Booz | Allen | Hamilton®

RESPONSE TO U.S. LEADERSHIP IN AI: A PLAN FOR FEDERAL ENGAGEMENT IN DEVELOPING TECHNICAL STANDARDS AND RELATED TOOLS

SUBMITTED BY
Booz Allen Hamilton
8283 Greensboro Drive
McLean, VA 22102

1.0 INTRODUCTION

While the NIST Plan for Federal Engagement in AI Standards is a good first step towards defining the role of the federal government in developing and monitoring AI standards, there is an opportunity to add more clarity and specificity to some of the recommended actions. The document cites the fact that many areas of AI are still in "formative" stages, and that further research is needed to inform standards, however, it is imperative that NIST use its role to drive focus into several key areas: data management, explainability and auditability, robustness, and governance. While it is important to be somewhat flexible and allow for standards to evolve over time with the technology, consensus around these features will ensure the development of trustworthy AI. Although the draft plan is a good starting point for developing standards that will improve the development and utility of AI systems, caution should be taken to not further today's approach to development which is defined by individual approaches to building and deploying the technology. The final plan should prioritize consolidation and streamlining approaches to ensure consensus and utility of the AI tools developed.

As it stands, the current approach for developing AI systems exposes federal agencies to the risk of developing tools and technologies that cannot be shared across different groups supporting similar missions, which then creates further bottlenecks and inefficiencies, wasted or duplicative resources, and furthermore impacts the ability to deliver to the mission. Although there exists tension between the desire to allow some flexibility in standards development and a more prescriptive approach, NIST should leverage its role as a liaison to identify areas where there are likely to be duplicative efforts or groups working towards similar goals and set standards to ensure consistency in the quality of solutions developed. In addition, this provides an opportunity for better sharing and capturing lessons learned. The current scenario with a lack of standards will be felt by government organizations for years to come as the systems today become legacy systems; even if the standards today must be modified in the future, they will impact the government's ability to both assess systems it purchases, as well as maintain them over the procurement lifecycle.

NIST should add clarifying language around coordination of horizontal standards across agencies, particularly for those focused on similar missions, to enable the sharing of AI tools and insights across agencies to enable faster speed to mission. NIST has the opportunity to drive focus across the following crucial areas: data management, auditability and explainability, robustness, and governance. Although vertical standards will allow for further specification within specific industry areas, NIST should emphasize the importance of horizontal standards within these key areas to enable the creation of trustworthy AI.

2.0 DATA MANAGEMENT

NIST's draft plan includes strong recommendations for developing and evolving data standards to support and facilitate the development of AI systems, specifically with the recommendations around non-traditional collaborative models with the private sector and increasing Federal data discoverability. Crucial to ensuring the integrity and trustworthiness of AI systems is ensuring a level of consistency in the quality and provenance of training data utilized to build AI models and systems. As discussed in Section 4.0 of this document around robustness, developing standards for data provenance now will provide additional protection against the threat of data poisoning in the future. In conjunction with federal agencies, NIST should coordinate and facilitate improvements in data sharing to support building better AI systems.

While Section 3.3 contains language around how the government and private sector can explore non-traditional collaborative models such as open data initiatives, there is an opportunity to add more specificity and focus for different AI applications, such as identifying initial priority areas for focused efforts around dataset discoverability, identification, and shareability. Issues of data ownership and licensing are significant within the context of the federal government systems where many open source licenses prohibit the use of data in contexts where the output will not be made

public. One solution to this challenge could be to supplement data discoverability / shareability initiatives with the ability to search data by license to make this information more readily available. As NIST identifies ways to increase data discoverability and access to Federal government data, it should begin by identifying, organizing, and sharing datasets that align to priority areas and missions with a heavy AI focus. For example, given the Department of Defense and Joint AI Center's organization around the National Mission Initiatives (NMIs) primed for AI development, NIST could begin facilitating the development of datasets and appropriate standards for the curation of datasets that align to the NMIs around humanitarian assistance and disaster relief (HADR) and predictive maintenance. Additionally, with the interest and focus on building AI tools for healthcare or fraud, waste, and abuse applications, NIST should work with applicable agencies to identify, prioritize, and develop standards that are appropriate for those specific applications. This approach will ensure that the standards developed are appropriate for specific verticals while informing the evolution of broader horizontal standards.

Recommendation: NIST should consider supplementing data discoverability / shareability initiatives with the ability to search data by license.

NIST should identify and prioritize the curation and development of datasets, standards, and sharing practices that align to missions or verticals with heavy AI focus such as humanitarian assistance and disaster relief or predictive maintenance.

3.0 EXPLAINABILITY AND AUDITABILITY

Much of the language in the Draft Plan centers around elements that support the research and development of "trustworthy" AI. Trustworthiness standards, as defined in the draft plan, include accuracy, explainability, resilience, safety, security, and reliability; the ability to measure, track, and monitor such features is contingent on the ability to build auditable and explainable AI. Rather than discussing broadly how to conduct research or focus industry engagement on these specific areas, NIST should examine the intersection of these features and recognize that to achieve trustworthy AI, AI systems must be transparent and explainable. This should be tackled from the bottom-up; NIST should enact standards that require auditability be a feature that is built into AI systems, which can then elevate and inform the process for creating explainable and trustworthy AI. The level of trustworthiness or explainability should be tailored to the intended use case or scenario.

The opacity of AI systems is one of the greatest barriers to trustworthy AI. Despite heavy research investment in AI over the last decade, in many cases, models are still black boxes, meaning that it can be impossible to parse or understand why the model reached a specific conclusion or recommendation. Models lack transparency and explainability, but the first step to ensuring that they are built into systems is to ensure that AI models are built to be auditable, meaning it is possible to have insight into the training data, model information, and other system information that inform the end decision, recommendation, or output. Auditability is key to understanding how and why AI systems arrive at conclusions or recommendations, and can then inform the appropriate explanation for the recipient of the information. Pursuant to Section 3.3, NIST should support and expand public-private partnerships to understand how to best tackle the challenge of explainability across broad industry verticals, and use the insights derived to inform horizontal standards

Section 1E of the Draft Plan calls out the need for tools for accountability and auditing to examine system output, traceability, or a record of events, but NIST has the opportunity to establish standards that influence the build of systems themselves that are auditable, rather than separate tools to perform this function. If AI systems are built to be auditable from the start, these additional tools for accountability and auditing will only be needed for independent verification and validation, which will ensure another layer of protection and improved performance. Building auditability into

systems as a part of the fundamental design ensures that there is transparency into how systems make decisions. Auditable AI systems should contain key metrics and information around system build and performance, which can be sourced by engaging with the appropriate agency and industry groups to identify the most salient features for agency/industry verticals. Like version history in software releases, AI audit trails should at a minimum contain information such as model version (source of the original model, the technique used to train it, performance metrics around accuracy, when it was last tuned), and data provenance (source of the training data). At an even more granular level, systems should be built so that this information can be accessed / tracked in real-time, so that it is possible to parse this information at the time of inference. Although many of these standards with respect to model versioning provenance, audit trails, etc. are likely already part of organizations' internal software development practices and aren't necessarily shared publicly, NIST should take on the responsibility of identifying, validating, and communicating best practices through its relationships with the broad machine learning research community, as well as its own research. Connecting back to the recommendations around data management in the previous section, the ability to build auditable systems is highly dependent on the degree of data maturity, which then influences the repeatability of the AI model development process.

By approaching the trustworthiness issue from the system-level, NIST can drive agencies and industry to common standards and metrics around how AI systems should be built, what information needs to be captured to establish performance evaluation metrics and benchmarks, and ultimately, stronger risk management and mitigation procedures. The all or none nature of ecosystem ownership especially with respect to data access and the sensitivity of the models means that private organizations may not be highly incentivized to produce easily auditable standards for their models, but it is up to organizations such as NIST to drive progress against goals around trustworthy AI. NIST should work with industry to discover and evaluate value creation from making trustworthy AI. If organizations can recognize additional value from market-driven acceptance of AI because AI trustworthiness is perceived as having more utility, then those organizations will be more apt to pursue it on their own, which will lead to more and better approaches for tackling the challenge.

Recommendation: Add language into section 3.2 around research into defining and building auditability into AI systems to promote explainability and further increase AI system trustworthiness.

NIST should take on the responsibility of identifying, validating, and communicating best practices through its relationships with the broad machine learning research community, as well as its own research; work with industry to discovery and evaluate value creation from making trustworthy AI.

Adversarial Attacks

Adversarial threats can occur during the training phase of building an AI model, as well as the inference phase, both of which can impact the performance and output of the solution. In the training phase, data poisoning occurs when engineered bad data is introduced into the training set and impacts model development and ultimately, performance, if it is deployed. In the inference phase, white box or black box attacks occur when attackers manipulate input data to fool a trained model, or can use the model output to reverse-engineer and produce adversarial input data that impacts model performance.

4.0 ROBUSTNESS

The Draft Plan is lacking with respect to addressing the unique challenges associated with AI security. Due to the lack of standards and governance in how they are built, AI systems are susceptible to adversarial threats, which can seriously threaten the integrity of the data and models that inform system output. If AI systems aren't robust and secure against adversarial attacks, they cannot be trustworthy, regardless of any efforts taken to build features that enable auditability or explainability. Both types of adversarial attacks (referenced in the text box) can have serious consequences when deployed, because the information / recommendation provided will be informed by false data, or a faulty model. Section 2A of

the Draft Plan highlights important standards characteristics that warrant Federal government consideration, and while several of the areas address components of the adversarial AI challenge, none specifically speak to the importance of building secure, robust systems that can withstand adversarial attacks. Pursuant to the recommended actions in Section 3.3, NIST should add language around engaging industry to conduct research and identify ways to build proactive defensive measures against adversarial attacks into AI systems.

NIST has a crucial role in ensuring the integrity and ongoing security of AI systems as they evolve. Although the risks are relatively low today as there are still challenges with deploying and operationalizing AI to the enterprise, the challenges with respect to preventing and safeguarding these systems against attacks should not be taken lightly. It is important to build the foundation now to proactively address adversarial attacks by understanding best practices to safeguard against data poisoning and model tampering. Many of these approaches are still nascent as research is still ongoing to create methods to identify tampering or establish proactive safeguards and measures. NIST can use its role as the arbiter of AI standards to both encourage collaboration between agencies and industry to develop the best methods to approach this challenge, while taking the lessons learned from this collaboration to inform the development of AI security standards. Ultimately, as AI systems become more embedded in the tools leveraged everyday, safeguarding against adversarial attacks will become paramount to both ensuring the trustworthiness and integrity of our AI systems, while also safeguarding our national security interests. For examples of AI benchmark programs for adversarial AI-specific threats, NIST should look to DARPA's Guaranteeing AI Robustness Against Deception (GARD) program aimed at producing a government-owned set of benchmarks and a test harness for measuring a system's ability to defend against adversarial attacks.

Recommendation: NIST should add language into Section 3.3 around engaging industry to conduct research and identify ways to build proactive defensive measures against adversarial attacks into AI systems.

5.0 GOVERNANCE

Pursuant to the recommendations in Section 3.1, crucial to bolstering AI standards-related knowledge, leadership, and coordination among agencies is establishing common governance standards for AI tools and solutions. As part of the National Science and Technology Council (NSTC) Machine Learning / Artificial Intelligence (ML/AI) Subcommittee's efforts to gather and share AI standards-related needs, it should examine and identify needs for common governance controls and practices. The current process for developing and deploying AI models is largely piecemeal, with bespoke models developed and deployed for specific data or problem sets. Rather than considering how to best develop and deploy AI that can be operationalized at enterprise scale, groups operate independently of one another to create AI models and tools that work for their problems. This approach, defined by a lack of coordination and oversight, exposes organizations to serious risk, because there is no way to establish common governance controls and procedures. By establishing common standards around governance that can be adapted for vertical standards to support individual agency or industries' needs, NIST can ensure the appropriate level of governance is adopted commensurate to the deployment, thereby reducing risk. NIST should add language to Section 3.1 to define its role in oversight of the establishment of common AI governance controls.

Establishing common governance controls for AI systems will depend on many of the recommended actions contained in Section 3.2. For example, by focusing research on developing metrics to assess the trustworthiness of AI systems and to inform risk management strategies including monitoring and mitigating risks, agencies will likely come to consensus around the appropriate attributes that should define and inform governance controls and procedures, which should be discussed and shared to enable better coordination and information sharing. While vertical standards should be

left to the individual agency level to coordinate as appropriate, NIST should act as the coordinator and arbiter of horizontal standards that inform governance controls. The areas mentioned earlier, auditability, explainability, and robustness will be crucial to building trustworthy AI systems, and also to identifying technical approaches to implement responsible behaviors. NIST can facilitate the sharing of such methods to inform common governance controls and approaches to ensure that the approach to developing and deploying trustworthy AI is one with minimal risks. By stipulating a more coordinated approach to AI governance, NIST can ensure that agencies accelerate speed to mission, which further allows us to maintain our success and positioning as the global AI leader.

Recommendation: NIST should add language to Section 3.1 to define its role in oversight of the establishment of common AI governance controls.

6.0 CONCLUSION AND RECOMMENDATIONS

NIST's draft plan for AI standards is the first step towards closing the gap between current AI solutions and standards in the U.S., as well as filling in the missing voice in the international arena. The European Union, OECD, and China have recently announced standards and guidelines to promote the development of responsible AI, but NIST's draft plan can ensure congruence in the approach to developing trustworthy AI systems and maintain U.S. leadership in AI. The longer we wait to address these issues, the more technical debt the government will accrue, which will only create more rework in the long run. To add more clarity to the draft, NIST should add further specificity and clarity to the following components essential to building and ensuring trustworthy AI:

AREA	RECOMMENDATIONS
Data Management	• Supplement data discoverability / shareability initiatives with the ability to search data by license • Identify and prioritize the curation and development of datasets, standards, and sharing practices that align to missions or verticals with heavy AI focus such as humanitarian assistance and disaster relief or predictive maintenance
Auditability and Explainability	• Add language into section 3.2 around research into defining and building auditability into AI systems to promote explainability and further increase AI system trustworthiness • Take on the responsibility of identifying, validating, and communicating best practices through its relationships with the broad machine learning research community, as well as its own research • Work with industry to discovery and evaluate value creation from making trustworthy AI
Robustness	• Add language into Section 3.3 around engaging industry to conduct research and identify ways to build proactive defensive measures against adversarial attacks into AI systems
Governance	• Add language to Section 3.1 to define its role in oversight of the establishment of common AI governance controls