CIfAI

Cognitive Insights for Artificial Intelligence

Request for Information (RFI) Related to NIST's Assignments under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (AI) by NIST, Department of Commerce.

Submitted by Monica Lopez, PhD and Irene Gonzalez, PhD
Organization, Cognitive Insights for Artificial Intelligence (CIfAI)
January 31, 2024

On behalf of Cognitive Insights for Artificial Intelligence (CIfAI), we write in response to the National Institute of Standards and Technology (NIST) seeking information to assist in carrying out several of its responsibilities under the Executive Order on *Safe, Secure, and Trustworthy Development and Use of AI* issued on October 30, 2023. We support NIST's efforts in seeking stakeholder input to undertake an initiative for evaluating and auditing capabilities relating to AI technologies and to develop a variety of guidelines to enable the deployment of safe, secure, and trustworthy systems.

We at CIfAI provide strategic research-based solutions from a human-centered perspective to ensure the safe and ethical design, development, deployment, and management of AI-enabled autonomous systems across various industries. Our values-based approach is founded on accuracy, consistency, and context-dependency, and supports trusted data across every phase of the AI lifecycle to achieve confident and fair decision making.

Background

Artificial Intelligence (AI) is the most named and hottest field today. It is being incorporated in every industry (agriculture, oil, fashion, healthcare, finance, food, real estate, manufacturing, transportation, construction, retail, military, media, education to name a few).[1] An extension of AI is generative AI, a category of AI capable of generating new content. It is built on existing technologies like large language models (LLMs) which are trained on large amount of text from the Internet and learn to predict the next word in a phrase. Generative AI (GenAI) systems not only generate text but other types of content like images, audio, Video and code. Unfortunately, GenAI can be misused by malicious actors to generate deepfakes to cause personal harm to

[1] Market Trends, May 27,2021. Top 50 Use Cases of Artificial Intelligence in Diverse Sectors, Analytics Insight. https://www.analyticsinsight.net/top-50-use-cases-of-artificial-intelligence-in-diverse-sectors/

individuals and contribute to the spread of fake news and disinformation. Moreover, GenAI systems can unintentionally be used to generate misinformation, confident sounding made-up facts and confabulations, and lead to copyright infringement. GenAI applications can also lead to other legal and regulatory risks, including privacy and security concerns, as has been seen with incidents involving, for example, the disclosure of corporate source code. Like other AI models, GenAI is vulnerable to biased output and abuse. Also, given the limited availability of expertise and computational resources to address GenAI vulnerabilities, its use poses a business risk. If GenAI systems are not adequately secured, they could become a target of cyberattacks, creating additional data security concerns.

Due to the proliferation of AI-generated content and increasing risks associated with such, AI digital watermarking has become a way to identify if digital content was generated using GenAI. Although watermarking is well suited for generated pixel art, video and audio, it is not suited for text-based GenAI content to check for tampered text against different watermark detection tools. In fact, current AI-detection tools, without watermarking, may not even recognize AI-generated text that has been modified by a person.

We appreciate the opportunity to provide written comments that cover some of the issues presented above. We provide answers to three selected questions under **Topic 1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security**.

Question: **The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI, and what roles individuals bringing such knowledge could serve**.

Answer: Governing GenAI necessitates a multi-/interdisciplinary approach because AI-based systems have cross-sectoral and cross-societal consequences. What matters fundamentally is (a) the use case of the system –its context of use, its uniqueness of use— and (b) the user of the system –their understanding of it (i.e., agency over the system) and their use of it (i.e., in alignment or misalignment with the system's intended purpose). Governing these core issues requires a comprehensive mindset to consider the various elements of each, requiring individuals with expertise in thinking about and applying big picture thinking to find solutions that directly address the above.

Recommendations

1) That the role of an AI Governance Officer, for example, should be prioritized in every organization and considered equal to other officers in the C-suite.
2) This role specifically requires cross-domain expertise in the science, ethics, business, law, and geopolitics of AI.
3) Important skills should include the following:
    - foundational knowledge of AI systems and their lifecycle,
    - empirical work in any of the areas of AI system reliability, safety, etc.,

- awareness of the social impacts of AI-enabled systems and how risks are managed,
- understanding and implementation of emerging responsible AI principles and AI governance frameworks,
- understanding of current and emerging laws applicable to AI-enabled systems, and
- awareness of concerns and debates surrounding AI governance regimes for various use cases.

Question: **Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users).**

Answer: Everyone has a role to play in responsibly addressing the harms of GenAI because everyone is a user of these systems in some form or another. Moreover, behavior of responsibility and its consequences are interconnected; no one is immune to the social impacts of AI system use.

Recommendations

1. AI system developers need to reflect upon, design, implement, monitor and record explicit, empirically tested guardrails in their AI-based systems. Building protections should include the

- mitigation of unfair biases,
- implementation of policies that ban illegal, inappropriate, misleading and harmful content, and mechanisms of enforcement of said policies,
- creation of safeguards for minors, and
- infrastructure for the protection of identified copyrighted material.

These are key to implementing ethically-aligned and soon-to-be legally compliant AI-enabled technologies.

2. AI system deployers are essentially companies that use an AI system developed in house (e.g., cybersecurity companies to monitor network traffic and transactions from any industry) or third parties that buy an already developed AI system (e.g., banks to carry out financial transactions, loans and account services). Third-party AI-enabled systems that can be purchased, licensed or accessed pose increasing risk for organizations. As a result, the following minimum requirements should be endorsed:

- Deployers need to understand, check for, validate and explain the integrated guardrails of such AI systems to minimize privacy and security risks, detect fraud, and respond effectively to potential cyberattacks.
- Guardrails can be assessed through an appropriate risk management strategy and the evaluation of system audit reports that include impact assessments of clearly defined and identified harms from each use case.
- An effective risk management strategy should include the analysis of data used by the AI system and the system's resulting limitations, complying with regulatory requirements by producing system maintenance reports as more data are acquired and used to train the AI

system and/or modifications made to the model, and quickly identifying and addressing risks via real-time monitoring of the AI-enabled system's activity in the field.

- Company boards must recommend and support robust governance and controls of high risk AI-enabled systems, including GenAI systems given their current prevalence across many use cases.
- Companies must engage with regulators to discuss regulatory constrains, establish standards, and normalize the generation of system reports for AI/ML usage.

3. AI system end users are essentially the general public, professionals from all industries (from the technical to the creative), and students. The understanding of how AI models behave and their outputs, including their societal impact, is definitely a challenge for both non-technical and technical end users. While there have been considerable efforts in explaining AI systems to end users, it remains an empirical challenge to find a single best solution that fits all cases.[2] As such:

- It is important that end users be aware of, understand, and be empowered to respond, for example, in the event of system failure, to guardrails implemented within the system they are using. A mechanism for end user reporting should be user-friendly and immediately checked for validity.
- Provide end users clear and honest explanations of the AI system by focusing on the explainability of key functionalities rather than just explaining how the entire AI system works in a general sense. This could be done by offering various explanations via visualizations, data accessing, linguistic metaphors, graphics, etc.
- End users with technical and know-how expertise who use an AI system for good purposes, such as ethical hackers or red-teaming groups, can be assigned to attempt unethical and illegal intrusions of an organization's AI system under the direction of such organization to uncover vulnerabilities in the software and report back such vulnerabilities to the organization in order to improve their defenses. Incentives could be determined for the crowdsourcing of such, for example.
- End users who gain illegal access to an AI system for nefarious purposes (e.g., malicious hackers) should be deterred with strong guardrails and a variety of defense approaches that have been implemented[3] to keep the AI systems safe.

A series of papers providing the current landscape for end-user development for AI, which explains how users, even without AI and/or programming skills, can customize the AI behavior to their needs is available for perusal in the *Proceedings of the 9th International Symposium on End-User Development*.[4]

[2] Laato S. et al. How to explain AI systems to end users: a systematic literature review and research agenda. Internet Research Vol. 32 No. 7, 2022 pp. 1-31. Emerald Publishing Limited 1066-2243 DOI 10.1108/INTR-08-2021-0600

[3] Andrew J. Lohn AJ. December 2020. Hacking AI. A PRIMER FOR POLICYMAKERS ON MACHINE LEARNING CYBERSECURITY. CSET. Center for Security and Emerging Technology. doi: 10.51593/2020CA006. https://cset.georgetown.edu/publication/hacking-ai/

[4] Spano LD, Schmidt A, Santoro C and Stumpf, S. (Eds.) (2023). End-User Development IS-EUD 2023. Lecture Notes in Computer Science, vol 13917. Springer, Cham. https://doi.org/10.1007/978-3-031-34433-6_2

***Question:* Economic and security implications of watermarking, provenance tracking, and other content authentication tools.**

Answer: Implemented by Italian paper manufacturers in the 13th century, watermarking has been in use for centuries to identify ownership, thwart forgery, and prove authenticity of an asset. Currently, watermarking technology has considerably advanced and is used to assert intellectual property ownership, to protect confidential information, to indicate the validity of a legal document, to prove authenticity of electronic data, or to help prevent counterfeiting, among others. Watermarking is also used to market and protect all kinds of images from theft. Unfortunately, the rapid development of low-cost and powerful editing technologies have made easier the tampering and forgery of digital media via removal of watermarks and the manipulation of images, objects, films, photos, text, videos, audio files, etc. To our dismay, TV and film piracy comprise nearly 60% of all digital intellectual property theft, making it an especially urgent problem for filmmakers and video professionals to address. This has cost at least 290,000 jobs and $29 billion in lost revenue in the film and television industry alone.[5] This is a threat to the economy more generally and the current workforce in particular. Also, the rise of sophisticated and complex software algorithms makes removal of watermarks easy, becoming the biggest threat to watermarks. To make things more complicated, advances in AI systems, and in particular GenAI, have produced an enormous amount of dangerous AI-generated content via the proliferation of sophisticated fake audio and video —as evidenced through voice, face and body manipulation— and thus posing a threat to society. To mitigate these threats, AI watermarking and AI detection tools for such have been addressed by the Biden Administration; through an Executive Order, NIST of the Department of Commerce is charged with creating authentication and watermarking standards for GenAI systems.[6]

Unfortunately, trying to implement tamper-proof methods to addressing the harms of AI-generated content using AI watermarking has not worked because AI detection tools can be manipulated with a number of scenarios to detect, remove or manufacture different watermarks without human intervention. Essentially, watermarking is vulnerable to being tampered with, which can trigger false positives and false negatives. This has already been tested by researchers who have been able to evade current watermarking to add fake watermarks to images.[7] Thus, the approach is to create AI-detection tools that, while not full-proof, can protect digital media from exploitation and piracy. One example is a tool for watermarking and identifying AI-generated images created by Google DeepMind (i.e., SynthID). This technology is based on embedding the

[5] 2021 DPE Factsheet. Intellectual Property Theft. A threat to working people and the economy. Department for Professional Employees. https://www.dpeaflcio.org/factsheets/intellectual-property-theft-a-threat-to-working-people-and-the-economy

[6] 2023 Dec 21, 2023. Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11) https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the

[7] Knobs K. Oct 3, 2023. Researchers Tested AI Watermarks—and Broke All of Them. Wired. https://www.wired.com/story/artificial-intelligence-watermarking-issues/

watermark in the pixels of an image and remaining detectable even in the case that metadata are lost by file manipulation. It uses two deep learning models —for watermarking and identifying— that are trained together on different sets of images. The combined model is optimized on a range of objectives, including correctly identifying watermarked content and improving imperceptibility by visually aligning the watermark to the original content.[8]

Another example is by a team from the University of Maryland who showed that a watermark is almost impossible to remove unless the model parameters are changed by a certain threshold. The watermark is also certifiable and empirically more robust compared to previous watermarking methods. Essentially, the technology reveals that the use of a randomized-smoothing-based training scheme is useful to generate an unremovable and certifiable neural network watermark.[9]

One additional example is by another team, also from the University of Maryland. The researchers evaluated the reliability of watermarks as a mechanism for the documentation and detection of machine-generated text. They showed that watermark reliability as a function of text length turns out to be a strong property of watermarking. This reliability is independent of text length and produces a rigorous and interpretable P-value that the user can leverage to control the false positive rate.[10]

Although such examples provide some level of reliability against the tampering of watermarks, it has been found that simply changing image files into different formats (HEIC, TIFF, JPG, JPEG, etc) can lead to watermarks being rendered useless or removed. Also, there are apps that remove watermarks with a few clicks. Furthermore, in the case of images, any cropping, re-sizing and overall editing will interfere with the process of an AI detection tool to scan or read the watermark. Moreover, with respect to watermark functioning and features, there is no consistency and accuracy between different watermarking technologies. It is currently impossible at Internet scale and speed to generate a forgery-free watermark as a result of content manipulation software increasingly becoming more sophisticated and metadata easily being manipulated and providing no proof of its origins.

Regarding the origins (provenance) of a piece of digital content (image, video, audio recording, code or document), the C2PA (Coalition for Content Provenance and Authenticity) standard has already been put to work via an extension in a web browser. An open source Google Chrome web browser extension that validates digital assets based on the C2PA standard was launched. The extension makes it easy to check images for C2PA manifests and adds a Content Credentials

[8] Goal S. and Holly P. 2023. Identifying AI-generated images with SynthID. Google DeepMind. https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/

[9] Bansal A. Et al. 2022. Certified Neural Network Watermarks with Randomized Smoothing. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162

[10] Kirchenbauer J. Et al. 2023. On the Reliability of Watermarks for Large Language Models. Preprint under review. https://arxiv.org/pdf/2306.04634.pdf

‘pin' to the image, if a manifest is present.[11] Content provenance seem to be a great solution in detecting whether or not digital content is fake. C2PA has already delivered several versions (1.0 and 1.1 in 2021; 1.2 in 2022) of the technical standard. New file formats aligning with GenAI, live video and audio are in the works.[12] Details of the 2022 version of C2PA Implementation Guidance is available for perusal.[13] This technical standard has the purpose of providing publishers, creators, and consumers the ability to trace the origin of different types of media.

Another way to protect content provenance is the use of ZIRCON, a novel zero-watermarking approach to establish end-to-end data trustworthiness in an IoT network. Provenance information is stored in a tamper-proof centralized network database through watermarks, generated at source node before transmission. The system is robust against several attacks, lightweight, storage efficient, and better in energy utilization and bandwidth consumption, compared to prior art.[14] Implementation of this approach could be another way to ensure data integrity and its secure transmission.

Another alternative way to protect digital content is SAFE™. It is a digital watermark embedding and detection tool for digital assets. This tool is used to embed and detect digital watermarks in a device (e.g., desk computer, laptop, tablet, smart phone, camera). SAFE™ digital watermarks communicate content provenance, authenticity, and copyright information about a digital asset in a way that is both secure and inextricably linked to the asset itself, and seems to be a trustworthy and reliable digital watermark.[15]

Given the constraints and system challenges of watermarking, in spite of the technological advances, it has been very difficult to create a tamper-proof and indestructible watermark.

Recommendations to help in the protection of watermarking and content provenance

- The AI developer should offer AI watermarking features to users, but users should ultimately determine its use for content generated by their prompts to protect their identity and fundamental rights.

[11] Digimarc Blog. Jan 2024. Offering Free Digital Watermark Embedding and Detection Tools to Device and Chip Manufacturers and Content Creation Platforms. https://www.digimarc.com/blog/offering-free-digital-watermark-embedding-and-detection-tools-device-and-chip-manufacturers

[12] Ref. 2023. Coalition Content Provenance and Authenticity, C2PA. https://c2pa.org/faq/

[13] Ref. 2022. C2PA Implementation Guidance. https://c2pa.org/specifications/specifications/1.0/guidance/Guidance.html

[14] Farah O. Et al. 2023. ZIRCON: Zero-watermarking-based approach for data integrity and secure provenance in IoT networks. Preprint under review. https://arxiv.org/pdf/2305.00266.pdf

[15] Digimarc Blog. Jan 2024. Offering Free Digital Watermark Embedding and Detection Tools to Device and Chip Manufacturers and Content Creation Platforms. https://www.digimarc.com/blog/offering-free-digital-watermark-embedding-and-detection-tools-device-and-chip-manufacturers

- The watermark must help people to identify any AI-generated content in a reliable, consistent and trustworthy manner.
- An AI tool cannot be misused by applying or removing watermarks in non-AI or AI-generated content, respectively. There is need to establish guardrails.
- The watermark could be invisible (desynchronization) or hidden and cover part or the entire image or object. These watermarks are almost impossible to completely completely since doing so will alter the consistency or underlying property of the object/image's authenticity for non-AI or AI-generated content.
- The watermark could be bound permanently to the object/image such that any attempt to remove it will also remove the area of the object/image where the watermark is bound.
- Alternatively, utilize the service of a watermarking company to search for whether the image/object is being misused, or used fraudulently, or is generated by AI. Although expensive, it can provide proof for legal action.
- As another alternative option, it is suggested to use a tool like SAFE™ to protect digital content at the device level (e.g., desk computer, laptop, tablet, smart phone, camera), rather than protecting particular media content.
- For content provenance, ZIRCON could be used to ensure its integrity by verifying the origin and authenticity of images, objects, videos, podcasts, code or news articles.
- Ideally, implementation of the C2PA standards to allow for consistency of detection watermarks via AI-tools is advisable, given the lack of AI regulations for GenAI and the difficulty in watermarking all GenAI content.
- Ultimately, regulation and oversight is necessary to oblige companies to institute the use of watermarking their GenAI outputs.

In regard to watermarking economics, the digital watermarking surge has been largely fueled by the rapid growth of the Internet and creation and proliferation of digital media, which has raised significant concerns about copyright infringement and theft of intellectual property. Thus, digital watermarking growth underscores its importance to protect and safeguard the intellectual property and copyrights of content-creating professionals from all industries and to ensure the integrity of digital content (images, pictures, video, audio, code, documents, etc.).

Watermarking potential is vast and can bring novel uses for content authentication as digital technologies continue advancing. Given this growth, the global Digital Watermark Technology market size was valued at USD 47.02 million in 2022. It is expected to expand at a CAGR (Compound Annual Grow Rate) of 8.45% during 2022-2028, and will reach USD 76.5 million in 2028[16]. As can be seen, this is an industry of considerable and continued market growth.

[16] Trending Reports. 2024. 2031 "Digital Watermark Technology Market Size" / Major Downstream Customers Analysis. https://www.linkedin.com/pulse/2031-digital-watermark-technology-market-size-majordownstreamcustomersanalysis-67y1c/