



| Brief |

Executive Order Outlines U.S. Priorities on AI Governance, Risk Mitigation

B-2023-11-08b

Classification: CLEAR

Intelligence Requirements: Synthetic Media, Deepfakes

Scope Note

ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were **identified prior to 12:00 PM (EST) on November 7, 2023**; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

Brief | Executive Order Outlines U.S. Priorities on AI Governance, Risk Mitigation

> Key Findings

- U.S. President Joe Biden's latest Executive Order seeks to outline a policy roadmap for artificial intelligence (AI) including measures to safeguard against emerging risks and to orchestrate a framework for the responsible development and deployment of AI.
- The Order also reflects the Biden administration's efforts to make AI more secure, to align U.S. policy on this issue with similar emerging policies in the European Union and United Kingdom, and to respond to growing competition in AI development from China.
- The Order directs the National Institute of Standards and Technology (NIST) to take the lead to introduce industry standards for safe, secure, and trustworthy AI systems.
- The Order further seeks to protect US-based data and places the burden for enhanced regulatory oversight on the Secretary of Commerce to improve transparency regarding how AI data is used by non-U.S. entities that may facilitate malicious cyber activity.

| Analyst Commentary

U.S. President Joe Biden's latest executive order seeks to outline a policy roadmap for artificial intelligence (AI) including measures to safeguard against emerging risks and to orchestrate a framework for the responsible development and deployment of AI. On October 30, 2023, President Biden issued [Executive Order 14110](#) on the "Safe, Secure and Trustworthy Development and Use of Artificial Intelligence" to foster and fund the safe development of trustworthy AI, to guide US federal agencies on AI regulatory principles, and to align private industry efforts with national defense priorities.

- The Order also reflects the Biden administration's efforts to make AI more secure and to align US policy on this issue with allies in the European Union and United Kingdom and to respond to growing competition in AI development from China.

Encompassing a broad range of industries and sectors, the directive assumes that maximizing the positive potential and multitude of advantages of AI necessitates curbing its significant risks. To that end, the Order seeks to advance and govern the development and use of AI in accordance with eight guiding principles and priorities:

- Artificial Intelligence must be safe and secure;
- Promoting responsible innovation, competition, and collaboration to unlock the technology's potential to solve some of society's most difficult challenges;
- The responsible development and use of AI require a commitment to supporting American workers;
- Artificial Intelligence policies must be consistent advancing equity and civil rights;
- The interests of Americans who increasingly use, interact with, or purchase AI and AI-enabled products in their daily lives must be protected;
- Americans' privacy and civil liberties must be protected as AI continues advancing;
- Manage the risks from the Federal Government's own use of AI and increase its internal capacity to regulate, govern, and support responsible use of AI to deliver better results;
- The Federal Government should lead the way to global societal, economic, and technological progress, as the United States has in previous eras of disruptive innovation and change.

Mitigation

B-2023-11-08b

TLP: CLEAR

The Order directs NIST to take the lead to introduce industry standards for safe, secure, and trustworthy AI systems. NIST has been directed to create three resources within 270 days from the issue date of the Order:

- A generative AI risk management framework intended to accompany NIST's AI Risk Management Framework;
- A framework to apply secure development practices to generative AI and dual-use foundation models;
- An initiative to create guidance and benchmarks for evaluating and auditing AI capabilities. This will focus on capabilities through which AI could cause harm, such as cybersecurity and biosecurity.

The Order also outlines mandates for government agencies and private firms to report training and safety test results to the U.S government, requirements which are based on the size of models and their capabilities, reflecting a concern that unchecked proliferation of GPT-like models could lead to a heretofore unseen malicious use by bad actors, against which it would be hard to defend. Reporting requirements apply to large computing clusters and models trained using a quantity of computing power at the level of ~25-50K clusters of H100 GPUs—a massive scale of computational resources, indicating that each cluster consists of multiple high-performance H100 graphics processing units involved in the training of AI models, which denotes a very high level of computing power.

- These parameters can change at the discretion of the U.S. Secretary of Commerce, but the specified size and interconnection measures are intended to bring only the most advanced “frontier” models into the scope of future reporting and risk assessment.
- Compute-based limits are very unlikely to effectively mitigate the harms that can be achieved by smaller models, including the propagation of online mis-, dis-, and malinformation, as well as model hallucinations.

The Order further seeks to protect U.S.-based data and places the burden for enhanced regulatory oversight on the Secretary of Commerce to improve transparency regarding how AI data is used by non-U.S. entities that may facilitate malicious cyber activity. Within 90 days of the Order's publication, the Secretary of Commerce is tasked with drafting regulations that will:

Mitigation

B-2023-11-08b

TLP: CLEAR

- Obligate U.S. Infrastructure-as-a-Service (IaaS) providers to report to the Secretary of Commerce any instance in which a foreign entity engages their services to train a substantial AI model that could be employed for malevolent cyber activities (referred to as a "training run").
- Impose a requirement on U.S. IaaS providers that their foreign resellers must report to the Secretary of Commerce when a foreign entity purchases and utilizes their product for a training run. This regulation is designed to prevent foreign resellers from offering services from U.S. IaaS providers unless they agree to monitor and report these transactions for U.S. regulatory purposes.

The Order represents a positive step by the U.S. government in acknowledging and preparing to mitigate potential risks arising from AI, particularly in relation to privacy, discrimination, open competition and security. International collaboration on safety standards and applications will be crucial to this effort even as the US seeks to lead on this issue.

- Watermark requirements for generative AI-created content outlined in the Order are likely to be difficult to implement. Research has shown that watermarks can be easy to remove or obfuscate, particularly on social media. The success of such a policy will likely rely on innovative ways to make the removal of watermarks more difficult, as well as positive engagement with the policy by online platforms.
- The burdens of testing large AI models are likely to be disproportionately borne by smaller organizations with fewer resources, which could reinforce monopolies and stifle competition.

Appendix A: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	Sources may use TLP:RED when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	Sources may use TLP:AMBER when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	Recipients may NOT share TLP:RED with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	Recipients may ONLY share TLP:AMBER information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. Note that TLP:AMBER+STRICT restricts sharing to the organization only.
	Green	Clear
WHEN SHOULD IT BE USED?	Sources may use TLP:GREEN when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	Sources may use TLP:CLEAR when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	Recipients may share TLP:GREEN information with peers and partner organizations within their sector or community but not via publicly accessible channels.	Recipients may share TLP:CLEAR information without restriction, subject to copyright controls.

Appendix B: ZeroFox Intelligence Probability Scale

All ZeroFox intelligence products leverage probabilistic assessment language in analytic judgments. Qualitative statements used in these judgments refer to associated probability ranges, which state the likelihood of occurrence of an event or development. Ranges are used to avoid a false impression of accuracy. This scale is a standard that aligns with how readers should interpret such terms.

Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain
1-5%	5-20%	20-45%	45-55%	55-80%	80-95%	95-99%