7] In addition to ethical, legal, and societal concerns, research related to the politics, justice, equity, and environmental implications of AI has flourished in recent years, but needs greater investment to ensure the insights from these fields can thoughtfully inform and be integrated from design through deployment and monitoring. This includes forming technology and governance oversight strategies that can be implemented throughout the AI system's lifecycle. Transparency will be critical here as value judgments will be incorporated into how technologists define and encode "ethical doctrine" (see p. 22 in AI R&D Strategic Plan). Additional research on how the human

[3] Milagros Miceli, Martin Schuessler, and Tianling Yang, "Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision," *Proc. ACM Hum.-Comput. Interact. 4*, CSCW2, Article 115 (October 2020), https://doi.org/10.1145/3415186.
[4] Richmond Wong, "Tactics of Soft Resistance in User Experience Professionals' Values Work," *Proceedings of the ACM on Human-Computer Interaction*, (October 2021): 1–28, https://doi.org/10.1145/3479499.
[5] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong, "This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology." *Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 119* (November 2019), https://doi.org/10.1145/3359221.
[6] Jessica Newman, "Explainability won't save AI," *Brookings TechStream*. (May 19, 2021). www.brookings.edu/techstream/explainability-wont-save-ai/.
[7] Nicole Chi, Emma Lurie, and Deirdre K. Mulligan, (July 2021). "Reconfiguring Diversity and Inclusion for AI Ethics," AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. https://dl.acm.org/doi/10.1145/3461702.3462622.

rights legal framework and norms can be used to guide ethical AI development and deployment is also needed.[8]

Lastly, an additional research challenge in this area that urgently requires greater investment is the detection of malicious uses of AI including the use of synthetic content for manipulation, harassment, financial, and political gain.[9]

Strategy 4: Ensure the safety and security of AI systems.

We encourage strengthened research on how to manage and prevent safety and security challenges from increasing as AI systems become more advanced and multiply their capabilities, including the role of greater transparency and public awareness.

Technical solutions to prominent AI safety and security problems remain elusive and are a critical issue that requires federal R&D investments along with collaborative efforts among government, industry, academia, and civil society. It is imperative that the National Institute of Standards and Technology (NIST) adheres to the legal guidance in the National AI Initiative Act to support research on "safety and robustness of artificial intelligence systems, including assurance, verification, validation, security, control, and the ability for artificial intelligence systems to withstand unexpected inputs and adversarial attacks."[10] As stated in the 2019 Update, state-of-the-art AI systems today can still "be made to do the wrong thing, learn the wrong thing, or reveal the wrong thing, for example, through adversarial examples, data poisoning, and model inversion, respectively." This is particularly pressing for the application of AI technologies in critical infrastructure, defense, and safety-critical systems.

Moreover, we agree that as AI systems continue to grow in capabilities, they will likely grow in complexity, making it ever harder for correct and desirable performance to be verified and validated.[11] AI safety and value alignment remain critical research challenges, especially for

---

[8] David Kaye, special rapporteur on the promotion and protection of the right to freedom of opinion and expression, "Report on artificial intelligence technologies and implications for freedom of expression and the information environment." *United Nations Office of the High Commissioner for Human Rights. https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx*.

[9] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting world leaders against deep fakes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. (2019) https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf.

[10] National Defense Authorization Act for Fiscal year 2021. HR 6395. Division E - National Artificial Intelligence Initiative Act of 2020. https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210.

[11] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Kohd PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, and Liang P, "On the Opportunities and Risks of Foundation Models," *arXiv*, (2021), https://arxiv.org/abs/2108.07258.

multi-purpose or general-purpose AI systems,[12,13] as stated in both the 2016 Plan and 2019 Update. We expect these challenges will increase in the near future, as AI systems become more advanced and multiply their capabilities, with both greater beneficial opportunities and risks in case of misuse or failures of safety or security controls.

We believe that greater transparency and public awareness are needed to support AI safety and security. End-users should have an understanding of the safety and security of AI systems and supporting accountability mechanisms, including clear steps for redress. Research is necessary on how to do this effectively. We also advocate for studying the kinds of vulnerabilities and failures that are likely to arise from real-world threat scenarios, and from software vulnerabilities in the AI supply chain.

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

We encourage research on how to reduce energy and carbon footprints for AI development and operation, and the role of public training and testing environments in that reduction.

The trend toward larger and more complex AI models, requiring larger training datasets and significant computing resources, has increased in recent years. This trend typically benefits already powerful companies and institutions, and comes with a significant and often-unsustainable environmental cost.[14,15] More research is needed to better understand how to reduce energy and carbon footprints for AI development and operation, and the role of public training and testing environments in that reduction.[16]

Shared public datasets and secure environments for AI training and testing are an important way to ensure that progress in AI meets the needs of a diverse spectrum of AI interests and applications and can support the public good. Public datasets and environments for AI training and testing can also offer secure software sandboxes, regulatory sandboxes, and testing servers. By creating shared datasets and secure environments for cross-institutional testing, a greater diversity of innovators, entrepreneurs, SMEs in various sectors, and researchers from varying epistemological approaches may be supported.

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

---

[12] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt, "Unsolved Problems in ML Safety," *arXiv*, https://arxiv.org/abs/2109.13916.
[13] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control.* (New York: Viking, 2020).
[14] Emma Strubell, Ananya Ganesh, Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP," In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. (July 2019). arxiv.org/abs/1906.02243.
[15] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* (March 2021). dl.acm.org/doi/abs/10.1145/3442188.3445922.
[16] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Hung Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeffrey Dean. (2022), "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," *TechRxiv*. Preprint. https://doi.org/10.36227/techrxiv.19139645.v2.

We encourage research that investigates how standards, benchmarks, and testing requirements for a broad set of quality controls will inform evolving AI development and deployment, and how to encourage adoption.

Ongoing efforts to measure and evaluate AI technologies through standards and benchmarks, for example by the National Institute of Standards and Technology (NIST), the International Organization for Standardization (ISO), and the Institute of Electrical and Electronics Engineers (IEEE), are extremely valuable. However, as noted in the 2019 Update, we agree that benchmarks, metrics, and testing requirements for a broad set of quality controls are still lacking and require greater research investment. Specifically, we argue that while benchmark datasets are important, they should be alive – in the sense that they need to be enhanced by new data and connected to domain problems with human committees and evaluators, not just metric numbers, which should serve as supporting information.

We also caution that establishing standards and benchmarks can lead to lock-in and path dependencies for AI system development and deployment that will be difficult to circumvent. If these processes are too rigid and resource intensive, they may lead to workarounds, lack of compliance, and other harmful spillover effects. It is therefore of great importance that standards and benchmarks are robust yet flexible enough to adapt to changing norms and needs, including how to draw upon a human rights legal framework, which puts humans' civil, political, economic, and social wellbeing—as opposed to institutional benefits—at the center of development.[17] We encourage funding allocation to support research in this space, especially the utility of the NIST AI Risk Management Framework (NIST AI RMF) to better ensure its long-term success. While NIST's AI RMF is voluntary, it would benefit from research on how AI governance testbeds may be used to evaluate its effectiveness at shaping AI development and deployment.

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

We emphasize the need to not only broaden participation in computing and engineering fields, but also to provide educational opportunities to train computer scientists and engineers to be fluent in social and ethical impact, and in professional responsibility.

As noted in the 2019 Update, we agree that multidisciplinary teams are essential to a thriving AI R&D workforce, and that "it is imperative to broaden the participation among groups traditionally underrepresented in computing and related fields." We emphasize that many different definitions of "underrepresented" may be in scope here–the inclusion of foreign

[17] Brandie Nonnecke and Phil Dawson, "Human rights implications of algorithmic impact assessments: Priority considerations to guide effective development and use". Harvard Carr Center Discussion Paper Series. (Oct. 21, 2021), https://carrcenter.hks.harvard.edu/publications/human-rights-implications-algorithmic-impact-assessments-priority-considerations.

researchers, ethnic minorities, women, representatives of the LGBTQ+ community, the differently abled, and other groups historically marginalized within the disciplinary culture of computer science and engineering. The integration of feedback from diverse stakeholder groups at multiple points of AI development is a recognized path to system reliability and safety.[18] In addition to providing education in computational thinking at all levels across disciplines, we emphasize the need for educational materials and opportunities to help train computer and information scientists and engineers to be fluent in social and ethical impact, and in professional responsibility.[19,20]

Strategy 8: Expand public-private partnerships to accelerate advances in AI.

We encourage increased focus on international cooperation and coordination on AI research as well as support for research partnerships that include civil society and impacted communities.

International cooperation and coordination on AI is increasingly critical and we advocate maintaining and expanding this emphasis. For example, further research is needed to advance opportunities for collaboration with allies to improve information sharing, reduce potential "race-to-the-bottom" dynamics, and design Track I, 1.5, and II diplomacy mechanisms.

In addition to partnerships with academia and industry that generate technological breakthroughs in AI, we also recommend the inclusion of partnerships with civil society and impacted communities to ensure applications of AI achieve their aims and do not cause unexpected or disproportionate harm.

[New] Strategy 9: Support transparency and documentation of AI systems and applications.

We encourage support for research that identifies effective mechanisms for transparency and documentation of AI systems and applications.

We argue an additional strategic aim is warranted and therefore propose a ninth strategy to support transparency and documentation of AI systems and applications. The need for ongoing research on the transparency and effective explainability of AI systems is already discussed in both Strategy 3 and Strategy 4. However, research into how to document and share the characteristics of AI systems is a current gap in the R&D Plan. While there has been critical

[18]Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz, "Hard choices in artificial intelligence," *Artificial Intelligence* 300 (2021): 103555. https://doi.org/10.1016/j.artint.2021.103555.

[19] Barbara J. Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo, "Embedded EthiCS: integrating ethics across CS education," *Communications of the ACM, 62*, no. 8, (Oct. 29, 2019): 54-61. https://doi.org/10.1145/3330794.

[20] Amy J. Ko, Alannah Oleson, Neil Ryan, Yim Register, Benjamin Xie, Mina Tari, Matthew Davidson, Stefania Druga, and Dastyni Loksa, "It is time for more critical CS education," *Communications of the ACM, 63*, vol. 11 (2020): 31-33. https://doi.org/10.1145/3424000.

research in this space in recent years,[21,22,23,24] there is ongoing need for research on how best to carry out and facilitate standardized descriptions of features of AI systems. Some of the types of descriptions that may be relevant are characteristics about the AI system, its performance metrics, and its outcomes including expected behaviors, limitations, evaluation across varying conditions and populations, information about which datasets and training environments have been used and why, as well as human-interpretable logging of a system's activity, metadata, and impacts. Further research is also needed to explore processes that support these activities, which include verification of the characteristics over time, internal reviews, and reporting mechanisms. Improving classification and documentation of AI systems and applications should be a research priority because the current lack of standardization contributes to the dearth of trust in AI development, preventing increased discovery and adoption.[25,26,27] Moreover, this is an area that would benefit from federal investment because industry is unlikely to address this on its own and because it may facilitate greater coordination and communication between organizations, disciplines, and sectors.

We understand that the National AI R&D Strategic Plan is, by design, solely concerned with addressing the research and development priorities associated with advancing AI technologies, and does not describe or recommend policy or regulatory actions related to the governance or deployment of AI. The call for increased focus on transparency and documentation of AI systems is oriented toward supporting research and development. Without institutionalized mechanisms for sharing the types of tools being built and used for different purposes, it is more challenging to share knowledge and learn from the experiences of others.

The 2019 Executive Order on Maintaining American Leadership in Artificial Intelligence called on federal agencies to improve their data and model inventory documentation to enable discovery and usability, and the 2019 Update emphasized in Strategy 5 that, “development and adoption of best practices and standards in documenting dataset and model provenance will enhance trustworthiness and responsible use of AI technologies.” However, Strategy 5 is primarily focused on improving access to datasets and training environments rather than documenting the characteristics and uses of AI systems. Adding a new strategy to support transparency and documentation of AI systems and applications will not only accelerate research

[21] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, "Model Cards for Model Reporting," *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), https://dl.acm.org/doi/10.1145/3287560.3287596.

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. “Datasheets for Datasets,” arxiv.org/abs/1803.09010.

[23] Bin Yu and Karl Kumbier, “Veridical data science,” *PNAS*. 117, no. 8 (Feb. 25, 2020): 3920-3929. https://doi.org/10.1073/pnas.1901326117.

[24] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu, “Definitions, methods, and applications in interpretable machine learning,” *PNAS*, 116, no. 44, (Oct. 29, 2019): 22071-22080. https://doi.org/10.1073/pnas.1900654116.

[25] "OECD Framework for Classification of AI Systems: a tool for effective AI policies," OECD Digital Economy Papers. (Feb. 2022.) https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.

[26] Catherine Aiken, "Classifying AI Systems," CSET Data Brief. (Nov. 2021.) https://cset.georgetown.edu/publication/classifying-ai-systems/.

[27] Thomas Krendl Gilbert, Sarah Dean, Tom Zick, and Nathan Lambert. (Feb. 2022), "Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems," *Center for Long-Term Cybersecurity White Paper Series*. https://cltc.berkeley.edu/reward-reports/.

in this critical area, but also advance the aims of the other eight strategies by contributing to knowledge of the AI landscape.

Contact

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

Our best,

Anthony M. Barrett, Ph.D., PMP, Visiting Scholar, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Ann Cleaveland, Executive Director, Center for Long-Term Cybersecurity, UC Berkeley

Camille Crittenden, Ph.D., Executive Director, CITRIS and the Banatao Institute, UC Berkeley; Co-Founder, CITRIS Policy Lab and EDGE in Tech Initiative at UC

Samuel Curtis, AI Policy Researcher & Project Manager, The Future Society

Jordan Famularo, Ph.D., Postdoctoral Scholar, Center for Long-Term Cybersecurity, UC Berkeley

Hany Farid, Ph.D., Professor, Electrical Engineering and Computer Sciences and the School of Information, UC Berkeley

Thomas Krendl Gilbert, Ph.D., Research Affiliate, Center for Human-Compatible AI, UC Berkeley; Digital Life Initiative, Cornell Tech

Ken Goldberg, Ph.D., Professor, Industrial Engineering and Operations Research William S. Floyd Jr. Distinguished Chair in Engineering, UC Berkeley

Carlos Ignacio Gutierrez, AI Policy Researcher, Future of Life Institute

Dan Hendrycks, Ph.D. Candidate, Berkeley AI Research Lab, UC Berkeley

Niki Iliadis, Senior AI Policy Researcher, The Future Society

Alexa Koenig, J.D., Ph.D., Executive Director, Human Rights Center, UC Berkeley School of Law; Co-Founder, Human Rights Investigations Lab

Yolanda Lannquist, Head of Research & Advisory, The Future Society

Richard Mallah, Director of AI Projects, Future of Life Institute

Nicolas Miailhe, Founder & President, The Future Society

Nicolas Moës, Head of Operations & AI Policy Researcher, The Future Society

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

Brandie Nonnecke, Ph.D., Director, CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley; Co-Director, AI Policy Hub

Ifejesu Ogunleye, Researcher, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Andrew W. Reddie, Ph.D., Assistant Professor of Practice, School of Information, UC Berkeley

Stuart Russell, Ph.D., Professor Computer Science and Smith-Zadeh Professor in Engineering, UC Berkeley

Charis Thompson, Ph.D., Chancellor's Professor and Associate Dean, Computing, Data Science, and Society, UC Berkeley

Richmond Y. Wong, Ph.D., Postdoctoral Scholar, Center for Long-Term Cybersecurity, UC Berkeley

Bin Yu, Ph.D., Chancellor's Distinguished Professor, Departments of Statistics and Electrical Engineering and Computer Sciences, Class of 1936 Second Chair, L&S, UC Berkeley

Rebecca Wexler, J.D., Assistant Professor, School of Law, UC Berkeley

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

University of California, Irvine School of Medicine and UC Irvine Health

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.

UNIVERSITY OF CALIFORNIA, IRVINE

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

UC Irvine Health
101 The City Dr S,
Orange, CA 92868

RFI Response: National Artificial Intelligence Research and Development Strategic Plan

To whom it may concern, March 4th 2022

Please see below for our response on behalf of the University of California, Irvine School of Medicine and UC Irvine Health. Thank you.

Very respectfully yours,

William H. Yong MD, FCAP
Vice-Chair, Integration and Strategic Initiatives
Chief of Neuropathology and Professor
Department of Pathology and Laboratory Medicine
University of California, Irvine School of Medicine

Edwin Monuki MD, PhD
Chair and Warren L. Bostick Professor
Department of Pathology and Laboratory Medicine
University of California, Irvine School of Medicine

Leslie M. Thompson PhD
Co-Director, Institute for Precision Health
Donald Bren and Chancellor's Professor,
Department of Psychiatry & Human Behavior
University of California, Irvine School of Medicine
Department of Neurobiology and Behavior,
University of California, Irvine School of Biological Sciences

Michael J. Stamos, MD, FACS, FASCRS
Dean and Professor of Surgery
University of California, Irvine School of Medicine

Tom Andriola
Vice Chancellor, Information, Technology and Data
Chief Data Officer

Co-Director, Institute for Precision Health
University of California, Irvine and UCI Health

Steve A. N. Goldstein MD, PhD
Vice Chancellor, Health Affairs and Distinguished Professor
University of California, Irvine

RFI Response: National Artificial Intelligence Research and Development Strategic Plan

Healthcare expenditures are a large and growing part of the U.S. economy reaching $4.1 trillion and 19.7% of GDP[1]. The basis for the increased expenditures includes health care service price and intensity, inefficiency of care delivery as well as an aging population[2]. Waste may be a quarter or more of healthcare expenditures but Artificial Intelligence (AI) promises to improve accuracy and speed of diagnosis as well as to guide therapy potentially reducing waste and lowering costs in one estimate by 150 billion dollars per year[3]. We welcome the National Artificial Intelligence Research and Development Strategic Plan and offer the below suggestions for matters relevant to biomedical and healthcare AI.

a. Develop a standardization roadmap[4]. Standardization is essential for safe and reliable AI performance[4]. Standardization should be done cautiously and may evolve over time as technologies advance. Standardization may facilitate collaborative work by standardizing vocabulary, developing ontology and defining terms for the biomedical AI community including for genetic, radiology, pathology, and other clinical uses. Developing and encouraging the use of standard file formats for data elements such as images may facilitate research and implementation. Organizations with relevant expertise such as the American Medical Informatics Association and Association for Pathology Informatics among others should be engaged. A national committee for AI standardization should work closely with international partners such as the International Organization for Standardization (ISO)[5] to enhance interoperability within and across borders.

b. Develop a STEM- and AI-competent work force. Training of STEM students should begin in elementary school and continue through high school, college and graduate school. Funding university and involving private sector information technology (IT) companies in this developmental process is critical to developing an IT/AI competent-work force. Programs for females, underrepresented minorities, and first-in-family college attendees should be a part of the pipeline from childhood on. Training programs for the existing healthcare workforce should be incentivized to facilitate AI technology adoption[6].

c. Fund artificial intelligence research institutes and clinical trial networks at universities and health systems. Incentives and perhaps even requirements for private industry or other public health entities to work with these academic research institutes could be created. Veterans Administration hospitals, children's hospitals and public health departments should be included in the networks. Most clinical trials are underrepresented in terms of low income and minority patients[7]. Academic Medical Centers or healthcare networks with substantial underrepresented populations (e.g. Safety Net) should be prioritized in AI research and care networks.

d. Fund or subsidize low-cost scalable, redundant storage and high bandwidth data networks nationally and make it available at middle and high schools, universities and medical

schools. Robust, environmentally-friendly power sources and networks should be developed in concert with the data networks.

e. Support digitization of stained patients' histopathology slides. Millions of diseased patient tissue samples are stored as stained or immunostained glass slides that are the current standard for diagnosis and guiding therapy. These histologic slides require digitization into whole slide images for image analysis using AI. These digital images can then be analyzed together with other patient disease parameters such as sequencing data, radiologic images etc. Developing and supporting whole slide digital imaging infrastructure will facilitate research and clinical adoption of AI for diagnosis and therapy decisions[8].

f. Develop improved models for translating AI tools and technologies into the care setting to benefit patients, including those in Safety Net. Of particular importance is adequate staffing and funding of the Food and Drug Administration (FDA) for the evaluation and approval of AI-based medical devices and technologies commensurate to their rate of growth. Mechanisms to follow, re-evaluate, or re-validate AI-infused tools periodically for safety and benefit are essential. Experts including in ethics from research universities and academic medical centers should be integrated into this process. In addition, existing laboratory inspection and accreditation mechanisms such as those provided by the College of American Pathologists may be adapted to ensure robust workflows in the clinical care setting.

References

1. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical

2. Dieleman JL, Squires E, Bui AL, et al. Factors Associated With Increases in US Health Care Spending, 1996-2013. *JAMA.* 2017;318(17):1668–1678. doi:10.1001/jama.2017.15927

3. https://healthitanalytics.com/news/future-ai-opportunities-for-improving-care-delivery-cost-and-efficacy

4. https://www.dke.de/en/areas-of-work/core-safety/standardization-roadmap-ai

5. https://www.iso.org/committee/6794475.html

6. https://www.healthcareitnews.com/ai-powered-healthcare/workforce-training-key-ai-success-healthcare

7. Jiang S, Hong YA. Clinical trial participation in America: The roles of eHealth engagement and patient-provider communication. Digit Health. 2021 Dec 14;7:20552076211067658. doi: 10.1177/20552076211067658.

8. https://jcp.bmj.com/content/jclinpath/74/7/409.full.pdf

Federal Register Notice 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and, February 2, 2022.

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

World Privacy Forum

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.

Comments of the World Privacy Forum to the Office of Science and Technology regarding RFI Response: National Artificial Intelligence Research and Development Strategic Plan

AI R&D RFI Response Team
Attn :
NCO
2415 Eisenhower Avenue
Alexandria, VA 22314

Thank you for the opportunity to comment on the RFI for the update of the *National Artificial Intelligence R & D Strategic Plan*, 87 FR 5876, https://www.federalregister.gov/documents/2022/02/02/2022-02161/request-for-information-to-the-update-of-the-national-artificial-intelligence-research-and .

The World Privacy Forum is a nonprofit, non-partisan 501(c)(3) public interest research group. We have published many research papers and policy comments focused on privacy and data governance. Much of our work explores data governance of complex ecosystems, particularly in the areas of health, identity, AI and machine learning ecosystems. We have conducted significant field research in biometrics for a peer-reviewed study on the Aadhaar biometric ecosystem in India. (Pam Dixon, *A Failure to Do No Harm: India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the U.S., S*pringer Nature, Health Technology. DOI 10.1007/s12553-017-0202-6. http://rdcu.be/tsWv. Open Access via Harvard- Based Technology Science: https://techscience.org/a/2017082901/.) For more about our work and our reports, data visualizations, Congressional testimony, and consumer guides, see http://www.worldprivacyforum.org.

Our comments focus on Strategy 1 and Strategy 3. First, we support long-term investments in AI research. Second, it has become clear that a portion of these investments need to be joined with Strategy 3, *Understand and address the ethical, legal, and societal implications of AI*. Our discussion in these comments is centered in the area of identity ecosystems that utilize AI. Much more work needs to be done to research what kinds of policies, laws, and regulations are the best fit for these AI systems, and what benchmarking and testing reveals about the effectiveness of the regulatory models. In these comments, we discuss and compare regulatory models for

biometric forms of identification, and we propose a model worth much more focused study and work. We also briefly discuss the biometrics approaches in the current version of the *National Artificial Intelligence R & D Strategic Plan*.

I. Discussion of biometric benchmarking programs mentioned in the 2019 *Update of the National Artificial Intelligence R & D Strategic Plan*.

Regarding the 2019 *National Artificial Intelligence R & D Strategic Plan*, page 35 of the plan mentions NIST's Facial Recognition Vendor Testing, FRVT, one of the world's leading benchmarking programs in biometrics, particularly its benchmarking of face recognition and other biometric modalities utilizing machine vision. This is an important program, and we support its continuing inclusion and discussion in the report as a key exemplar of an outstanding R & D program.

One particular item we want to highlight in the NIST program is that the FRVT has moved to a "living document" format, in that the benchmarking studies have moved to *ongoing testing*, rather than *periodic testing*. We see this approach as an important implementation to modern benchmarking. It is worth including "living documents" as a best practice in other programs as appropriate.

II. The need for formal research into biometric regulatory approaches, and a suggested model that deserves further research and development

Current approaches to biometric regulation thus far have been insufficiently constructed, and do not address the full ecosystem of biometric modalities[1] and the risks that can arise from them. With the increased utilization of biometric systems, the risk level is high enough that commonly-used regulatory controls such as simple, checkbox-types of consent mechanisms or indirect or passive consent are no longer practicable for addressing the range and level of risks that biometric technologies present.[2] Most of the existing approaches to biometric policy in the US rely on models that do not have good potential for solving the complex policy issues biometrics ecosystems raise.

This being said, there are existing regulatory models – namely chemical safety regulation models -- that provide an important set of administrative and procedural controls that could be utilized for biometric regulation, as they allow breadth of regulation while at the same time facilitating granular approaches fit for each particular biometric modality. Robust procedural and administrative controls used in the chemical safety models form a much richer safety net than checkbox approaches. The model that existing safety regulations provides is a possible pathway forward toward approaching biometric regulation in a more thoughtful, comprehensive, and globally harmonized and administrable approach.

[1] Biometric modalities include: DNA, face, fingerprint, speech, voice, iris, retina, periocular, ear, gait, tattoo, heartbeat, hand geometry, odor, behavioral, typing recognition, vein recognition, for example. See Biometrics Institute: https://www.biometricsinstitute.org/what-is-biometrics/types-of-biometrics/.

[2] For a thorough discussion of biometric risks, see: *The dark side of identity systems*, *Trilogy*. ID4Africa, Fall/Winter 2021. Episodes 23, 24, 25. https://id4africa.com/livecasts/.

In recent years, biometrics have been advanced by deep learning architectures and techniques, and the advances far outpace policy adaptations.[3] The full biometric ecosystem -- from data collection to algorithmic quality, from the hardware of biometric systems, to the implementation of the full systems – has components that create, or can create, meaningful risk. These risks have been rigorously documented, and are no longer theoretical. makes a difference in accuracy.[4] The primary controversies around face recognition systems[5] arise largely from the well-documented potential for racial, gender, and age[6] bias, as well as politically-driven utilizations of such systems.[7] Additionally, some face recognition systems rely on unconsented data collections, which is also controversial.[8]

As tempting as it is to discuss biometric systems and risks in isolation, for example, by focusing on a single modality such as face recognition or DNA or even on a particular use case, it is essential to contextualize biometric systems in the broader context of all biometric modalities, and view biometrics as a complete ecosystem of multiple biometrics. Additionally, biometric modalities are often layered in what is called a multi-modal or multi-biometric approach, combining face and iris, or iris and fingerprint, and so on. [9] In law enforcement contexts, multi-biometrics are extremely common.

As discussed, current legal frameworks do not yet effectively address the full range of biometric modalities and the risks associated with them, resulting in incomplete policy protections and fragmentation. Fragmented approaches toward the regulation of biometrics is a category risk in and of itself, and ultimately, fragmented approaches are not a sustainable solution for the types of meaningful risks biometrics can pose within jurisdictions, and across jurisdictional boundaries. The US can do much more to conduct research into what kinds of legal and

---

[3] Certain modalities of biometrics, particularly face recognition systems, have been profoundly changed by advances in certain AI architectures, specifically, Convolutional Neural Networks (CNNs). See, Mayank Vatsa, Richa Singh, Angshul Majumdar, Ed. Deep Learning for Biometrics, CRC Press, 2018.

[4] J. Libert, J. Grantham, B. Bandini, K. Ko, S. Orandi, C. Watson, *NIST Report (NISTIR) 8307: Interoperability Assessment 2019: Contactless-to-Contact Fingerprint Capture*, NIST, 2019. https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8307.pdf

[5] European Parliament Resolution A9-0232/2021: *Artificial Intelligence in Criminal Law and its Use by Police and Judicial Authorities in Criminal Matters*, (2020/2016/INI), adopted 6 October 2021, Strasbourg, France. https://www.europarl.europa.eu/doceo/document/TA-9-2021-0405_EN.html, sections 25-31 in particular.

[6] Age bias in face recognition occurs in both younger and older individuals. See Anil Jain, *Biometric Recognition of Children, Challenges and Opportunities*. Michigan State University (June 7, 2016). http://biometrics.cse.msu.edu/Presentations/AnilJain_UIDAI_June7_2016.pdf. See also Patrick Grother, Mei Ngan, Kayee Hanaoka. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects in Facial Systems*, NIST (Dec. 2019), https://nvlpubs.nist.gov/nistpubs/ir/2019/ NIST.IR.8280.pdf.

[7] Patrick Grother, Mei Ngan, Kayee Hanaoka. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects in Facial Systems*, NIST (Dec. 2019), https://nvlpubs.nist.gov/nistpubs/ir/2019/ NIST.IR.8280.pdf. Regarding political risks, see: Teki Falconer and S. Okedara, *National Security Exemptions*, The Dark Side of Identity Systems, Episode 24, Segment 3. https://id4africa.com/livecast-ep24-the-dark-side-of-identity-part-2/

[8] *Facial Recognition : the CNIL orders Clearview AI to stop reusing photographs available on the Internet*, CNIL France, 16 December 2021. https://www.cnil.fr/en/facial-recognition-cnil-orders-clearview-ai-stop-reusing-photographs-available-internet.

[9] Multimodal biometrics and biometric fusion are instances when one or more biometric attributes are used together. The term *multibiometric* may also be used. See: Maneet Singh, Richa Singh, Arun Ross. *A comprehensive overview of biometric fusion*, Information Fusion 52 (2019), 187-205. https://www.cse.msu.edu/~rossarun/pubs/SinghRossBiometricFusion_INFFUS2019.pdf

regulatory frameworks are effective. Without this work, policymakers will not have the full data they need to make the best decisions about AI regulations in this particular ecosystem.

Consider general data protection regulatory frameworks and their interactions with biometric ecosystems. Currently, more than 145 jurisdictions have passed comprehensive national-level data governance and protection frameworks.[10] Because these frameworks often use generalized language regarding protections for biometric data, the existing rules cover biometrics, but often lack specifics tied to any particular modality of biometric. In the European Union (EU), the EU General Data Protection Regulation (GDPR),[11] covers biometrics as a sensitive data category. The GDPR does not address specific face recognition, iris, or multi-modal concerns, and does not address face recognition uses for some important use cases, for example, national security uses of biometrics are not covered under the GDPR. Beyond the GDPR, in multiple jurisdictions, regulatory solutions for biometric risks often have a strong emphasis on the modality of face recognition, iris, fingerprint, or DNA, and do not reflect a more mature legislative model.[12] The current approaches to biometric regulation are well-meaning but incomplete, and have created meaningful gaps in protections. On one hand, comprehensive legislation is too broad to be specific enough. On the other, a focus on single-modality biometric legislation is too narrow, and leaves gaps regarding the other modalities, and gaps regarding multi-modal systems.

For example, subnational legislation in the United States tends to focus on single biometric modalities, such as face recognition, and consent is often the primary tool utilized for protection.[13] Consent, when utilized by itself without any other administrative or procedural controls or underlying protections, provides quite poor protections in the biometric context.[14] The failure to conceptualize biometric guardrails in a more sophisticated ecosystem approach will lock in fragmented approaches, which over time, promises to create regulatory havoc and gaps in protections. It is an unsustainable policy strategy for national or sub-national legislatures to craft multiple, stand-alone, and possibly rivalrous biometric policies for separate biometric modalities and selected biometric use cases.

The chemical safety model, because it is both broad and granular, reduces policy fragmentation

---

[10] Greenleaf, Graham, *Global Data Privacy Laws 2021: Uncertain Paths for International Standards* (February 11, 2021). (2021) 169 Privacy Laws & Business International Report 23-27, Available at SSRN: https://ssrn.com/abstract=3836408

[11] EU General Data Protection Regulation, http://www.privacy-regulation.eu/en/index.htm. The GDPR went into effect May 25, 2018. See in particular Article 9.4.

[12] Biometric Information Protection Act (760 ILCS 14) (Illinois) https://ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004

[13] Biometric Information Protection Act (760 ILCS 14) (Illinois) https://ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004, Capture or use of Biometric Identifiers, Sec. 503.001. Title 11. (Texas) https://statutes.capitol.texas.gov/Docs/BC/htm/BC.503.htm.

[14] Consent has important uses, particularly in human subject research, where consent must be meaningful. International norms have developed around this context for consent. See: 21 CFR 50.20 *General Requirements for Informed Consent*. See also: 45 CFR part 46 and HHS, *Federal Policy for the Protections of Human Subjects ('Common Rule')* https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html. Consent as a primary tool for all data protection and privacy contexts, however, has been replaced by policies around legitimate basis for processing, among other protections. See *Click to Consent? Not good enough anymore.* Privacy Commissioner of New Zealand, 2019. https://privacy.org.nz/blog/click-to-consent-not-good-enough-anymore/.

with meaningful, measurable, and well-understood and documented risk evaluations and mitigations. The global body of chemical safety regulations exists in most countries to monitor chemicals that pose dangers to people and to protect people from a complex range of chemicals.[15] Chemical safety policies that are crafted at the national and subnational levels are built according to a common framework, use the same definitions, are fit for each jurisdiction, and are also harmonized globally while respecting jurisdictional contexts. Biometrics, as an equally complex data ecosystem with overarching risks and particular risks attached to specific modalities, could be similarly regulated under an umbrella of controls derived from safety and risk-oriented regulatory models, with each biometric modality receiving appropriate and granular regulatory attention befitting the modality being considered. Additional risks that arise from mandatory biometric enrolment in some systems is an additional type of risk to consider and mitigate, among other types of risks.

The most salient hallmark of chemical safety regulation models is the robust procedural and administrative controls they use, joined with highly granular applicability. For example, the regulations present an umbrella under which many types of chemicals can be individually regulated. Lead and arsenic, for example, are regulated under the same overarching framework, but lead and arsenic have differing protections and cutoff points, as is appropriate given the toxicological differences. In chemical safety regulations, the country-level legal frameworks are then harmonized by two multilateral institutions, the World Health Organization and the United Nations. The UN has a program called the *Globally Harmonized System of Classification and Labelling of Chemicals* (GHS), which is regularly updated.[16] The idea of the UN GHS is to bring a global, standardized approach to chemical safety across all jurisdictions.[17] Labeling is to be the same, level or grade of risk is be the same, and risk mitigation strategies would be similarly harmonized internationally. The UN GHS plan is part of the implementation of the Sustainable Development Goals (SDGs).

The key controls in chemical safety models include:

- Pre-market safety, quality, and other risk assessments and requirements
- Registration of products
- Ongoing product documentation
- Audits
- Post-implementation surveillance (observation) and documentation
- Compliance labeling
- Safety certifications
- Technological proof of compliance and risk mitigation
- Ongoing review, oversight, and multistakeholder feedback (and complaint mechanisms).

[15] United Nations, GHS, https://www.unece.org/trans/danger/publi/ghs/ghs_welcome_e.html. The World Health Organization operates the International Programme on Chemical Safety (IPCS.) https://www.who.int/health-topics/chemical-safety#tab=tab_1.
[16] GHS, Rev. 8, 2019. United Nations. https://unece.org/ghs-rev8-2019
[17] United Nations, GHS, https://www.unece.org/trans/danger/publi/ghs/ghs_welcome_e.html..

These administrative and procedural controls provide a toolbox of options that go beyond best practices, simple consent structures, and narrow bans, and could readily be utilized in biometrics. Under this type of framework, all biometric modalities would be regulated under one umbrella. Each modality would be subject to meaningful administrative and procedural controls. All biometrics -- DNA and face recognition, along with the other biometric modalities -- would be evaluated under the auspices of the regulation, with their own measures for accuracy and pre-market fitness.

In practice, this would mean that before a biometric product could be put out in the market, it would have to be assessed for pre-market safety, quality, and other risks. For face recognition systems, this would mean that the product could not be discriminatory in its operations, and the risk points would have rigorous testing. For example, age, gender, and race biases would be tested. Each biometric product, after passing the assessment, would be registered, labelled, and would be required to submit documentation to regulators. Regulators would be able to conduct audits, and there would be a post-implementation market surveillance and documentation program. Safety certifications would need to be met, and biometric products would need to proactively provide proof of compliance and mitigate known risks. And finally, there would be ongoing review of the biometric products, consumer and end-user feedback, as well as formal complaint mechanisms.

In considering how this system could be worked out in practice, it is useful to review some exemplar implementations regarding chemical safety regulatory models. The regulations are notable, in that they are harmonized across borders, yet the regulations are also fit for each country-level context of economic and technological development. This is important, and would need to be present for any harmonized biometric approaches.

The EU has two significant member state-wide regulatory models in the area of chemical safety. Both regulations offer excellent tools for mitigating harms. REACH[18] is the European Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals. It entered into force in 2007, replacing the former legislative framework for chemicals in the EU. This important and precedent-setting regulation applies to essentially every chemical product manufactured, imported, or sold within the EU. Manufacturers and importers must register all substances produced above a set yearly volume, and also must identify risks associated with the substances they produce, demonstrate compliance in mitigating the risks, and establish safe use guidelines for the product so that the use of the substance does not pose a health threat.

Another precedent-setting regulation, RoHS,[19] applies to any business that sells electrical or electronic products, equipment, sub-assemblies, cables, components, or spare parts directly to RoHS-directed countries. Products must be cleared for market prior to launch All parties in supply chain must provide documentation/recordkeeping, regularly update information, Mandatory compliance labeling. All of these features could be helpful in regulating biometric products. Other countries that have enacted RoHS include Japan, Korea, and China. In the U.S.,

[18] European Commission, REACH, https://ec.europa.eu/growth/sectors/chemicals/reach_en.
[19] European Commission, RoHs Directive, Current: (2011/ 65/ EU). First RoHS Directive: (2002/95/EC), https://eur-lex.europa.eu/legal-content/en/TXT/?uri=celex%3A32011L0065

the states of California, Colorado, Illinois, Indiana, Minnesota, New Mexico, New York, Rhode Island, and Wisconsin, among others, have enacted RoHS-like and e-waste regulations.

In the US, a federal statute, the Chemical Safety for the 21st Century Act,[20] regulates chemical substances of concern. The statute has these compliance requirements: pre-manufacture notification for new chemical substances prior to manufacture, where risks are found (after risk assessment), testing by manufacturers, importers, and processors, Certification compliance, Reporting and record keeping. If a substance presents a substantial risk of injury to health or the environment the party must immediately inform the EPA. As mentioned earlier, in addition to the Chemical Safety for the 21st Century Act, some states adopted additional EU-style regulations after the European RoHS model.

Most countries in Africa already have regulations in place that assert legally binding controls on toxic substances. Lead is one example of a regulated toxic substance under a variety of such laws in African countries. In Algeria, for example, Arrêté No. 004/MINEPDED/CAB of 21 September 2017, modifies and completes the list of chemicals in Décret No. 2011/2581/PM of 23 August 2011, which regulates dangerous chemicals. Among other controls, the regulations prohibit the manufacture, sale and import of paints containing more than 90 ppm of lead (10/8/17). Algeria, Cameroon, Ethiopia, Kenya, South Africa, and Tanzania are among the African countries that have similar regulations.

India has adopted the National Action Plan for Chemicals, India's version of safety regulations for hazardous chemicals.[21] In the past, India's regulations were not modeled after "REACH," the EU regulation. In late 2019 and continuing into 2020, however, India embarked on the creation a National Action Plan for Chemicals (NAPC) to move into a more REACH-like system.[22] The idea is to create a harmonized system of classification of toxic chemicals that complies with the UN's Global Harmonization Strategy for chemical safety. Helping this effort is India's standing committee for chemical safety legislation, the National Coordination Committee (NCC) under the Ministry of Environment, Forest and Climate Change (MoEF&CC).[23] The late 2019 draft National Action Plan for chemical safety for India recommendations to compile a national chemicals inventory; analyse and assess the risks of those chemicals; implement the UN Global Harmonization Strategy (GHS); and develop risk mitigation strategies, policies and regulations.

Biometrics, as a technology of concern, merits high levels of attention to administrative and procedural controls, as well as a focus on harmonization on key aspects of regulation, such as agreement on definitions. However, if there is not specific R & D funding for investigating and studying regulatory models for biometrics, documenting their effectiveness, learning how and where to improve the models – then this work becomes extremely difficult for national legislatures to do on their own. As OSTP already knows, AI policy is complex, nuanced, and

---

[20] U.S. Environmental Protection Agency, *Chemical Safety for the 21st Century Act*, https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/frank-r-lautenberg-chemical-safety-21st-century-act.

[21] National Disaster Management Authority of India, *Chemical Disaster Page*, https://ndma.gov.in/en/2013-05-03-08-06-02/disaster/man-made-disaster/chemical.html.

[22] Chemical Watch, *India's draft national plan includes inventory and registration* (Jan. 6, 2020), https://chemicalwatch.com/86343/indias-draft-national-chemical-plan-includes-inventory-and-registration

[23] Government of India, Ministry of Environment, Forest, and Climate Change, http://moef.gov.in.

requires rigor. The World Privacy Forum encourages the OSTP to consider adding to the updated report some mention of the role of R & D in studying policy and regulatory solutions for focused areas of AI implementations.

The World Privacy Forum stands ready to assist, and to answer any questions you may have.

Respectfully submitted,

Pam Dixon
Founder & Executive Director,
World Privacy Forum
www.worldprivacyforum.org