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.

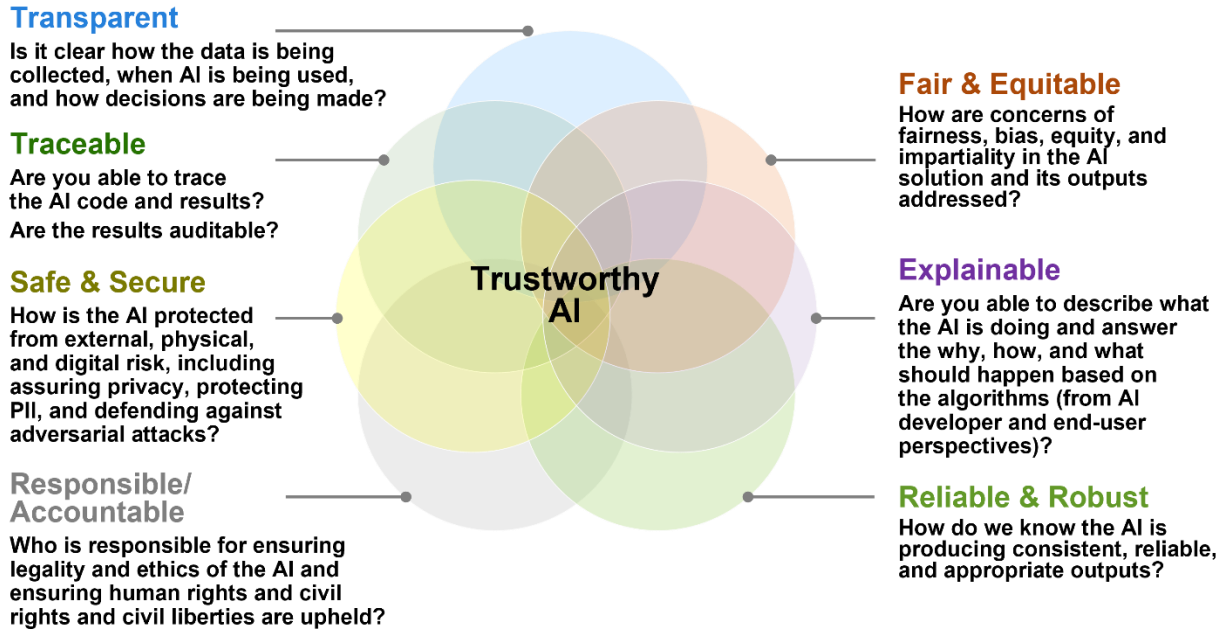

Figure 1 - Seven Connected Elements of "Trustworthy AI"

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.