

Morgan Lewis

LAWFLASH

EU AI ACT, US NIST TARGET CYBERATTACKS ON AI SYSTEMS— GUIDANCE AND REPORTING OBLIGATIONS

July 16, 2024

AUTHORS

Elizabeth B. Herrington, Vishnu Shankar, Phillip J. Wiese

The European Union published on July 12, 2024 the final text of its Artificial Intelligence (AI) Act, in force on August 1, 2024, which will implement material cybersecurity and incident reporting requirements, among other requirements, for companies in response to increasing cyberattacks on AI systems and models. These regulatory obligations mirror initiatives taken by other governments to address cyberattacks on AI systems, notably, the United States' National Institute of Standards and Technology (NIST) releasing guidelines earlier this year on preventing and mitigating such incidents. As governments intensify their efforts against these attacks, organizations should consider maintaining robust information governance and security policies and assessing the regulatory obligations and legal risks associated with cyberattacks on AI systems and models.

As AI and machine learning (collectively, AI) systems and models become more ubiquitous in the marketplace, they are more frequently the target of cyberattacks. These technologies can be lucrative targets because they often contain vast troves of data, some of which may be commercially sensitive or personal. Attackers may target AI models to gain access to underlying information or to disrupt the model's processes. Many leading developers of AI systems and models are taking these risks seriously.

Notably, OpenAI Inc. announced on June 13, 2024 that US General Paul Nakasone, former leader of US Cyber Command and former US National Security Agency Director, was joining its board of directors, which "underscore[d] the growing significance of cybersecurity as the impact of AI technology continues to grow."

To begin addressing AI-related cybersecurity concerns, the US Department of Commerce's National Institute of Standards and Technology (NIST) published guidance on January 4, 2024 that identified four specific types of cyberattacks and offered ways for companies to prevent or mitigate the impact of those attacks. This LawFlash focuses on that guidance in addition to the cybersecurity and incident reporting obligations under the EU's new AI Act, which comes into force on August 1, 2024. The final text of the act was published on July 12, 2024.

THE US EXECUTIVE ORDER AND THE NIST AI FRAMEWORK

On October 30, 2023, President Joseph Biden issued an executive order to establish standards around the use of AI.[1] Titled Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, the order, among other things, directed NIST to develop guidelines and best practices for companies to follow to mitigate or avoid cyberattacks, with the stated goal of promoting “consensus industry standards” to protect AI systems.

In early 2024, NIST published Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, which identified four primary areas of adversarial machine learning (ML) attacks: (1) data “poisoning” attacks, (2) data “abuse” attacks, (3) privacy attacks, and (4) “evasion” attacks.[2] As NIST noted, “[t]he spectrum of effective attacks against ML is wide, rapidly evolving, and covers all phases of the ML lifecycle,” but these four attacks are currently prevalent and important to understand.

- **Data poisoning attacks:** These attacks occur during the training stage of the ML algorithm, whereby the attacker corrupts the data that is input into the algorithm so that the AI model learns from “poisoned” data. NIST described poisoning attacks as “the most critical vulnerability” to ML systems.
- **Data abuse attacks:** These attacks involve the provision of incorrect information from a legitimate but compromised source. In these attacks, an attacker may repurpose a system’s intended use to achieve their own objectives by using indirect prompt injection. For example, search chatbots on an organization’s website can be prompted to generate disinformation or hide certain information or sources from users such that the responses are inaccurate or incomplete.
- **Privacy attacks:** These attacks involve reverse engineering information from the AI model to learn sensitive information about individual users, which if successful, may have privacy implications depending on the type of information revealed. These attacks may also involve reverse engineering data to learn about sensitive critical architecture, which could lead to the reconstruction of the AI model for illegitimate purposes.
- **Evasion attacks:** Once an AI model is deployed, an attacker may change the model’s inputs to alter its response. This can include digitally modifying an image so that it is not successfully identified or physically modifying an objection. An example of the latter would be an attacker modifying a driverless vehicle’s road sign detection classifier by adding black and white stickers to a stop sign, leading the vehicle to ignore the sign and not stop.

Each of these attacks may be easy to mount. NIST cautions that poisoning attacks, for example, can be mounted by controlling a few dozen training samples—a small percentage of an entire training dataset. Going forward, companies involved in developing or implementing AI systems may wish to consider monitoring these types of cyberattacks, as well as novel attacks that arise, as cyberattacks in the AI and ML space are continually evolving.

NEW EU AI ACT: CYBERSECURITY AND INCIDENT REPORTING OBLIGATIONS

Reflecting the seriousness of the risks identified by NIST, the EU AI Act specifically acknowledges certain of these risks—including “data poisoning,” “model evasion,” and “adversarial” attacks (recital 77; article 15(5))—and the consequences of these risks arising (recital 110). This could include the loss of human control, interferences with critical infrastructure, disinformation, harmful bias and discrimination, and societal risks. In turn, like the EU General Data Protection Regulation (GDPR), the AI Act imposes cybersecurity and incident reporting obligations; these obligations are different and run in parallel to the GDPR and EU sector-specific laws.

Cybersecurity Obligations Under the AI Act

Notably, the AI Act requires “providers” of “high-risk” “AI systems” and “General Purpose AI” (GPAI) models to implement security and resilience measures, including as described below:

- **High-risk AI systems:** Notably, high-risk AI systems are to be “designed and developed” in such a way that it achieves “an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle” (article 15(1)). The European Commission is required to “encourage” the development (in cooperation with “relevant stakeholders and organisations”) “benchmarks and measurement methodologies” relative to these requirements (article 15(2)).
 - High-risk systems are also to be as “resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates,” including in “interactions with natural persons or other systems” (article 15(4)) and “against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities” (article 15(5)). Providers of high-risk AI systems are also required to disclose to “deployers” of such systems the “characteristics, capabilities and limitations” of these systems, including with respect to “any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity” (article 13(3)(b)(ii)).
- **GPAI models:** The act recognizes that cyber threats to GPAI models with “systemic risks” include “accidental model leakage, unauthorised releases, circumvention of safety measures, and defence against cyberattacks, unauthorised access or model theft” (recital 115). In turn, providers of such GPAI models are to “ensure an adequate level of cybersecurity protection” and the “the physical infrastructure of the model” (article 55(1)(d)). Notably, cybersecurity obligations are applicable to such providers “regardless of whether [the model] is provided as a standalone model or embedded in an AI system or a product” (recital 114). The act also states that “it should require” providers of such GPAI models to “perform the necessary model evaluations” to document “adversarial testing” and “continuously assess and mitigate systemic risks,” including “risk management policies” (recital 114).

Incident Reporting Obligations Under the AI Act

The AI Act also imposes incident reporting obligations on both providers and “deployers” of AI systems and GPAI models, even in certain circumstances where AI systems are being tested (article 60(7)). Notably, providers (and in certain circumstances, deployers) of high-risk AI systems and GPAI models (which present systemic risks) must report “serious incidents” to the appropriate governmental authorities, and in certain circumstances, relevant participants in the AI chain (article 55(1)(c); 73; 26(5)). Serious incidents may include death or serious harm to a person, serious and irreversible disruption to critical infrastructure, serious harm to property or the environment, or infringement of fundamental rights laws (article 3(49)).

Importantly, the timeframes for reporting incidents under the EU AI Act are *tight*, even relative to those under the GDPR. However, the timeframe will depend on the circumstances. For example, if a causal link is established between the AI system and the serious incident, the incident must be reported *immediately*.

Significantly, the act contains specific time limits on reporting timeframes that depend on the seriousness and impact of the incident. For example, a serious incident with “widespread infringement” (which could involve cross-border or critical infrastructure impacts) must be reported “immediately” but not later than *two days* following “awareness” (article 73(3)). Similar to reporting “personal data breaches” under the GDPR, the initial report may be “incomplete” and thereafter followed by a “complete report” (article 73(5)).

KEY TAKEAWAYS: MITIGATING TECHNICAL, REGULATORY, AND LEGAL RISKS

While, according to NIST, defenses to cyberattacks on AI models are “incomplete at best,”^[3] organizations may wish to consider taking—in addition to their existing information security plans and policies—measures such as the following:

- **Maintain control over training datasets:** When an AI model uses datasets outside the organization’s control, the quality of the inputs may suffer and could increase the risk of the data being poisoned or corrupted. Depending on the circumstances, an organization may be able to better track the origin of internally sourced datasets and sanitize it more effectively. To the extent a third-party dataset is used for AI training, an organization may wish to consider performing appropriate due diligence.
- **Regularly test the AI models:** Dataset integrity may be maintained by regularly testing and training AI models, ideally through human feedback rather than AI. Such training may mitigate poisoning and evasion attacks.
- **Maintain a strong information governance and security policies and plan:** Routinely documenting the processes, analysis, and decisions made for the AI systems. Even if an attack occurs, this documentation may help demonstrate the organization took appropriate steps to safeguard its information and AI systems.

The EU AI Act also sets out (illustratively) information security measures which may be undertaken. For example, the AI Act suggests with respect to GPAI models with systemic risks “securing model weights, algorithms, servers, and data sets, such as through operational security measures for information security, specific cybersecurity policies, adequate technical and established solutions, and cyber and physical access controls, appropriate to the relevant circumstances and the risks involved” (recital 115).

Organizations may wish to carefully consider the regulatory and legal risks—which could include both regulatory enforcement and private litigation—from successful and unsuccessful cyberattacks to AI systems and models, including those that arise from:

- Product liability, data privacy, cybersecurity, and other regulatory laws (such as the AI Act and the GDPR)
- Contractual and tortious liability
- Intellectual property risks (such as trade secret and copyright-related risk)

In addition, organizations may be subject to notable AI-related cyber incident reporting obligations arising from longstanding laws like the GDPR and newer AI laws. In turn, organizations may need to update their existing information security and incident response plans, and conduct AI-focused cybersecurity “tabletop exercises” to reflect the unique cybersecurity risks relating to AI systems and models.

HOW WE CAN HELP

Morgan Lewis Lawyers are well suited to help companies navigate AI-related enforcement and litigation matters in the European Union and United States. Our team stands ready to assist companies designing, developing, or using AI navigating this evolving and challenging cyber threat landscape.

CONTACTS

If you have any questions or would like more information on the issues discussed in this LawFlash, please contact any of the following:

Authors

Elizabeth B. Herrington
Vishnu Shankar
Phillip J. Wiese

Boston

Doneld G. Shelkey

Brussels

Vishnu Shankar

Chicago

Elizabeth B. Herrington

Los Angeles

Megan A. Suehiro

London

Vishnu Shankar
Chris Warren-Smith
Mike Pierides

Munich

Daja Apetz-Dreier

Philadelphia

Ezra D. Church
Kristin M. Hadgis

Princeton

Gregory T. Parks

Silicon Valley

Dion M. Bregman
Andrew J. Gray IV

Washington, DC

Dr. Axel Spies

[1] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Oct. 30, 2023).

[2] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, National Institute of Standards and Technology (Jan. 4, 2024).

[3] NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems (Jan. 4, 2024).