RE: Response to the National Institute of Standards and Technology ("NIST") on "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)" (NIST–2023–0009)

To Whom It May Concern:

Palantir Technologies Inc. ("Palantir") is a U.S.-based software company that provides a Platform which enables public, private, and non-governmental organizations to integrate, analyze, and collaborate on their data in a secure and privacy-protective way. We are proud to make software that enables the institutions that serve our societies to use their data responsibly and effectively. Palantir was founded in 2003 on the conviction that it is essential to preserve fundamental principles of privacy and civil liberties while using data. It is for this reason that Palantir established one of the world's first Privacy & Civil Liberties ("PCL") Engineering teams more than a decade ago, specifically to focus on the development of privacy-protective technologies and to foster a culture of responsibility around their development and use.

Our response to this Request for Information ("RFI") is based on insights gathered over 20 years of experience building technology to uphold and enforce ethical and accountable practices in the use of our software products, including Artificial Intelligence ("AI") enablement tools and platforms. Palantir has contributed extensively to the conversation on the rise and appropriate use of AI technology through multiple public forums and responses to other RFIs to the Federal Trade Commission, National Telecommunication and Information Administration, Office of Science and Technology Policy, Office of Management and Budget, as well as to previous NIST RFIs. In addition to these responses, we have also contributed oral and written responses to the United Kingdom's House of Lords in their Inquiry on AI in Weapon Systems, the U.S. Senate Committee on Armed Services Subcommittee on Cybersecurity, and the U.S. Senate Bipartisan AI Insight Forum.

We are grateful to NIST for the opportunity to contribute to this important policy discussion. We welcome any request for clarification and look forward to the final memorandum on these critical issues.

Sincerely,

Anthony Bak, Head of AI Implementation, Palantir Technologies

Courtney Bowman, Global Director of Privacy and Civil Liberties Engineering, Palantir Technologies

Arnav Jagasia, Privacy and Civil Liberties Engineering Lead, Palantir Technologies

Carmen Jenkins, Solutions Lead, Commerce, Palantir Technologies

Morgan Kaplan, Senior Policy & Communications Lead, Palantir Technologies

Responses to Request for Information

As part of this request for information, we are pleased to share our perspective on two of the three main areas of interest identified: (Section 1) "Developing Guidelines, Standards, and Best Practices for AI Safety and Security," and (Section 3) "Advancing Responsible Global Technical Standards for AI Development."

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

As NIST develops guidelines and best practices for the responsible development, deployment, and use of AI systems, we encourage an emphasis on a few key factors: (1) Foundational Investments in Digital Infrastructure for AI; (2) General Testing & Evaluation Principles (that are inclusive of, but more expansive than red-teaming alone); and (3) Privacy-Enhancing Technology Development for AI.

Foundational Investments in Digital Infrastructure for AI

Based on our experience building software platforms for AI development, evaluation, deployment, and use, we have seen first-hand the importance of foundational investments in the digital infrastructure of AI systems, for both traditional machine learning ("ML") applications and for Generative AI. Such digital infrastructure should include high-quality data integration, pipeline management, data quality checks, system audit logs, granular access controls, the ability to version and branch data, and collaboration features to annotate datasets and identify addressable issues over time. From our experience, this kind of digital infrastructure is an essential enabling technology for the adoption of AI broadly, providing the software toolkit necessary to train and test AI models on relevant data, deploy models securely to end users, and capture decisions made with models to improve future iterations of the model lifecycle.

Furthermore, many of the objectives of federal AI policymaking focus on functions provided most dependably through digital infrastructure. For example, the Office of Management and Budget's draft memorandum on Guidance for Regulation of Artificial Intelligence Applications encourage agencies to adopt various release management practices, beyond the development and evaluation of the model itself.[1] Such capabilities should be part and parcel to the digital infrastructure in which an AI system is deployed. Moreover, the "checks and controls" suggested in this RFI would also be implemented in the digital infrastructure, which among other critical information management functions, ultimately also hosts an AI system.[2] As such, best practices

[1] For example, Section 5(c)(iv)(B) of the draft memorandum encouraged federal government agencies to "to leverage pilots and limited releases, with strong monitoring, evaluation, and safeguards in place, to carry out the final stages of testing before a wider release", https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

[2] Section 1(a)(1), which described topics related to NIST's assignment, includes "[t]he possibility for checks and controls before applications are presented forward for public consumption."

for the safe, secure, and trustworthy use of AI must include guidance about the digital infrastructure that will ground and enable the responsible use of those models.

At Palantir, we have experience building and fielding the digital infrastructure that has enabled our customers to use AI responsibly to support their most critical decision-making in a variety of environments, domains, and jurisdictions. Based on this experience, we can offer the following requirements for building safe, transparent, secure, and effective digital infrastructure for AI systems:

- **Security:** The digital infrastructure of an AI system provides the foundational security required for deploying an AI model in practical settings. For example, robust access controls are necessary to ensure that only authorized users can release a new version of a model or appropriate classification-based access controls are in place when training a model on new data. These security measures can govern access to training data or model execution and extend to new facets and potential applications of AI systems, such as access to "tools" used by a Generative AI system or georestrictions based on data and model residency requirements.

- **Data protection & privacy:** Capabilities for data protection and privacy are also critical for the digital infrastructure of an AI system. Federal government policy on AI governance has repeatedly underscored this principle.[3] As such, digital infrastructure must provide the capabilities necessary to uphold these essential and often nuanced policy requirements. This could include techniques for management of PII, data minimization, purpose limitation, and granular deletion, among others, as we discuss below in our response.

- **Data governance:** Data governance is a critical requirement of any digital infrastructure used to develop or interact with AI models. Data governance itself encompasses a broad category of capabilities, from tracking metadata and provenance, to release management to data quality checks. While these are certainly important for the digital infrastructure of any mission-critical application – regardless of whether AI is used – they are critical for establishing trust when working with a data-driven AI system. For example, data provenance capabilities can help evaluators understand the representativeness of data used to train a model or provided as input to a model. As another example, metadata about the data represented in a dataset, such as the distribution of certain attributes of the data or the recency of the data, are crucial for understanding whether the data might reflect an unwanted bias. Similarly, capabilities for continuously monitoring data quality are necessary for understanding how an AI system is actually operating in production environments and ensuring that its performance does not erode or succumb to known vulnerabilities of AI/ML brittleness.

[3] See Office of Science and Technology Policy, *Blueprint for an AI Bill of Rights*; Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence; Office of Management and Budget, *Memorandum for the Heads of Executive Departments and Agencies on Guidance for Regulation of Artificial Intelligence Applications*; National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework*.

- **Multi-stakeholder collaboration:** Building an AI system requires an interdisciplinary process wherein engineers, social scientists, domain-experts, representatives of the parties impacted by the AI system, and other relevant stakeholders are enabled to exchange insights from their respective areas of expertise, jointly address challenges, and constructively iterate on solutions throughout the AI lifecycle. Collaborating with a diverse set of stakeholders can increase trust in the AI system and can better assure that the AI system, as designed, will facilitate its intended impact. Importantly, this form of multi-stakeholder engagement also requires a guarantee that partners can maintain security and control over any data they choose to contribute to ensure its proper use.

- **Iterative development for decision support:** The digital infrastructure of an AI system can also allow organizations using AI to better capture the outputs of an AI model. When AI is used for decision-making, understanding how the suggestion (or output) of an AI system contributes to the end user making a real decision – and the extent and impact of that decision – is critical for transparency, retrospective audits, reproducibility, and contestability. At Palantir, we have seen firsthand how connecting AI to a data and decision model in our software platforms enables customers to use AI for decision-making, with the rails required to do so safely and in a human-centric manner.[4]

Red-teaming and Beyond: T&E Strategies for Assessing Generative AI Use Cases

A robust testing and evaluation ("T&E") process is essential for ensuring that AI systems and models are safe for deployment and operate as intended. With the increasing prominence of Generative AI applications, red-teaming has recently come into focus as a particularly important component of any reasonable T&E strategy. Red-teaming can help evaluators (e.g., supervisors, safety engineers, subject matter experts, ethicists, etc.) understand the specific failure modes of a given AI system, where an AI system may be prone to adversarial exploitation, and what safeguards developers can institute to ensure better outcomes.

While we agree that red-teaming is essential to many effective T&E processes, we also find it necessary to emphasize that red-teaming is just one method of evaluation in an entire suite of testing, evaluation, monitoring, maintenance, and risk management processes. For red-teaming to be most impactful, it should be used in conjunction with a broader risk management framework that may include methods and strategies drawn from a diverse range of T&E techniques that each have their own context-specific strengths and weaknesses. As such, we encourage NIST to adopt a wide lens for T&E best practices, situating red-teaming as one important component of an entire ecosystem of other complementary — and in some cases alternative — approaches.

Given that one of the goals of this RFI is to help inform a Generative AI "companion resource" to the AI Risk Management Framework, we describe in more detail below the broader spectrum of approaches that, together, can create robust T&E strategies. These strategies are, for the most part, generalizable to other classes of Generative AI (i.e., image generation, multimodal, etc.); however, for the sake of concrete illustration, we focus this exegesis on the most prominent Generative AI subclass, Large Language Models ("LLMs") and specifically LLM-powered

[4] For more, see our blog post, *Connecting AI to Decisions with the Palantir Ontology*.

workflows. In particular, we provide guidance on: (a) Basic-level T&E strategies; (b) Advanced-level T&E strategies; (c) Operational testing strategies; and (d) Scenario and simulation testing strategies. We recommend that NIST similarly share such guidance and best practices to relevant stakeholders considering how to evaluate the safety, security, and deployability of AI systems – and particularly, AI systems employing LLMs – in real-world environments.

The below documented strategies should also be considered in addition to what is typically required for the T&E process of virtually any type of machine learning or statistical model (e.g., standard metrics tracking between model versions over time and across data segments; being able to approve and recall models from production; model review and compliance workflows that facilitate discussion between AI experts and other stakeholders; ongoing monitoring for model and system drift; etc.).

Basic-level T&E Strategies

- **Reduction to Supervised Learning:** The easiest and perhaps most common T&E strategy is to reduce the problem to look like a classic supervised learning problem and use existing and well-understood T&E methodologies.

  To use an example, consider modeling the problem of extracting factory locations from incoming problem reports as named entity recognition, or modeling the problem of an LLM choosing among a selection of tools to run for text classification.

  When using an LLM to solve one of these problems, it is possible to start with very little, or even no labeled data. The first task then is to gather requisite testing data so that future T&E can be performed using standard methods. In low-risk scenarios, after trying several examples, one can gather data using a human-in-the-loop review process, either through an explicit label generation process or in benign circumstances, in production directly. To bootstrap and accelerate the creation of labeled data, it can be useful to use a data generation and labeling process with an LLM employed to produce test examples, which are then validated by a human. When taking this approach, users should always record which examples are generated by a human and which are generated by an LLM as performance may vary due to (perhaps subtle) differences in the data. This validated, labeled dataset can then be used as a test set to quantify model performance and validate continued performance in case of model changes.

  After a system has been running in this configuration, users may have enough training data to build a more traditional supervised model. The additional costs of training and sustaining a supervised model (i.e., in the form of compute, time, talent, and data) should be weighed against the benefits of basing a system around an LLM. One benefit includes the flexibility to easily adjust and reformulate the model; for instance, in our ongoing example, changing data extraction specifications from just “factory locations” to “factory and retail store locations” is potentially just a matter of inputting one or two lines in the prompt.

- **Track Prompts as Hyper-parameters**: Prompts into LLMs have significant impact on model output due to variations in prompt structure, wording, length, examples, and many other factors. Prompt engineering and other methods for steering model outputs towards desired outcomes has thus become the subject of much experimentation. Yet it remains challenging to identify patterns for *why* certain prompting strategies produce results through evaluating model characteristics alone. Evaluating empirical results from experimentation is currently the only reliable method for making such assessments.

As such, it may be worthwhile to employ model debugging and improvement tactics to decompose prompts into additional metadata, for instance, to track which embedded examples (if any) are used. In more complicated situations, the generation of the prompt (potentially using AI) is dynamic and should no longer be considered as a hyper-parameter, but the generation process itself should undergo a T&E process.

Advanced-level T&E Strategies

- **Decompose Tasks**: LLM tasks, unlike tasks from traditional ML models, can consist of performing multiple complex actions with each inference. As such, to evaluate such tasks, it is best practice to decompose the output into components that are both understandable and represent task targets to develop against.

    - **Generated Content**: In the abstract, one can decompose generated content and independently measure if the component parts are syntactically and semantically correct. In the case of free text, these characteristics may be difficult to evaluate and require "human evaluation" or "red-teaming" (see sections below for more detail). However, in many use cases, when using an LLM to form code or hit tool API endpoints, one can separately evaluate syntax (does the code parse as valid) from semantics (does the code do the requested task).

      Further decomposition isn't definable in the abstract and thus should be considered at the use case level and on a case-by-case basis. For example, in the case of asking for a medical summary of doctors' notes, we can separately ask about syntax (is the output correct language), semantics (did it capture the most important elements), and add in time (did it link the event to the proper event date), event sequencing order, etc.

- **Model State Space**: Although there are many kinds of agents/agent systems, one frequently observed involves the use of agents modeling a workflow as a Markov Decision Process and controlling state space transitions between sub-workflow elements. These systems can be difficult to evaluate even when combined with other methods outlined here (such as task decomposition, looking at overall workflow metrics, etc.). Simulations and scenario analysis can be useful for understanding how interconnected pieces behave. Evaluation in more generality is beyond the scope of this response.

- **Workflow KPIs as Metrics**: Some workflows do not have easy statistical measurements that can be applied directly to the output of the model. However, even in those cases, one should track model impact on the workflow as a whole via a designated KPI. For example, consider an LLM to assist with call center transcript generation. While it can be difficult to understand the effectiveness of a particular generated talk track, it is still possible to track system impact at the metric of “speed-to-resolution” and “aggregate cost.” Note that this information may not be immediately available at inference time and may only be understood in aggregate.

- **Human Evaluation**: Human evaluation (i.e., observation, approval, disapproval), much like data labeling, is an intensive process that in the best case is integrated seamlessly into the workflow, but in other cases may require both training and domain expertise, as well as an ancillary “oversight” workflow. In many workflows, human oversight, with implicit evaluation of model performance, is both the most important safeguard and source of evaluation data.

- **Red-teaming**: Although it has a longer history in other domains such as cybersecurity, red-teaming is now a developing area for the testing of AI systems. Unlike human evaluation, in red-teaming, humans take an active role in probing system limitations. But even as red-teaming standards are being developed and refined, active efforts to conduct red-teaming should take care to include subject matter expertise in the area of LLM application, which may include bringing in outside experts to assist with the red-teaming process.

- **Adversarial Evaluation:** Models can be evaluated with respect to a (growing) list of documented adversarial attack techniques. The specific details of the deployment determine what models of attack should be considered for evaluation. For example, while some models may not directly receive input from users (via chat interface), they may process information pulled from a database, which represents a distinct source of adversarial risk.

- **LLMs to Evaluate LLMs**: Increasingly, LLMs are being used to validate outputs created by a different LLM. This technique can be used to both reduce errors and estimate accuracy. However, this approach risks introducing a “turtles all the way down” scenario where the evaluation procedure — itself a model — needs an evaluation procedure. From the practical perspective of moderating errors, this approach has its merits. However, it is not a robust stand-alone T&E process, can introduce additional brittleness and errors, and should be discouraged as a singular approach, especially in high-risk application environments. Human evaluators and human/expert-driven evaluation approaches should be preferred in highly consequential application contexts.

Operational Testing Strategies

Operational testing strategies involve the use of realistic operational conditions to validate the performance of AI systems and is achieved for all systems (sometimes unintentionally) when they are deployed for actual usage. As a general principle, operational approaches are both riskier and costlier than standard supervised learning model T&E processes. Solely relying on operational testing — in lieu of other kinds of T&E methods — may seem unusual for individuals coming from an AI background, but most other fields of engineering are not structurally setup to have a zero-risk T&E infrastructure, which we usually have for AI models. In many fields, operational testing is not just the norm, but the only way that ideas can be vetted.

For example, operational testing in the form of clinical trials is central to drug development, which is a high-risk and high-consequence domain.

Extending beyond traditional model testing, operational testing provides the opportunity to gather system-level KPIs, such as time to execute, cost, and comparisons relative to the prior workflow (assuming the existence of antecedent, sufficiently comparable workflows). These KPIs give a more holistic understanding of a new AI system's value and can complement (and at times substitute for) model-specific T&E by capturing more realistic model impact and risks through actual usage. Furthermore, operational testing can reveal which errors are most impactful on workflow outcomes and even sometimes show that improved model performance can lead to negative outcomes under operational use (e.g., if users start to pay less attention because of boredom, or other reasons, as they rely on a well-performing, but imperfect model). Operational testing is also useful for making programmatic decisions (e.g., evaluating whether the AI-powered workflow is delivering value).

There are some conditions in which traditional T&E methods are insufficient, making operational testing necessary. For example, some use cases require operational testing because the workflow has delayed feedback elements or includes sequential interactions with other complex systems or people. Many LLM-driven workflows fall into this category, such as marketing campaigns or call center script development. More broadly, many autonomous systems require operational testing because the nature and variability of the real world involves greater complexity (i.e., more degrees of freedom) than what can be captured in static data sets and simulations.

Key techniques for safe and effective Operational Testing include:

- **Human Review**: Manual review of LLM workflows and outputs to ensure the AI system is functioning as expected. Human review can include both basic testing and dynamic "red-teaming" of LLM and system capabilities.

- **Safety/Unit Tests**: These tests are "sanity checks" where good results are known to the person running the test, but exist in small enough numbers that prohibit meaningful statistical testing. The difference between human review and unit tests is that while human review is subject to feedback loops where the human can interrogate different aspects of the system and respond to system behaviors, unit tests are automated. Unit tests are akin to in vitro toxicology tests that are done before starting Phase 1 of clinical trials.

- **Incremental Release**: This technique involves releasing new versions of an LLM-powered workflow to a small group of users — who know that they are using a new model — in order to do basic sanity and safety testing before approving a broader release. This is also akin to a Phase 1 trial in drug development, where the objective is to determine toxic effects before proceeding with larger studies.

Once the basic safety and performance of a system is established, complete end-to-end testing can commence in as realistic (operational) situations as possible, including:

- **A/B Testing**: With A/B testing, evaluators can compare the new model to an existing model with well-defined performance characteristics. The model being evaluated is usually the current production model, but you could consider another baseline. A/B testing is akin to Phases 2-4 in clinical trials, wherein researchers independently attempt to verify efficacy, compare to existing standards of care, and identify long-term effects in discrete trials. In most cases, this level of statistical rigor is not applied to AI systems.

- **Shadow Testing**: While not available to all LLM-powered workflows, one can sometimes deploy and evaluate a new model as a hypothetical alternative to an original model, whereby the original model is still used for the actual workflow, but the alternative choice of the new model is recorded and evaluated through post-hoc analytics.

- **Monitoring**: As one moves away from canned testing to realistic and open-ended usage, the on-going monitoring of system performance is critical to ensure that a model is not deviating from expected performance. For example, in clinical trials, this is called post-trial monitoring and is critical since the true variance of the population is difficult to capture under clinical trial conditions.

Scenario and Simulation Testing Strategies

Scenario and simulation strategies are the “digital twin” equivalent to running an autonomous system in a world simulation. They represent a middle ground between traditional model-focused testing and the real world, user-focused testing described as “Operational Testing” above. Scenarios are ultimately a combination of simulations, heuristics, programmatic interventions, and real-world data. In a scenario, a new model is run in its operational context, but potentially alongside simulations of other aspects of the system in order to test and characterize its operating characteristics.

Scenarios are most commonly used to understand complex systems that are difficult to reason through in the abstract. They provide a safe environment to systematically explore counterfactuals and enable optimization of difficult-to-reason-through parameters.

For LLMs in particular, scenarios and simulations can be used to measure how an LLM interacts with the system in which it is embedded, and from a T&E perspective, can add important safety and performance information that is otherwise inaccessible in the absence of operational testing.

Privacy-Enhancing Technologies for AI

The Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (“AI Executive Order” or “AI EO”) specifically highlights that “Americans’ privacy and civil liberties must be protected as AI continues advancing” and instructs agencies to use tools including privacy-enhancing technologies (“PETs”) “to protect privacy and to combat the broader legal and societal risks” of the improper collection and use of data.[5] Palantir established one of the world’s first Privacy & Civil Liberties Engineering teams, specifically to focus on the

[5] Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 2(f).

development of these types of privacy-protective technologies. Based on this expertise, we encourage NIST to adopt a broad definition of privacy-enhancing technologies in order to encompass core data protection functions such as data minimization, purpose/use limitation, storage limitation, and management of sensitive data, all which are highly effective at protecting privacy in practical applications of AI.

In Section 3(z) of the AI Executive Order, "privacy-enhancing technologies" is broadly – and thus, we believe, appropriately – defined as "any software or hardware solution, technical process, technique, or other technological means of mitigating privacy risks arising from data processing." The AI Executive Order, however, emphasizes specific approaches for privacy protection that are more common in the academic literature as promising research avenues, but does not include the full spectrum of approaches that are widely adopted in real-world, operational settings.[6] For example, the Executive Order places special emphasis on the techniques of differential privacy, homomorphic encryption, and secure multi-party computation.[7] These are certainly valuable privacy-enhancing technologies to explore, and guidance and best practices for deploying them in practice will help industry and government agencies adopt them more effectively in settings where they are truly applicable. In our experience, however, these technologies are still in their nascency and likely have limited applications. We therefore advise that continued investments in more foundational data protection capabilities will be equally, if not more important at protecting Americans' privacy in practice.

In addition to the specific privacy-enhancing technologies listed in the AI Executive Order, we encourage NIST to similarly emphasize and provide guidance for data protection technologies as part of a basic category of first-order PETs.[8] From our own experience developing software with industry-leading data protection capabilities, we can offer the following best practices for employing such technologies:

- **Deletion:** Deletion is an important remedy for privacy harms – even more so in data-driven AI systems. NIST can encourage that data used for AI systems should not be retained for longer than necessary. Capabilities for both the scheduled deletion of data and setting short retention periods by default can help uphold privacy by design in an AI system. Moreover, deletion capabilities should also be lineage-aware: that is, when data is deleted, that deletion action should propagate to all data derived from the original data source. This granular, lineage-aware approach to deletion ensures that deletion is comprehensive, preventing inadvertent re-use of data that should have been deleted.[9]

[6] Josep Domingo-Ferrer et al., *The Limits of Differential Privacy (and its Misuse in Data Release and Machine Learning),* Communications of the ACM, Volume 64, Issue 7 (2021), https://dl.acm.org/doi/10.1145/3433638.

[7] Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Sections 3(z) and 9(b).

[8] For more on how we draw a distinction between "first-order" or "basic" PETs and more nascent, speculative, or "exotic" PETs, see our blog post, *Privacy-Enhancing Technologies (PETs): An adoption guide (Palantir RFx Blog Series, #6)*

[9] For more, see our blog post, *Designing for Deletion (Palantir Explained, #6): Tactics for building an effective framework*.

- **Purpose Limitation & Use Limitation:** Enforcing the intended use of data or an AI model is critical for respecting the privacy of data subjects represented in the underlying data. Repurposing data and models for other uses may lead to privacy harms and violate the conditions under which the data or model was originally acquired. At the very least, it may lead to poor results from the AI model if the data is not representative of the new task. As such, we recommend that NIST encourage that digital infrastructures for developing and deploying AI systems include capabilities for purpose and use limitation. Purpose limitation frameworks – including Purpose-Based Access Controls[10] – can help ensure that data and models are only used for their intended purpose, protecting the data and privacy of data subjects.

- **Data Minimization:** Not all data in a dataset will be germane for model training, testing, or use. As such, we recommend that NIST provide best practices on how model developers, evaluators, and users can handle sensitive data when working with AI models. For example, sensitive attributes might be helpful for model evaluation to understand disparate impact of the model, but it might not be necessary – or even permitted – for model training. Coupled with techniques for use limitation, data minimization capabilities can better assure that sensitive data is only used when strictly required in the AI lifecycle.

3. Advance Responsible Global Technical Standards for AI Development

AI systems, from their development, to deployment, to use, are often not constrained by international borders. The scientific research that drives advancements in AI technology is an inherently global process; the commercial firms, governments, and organizations that collaborate toward a shared AI-enabled future are multi-national in nature; and the most common and overarching societal benefits and risks that come with the proliferation of AI will be apparent across national boundaries.

However, in addition to these shared benefits and risks, the increasingly ubiquitous use of AI technologies will implicate unique country-, sector-, and community-specific challenges that will require more flexible and tailored approaches to accountability mechanisms. As such, there is a need for balancing both country- and domain-specific regulatory standards, as well as global technical standards that can address shared opportunities and risks that transcend national borders.[11] Below, as part of this RFI's explicit call, we discuss the importance of this latter need to create responsible global standards for AI development and provide some suggestions on how to conduct this important and formidable pursuit.

[10] For more, see our blog post, *Purpose-Based Access Controls at Palantir (Palantir Explained, #2)*.

[11] See Peter Cihon et al., *Should artificial intelligence governance be centralised? Design lessons from history*, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020), https://www.fhi.ox.ac.uk/wp-content/uploads/Should-Artificial-Intelligence-Governance.pdf, for an interesting discussion on the delicate balance between centralized and fragmented approaches to international AI governance.

How to Create Beneficial and Responsible Global Technical Standards

A guiding consideration of any attempt to establish global AI technical standards should be a recognition of the manifest challenge of this undertaking. Namely, such standards will need to translate generally accepted principles — such as they can be established — into sufficiently meaningful normative practices that can be reasonably applied across the widest breadth of countries, cultures, jurisdictions, and societies.

Alternatively, the risks of failing to establish global standards are daunting. A contrasting approach that entails multiple, competing[12] (and in the worst case, fundamentally incompatible) standards would be liable to surface at least three major structural challenges, with implications to international trade, competition, and ultimately the safety and security of AI products:

- First, the need to navigate, translate, and adhere to multiple AI standards can create regulatory confusion for multi-national firms and actors seeking compliance, increasing the likelihood for inadvertent compliance errors that will ultimately lower AI system safety, security, and accountability outcomes.

- Second, the establishment of numerous competing standards may create unnecessarily complex barriers to entry for early-stage AI developers seeking to break into global markets. This outcome would stifle important sources of innovation, reinforce existing AI monopolies, and ultimately weaken market-based incentives for larger AI companies to innovate towards increasingly safe, secure, and responsible products. By consolidating the number of competing global standards, we can maximize regulatory legibility for more commercial firms around the world, decrease the costs of compliance for a broader ecosystem of innovators, and empower a more resilient commercial marketplace for LLM development.[13]

- Third, for those firms who do have the resources to navigate multiple unique standards, they may be able to engage in a type of “forum shopping” or “standard shopping” by which they will selectively choose to adhere to a standard that more closely aligns with that firm’s business model, or they will select out of compliance-based markets that are perceived to be the most stringent.[14]

As such, we advise an approach to global AI technical standards setting that commences on a base of foundational principles recognized by the international community — e.g., a statement of fundamental rights — in order to achieve a greater degree of legitimacy across partner nations and serve as a credible check to misaligned or non-rights-based standards that may emerge from

[12] See Charles Mok, *Global Competition for AI Regulation, or a Framework for AI Diplomacy?*, The Diplomat (2023), https://thediplomat.com/2023/11/global-competition-for-ai-regulation-or-a-framework-for-ai-diplomacy/

[13] For more on these dynamics, see Cameron F. Kerry et al., *Strengthening international cooperation on AI*, The Brookings Institution (2021), https://www.brookings.edu/articles/strengthening-international-cooperation-on-ai/.

[14] For more on AI forum shopping in the EU context, see, e.g., Mauritz Kop, *EU Artificial Intelligence Act: The European Approach to AI*, Transatlantic Antitrust and IPR Developments (2021), https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/.

illiberal countries. From there, the approach should focus on adopting and working through an agreed upon *methodology* or *process* over any explicit substantive requirements.

To help incentivize and facilitate the ultimate adoption and adherence of global norms and standards for AI development, as well as disincentivize adverse behavior that will undermine responsible AI development, we can recommend the following measures and procedural considerations:

- That organizations seeking to establish global standards provide inclusive, fair, and equitable compliance resources to adherents from across the globe.

- That organizations seeking to establish global standards coordinate with one another to establish and use common language, definitions, and technical references, to make it easier for global compliance seekers to understand where different standards overlap (or deviate) from one another in practice, as opposed to in rhetoric.

- As the United States is currently the market leader in global AI innovation — and per Section 11 of the AI Executive Order, "Strengthening American Leadership Abroad" — the U.S. Government can play a proactive role in both setting global standards and serving as facilitator for coordination across international compliance bodies. In particular, NIST and the newly established U.S. AI Safety Institute can help provide a high-level consolidation of competing international standards, and further promote global adherence to the safety and rights-based global standards set by the U.S. Government, in conjunction with industry and civil society leaders. This may ultimately ease both the burden and rate of compliance for international commercial technology providers, particularly those that are less-resourced to navigate numerous independent regulatory bodies.

- To increase stakeholder engagement at the international level with U.S.-established standards, NIST can create avenues for consistent and constructive dialogue with non-U.S. commercial firms, universities, and civil society actors with whom the U.S. has a vested interest in compliance with U.S.-based standards. Including international actors in the process of norms and standard creation will help establish global legitimacy for U.S.-initiated standards, as well as a mechanism for debate and adjudication where and when the domestic and international normative discussions become misaligned.

These measures, while not exhaustive,[15] can provide a firm basis for expanding the effort and work towards a gradual consensus on AI technical standards.

[15] For further discussion on how to achieve broad support for global standards, see, e.g., Var Shankar and Philip Dawson, *AI standards and certification programmes in a competitive global landscape*, Observer Research Foundation (2024), https://www.orfonline.org/expert-speak/ai-standards-and-certification-programmes-in-a-competitive-global-landscape.