Research agenda on risks from synthetic content

Canada

Australian Government

Department of Industry, Science and Resources

Overview

The International Network of AI Safety Institutes aims to advance AI safety and security science, research, testing, and guidance in collaboration with experts across industry, academia, and civil society. In pursuit of this mission, a research agenda has been developed to advance the understanding and mitigation of risks from synthetic content. This research agenda identifies priority areas, both for potential Network collaboration and to incentivise research and funding for those in the AI community working on understanding and mitigating the risks from synthetic content. It is supported by, and will be used as a reference in consideration of work planning and prioritization by, Australia, Canada, France, the European Commission, Japan, Kenya, Korea, Singapore, and the United Kingdom. This agenda is not binding on Network members, that independently determine their areas of focus and prioritization.

At the inaugural Network convening held in November 2024, the Network gathered input from members and external stakeholders to inform the further elaboration of the research agenda.

Scope of risks

With generative AI advancing rapidly, we are seeing an explosion of synthetic content across text, audio, image and video. The increasing volume of synthetic content and widely available access to generating this content at scale pose significant challenges. While synthetic content has important positive and benign uses, its widespread production and distribution risks undermining trust and causing harm to individuals, organisations, communities, and society. The primary ways in which synthetic content causes harms are:

- *Harmful content:* for example, content depicting child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII) harms children as well as people depicted in NCII.
- *Facilitation of fraud, impersonation and deception*: content that realistically simulates communications from real people or organisations can facilitate fraudulent activities, result in reputational harm and/or leverage known personalities for deceptive purposes
- *Undermining trust and individual autonomy:* the ability to easily disseminate undetectable content that simulates real-world events or human-generated work can undermine trust in institutions and the digital information environment, preventing people from being able to distinguish human-generated content from highly realistic AI-generated or manipulated content.

Current state of the science

Mitigating risks due to harmful content

There is broad agreement that AI-generated CSAM and NCII are unacceptable. A great deal of work has been done to assess the scale of the problem and mechanisms for identifying, assessing and mitigating the risk of CSAM and NCII.[1,2]

On the basis of this work, a coalition of non-government and private sector partners have come together to lay out key measures needed to ensure safety by design across the synthetic content lifecycle.[3] These include:

- measures taken during training and development of generative AI models and systems
- safeguards implemented prior to and during deployment of those models and systems
- combatting the dissemination of AI-generated CSAM content and its harmful effects.

However, further work is needed to understand the propensity of models to produce harmful content and improve the effectiveness of safeguards.

Risk mitigation through content transparency

Digital content transparency (DCT) techniques are an important class of mechanisms for managing risks arising from the difficulty in distinguishing AI-generated content from human-generated content. These techniques help to distinguish synthetic from non-synthetic content by providing information and transparency regarding the origin and history of content. Some of these techniques are promising but not yet widely adopted, while others are not yet reliable. Currently, there is no single technique or mitigation that can reliably protect against the risks associated with unidentified synthetic content.

The efficacy of current DCT techniques varies by the specific technique, its application and the modality of content to which it is applied, amongst many other factors. Industry actors and researchers have not reached consensus on how well specific techniques work for their intended purpose, such as in providing trustworthy signals to users about the origin of a given piece of content.

---

[1] Internet Watch Foundation 'How AI is being abused to create child sexual abuse imagery', October 2023.
[2] D Thiel, M Stroebel, R Portnof, 'Generative ML and CSAM: Implications and Mitigations', Thorn and Stanford Internet Observatory, June 2023.
[3] Thorn, 'Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments', July 2024.

Provenance data tracking techniques, which record the origins and history of digital content, can be used to establish whether content is synthetic or non-synthetic across audiences.[4] Current research of these techniques shows mixed reviews and a lack of consensus. For example, even the 'strongest' watermarks can be attacked,[5] including for black box language model outputs.[6] There are also implementation challenges for secure metadata specifications,[7] including issues related to retaining edits made to content by different entities and adapting traditional public-key infrastructure (PKI) to incorporate cryptographic and 'soft bindings' (which can include digital fingerprints or watermarks) in content that is disseminated across platforms. Broadly, achieving interoperability across formats, and maintaining the security and privacy of metadata, watermarks and content itself across hardware, software and digital platforms remains a challenge. At the same time, other researchers have identified the positive effects of digital watermarks and other solutions that mark and detect the content produced by generative AI models and systems as important mitigating measures that can be reliable in responding to less sophisticated adversaries and reducing harms at scale.[8]

A multitude of synthetic content detection tools and techniques are available. Many of these are likely to be more suitable for use by analysts and experts (e.g., by social media platforms or forensic investigators). Moreover, because detection results typically generate results in probabilistic terms, they may be difficult to interpret,[9] particularly if they are not supported by plain language explanations as to how the results were generated.

Some detectors are developed to identify provenance signals, such as the existence of watermarks attached to digital content. These detectors are useful for AI developers to track the source of content generated by specific models, or for other actors, including platforms, law enforcement authorities, journalists and the public to determine the origin of content. Effective watermarks are in significant part reliant on effective detectors and policy levers.

---

[4] B Chandra, J Dunietz, K Roberts, 'Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency' NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, April 2024, https://doi.org/10.6028/NIST.AI.100-4.
[5] H Zhang, B Edelman, D Francati, D Venturi, G Ateniese and B Barak, 'Impossibility of strong watermarking for generative models', July 2024, doi.org/10.48550/arXiv.2311.04378.
[6] D Bahri, J Wieting, D Alon and D Metzler, 'A watermark for black-box language models', October 2024, doi.org/10.48550/arXiv.2410.02099.
[7] For example, The Coalition for Content Provenance and Authenticity's (C2PA) specification.
[8] For example, A Knott, D Pedreschi, R Chatila, T Chakraborti, S Leavy, R Baeza-Yates, D Eyers, A Trotman, P Teal, P Biecek, S Russel and Y Bengio, 'Generative AI models should include detection mechanisms as a condition for public release', Ethics and Information Technology, 2023, 25(55), https://doi.org/10.1007/s10676-023-09728-4; H Farid, 'Watermarking ChatGPT, DALL-E and Other Generative AIs Could Help Protect Against Fraud and Misinformation', *The Conversation,* 27 Mar 2023.
[9] S Gregory, 'Pre-Empting a Crisis: Deepfake Detection Skills + Global Access to Media Forensics Tools', *WITNESS Blog*, 14 July 2021.

By contrast, in the absence of watermarks or provenance metadata, other methods rely on signals in the content of the AI output, such as statistical regularities in generated text, or image features in generated images. Such ex-post detection techniques are criticised for being reactive and inaccurate,[10] as well as unreliable in many real-world use cases, such as in scenarios when teachers must determine whether student-submitted text is AI-generated.[11] There are also important gaps in the data that can be analysed by such detectors. For example, to date, most research on synthetic text detection has focused on English or other high-resource languages.

[10] D Kovtun, 'Testing AI or Not: How Well Does an AI Image Detector Do Its Job?,' *bellingcat*, 11 September 2023.
[11] S Vinu, A Kumar, S Balasubramanian, W Wang and S Feizi, 'Can AI-generated text be reliably detected?', February 2024, doi.org/10.48550/arXiv.2303.11156.

Research priorities

Many research questions related to mitigating risks from synthetic content remain open, with little consensus on what research areas to prioritise. The Network has identified **4 areas** with multiple subtopics where further research is needed to advance the state of the science for understanding and mitigating risks from synthetic content. The below list is neither exhaustive nor ranked by importance.

1. How can safeguards built into AI models and systems to reduce harmful outputs be evaluated and improved?

Safeguards that are built into AI models and systems are a growing area of research that can help limit the creation of harmful synthetic content. Research and testing on the generation of such outputs, particularly for the most harmful categories of content (e.g., AI-generated child sexual abuse material or non-consensual intimate imagery), could facilitate the adoption of improved mitigations by AI developers. In particular, the application of multi-layered safeguards to prevent generative AI systems from producing harmful categories of content is a ripe area for research.[12]

Potential sub-topic area: Methods to prevent generative models and systems from generating harmful categories of content, such as non-consensual intimate imagery

This sub-topic area covers one of the most salient and egregious harms from generative AI models, AI-generated non-consensual intimate imagery (AIG-NCII). Technical mitigations for this area of harm are discussed at length in a variety of publications, however, much of the work on safeguards is done by companies with significant resources at their disposal, rather than by independent researchers. Expanding independent academic research on technical model and system safeguards to prevent or disrupt AIG-NCII generations, especially for more advanced and multimodal models, such as multimodal safety classifiers for model outputs and methods such as 'concept erasing' via fine-tuning to suppress sexual outputs while maintaining generation quality,[13] will be crucial to reduce and control downstream harms from AI model outputs.

[12] M Shamsujjoha, Q Lu, D Zhao, and L Zhu, 'Designing Multi-layered Runtime Guardrails for Foundation Model Based Agents: Swiss Cheese Model for AI Safety by Design' August 2024, doi.org/10.48550/arXiv.2408.02205.
[13] S Hong, J Lee, and S Woo. 'All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models' Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(19), doi.org/10.1609/aaai.v38i19.30107.

Potential sub-topic area: Testing the robustness of model safeguards to fine-tuning

Fine-tuning models has become a standard practice for specialising models for use for particular applications. However, there has been initial research showing that fine-tuning, even in the form of minimal, seemingly benign adjustments, can erode preexisting safeguards and undermine alignment strategies. Stanford Human-Centered AI (HAI) has highlighted tradeoffs between fine-tuning customisations and model safety and noted that interventions to address safety issues are still nascent and not foolproof.[14]

Further research into the impacts of fine-tuning on the effectiveness of model safeguards, including benchmarks for evaluation, is necessary for both open-source and closed-source models that can be fine-tuned by users. This includes detecting and verifying model integrity – establishing that safeguards implemented throughout the lifecycle have not been secretly removed or compromised.

Potential sub-topic area: Improving safeguards in multimodal contexts

Generative AI models are increasingly capable of accepting multimodal input and generating multimodal output (text, image, video, etc.). As a result, improving safeguards for multimodal contexts is becoming increasingly complex. For example, pairing a seemingly benign image with a seemingly benign textual prompt could result in a malign output if preexisting safeguards are unable to persist with combinations of text, image and other input types. Further research on improving safeguards within individual modalities, as well as across combinations of modalities, is necessary as multimodal-capable industry models become more widely adopted.

2. How can current DCT techniques and their implementations be evaluated?

Ensuring the security, reliability, privacy and accessibility[15] of any DCT technique's implementation is essential and provides a crucial foundation for mitigating the risks from synthetic content that might otherwise be mistaken for non-synthetic.

The utility of any technique is limited if the implementation of techniques can be exploited by adversarial actors. For example, a threat actor could assert false or spoofed provenance. Moreover, provenance could be easily lost after minor and/or seemingly benign modifications are made to content due to robustness issues, or the technique itself could compromise the privacy of the creator or user. If a watermark is robust, it will generally be difficult to remove; this could also result in the watermark and any information that it contains being tracked by different entities, without the consent of individuals whose

[14] P Henderson, X Qi, Y Zeng, T Xie, P Chen, R Jia and P Mittal, 'Safety Risks from Customizing Foundation Models via Fine-tuning', Stanford Human-Centered AI, January 2024.
[15] In cases of accessibility, cultural as well as multilingual contexts must also be considered.

personal information it contains.[16] This may be a more substantial privacy issue with metadata as well, as metadata typically contains more information about the content, including modifications and edits, as well as its creator.

Potential sub-topic area: Ecosystem mapping of current content authentication standards and protocols, including interactions between tools, systems, platforms, jurisdictions and potential security and privacy vulnerabilities at the ecosystem level

Security, robustness and privacy issues with the implementation of current content authentication methods often reveal themselves at the wider ecosystem level, as digital content is created and disseminated across different hardware, software, platforms and online services. Emerging protocols for content authentication and labelling include:

- C2PA's well-known specification
- Secure Evidence Attribution Label (SEAL), a completely open-source permissively licensed specification
- the Numbers protocol, a decentralised blockchain solution that also utilises the C2PA
- the ISCC standard for content identifiers, a digital fingerprint
- the use of other techniques such as digital watermarks.

Adoption of these standards is nascent, though growing.

Evaluating how these emerging visible or invisible watermark standards are being implemented, and where specific security and privacy vulnerabilities may occur at an ecosystem level, can help organisations course-correct, improve their implementations and coordinate on fixing these issues. It can also inform design of improved protocols, which is itself a promising research direction.

[16] *Center for Democracy and Technology*, 'Privacy Principles for Digital Watermarking,' 2 June 2008.

Potential sub-topic area: Improving benchmarks to test the adversarial removal, tampering and forging of watermarks across different modalities of content

A major problem for digital watermarking is robustness and security in the face of seemingly benign edits or adversarial attempts to remove the watermark. Even watermarks that are robust to specific classes of perturbations can be vulnerable to adversarial attacks. Overt watermarks applied to small portions of a piece of synthetic or non-synthetic content can easily be edited out, sometimes even by accident, and it is often similarly straightforward to deliberately remove covert watermarks if they are not robustly applied. Further benchmarking research in this domain is necessary to improve the security of watermarking across different modalities of content (image, audio, video, text).

Another idea to explore is whether modifications to content can be tracked using watermarks. That is, if watermarked content is modified or edited, how should that be reflected in the watermark?

Potential sub-topic area: Improving the evaluation of DCT techniques

Currently, there is a notable lack of systematic, realistic and standardised evaluations and benchmarks for various DCT techniques, including both provenance data tracking as well as detection methods, revealing gaps in understanding how current implementations perform in terms of security, privacy, reliability, interoperability and accessibility. Public benchmarks for specific techniques can be developed to help align implementations to specific and recognised standards.

How can we develop realistic benchmark datasets that more closely match how AI-generated content appears in the real world? What qualities, in addition to standard performance metrics, should a high-quality evaluation framework test for (e.g., reliability, utility, transparency, etc.)? How can we develop more cross-lingual evaluations?

3. How is synthetic content spreading in the information environment and what systemic impacts is it having (e.g., with regard to public trust)? How are DCT techniques being adopted and used in the real world and what are their potential systemic impacts?

The research community should prioritise investigations related to the broader systemic impacts of AI systems and their outputs.[17] When AI systems are adopted at scale, reach across borders and affect the lives of billions of users, we must be attentive to second-order effects and negative externalities, in particular the effects on public trust from the widespread production and distribution of synthetic content. In addition, it is essential to study the effects of adopting various content transparency techniques to ensure that any unintended risks introduced by particular sets of mitigations are also addressed.

Potential sub-topic area: The impact of synthetic content on global information ecosystems and public trust, and what can be addressed by content transparency techniques

Synthetic content is rapidly spreading across internet ecosystems, and we have very little understanding as to how it may be impacting public trust in information. It would be useful to develop a framework to categorise different types of synthetic content and their varying impacts on public trust, and to identify the threat models that can be addressed by content transparency techniques and their respective stakeholders. This can also inform what kinds of novel mitigations and interventions may be needed to increase public trust.

Potential sub-topic area: The adoption and usability of content transparency techniques

DCT techniques are only useful if a large portion of the population is using them, hence it is important to understand how they are currently being used 'in the real world' and what may be hindering their adoption: how do users perceive, interpret and respond to content disclosure? Can they reason through the implications? What provenance information is deemed useful and who should it be visible to? How does it influence their behavior and impact their level of trust? How dependent is the tools' effectiveness on context, culture and other factors? Can we anticipate how well they'll work in advance of widespread deployment? When should it be possible to opt out?

[17] L Weidinger, M Rauh, N Marchal, A Manzini, L Hendricks, J Mateos-Garcia, S Bergman, J Kay, C Griffin, B Bariach, I Gabriel, V Rieser and W Isaac, 'Sociotechnical Safety Evaluations of Generative AI Models', Google Deepmind, October 2023, doi.org/10.48550/arXiv.2310.11986.

Potential sub-topic area: The impact of the adoption of techniques and mitigations for synthetic content on global information and media ecosystems, particularly in the global majority

The impact of content transparency techniques is not well understood. For instance, the implementation of some mitigations can entrench pre-existing digital divides across populations and specific demographic groups. Initial investigations have been done by civil society organisations, such as WITNESS, to understand where content authentication implementations could violate or negatively impact human rights.[18] Much more empirical research is needed to understand impacts of techniques and mitigations, both positive and negative, on media ecosystems that operate outside of the Global North.

Potential sub-topic area: Validating real content and dealing with the spectrum between original and AI-generated

Content transparency techniques can be used to authenticate real content as opposed to flagging AI-generated content, hence it would be useful to understand the impacts of this labelling paradigm and potential false positives. Can such systems be used adversarially, e.g., to discredit real human content? What level of AI modification shifts the label from real to AI-generated?

4. What technical or non-technical approaches can be used to support or advance DCT techniques?

Emerging approaches capable of advancing risk mitigation efforts exist but are nascent in comparison to current methods to authenticate and verify the origin of content. These methods are crucial to develop as AI capabilities become more advanced, adversarial use of AI models becomes more sophisticated, and synthetic content itself becomes more complex.

Potential sub-topic area: Designing text watermarks that survive translation

Text watermarks inserted by existing methods do not survive translation from one language to another.[19] This lack of 'cross-lingual consistency' poses challenges for the robustness of watermarks in both benign and adversarial settings. Further research in this area is needed to ensure robustness and reliability of watermarks during language translation.

[18] J Castellanos and S Gregory, 'WITNESS and the C2PA Harms and Misuse Assessment Process', *WITNESS Blog*, 2 December 2021.
[19] Z He, B Zhou, H Hao, A Liu, X Wang, Z Tu, Z Zhang and R Wang, 'Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models', June 2024, doi.org/10.48550/arXiv.2402.14007.

Potential sub-topic area: Developing a benchmark and improving techniques for content-based detection methods

Initial research suggests that essays written in English by second-language learners are more likely to be flagged as AI-generated by detection tools[20]. Research has also identified heterogeneous performance of generative AI across languages and cultures. Further research is needed to better understand, measure and improve the performance of techniques for detecting synthetic content, including mitigating systematic occurrence of false positive incidents.

Potential sub-topic area: Improving semantic understanding and attribution for multimodal detection of synthetic content

Current synthetic content detection methods often fall short in practical use cases. Advanced efforts that integrate other types of information are needed. This could include consideration of account network information, user activities, as well as patterns appearing across different content sources – moving beyond a single content-based approach to detecting synthetic content. A further challenge is how to address sophisticated networks of adversaries utilising both synthetic and non-synthetic content generation. Developing detection methods that incorporate these broader attribution signals will be crucial to improving robustness against advanced adversarial tactics.

Potential sub-topic area: Human collaboration with AI detection tools, their outputs, and interpretation of results

The interaction of humans and AI detection tools in specific, high-risk contexts, and subsequent interpretation of the results to take further action has been an under-researched area. Some of this work is taking place within civil society, including human rights and journalistic organisations. WITNESS, for example, through their Deepfakes Rapid Response Task Force, involves fact-checkers and journalists in the process of detecting AI-generated content. More research is needed into hybrid approaches that evaluate human interactions with detection tools for fact-checking, provenance tracking and mitigating harmful impersonations, among other applications, in order to examine the dissemination and impact of these tools in society.

[20] W Liang, M Yksengonul, Y Mao, E Wu and J Zou, 'GPT Detectors Are Biased against Non-Native English Writers', *ScienceDirect*, 10 July 2023.

Potential sub-topic area: Improve and develop novel DCT techniques that enhance the performance and security of current implementations and standards

Security, privacy, robustness, and efficiency issues in the implementation of current DCT techniques mainly arise because of weaknesses in the underlying algorithms, such as perceptual hashing, watermarking and content authentication algorithms. Identifying fundamental security, privacy and efficiency weaknesses can help with improving existing DCT techniques and with developing novel techniques to overcome these issues.

Glossary

Digital content transparency techniques: methods which facilitate access and exposure to information regarding the origin and/or history of digital content.

Digital watermarking: a technique to embed information into content (image, text, audio, video) while making it difficult to remove. Such watermarking can assist in verifying the authenticity of the content or characteristics of its provenance, modifications, or conveyance.[21]

Metadata: information describing the characteristics of data.[22]

Content authentication: utilises provenance data tracking methods (technical methods to track the origin and/or history of content, including watermarking, metadata, digital fingerprints) to determine the nature of the content.[23]

Provenance data tracking: techniques that record the origin and history of content.[24]

***Security*:** protecting digital content transparency techniques, and the systems used to execute those techniques from unauthorised access, use, disclosure, tampering, spoofing, disruption, or destruction.[25]

Reliability: ensuring that digital content transparency techniques are reliable, by being robust to benign edits and manipulations, can be effectively used to identify the origin of digital content, are applied and improved based on the current technical state of the art, and preserve the integrity of the content.

Disclosure: providing users with information about how content was created, modified, and/or published, as well information about how techniques and mitigations for synthetic content are applied.

Accessibility: providing individuals, organisations, and populations, particularly those that are lower-resourced, the opportunity to obtain the benefits of digital content transparency. Techniques that are applied unequally around the world could entrench existing or produce new digital disparities.

---

[21] Chandra, Dunietz and Roberts, 'Reducing Risks Posed by Synthetic Content'.
[22] C Johnson, M Badger, D Waltermire, J Snyder and C Skorupka, 'Guide to Cyber Threat Information Sharing', NIST Computer Security Resource Center, October 4 2016, doi.org/10.6028/NIST.SP.800-150.
[23] Chandra, Dunietz and Roberts, 'Reducing Risks Posed by Synthetic Content'.
[24] Ibid.
[25] *NIST Computer Security Resource Centre*, 'Infosec Glossary', n.d.