ALERT

Department of Commerce Releases Five Products to Help Guide AI Development

By The Honorable Jerry McNerney, Elizabeth Vella Moeller, Brooke L. Daniels, Benjamin J. Cote, Amaris Trozzo

TAKEAWAYS

- The Department of Commerce released three final guidance documents incorporating public comments from earlier this spring which provide recommendations for managing AI risk, securing AI software development processes, and developing global AI norms and standards.
- The AI Safety Institute has published draft guidance for public comment providing recommendations to mitigate threats specific to AI dual-use foundation models, which are those models that could be leveraged to create weapons or facilitate cyberattacks.
- The National Institute of Standards and Technology has also released an open-source software that can help test the resilience of systems to adversarial attacks.

08.16.24

Less than a year after the publication of the Executive Oder (EO) on the Safe, Secure, and Trustworthy Development of AI, the Department of Commerce has finalized three pieces of guidance to fulfill its obligations under the EO. In addition, the recently created AI Safety Institute (AISI) has provided draft guidance to help AI developers mitigate risk of dual-use foundation models. AISI is soliciting comments on this proposal, which are due by September 9. The National Institute of Standards and Technology (NIST) has released open-source software that can be used to test AI systems responses to adversarial attacks.

Managing the Risk of Misuse for Dual-Use Foundation Models

AISI has released its first draft document regarding mitigating risks specific to dual-use foundation models. As defined by the EO, foundation models are those that are trained on broad data; use self-supervision; are applicable across a wide range of contexts; and exhibit (or could be easily modified to exhibit) high levels of performance at tasks that could pose a threat to national security. Current models contain up to 10^{14} parameters, but while presenting great computational potential, can also be leveraged for the development of weapons and facilitate offensive cyberattacks. The guidance is primarily directed at developers; however, the objectives and recommendations can be leveraged by users across the model's life cycle. The guidance identifies seven objectives that will help mitigate risk and improve the safety of dual-use foundation models, including:

- anticipating potential misuse risk;
- establishing plans for managing misuse risk;
- managing the risk of model theft;
- measuring misuse risk;
- ensuring that misuse risk is managed before deploying foundation models;
- collect and responding to information about misuse after deployment; and
- providing appropriate transparency about misuse risk.

AISI provides methods to measure or track success towards achieving each objective. Each objective is accompanied by discrete practices that the developer (or other actor) can employ to achieve the objective along with recommendations on how to best carry out that practice. AISI identifies documentation that the developer can create and share to prove its alignment with each proposed practice and objective. For example, AISI is seeking comments from the public on how the risks of misuse, throughout the AI development cycle, can be further remediated. Each risk is provided with an accompanying objective and task to help alleviate the risk so that dual-use foundation models can be maximized for their potential without the risk of corruption by malicious actors. AISI is seeking input from public stakeholders to determine whether there are additional recommendations, practices or objectives it should consider in issuing its final guidance.

Comments will be accepted through **September 9, 2024.** *If you are interested in providing comments to the AISI, please reach out to Pillsbury who can support your comment drafting and submission efforts.*

Final Guidance

In April of 2024, NIST released its initial draft for public review and feedback. With these lessons-learned from the academic, nonprofit and industry stakeholders, NIST has now released three final documents addressing AI risk, AI software development and AI global standards development.

The AI Risk Management Framework (RMF) AI Profile is designed to help organizations identify and manage generative AI risks, including the potential use of their generative AI platforms to develop dangerous weapons, cite harmful content, produce false and erroneous results and reveal sensitive information. The guidance identifies a variety of risks that are specific to generative AI, noting that this list is not exhaustive, and provides examples of how trustworthiness can be further enhanced.

The Secure Software Development Practices for Generative AI and Dual-Use Foundation Models guidance provides recommendations to help reduce threats to, and better secure, the data underpinning AI systems and is meant to be used in conjunction with the Secure Software Development Framework. The guidance addresses the risk that is unique to AI models in that the interaction between the data and the configuration parameters can create closed loops that can be subject to manipulation. The guidance provides specific tasks that the AI developer can implement, in order of priority based on risk to the system, to secure their data from malicious actors.

Finally, the Plan for Global Engagement on AI Standards provides targeted guidance on the development of global AI norms and standards. This paper addresses how the United States can, by collaborating with standards developing organizations, industry, academia and other state governments, lead in developing AI standards related to transparency, access and the ethical deployment of AI. The plan addresses standards not only across the use of AI, but also provides assessments of sector-specific AI risks. The guidance identifies certain topics as being ready for international standard development, including developing standard terminology, setting measurement methods and metrics, increasing transparency so users understand when they interact with digital content, developing risk-based management processes for AI systems (like the NIST AI Risk Management Framework) and enhancing security and privacy.

NIST Software for Measuring Adversarial Attacks

NIST has released a software package to measure the impact of adversarial attacks. The software addresses a specific risk to AI systems—the potential for malicious actors and adversaries to feed poison data or harmful inquiries into the AI system to result in incorrect, and maybe even dangerous, outputs over time. The software will determine how resilient the AI system is in resisting such false information and maintaining its original mission and instruction. The software is one way that AI developers can test their products for free before broad deployment. The open-source software is available for download here.

Pillsbury's multidisciplinary team of AI thought leaders and legal and strategic advisors is an industry leader in strategic promotion of responsible and beneficial AI. Pillsbury is closely monitoring AI-related legislative and regulatory efforts. Our AI team helps startups, global corporations and government agencies navigate the landscape impacted by emerging developments in AI. For insights on these rapidly evolving topics, please visit our Artificial Intelligence practice page.

These and any accompanying materials are not legal advice, are not a complete summary of the subject matter, and are subject to the terms of use found at: https://www.pillsburylaw.com/en/terms-of-use.html. We recommend that you obtain separate legal advice.