Baker McKenzie.

Client Alert

December 2024

United States: AI Safety Institute releases its first synthetic content guidance report (NIST AI 100-4)

18 December 2024 (seven-minute read)

In brief

The US Artificial Intelligence Safety Institute (AISI), housed within the National Institute of Standards and Technology (NIST), announced on 20 November 2024 the release of its first synthetic content guidance report, NIST AI 100-4 Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency ("**NIST AI 100-4**"). "Synthetic content" is defined in President Biden's Executive Order on Safe, Secure, and Trustworthy AI ("**EO 14110**") as "information, such as images, videos, audio clips, and text, that has been significantly altered or generated by algorithms, including by AI."

Contents

NIST AI 100-4 examines the existing standards, tools, methods and practices, as well as the potential development of further science-backed standards and techniques to help manage and reduce risks related to synthetic content by: 1) recording and revealing the provenance of content, including its source and history of changes made to the content; 2) providing tools to label and identify AI-generated content; and 3) mitigating the production and dissemination of AI-generated child sex abuse materials ("**AIG-CSAM**") and non-consensual intimate imagery ("**AIG-NCII**") of real individuals. It reflects public feedback and consultations with diverse stakeholders who responded to NIST's Request for Information on 21 December 2023. Although compliance is voluntary, NIST AI 100-4 is expected to inform industry best practices for managing synthetic content risks.

Background

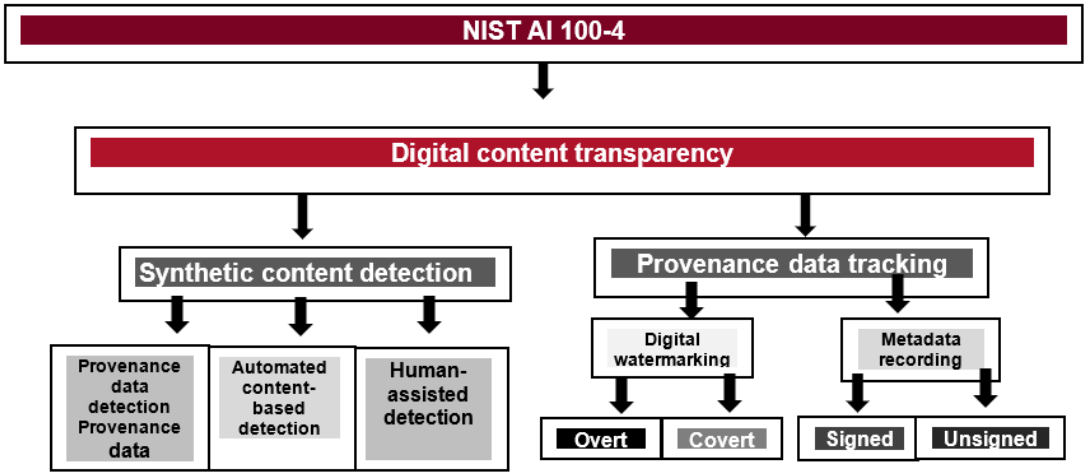
On 30 October 2023, President Biden signed EO 14110, establishing a government-wide effort to encourage responsible AI development and deployment through federal agency leadership, regulation of industry, and engagement with international partners. Specifically, it established the AISI to advance the science of AI safety and address the risks posed by advanced AI systems. The AISI is tasked with developing testing, evaluations and guidelines that will help accelerate safe AI innovation in the US and around the world through the International Network of AI Safety Institutes ("**Network**").

The US Department of Commerce and US Department of State co-launched the Network as part of a global effort to advance the science of AI safety and enable cooperation on research, best practices and evaluation to harness the benefits of AI and avoid a patchwork of global governance. AISI will serve as the inaugural chair of the Network, whose other initial members include Australia, Canada, the European Union, France, Japan, Kenya, the Republic of Korea, Singapore and the United Kingdom. During the Network's inaugural meeting in San Francisco on 20 November 2024, technical experts from member governments joined leading AI developers, academics, civil society leaders and scientists from non-Network governments to discuss key areas of collaboration on AI safety and lend their technical and scientific expertise to the Network's mission. NIST AI 100-4 was published on the same day as part of the Network's global research initiative.

NIST AI 100-4

NIST AI 100-4's main goal is to identify a series of voluntary approaches to address risk from AI-generated content, such as impersonation and fraud. To that end, NIST AI 100-4 outlines state-of-the-art methods for detecting and tracking synthetic media generated by AI:

1. **Synthetic content detection**: Synthetic content detection includes techniques, methods and tools to classify whether or not a given piece of content is synthetic. This detection involves ascertaining the existence of recorded provenance information such as metadata, digital watermarks or other characteristics to help determine whether content was generated, modified or manipulated by AI.
    a. **Provenance data detection**: Provenance data detection involves looking for origin and lifecycle journey data that have been tracked via digital watermarks (either overt or covert) or metadata. Provenance data helps with detection primarily by describing the source or history of the content, including modifications. If provenance data becomes widespread, the absence of provenance data in a situation where it would be expected could itself arouse suspicion.
    b. **Automated content-based detection**: Automated content-based detection techniques identify synthetic content after it has been generated or modified based on traces left during generation or processing (e.g., image pixel irregularities or inconsistencies). It can also look for traces that manipulation leaves in the file metadata.
    c. **Human-assisted detection**: Human-assisted detection refers to human-in-the-loop methods where individuals, including crowd-sourced workers, data labelers, and/or domain experts, augment or supplement automated tools to help identify synthetic content, such as by assessing or validating detection model outputs. The effectiveness of these methods is subject to the domain and individuals' expertise. Humans may find detection more difficult as synthetic content generation continues to increase in sophistication.
2. **Provenance data tracking**: Current methods for provenance data tracking can help establish the authenticity, integrity and credibility of digital content by recording information about the content's origins and history.
    a. **Digital watermarking (overt and covert)**: Digital watermarking entails embedding digital provenance data into the image, text, audio or video content, making it difficult to remove. Digital watermarks can indicate the material's content origins, ownership details and timestamps.
    b. **Metadata recording (signed and unsigned)**: Metadata refers to data that may indicate the content's origin, time and date of creation, author, ownership, location of creation, and editing history. This information is typically embedded with the data it describes or stored in an external repository and linked to the content via some form of identifier.



Given the prominent harms and ease of creation of AIG-CSAM and AIG-NCII, NIST AI 100-4 takes the opportunity to focus in on these types of content. The document outlines emerging best practices and potential mitigations for tackling both AIG-CSAM and AIG-NCII, but we note that these recommendations would be well applied to reducing risks associated with other types of harmful synthetic content. Recommended best practices include the following:

1. **Training data filtering**: Generative AI models are more likely to create harmful content like AIG-CSAM or AIG-NCII if such images are included in their training data. To prevent this, entities should clean datasets before training a model to reduce the risk of the model generating problematic content. For example, this could include training machine learning-based "safety classifiers" using vetted CSAM and NCII to determine precision and recall rates to identify and remove harmful content in training data. Alternatively, entities could filter out content from websites that are known to host unwanted content or by curating datasets of malicious website links. Lastly, entities could reduce harmful content generation by training models only on vetted data such as licensed, curated stock images.

2. **Input data filtering:** Entities can filter inputted data to block AIG-CSAM or other harmful content that a user is intentionally attempting to generate. One method is through text-to-image models that prompt users to input to a deployed model that can be filtered to prevent the generation of potentially harmful images. Second, entities can provide input filtering classifiers that are trained to classify text or images into distinct categories of violative content that a user may try to import or prompt for. Third, entities can use keyword filters to prevent generation of images. This approach would identify known egregious content such as commonly known CSAM technology. Lastly, entities can use warning messages within a product to redirect a user looking for harmful content.
3. **Output filtering:** Entities can use output filtering by blocking generated images deemed harmful or violative. Output filters' effectiveness depend on the following:
    a. How well the training data covers a wide range of sexual content
    b. How well training data is labeled, particularly for ambiguous cases
    c. How confidence thresholds for blocking are set
    d. How changes in malicious actors' behavior are incorporated
4. **Hashing:** Entities can use two types of hashing to prevent harmful content like AIG-CSAM and AIG-NCII: cryptographic hashes and perceptual hash algorithms. Cryptographic hashes are designed in a way that a slight alteration to the input data would produce a vastly different cryptographic hash. Cryptographic hashes allow for identification of exact matches to a CSAM or other image that has been logged in a hash-sharing database. Perceptual hash algorithms attempt to output similar hashes for input files that humans perceive as similar. The hash value stays the same if the content is not significantly changed, such as in compression, brightness, orientation or color.
5. **Provenance data tracking techniques:** Entities can use provenance data tracking techniques such as digital watermarks and metadata recording to reduce synthetic content harms. Provenance data tracking can help identify and track synthetic content like AIG-CSAM and AIG-NCII, distinguish authentic imagery that shows real victims, and understand patterns of misuse of generative AI tools.
6. **Red-teaming and testing:** EO 14110 refers to red-teaming as a "structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI." Entities can do this by scanning the internet and internal systems for known prompts used in attempts to generate harmful content like AIG-CSAM and AIG-NCII and seeing how systems respond to these prompts. AI tool developers can then develop initial assessments of a model's propensity to generate harmful content and build upon it, aiming to prevent queries that yield illegal, abusive or otherwise unwanted content from doing so in future iterations of the model or system. Furthermore, "red teams" could be paired with "blue teams," who work on building defensive measures to prevent and/or address misuse. An established protocol or set of guidelines for red-teaming for legal and reputational risks could assist with measuring models' safety with respect to synthetic content.

Each digital content transparency approach holds the promise of improving trust by clearly and readily indicating where AI techniques have been used to generate or modify digital content. However, none of these techniques is a standalone solution. For successful digital content transparency, social efforts and initiatives must be applied alongside provenance data tracking, synthetic content detection approaches, and technical methods for preventing and reducing harms from AIG-CSAM and AIG-NCII. NIST AI 100-4 is a resource to promote understanding and helps lay the groundwork for the development of additional, improved technical approaches, such as science-backed global standards, to advancing synthetic content provenance, detection, labeling and authentication.

What's next

NIST AI 100-4 highlights that coordinated research and testing of advanced AI models across critical national security and public safety domains, such as radiological and nuclear security, chemical and biological security, cybersecurity, critical infrastructure, and conventional military capabilities, are needed. Ahead of the inaugural meeting, the Network has already announced USD 11 million in global research funding to: 1) understand the security and robustness of current digital content transparency techniques; 2) explore novel digital content transparency methods; and 3) develop model safeguards to prevent the generation and distribution of harmful synthetic content. The research funding is expected to be made available through national AI safety institutes (like the AISI), other national science agencies, and philanthropic foundations in the private sector.

Less than a week before NIST AI 100-4's release, DHS debuted the first-of-its-kind AI safety Framework for critical infrastructure. Between the Framework, NIST's Artificial Intelligence Risk Management Framework: Generative AI Profile publication and this latest NIST AI publication, the US continues to flesh out voluntary AI safety and governance standards as well as actively promote cross-border engagement to facilitate broad adoption.

NIST AI 100-4 and the AISI were developed as a result of President Biden's EO 14110, and President-elect Trump has indicated he may repeal EO 14110. As a result, it is uncertain whether AISI will survive under the incoming administration, and in light of President-elect Trump's desire to cut the federal budget, it remains to be seen whether US AI research funding will continue at the planned rate. However, given the level of international involvement in the Network's AI research initiatives, it seems likely that the work will continue at a global level, even if no longer driven from the US.

Organizations are encouraged to review the learnings set out in NIST AI 100-4, map the recommendations to their existing AI risk management processes, and consider implementing any new or missing elements. If you have any questions regarding compliance or tailoring your company's AI, privacy or cybersecurity governance program, please contact your Baker McKenzie attorney or the authors below.

Contact Us

Adam Aft
Partner
adam.aft
@bakermckenzie.com

Keo McKenzie
Partner
keo.mckenzie
@bakermckenzie.com

Cristina Messerschmidt
Associate
cristina.messerschmidt
@bakermckenzie.com

Mercedes Subhani
Associate
mercedes.subhani
@bakermckenzie.com