DLA Piper LLP (US)
1900 North Pearl Street
Suite 2200
Dallas, TX 75201-2467
www.dlapiper.com

Danny Tobey
Danny.Tobey@us.dlapiper.com
T 214.743.4538
F 972.813.6275

February 2, 2024
VIA E-MAIL

Information Technology Laboratory
ATTN: AI E.O. RFI Comments
National Institute of Standards and Technology
100 Bureau Drive
Mail Stop 8900
Gaithersburg, MD 20899-8900

Re: Response to National Institute of Standards and Technology Request for Information (Docket Number 231218-039)

Dear National Institute of Standards and Technology Members:

We write in response to the National Institute of Standards and Technology's ("NIST") Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence ("RFI"), published in the Federal Register on December 21, 2023, Docket Number 231218-0309.

The following response is sent on behalf of the members of DLA Piper (US) LLP's ("DLA Piper") Artificial Intelligence and Data Analytics Practice.

Artificial Intelligence Risk Management and Governance

It is imperative for Artificial Intelligence ("AI") actors to reassess and fortify their governance frameworks to manage the risks of generative AI. The convergence of technological innovation with ethical, legal, and societal dimensions necessitates a holistic and dynamic governance approach. Emphasizing robust risk assessment protocols, transparency, AI literacy, and collaborative engagement not only mitigates potential risks but also ensures the responsible and beneficial deployment of generative AI technologies. The following recommendations aim to guide AI actors in navigating the complex terrain of generative AI, fostering an ecosystem where innovation thrives in harmony with ethical standards and societal expectations.

- *Developing Advanced Risk Assessment Protocols*: Establish sophisticated frameworks to identify, assess, and prioritize risks associated with generative AI. This involves not just the technological aspects but also ethical, legal, and societal implications. Regularly updating these frameworks ensures that emerging risks are promptly addressed.

- *Enhancing Transparency and Accountability Mechanisms*: Implement comprehensive documentation processes for AI development, including data sourcing, model training, and decision-making algorithms. This ensures that every aspect of the AI's operation can be reviewed and audited, promoting accountability and trust.

- *Strengthening AI Literacy and Specialist Training*: Invest in extensive training programs to ensure that all stakeholders, including developers, users, and decision-makers, are well-versed in the strengths, limitations, and ethical considerations of generative AI. This cultivates a responsible approach to AI usage and innovation.

- *Promoting Collaborative Engagement and Standardization*: Encourage active collaboration between industry, regulatory bodies, and academic institutions to share insights, challenges, and best practices. This collective effort can lead to the development of industry-wide standards and guidelines, promoting consistency and safety in the deployment of generative AI technologies.

- *Red Teaming*: Regularly simulate scenarios where generative AI systems are challenged across multiple dimensions, including security & privacy, bias & fairness, and values alignment. Challenging the system regularly and in diverse areas of application can anticipate and mitigate risk.

- *Dynamic Consent Protocols*: Given the changing landscape of data usage, consent mechanisms should be flexible and transparent, allowing users to understand and control how their data is used by AI systems. This requires continuous engagement with users and other affected parties, providing clarity and control over their data throughout the AI lifecycle.

- *Long-Term Impact Monitoring*: Establish systems to continuously observe and assess the broader effects of AI over time, including societal, economic, and environmental impacts. This involves not just initial impact assessments but ongoing monitoring to identify subtle, long-term effects that may not be immediately apparent.

- *Cross-Disciplinary AI Audits*: Involve experts from various fields in auditing AI systems. This interdisciplinary approach ensures that AI is evaluated from multiple perspectives, capturing a more comprehensive understanding of its implications on various aspects of life.

Current Standards Gaps

Because much of the standardization creation related to AI occurred prior to the mass adoption of generative AI systems such as large language models ("LLMs"), the current landscape of AI standards does not adequately address the specific challenges and nuances associated with generative AI technology. Generative AI models possess extensive capabilities, creating high quality text, images, and video. These recent AI advancements require targeted guidelines and regulations, particularly in areas like mitigating biases, ensuring the ethical use of generated content, and safeguarding against the misuse of these models in generating misleading or harmful content.

Furthermore, the opacity of generative AI systems calls for enhanced standards in transparency and explainability. Given their complexity and the often 'black box' nature of their operations, it's vital to develop standards that clarify how content is made, which factors influence results, and how outputs can be trusted by users. This transparency is essential not only for user trust but also for ongoing monitoring and improvement of the models.

Lastly, the integration and interoperability of generative AI with other systems and frameworks poses a significant challenge, largely unaddressed by current AI standards. As generative AI becomes more embedded in various technological ecosystems, standards need to evolve to ensure that these integrations do not compromise the integrity or the intended functionality of the systems involved. This includes establishing clear guidelines for data handling, model updates, and system-to-system interactions, ensuring that the incorporation of generative AI into larger systems is both effective and responsible.

Techniques for Model Validation and Verification

Validating and verifying AI models is a multifaceted and context-dependent process, crucial for ensuring the models perform as intended and can be trusted in real-world applications. This involves not just assessing performance metrics but also understanding the model's behavior under various conditions and ensuring it aligns with ethical and operational standards. Each of the techniques summarized below plays a vital role in painting a comprehensive picture of the model's reliability and robustness.

- *Identify relevant performance metrics*: When designing the AI system, consult with key stakeholders, including users, subject matter experts, and impacted parties, to identify key performance metrics for the system. Different metrics can be most relevant in different applications of AI. For example, in disease detection, the false negative rate is often most important, while false positive rate may be more salient in loan approval systems. Regardless of the metrics chosen, the goals for model performance must be clearly articulated prior to model building.

- *Assess performance across many possible deployment scenarios*: Once an AI system is deployed, it will likely be used in areas which do not represent the development environment. While not all of the possible use cases can be anticipated, it is crucial to assess the model's performance under different conditions to minimize potential deployment biases. Again, consultation with all stakeholders is necessary to enumerate different scenarios and identify model weaknesses. Lessons learned from this area of investigation into model performance can be used to guide the use of the AI system by users, deployers, purchasers, and other third parties. For example, a model may only be performant for predicting the loan default risk for borrowers aged 25 – 65, and appropriate guidance should be provided so that the system does not produce predictions for those under 25 or over 65.

- *Ongoing Monitoring & Testing*: AI systems must be continually monitored to ensure high levels of performance. Prior to deployment, there should be a clear plan in place to monitor the model across the previously defined relevant performance metrics and in different use cases. A proven way to assess the performance of a system is to periodically compare model results and human made results for a random sample of instances. For example, a medical expert reviews a random sample of x-ray images for signs of pneumonia on a regular basis while being blinded from the results of the AI system designed to automatically identify pneumonia from lung x-ray images. The human expert and AI system results are compared to determine if the AI system remains performant. In any case, deployers and users of the AI system must be prepared to decommission the system if the ongoing monitoring and testing proves it is no longer performant.

Beneficial Red-Teaming Use Cases

AI red teaming is an indispensable strategy for enhancing the resilience and reliability of AI systems, particularly in domains where the stakes are high. AI red teaming is especially crucial in sectors where AI-driven decisions can have profound impacts or where system failure can lead to significant consequences, ensuring these technologies operate within safe, ethical, and legal boundaries. The following are use cases where AI red teaming is particularly beneficial.

- *Legal Liability:* Red teaming generative AI systems is particularly crucial to better understand potential legal exposure. Generative AI has been, and will continue to be, deployed in industries which are governed by stringent legal and regulatory frameworks, such as healthcare, or finance. The legal frameworks in these industries were developed as a mechanism to protect individuals from harm. Therefore, it is a critical safeguard in the responsible deployment of AI systems to validate that these essential legal safeguards are not contravened.
- *Critical Infrastructure*: In critical sectors such as healthcare, finance, and utilities, the robustness of AI systems is non-negotiable. AI red teaming is crucial in these domains, as it rigorously tests AI applications against extreme scenarios, ensuring systems can withstand and adapt to unexpected disruptions. The objective is not just to prevent operational failures but to safeguard against potential cascading effects that could lead to economic turmoil, jeopardize public health, or compromise national security. By anticipating and preparing for the worst-case scenarios, stakeholders can ensure the continuity and reliability of essential services, reinforcing crucial infrastructures.
- *Autonomous Systems*: The reliance on AI in autonomous systems demands an uncompromising commitment to safety and reliability. Red teaming in this context involves a comprehensive evaluation of these systems' ability to handle a spectrum of operational challenges, including navigating unforeseen environmental conditions, reacting to erratic behavior from other entities, and maintaining operational integrity in the face of threats. The aim is to ensure that these autonomous systems consistently make safe, reliable decisions, thereby minimizing the risk of incidents and enhancing public trust in these rapidly evolving technologies.
- *Cybersecurity Defense*: In the realm of cybersecurity, the landscape is continually evolving, with adversaries constantly devising new and more sophisticated methods of attack. AI-driven security systems, therefore, must be dynamic and proactive. Red teaming plays a critical role in this sector by simulating advanced cyber-attacks and testing the resilience of AI defenses. This not only helps in identifying potential vulnerabilities but also in developing a more adaptive, responsive cybersecurity posture, ensuring the protection of sensitive data and critical infrastructure against the increasingly sophisticated threats posed by malicious actors.
- *High-Stakes Decision Making*: In areas such as judicial sentencing, credit scoring, or recruitment, where AI-driven decisions can significantly affect individual lives, the stakes are exceptionally high. Red teaming in these domains focuses on rigorously testing AI systems to uncover and mitigate biases, ensuring decisions are fair, transparent, and accountable. It's about safeguarding the ethical integrity of AI applications and maintaining public trust. This process is fundamental in ensuring these technologies are not just advanced in terms of capabilities but are also aligned with societal values, ethical norms, and legal standards.

Structured Mechanisms for Red Teaming

As discussed above, AI red teaming is a structured, analytical process designed to rigorously evaluate AI systems by simulating adversarial scenarios and attacks. This proactive approach is essential in identifying vulnerabilities and risks, enhancing system robustness, and ensuring AI solutions are resilient against potential threats. By systematically challenging AI systems through scenario planning, attack simulation, vulnerability assessment, and mitigation strategy development, organizations can preemptively address risks, fortify AI defenses, and maintain the integrity and reliability of their AI deployments. The following are essential mechanisms for red teaming AI systems.

- *Scenario Planning*: This involves the meticulous construction of detailed, plausible adversarial scenarios that the AI system may encounter. The goal is to anticipate a wide range of challenges, from data leakage to direct attacks on the AI infrastructure, ensuring a comprehensive assessment of the system's preparedness and response mechanisms.

- *Attack Simulation*: Conduct controlled, deliberate attacks on the AI system to observe its responses in real-time. This dynamic testing goes beyond theoretical analysis, exposing the system to practical threats to gauge its resilience and capacity to maintain functionality under duress.

- *Vulnerability Assessment*: Methodically scrutinize every component of the AI system to identify vulnerabilities, whether in the data pipeline, algorithmic structure, or operational framework. This thorough examination helps in pinpointing specific weaknesses that could be exploited by adversaries, forming the basis for targeted improvements.

- *Mitigation Strategy Development*: Based on the insights gained from the above steps, developing robust strategies to address identified vulnerabilities. This involves not just technical solutions, but also considering broader organizational policies, training protocols, and communication channels to ensure a holistic enhancement of the system's defensive capabilities.

The DLA Piper Artificial Intelligence and Data Analytics Practice appreciates the opportunity to contribute to this essential discourse. We trust that our insights, drawn from a blend of legal expertise and a deep understanding of AI's multifaceted impact, will aid in shaping robust, comprehensive safety mechanisms for the future. We remain committed to collaborating to usher in the responsible evolution of AI technologies.

Respectfully Submitted,

DLA Piper
Artificial Intelligence and Data Analytics Practice