LATHAM&WATKINS

Client Alert
Commentary

Latham & Watkins Artificial Intelligence Practice May 10, 2024 | Number 3263

US AI Standards Take Shape: Key 180-Day Developments Following White House AI Executive Order

Companies should consider how new AI risk standards may align to their operations and whether to comment on the draft standards to shape their development.

On April 29, 2024, the White House announced that several federal agencies have issued draft AI standards and rules per their requirement to do so within 180 days after the White House's October 30, 2023, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (the Executive Order). These standards and rules will impact organizations that use AI, particularly providers of generative AI (GenAI) tools and healthcare providers — though much remains in draft form and open to public comment.

As we analyzed in our Client Alert President Biden's Executive Order on Artificial Intelligence — Initial Analysis of Private Sector Implications, the Executive Order marked the beginning of an ambitious, whole-of-government strategy to regulate AI in the United States while fostering innovation. The Executive Order focuses on eight guiding principles, including, as relevant here: ensuring the safety and security of AI technology; protecting consumers, patients, passengers, and students; advancing equity and civil rights; promoting innovation and competition; and protecting privacy. The Executive Order calls on federal agencies, such as the Federal Trade Commission (FTC) and Department of Homeland Security (DHS) to aggressively regulate the use of AI in their respective areas of authority. It also directs the National Institute of Standards and Technology (NIST) and other agencies to promulgate standards for AI risk management.

Below are key 180-day developments:

- NIST issued draft guidance on GenAI risk management, AI synthetic content risks, and AI secure coding, which is expected to be finalized by the end of July.

- DHS issued guidelines for the use of AI by critical infrastructure owners and operators, which DHS can make mandatory beginning at the end of June.

- DHS established an AI safety and security board.

- The Department of Health and Human Services (HHS) and the Centers for Medicare & Medicaid Services (CMS) issued a final rule requiring covered entities to prevent discriminatory use of AI in health services, which will go into effect on May 1, 2025.

- The Patent and Trademark Office (USPTO) is soliciting comments on draft standards for the application of patent protection to AI-generated material and is expected to issue further guidance and recommend further Executive Orders on copyright and AI by the end of June.

These developments represent one of the major milestones in implementation of the Executive Order, though further implementation steps remain. Other than a rule for healthcare providers regarding discrimination in the use of AI, the standards and rules discussed here are not mandatory — but they can inform regulator expectations for reasonable controls. For example, the FTC touts the NIST Cybersecurity Framework as consistent with the FTC's data security regulation. As the FTC and other agencies increasingly investigate and regulate the use of AI, companies should consider how applicable standards align to their operations and whether to comment on draft standards to shape their development.

In addition, most of these developments are not final, with the exception of the guidelines for the use of AI by critical infrastructure owners and operators, the healthcare final rule, and the announcement of DHS's Artificial Intelligence Safety and Security Board (which is separate from boards the Executive Order required various agencies to create to manage each agency's use of AI). Otherwise, these developments are drafts that are open to public comment (deadlines are noted below). This comment period provides a meaningful opportunity for companies to engage as the pace of AI regulation quickens.

The next milestone dates will occur at the 240-day and 270-day marks (June 26 and July 26). Within 240 days, the Department of Commerce (DOC) is expected to issue a complementary report to the draft report issued in the first 180 days (and described below) on reducing synthetic content risks. The complementary report expected in June will identify existing, and potential, standards, tools, methods, and practices for: authenticating content; labeling and detecting synthetic content; and preventing certain uses of generative AI, e.g., non-consensual intimate imagery (NCII). Within 240 days, DHS is expected to make critical infrastructure guidelines mandatory, and the National Science Foundation will engage other agencies to develop privacy-enhancing technology for AI. The 270-day mark will see a number of further developments, including, as relevant here, finalization of some NIST standards and further guidance from the USPTO on the patentability of AI and recommendations for executive actions on copyright and AI.

NIST Guidance on GenAI Risk, Synthetic Content, and Secure Coding

NIST published three standards pursuant to the Executive Order that relate to risk management in AI development. These include standards for developing a GenAI risk management framework (NIST AI 600-1); guidance on reducing risks posed by synthetic content (NIST AI 100-4); and standards for secure software development (Special Publication (SP) 800-218A). The three publications reinforce the US government's focus on two key AI risk vectors: (1) the risks from generative outputs; and (2) security risks in model and product development.

GenAI Risk Management (NIST AI 600-1)

On April 29, 2024, NIST published an initial public draft of *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1) This GenAI profile complements NIST's *Artificial Intelligence Risk Management Framework* (NIST AI 100-1) published in January 2023. The January 2023 framework is designed to guide organizations designing, developing, deploying, or using AI systems to help manage AI risks and promote trustworthy and responsible development and use. The GenAI profile goes deeper, setting out 12 core risks specifically posed by GenAI products that an appropriate risk management framework should take into account.

The 12 core risks include:

1. Chemical, Biological, Radiological, and Nuclear Information
2. Confabulation (confidently stated but erroneous or false content, sometimes called "hallucinations")
3. Dangerous or Violent Recommendations
4. Data Privacy
5. Environmental (e.g., resource consumption for model development)
6. Human-AI Configuration
7. Information Integrity
8. Information Security
9. Intellectual Property
10. Obscene, Degrading, and/or Abusive Content
11. Toxicity, Bias, and Homogenization
12. Value Chain and Component Integration

NIST AI 600-1 describes these GenAI risks and then provides over 400 actions, organized across four categories, that organizations can take to manage risk. While not every action will make sense for every organization, the categories, described below, substantially overlap with regulator expectations and legal requirements. For example, the NIST standards emphasize AI governance, which is a central tenet of the EU AI Act, and managing bias, a subject of recent FTC enforcement.

- **Governance:** Including undertaking risk assessments for new AI models or products.
- **Mapping:** Including deploying fact-checking techniques to verify the accuracy of output.
- **Measuring:** For example, conducting security assessments and audits to measure the integrity of training data.
- **Managing:** Such as implementing incident response and recovery plans.

How to comment: NIST has put out a request for comments on NIST AI 600-1, which are due by June 2, 2024, via www.regulations.gov.

Synthetic Content (NIST AI 100-4)

NIST released for public comment a draft publication of *Reducing Risks Posed by Synthetic Content* (NIST AI 100-4). The publication is intended to inform a report by DOC at the end of June that will discuss processes for content authentication and synthetic content labeling, detection, and prevention. This DOC report is required under Section 4.5(a) of the Executive Order. The NIST publication surveys existing and foreseeable approaches to tracking data provenance and detecting content that has been "significantly altered or generated by" AI. The publication also identifies research opportunities to help address these kinds of data-transparency challenges. The NIST publication can help companies consider reasonable strategies for disclosing and detecting synthetic content. Companies may need to take some of these steps to comply with the EU AI Act and recently proposed FTC rules on impersonation.

The NIST publication examines how two overarching strategies can improve digital content transparency:

- **Provenance data tracking:** Digital watermarking and/or metadata that records the content's origin and history, which helps content recipients to establish the authenticity, integrity, and credibility of the content. Insights into data provenance can establish ownership or consent and allow content to be traced back to a particular system or source.

- **Synthetic content detection:** Systems designed to classify synthetic data based on the properties of the data.

The publication assesses how these strategies may be used, with varying degrees of efficacy, to mitigate harms across various use cases and modalities, including mitigating against mis- and disinformation, fraud/extortion, and distribution of AI-generated child sexual abuse material (CSAM) and NCII. But the publication notes that companies should not solely rely on these strategies. Effectively mitigating instances of CSAM, for example, will likely require a combination of measures occurring across the AI lifecycle, including dataset filtration, prompt filters, and output filters, in addition to provenance-tracking techniques.

The publication identifies a number of challenges to implementing these strategies, including:

- **Product utility:** The more robust a watermark, the more likely it is to distort the content to which it is applied and be perceptible to the naked eye. The publication also observes that mitigations like filtering mechanisms and assigning unique numeric identifiers to alleged NCII may be overinclusive and unduly impede certain use cases.

- **Accessibility:** Passive disclosures (i.e., disclosures not readily perceptible by the senses, such as single-pixel watermarking methods) may require special skills or tools to detect. As a result, the average person may not be able to easily discern the provenance of AI-generated information.

- **Adoption:** The publication observes that there must be some degree of standardization to provenance conventions so people can know how to check the provenance of a given piece of content. However, the more a protocol becomes known, the more likely it is that bad actors can figure out how to remove or alter it.

- **Privacy:** Metadata containing provenance information may also include private information generated by users, and this data could be exploited by bad actors. Thus, developers of systems that host metadata may need to implement privacy controls to prevent the retention of private information in metadata.

The publication contains appendices of resources to help developers and red teamers (i.e., persons who conduct adversarial testing on AI systems) to integrate provenance data tracking and synthetic content detection into their product safety protocols. For instance, Appendix B links to technical tools related to digital content transparency, such as classifiers for detecting synthetic content or CSAM, and Appendix D contains a list of content detection datasets.

How to comment: NIST requests comments on the completeness and clarity of the publication. Submission instructions are provided on page 3 of the publication. Comments must be received by June 2, 2024.

Secure Software Development (Special Publication (SP) 800-218A)

NIST published an initial public draft of *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models* (Special Publication (SP) 800-218A), which is designed to be an AI-focused companion to *Secure Software Development Framework (SSDF)* (Special Publication (SP) 800-218), which was published in February 2022. While SP 800-218 generally describes a set of fundamental practices for secure software development, SP 800-218A addresses AI model development, which includes data sourcing for designing, training, fine-tuning, and evaluating AI models, as well as incorporating and integrating AI models into other software.

The guidance is designed for three audiences and echoes the classification of entities under the EU AI Act:

- **AI model producers:** Organizations that are developing their own GenAI and dual-use foundation models.
- **AI system producers:** Organizations that are developing software that leverages a GenAI or dual-use foundation model.
- **AI system acquirers:** Organizations that are acquiring a product or service that utilizes one or more AI systems.

In relation to these parties, SP 800-218A identifies a number of risks. For example:

- Training datasets may be acquired from unknown, untrusted sources. Model weights and other training parameters can be susceptible to malicious tampering.
- Some models may be too complex to thoroughly inspect with ease, potentially allowing for undetectable execution of arbitrary code.
- User queries can be crafted to produce undesirable or objectionable output and — if not sanitized properly — can be leveraged for injection-style attacks.

The document then identifies practices to address these novel risks. SP 800-218A contains nearly nine pages of recommended practices, grouped into the following categories:

- **Prepare the organization:** Including defining security requirements for software development.
- **Produce well-secured software:** Such as designing software to meet security risks.
- **Protect software:** For example, protecting all forms of code from unauthorized access and tampering.
- **Respond to vulnerabilities:** Including identifying and confirming vulnerabilities on an ongoing basis.

How to comment: NIST has put out a request for comments on SP 800-218A, which are due by June 2, 2024. Instructions for submitting comments are provided in the document’s front matter (PDF page 4).

DHS Guidance on AI Risk for Critical Infrastructure

The Executive Order required the Secretary of DHS to publish safety and security guidelines for use by critical infrastructure owners and operators. On April 29, 2024, DHS published these guidelines, as well as a corresponding press release. While these guidelines are not binding, the Assistant to the President for National Security Affairs and Director of the Office of Management and Budget must, within 240 days of their publication, mandate the guidelines, or portions of the guidelines, through regulatory or other appropriate action. Similarly, independent regulatory agencies are encouraged to mandate the guidelines through regulatory action in their areas of authority.

The guidelines are organized around three overarching categories of system-level risk:

- **Attacks using AI:** The use of AI to enhance, plan, or scale physical attacks on, or cyber compromises of, critical infrastructure.
- **Attacks targeting AI systems:** Targeted attacks on AI systems supporting critical infrastructure.
- **Failures in AI design and implementation:** Deficiencies or inadequacies in the planning, structure, implementation, or execution of an AI tool or system leading to malfunctions or other unintended consequences that affect critical infrastructure operations.

To address these risks, DHS outlines a four-part mitigation strategy, building on NIST's AI Risk Management Framework, discussed above. Critical infrastructure owners and users should consider these mitigation strategies, as described below, for their risk management framework:

- **Govern:** Establish an organizational culture of AI risk management. Prioritize and take ownership of safety and security outcomes, embrace radical transparency, and build organizational structures that make security a top business priority.
- **Map:** Understand your individual AI use context and risk profile. Establish and understand the foundational context from which AI risks can be evaluated and mitigated.
- **Measure:** Develop systems to assess, analyze, and track AI risks. Identify repeatable methods and metrics for measuring and monitoring AI risks and impacts.
- **Manage:** Prioritize and act upon AI risks to safety and security. Implement and maintain identified risk management controls to maximize the benefits of AI systems while decreasing the likelihood of harmful safety and security impacts.

AI Safety and Security Board

DHS announced the establishment of an Artificial Intelligence Safety and Security Board (the Board). The Board's duties include advising the Secretary of DHS, the critical infrastructure community, other private-sector stakeholders, and the broader public on the safe and secure development and deployment of AI technology in critical infrastructure. The Board is separate from the internal AI governance boards that Section 10.1(b)(iii) of the Executive Order required various agencies, including DOC and DHS, to create.

The Board will develop recommendations to help critical infrastructure stakeholders, such as transportation service providers, pipeline and power grid operators, and internet service providers, to more responsibly leverage AI technologies. It will also develop recommendations to prevent and prepare for AI-related disruptions to critical services that impact national or economic security, public health, or safety. The Board includes notable figures from leading AI companies.

HHS Rule Prohibiting AI Discrimination in Healthcare

HHS and CMS issued a final rule under Section 1557 of the Affordable Care Act advancing protections against discrimination in healthcare. The rule is intended to limit discrimination within the healthcare industry. It specifically addresses the directives of the Executive Order by extending discrimination prohibitions to the use of AI-powered patient care decision-support tools, which the rule defines as those that are used to guide healthcare decision-making that affects the care that patients receive.

The rule has three key AI requirements:

- **Prohibition on discrimination via AI:** Healthcare providers are not permitted to use AI tools to discriminate against patients on the basis of race, color, national origin, sex, age, or disability.
- **Inventory of AI tools:** Healthcare providers are required to inventory their tools that employ input variables or factors that measure race, color, national origin, sex, age, or disability.
- **Risk mitigation:** A covered entity must make reasonable efforts to mitigate the risk of discrimination resulting from an AI tool's use in its health programs or activities.

Notably, the rule does not require providers to disclose the use of AI tools to patients, though HHS recommends that providers do so in commentary to the rule. These AI rule provisions will be enforceable beginning on May 1, 2025 (360 days after publication in the *Federal Register* on May 6, 2024).

USPTO Comment Solicitation on Patentability

On April 30, 2024, USPTO published a request for public comment (RFC) on three topics: (1) the impact of AI on prior art; (2) the impact of AI on a PHOSITA (a person having ordinary skill in the art); and (3) the implications of AI that could require updating examination guidance and/or legislative change. The USPTO garnered these topics, and the specific questions surrounding them, from years of dialogue with the innovation community and AI experts, particularly with the AI and Emerging Technologies Partnership — an ongoing cooperative effort between the USPTO and academia, independent inventors, small businesses, industry, other government agencies, nonprofits, and civil society interested in AI's effect on intellectual property.

The RFC observed that key AI issues in these areas include the following:

- **Prior art**
  - Does 35 U.S.C. § 102 presume or require that a prior art disclosure be authored/published by humans?
  - Does non-human authorship of a disclosure affect its availability as prior art under 35 U.S.C. § 102? For instance, at what point, if ever, could the volume of AI-generated prior art create an undue barrier to the patentability of inventions?
  - Should parties be required to disclose to the USPTO if they are aware that a prior art disclosure was AI-generated?
  - How does the fact that a disclosure is AI-generated impact other prior art considerations, such as operability, enablement, and public accessibility?
- **PHOSITA**
  - Does the term "person" in the PHOSITA assessment presume or require that the "person" is a natural person, i.e., a human?
  - How might the availability of AI as a tool affect the level of skill of a PHOSITA or what is deemed to be common knowledge in the art?

  - How does the availability of AI to a PHOSITA impact the evaluation of objective indicia of obviousness?

- **Implications requiring guidance or legislative change**
  - What guidance concerning the impact of AI on prior art and on the knowledge of a PHOSITA would be helpful?
  - Should patent law be amended to account for any of these considerations?
  - Are there any helpful laws or practices in other countries concerning these issues that could be adapted to fit within US patent law?

How to comment: Comments may be submitted here. The comment period will close on July 29, 2024.

Other Notable 180-Day Developments

Beyond the initiatives covered above, federal agencies have taken various other actions required by the Executive Order, including:

- NIST publishing an initial draft of NIST AI 100-5, "A Plan for Global Engagement on AI Standards," which calls for a coordinated effort to work with key international allies and partners on common AI definitions and standards.
- Establishing a framework for nucleic acid synthesis screening to help prevent the misuse of AI for engineering dangerous biological materials.
- Piloting new AI tools for identifying vulnerabilities in vital government software systems.
- Developing bedrock principles and practices for employers and developers to build and deploy AI safely and in ways that empower workers.
- Releasing guidance to assist federal contractors and employers comply with worker-protection laws as they deploy AI in the workplace.
- Releasing resources for job seekers, workers, and tech vendors and creators on how AI use could violate employment discrimination laws.
- Issuing guidance on AI's nondiscriminatory use in the housing sector.
- Publishing guidance and principles that set guardrails for the responsible and equitable use of AI in administering public benefits programs.
- Developing a strategy for ensuring the safety and effectiveness of AI deployed in the healthcare sector.
- Several government measures related to using AI for public good, for example, Department of Education funding opportunities.
- Several initiatives to bring employees with AI experience into government.

Practical Takeaways

Spurred by the Executive Order, federal agencies have taken concrete steps to establish standards and expectations for the use of AI. These developments have several important implications:

- Organizations should carefully study the standards and expectations envisioned by NIST and others because they echo emerging legal requirements in the US and EU and, in any event, inform regulator expectations.
- Entities regulated by HHS should carefully review their AI systems to ensure compliance with the final rule on discrimination.
- Critical infrastructure providers should carefully review the DHS guidelines on AI risk management and track how these rules are codified as legal requirements through June 26, 2024 (the 240-day period to complete this work).
- Organizations should consider commenting on draft guidelines to help shape these expectations before the comment periods close this summer.

If you have questions about this Client Alert, please contact one of the authors listed below or the Latham lawyer with whom you normally consult:

Heather B. Deixler
heather.deixler@lw.com
+1.415.395.8110
San Francisco

Michael H. Rubin
michael.rubin@lw.com
+1.415.395.8154
San Francisco

Jamie D. Underwood
jamie.underwood@lw.com
+1.202.637.3365
Washington, D.C.

Max G. Mazzelli
max.mazzelli@lw.com
+1.415.395.8040
San Francisco

Molly O'Malley Clarke
molly.clarke@lw.com
+1.213.891.8935
Los Angeles

Will Schildknecht
will.schildknecht@lw.com
+1.213.891.8136
Los Angeles

You Might Also Be Interested In

Webcast: Artificial Intelligence — Legislative and Regulatory Updates and Their Implications

FTC Sharpens Its AI Agenda With Novel Impersonation Rulemaking

USPTO Releases Guidance on AI and Inventorship

President Biden's Executive Order on Artificial Intelligence — Initial Analysis of Private Sector Implications

Client Alert is published by Latham & Watkins as a news reporting service to clients and other friends. The information contained in this publication should not be construed as legal advice. Should further analysis or explanation of the subject matter be required, please contact the lawyer with whom you normally consult. The invitation to contact is not a solicitation for legal work under the laws of any jurisdiction in which Latham lawyers are not authorized to practice. A complete list of Latham's Client Alerts can be found at www.lw.com. If you wish to update your contact details or customize the information you receive from Latham, visit our subscriber page.