Programme de recherche sur les risques liés au contenu synthétique

Canada

Australian Government
Department of Industry, Science and Resources

Cette publication est disponible en ligne et en format PDF à: https://ised-isde.canada.ca/site/isde/fr/institut-canadien-securite-lintelligence-artificielle/programme-recherche-risques-lies-contenu-synthetique

Le programme de recherche est également disponible en ligne sur le site Internet du ministère australien de l'Industrie, des Sciences et des Ressources, à l'adresse suivante : https://www.industry.gov.au/publications/research-agenda-risks-synthetic-content

Pour obtenir une copie de cette publication sur papier ou la demander dans un média substitut (p. ex. braille ou gros caractères), veuillez remplir le formulaire de demande de publications à https://ised-isde.canada.ca/site/publications/fr/formulaire/demande-depublication
ou communiquer avec le :

Centre des services Web
Innovation, Sciences et Développement économique Canada
Édifice C.D. Howe
235, rue Queen
Ottawa (Ontario) K1A 0H5
Canada

Téléphone (sans frais au Canada) : 1-800-328-6189
Téléphone (Ottawa) : 613-954-5031
ATS (pour les malentendants) : 1-866-694-8389
Heures d'ouverture : De 8 h 30 à 17 h (heure de l'Est)
Courriel : ISED@ised-isde.gc.ca

Iu37-53/2025F-PDF
ISBN 978-0-660-75908-1

This publication is also available in English under the title *Research agenda on risks from synthetic content*

Aperçu

Le Réseau international des instituts pour la sécurité de l'IA a pour objectif de faire avancer la science de la sécurité et sûreté de l'IA, la recherche, les tests et les conseils en collaboration avec des experts issus de l'industrie, du monde universitaire et de la société civile. Dans le cadre de cette mission, un programme de recherche a été élaboré visant à améliorer la compréhension et l'atténuation des risques liés au contenu synthétique. Ce programme de recherche identifie les domaines prioritaires, tant pour une éventuelle collaboration au sein du réseau que pour encourager la recherche et le financement de la communauté qui s'efforce de comprendre et d'atténuer les risques liés aux contenus synthétiques. Il est appuyé par l'Australie, le Canada, la France, la Commission européenne, le Japon, le Kenya, la Corée, Singapour et le Royaume-Uni, qui s'y référeront pour planifier et hiérarchiser leurs travaux. Ce programme n'est pas contraignant pour les membres du réseau, qui déterminent de manière indépendante leurs domaines d'intérêt et leurs priorités.

Lors de la première réunion du Réseau en novembre 2024, le Réseau a consulté les membres et les intervenants externes afin de nourrir l'élaboration plus approfondie du programme de recherche.

Portée des risques

L'essor fulgurant de l'IA générative induit une prolifération exponentielle du contenu synthétique sous forme de textes, de sons, d'images et de vidéos. L'augmentation du volume de contenu synthétique et l'accès largement répandu à des outils permettant de générer ce contenu à grande échelle soulèvent des enjeux majeurs. Bien que le contenu synthétique ait d'importantes utilisations positives et bénignes, sa production et sa distribution à grande échelle pourraient fragiliser la confiance et causer préjudice aux personnes, aux organisations, aux communautés et à la société. Voici les principales façons dont le contenu synthétique peut causer des préjudices :

- *Contenu préjudiciable* : par exemple, le contenu montrant l'exploitation sexuelle d'enfants et les images intimes non consensuelles portent préjudice aux enfants ainsi qu'aux personnes figurant sur ces images.
- *Facilitation de la fraude, de l'usurpation de l'identité et de la tromperie* : le contenu qui simule de façon réaliste des communications impliquant des personnes ou des organisations réelles peut faciliter des activités frauduleuses, entacher la réputation de celles-ci ou permettre de profiter de la notoriété de personnalités à des fins trompeuses.

- *Fragilisation de la confiance et de l'autonomie individuelle :* la capacité de diffuser facilement du contenu indétectable, qui simule des événements réels ou du travail généré par l'être humain, peut nuire à la confiance dans les institutions et dans l'environnement d'information numérique, dans lequel les gens sont incapables de distinguer le contenu généré par l'être humain du contenu extrêmement réaliste généré ou manipulé par l'IA.

État actuel de la science

Atténuation des risques liés au contenu préjudiciable

Il est généralement admis que le contenu montrant l'exploitation sexuelle d'enfants et les images intimes non consensuelles générés par l'IA sont inacceptables. Un travail conséquent a été réalisé pour évaluer l'étendue du problème et les mécanismes permettant d'identifier, d'évaluer et d'atténuer les risques associés à ce genre de contenu[1,2].

À partir de ces travaux, une coalition de partenaires issus des secteurs privé et non gouvernemental s'est réunie pour définir les mesures clés nécessaires à la sécurité intégrée tout au long du cycle de vie du contenu synthétique[3]. Voici quelques exemples :

- Des mesures sont prises pendant la formation et l'élaboration de modèles et de systèmes d'IA générative;
- Des mesures de protection sont mises en œuvre avant et pendant le déploiement de ces modèles et systèmes;
- On combat la diffusion du contenu montrant l'exploitation sexuelle d'enfants généré par l'IA et ses effets préjudiciables.

Cependant, d'autres travaux sont nécessaires afin de comprendre la propension des modèles à produire du contenu préjudiciable et d'améliorer l'efficacité des mesures de protection.

---

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

Atténuation des risques grâce à la transparence du contenu

Les techniques de transparence du contenu numérique sont un élément essentiel des mécanismes de gestion des risques découlant de la difficulté à différencier le contenu généré par l'IA du contenu généré par l'être humain. Ces techniques aident à distinguer le contenu synthétique du contenu non synthétique en fournissant des renseignements et de la transparence concernant l'origine et l'historique du contenu. Certaines de ces techniques sont prometteuses, mais ne sont pas encore largement adoptées, tandis que la fiabilité d'autres techniques doit encore être démontrée. À l'heure actuelle, il n'existe pas de technique ou de mesure d'atténuation unique permettant d'assurer une protection fiable contre les risques associés au contenu synthétique non identifié.

L'efficacité des techniques actuelles de transparence du contenu numérique repose sur de nombreux facteurs, comme la technique en elle-même, son application ou la modalité d'application au contenu. Les acteurs et les chercheurs de l'industrie ne sont pas parvenus à un consensus sur l'efficacité de certaines techniques par rapport à l'objectif, par exemple en fournissant aux utilisateurs des signaux fiables sur l'origine d'un contenu donné.

Les méthodes de suivi de la provenance des données, qui permettent de retracer l'origine et l'historique du contenu numérique, peuvent être utilisées pour déterminer si le contenu est synthétique ou non parmi les auditoires[4]. Selon les recherches actuelles sur ces méthodes, les avis sont partagés et il y a absence d'un consensus. Par exemple, même les filigranes « les plus robustes » peuvent être compromis[5], y compris pour les sorties d'un modèle de langage en boîte noire[6]. Par ailleurs, des difficultés sont à signaler concernant la mise en œuvre des spécifications de métadonnées sécurisées[7], notamment pour la conservation des modifications apportées au contenu par différentes entités et l'adaptation de l'infrastructure à clés publiques traditionnelle afin d'intégrer les reliures cryptographiques et « souples » (qui peuvent comprendre les empreintes ou les filigranes numériques) dans le contenu diffusé sur les plateformes. De façon générale, l'interopérabilité entre les formats et le maintien de la sécurité et de la confidentialité des métadonnées, des filigranes numériques et du contenu lui-même sur le matériel, les logiciels et les plateformes numériques restent un défi. Parallèlement, d'autres chercheurs ont recensé les effets positifs des filigranes numériques et d'autres solutions permettant de

---

[4] Chandra, Dunietz et Roberts, "Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency", NIST AI, 20 novembre 2024.

[5] H Zhang, B Edelman, D Francati, D Venturi, G Ateniese et B Barak, "Impossibility of strong watermarking for generative models", 23 juillet 2024.

[6] D Bahri, J Wieting, D Alon et D Metzler, "A watermark for black-box language models", 2 octobre 2024.

[7] Par exemple, la spécification de la coalition pour la provenance et l'authenticité des contenus (C2PA).

marquer et de détecter le contenu produit par les modèles et systèmes génératifs de l'IA. Il s'agit de mesures d'atténuation importantes qui peuvent être fiables dans la lutte contre des adversaires moins sophistiqués et pour réduire les préjudices à grande échelle[8].

Une multitude d'outils et de techniques de détection de contenu synthétique sont disponibles. La plupart d'entre eux sont plus susceptibles d'être utilisés par des analystes et des experts (p. ex. plateformes de médias sociaux ou enquêteurs en médecine légale). De plus, les résultats de la détection peuvent être difficiles à interpréter[9], puisqu'ils sont souvent exprimés en termes probabilistes, surtout s'ils ne sont pas accompagnés d'explications en langage clair sur leur mode de génération.

Certains détecteurs sont conçus pour identifier les signaux de provenance, comme l'existence de filigranes numériques rattachés au contenu numérique. Ces détecteurs s'avèrent utiles pour les développeurs d'IA afin de suivre la source du contenu généré par des modèles spécifiques ou pour d'autres acteurs, y compris les plateformes, les autorités chargées de l'application de la loi, les journalistes et le public, qui peuvent ainsi déterminer l'origine du contenu. L'efficacité des filigranes numériques est étroitement liée à la qualité des détecteurs et des leviers politiques déployés.

À l'inverse, en l'absence de filigranes numériques ou de métadonnées sur la provenance, d'autres méthodes s'appuient sur des signaux présents dans le contenu généré par l'IA, comme des régularités statistiques dans les textes générés ou des caractéristiques visuelles dans les images générées. Ces techniques de détection a posteriori sont critiquées pour leur caractère réactif et leur imprécision[10], ainsi que pour leur manque de fiabilité dans de nombreux cas d'utilisation réels, comme lorsqu'un enseignant doit déterminer si un texte soumis par un élève a été généré par l'IA[11]. Il existe également des lacunes importantes dans les données accessibles à ces détecteurs. Par exemple, à ce jour, la plupart des recherches sur la détection de textes synthétiques ont porté sur l'anglais ou d'autres langues à ressources élevées.

---

[8] Par exemple, A Knott, D Pedreschi, R Chatila, T Chakraborti, S Leavy, R Baeza-Yates, D Eyers, A Trotman, P Teal, P Biecek, S Russel et Y Bengio, "Generative AI models should include detection mechanisms as a condition for public release". Ethics and Information Technology 25, article numéro 55, 28 octobre 2023 ; H Farid, "Watermarking ChatGPT, DALL-E and Other Generative AIs Could Help Protect Against Fraud and Misinformation", The conversation, 27 mars 2023.

[9] S Gregory, "Pre-Empting a Crisis: Deepfake Detection Skills + Global Access to Media Forensics Tools" WITNESS Blog, 14 juillet 2021.

[10] D Kovtun , "Testing AI or Not: How Well Does an AI Image Detector Do Its Job?", bellingcat, 11 septembre 2023.

[11] S Vinu, A Kumar, S Balasubramanian, W Wang et S Feizi, "Can AI-generated text be reliably detected?", 19 février 2024.

Priorités de recherche

De nombreuses questions de recherche liées à l'atténuation des risques liés au contenu synthétique restent en suspens, et il n'y a guère de consensus sur les domaines de recherche à privilégier. Le Réseau a cerné **quatre domaines** comportant plusieurs sous-thèmes pour lesquels des recherches approfondies sont nécessaires afin de faire progresser l'état de la science en vue de comprendre et d'atténuer les risques liés au contenu synthétique. La liste ci-dessous est non exhaustive et est présentée sans ordre de priorité.

1. Comment évaluer et améliorer les mesures de protection intégrées aux modèles et systèmes d'IA afin de réduire les résultats préjudiciables?

Les mesures de protection intégrées dans les modèles et systèmes d'IA constituent un domaine de recherche en plein développement et peuvent aider à limiter la création de contenu synthétique nuisible. La recherche et les essais sur la génération de contenu, en particulier pour les catégories de contenu les plus préjudiciables (p. ex. contenu montrant l'exploitation sexuelle d'enfants et images intimes non consensuelles générés par l'IA), pourraient faciliter l'adoption de meilleures pratiques d'atténuation par les concepteurs d'IA. L'application de mesures de protection à plusieurs niveaux pour prévenir la production de contenu préjudiciable par les systèmes d'IA générative représente un domaine de recherche particulièrement prometteur[12].

Sous-thème potentiel : Méthodes visant à empêcher la production de contenus préjudiciables par les modèles et systèmes d'IA générative, comme des images intimes non consensuelles

Ce sous-thème aborde l'un des préjudices les plus importants et graves causés par les modèles d'IA générative : les images intimes non consensuelles générées par IA. Les mesures techniques d'atténuation de ce préjudice sont examinées en détail dans diverses publications. Cependant, la majeure partie du travail sur les mesures de protection est réalisée par des entreprises disposant de ressources importantes, plutôt que par des chercheurs indépendants. L'expansion de la recherche universitaire indépendante sur les modèles techniques et les mécanismes de protection des systèmes pour prévenir ou perturber la génération d'images intimes non consensuelles par l'IA, en particulier pour les modèles multimodaux et plus avancés, comme les classificateurs de sécurité multimodaux pour les sorties des modèles et les méthodes comme l'« effacement des concepts » par un réglage fin visant à supprimer tout le contenu sexuel tout en préservant la qualité de la

[12] M Shamsujjoha, Q Lu, D Zhao, et L Zhu,"Designing Multi-layered Runtime Guardrails for Foundation Model Based Agents: Swiss Cheese Model for AI Safety by Design", 5 août 2024.

génération[13], sera essentielle pour réduire et empêcher les préjudices en aval découlant des images intimes non consensuelles générées par l'IA.

Sous-thème potentiel : Mise à l'essai de la robustesse des mesures de protection du modèle au réglage fin

Le réglage fin des modèles est devenu une pratique courante pour spécialiser les modèles en vue d'applications particulières. Cependant, des recherches initiales ont montré que même sous la forme d'ajustements minimes et apparemment anodins, le réglage fin peut affaiblir les mesures de protection préexistantes et compromettre les stratégies d'harmonisation. Le Stanford Human-Centered AI (HAI) Institute a mis en évidence les compromis entre la personnalisation par réglage fin et la sécurité des modèles, et a noté que les interventions visant à résoudre les problèmes de sécurité sont encore embryonnaires et ne sont pas infaillibles[14].

Il est primordial de poursuivre les recherches concernant les répercussions du réglage fin sur l'efficacité des mesures de protection des modèles, qu'il s'agisse de modèles à source ouverte ou à source fermée que les utilisateurs peuvent peaufiner. Cela englobe la détection et la vérification de l'intégrité du modèle, c.-à-d. s'assurer que les mesures de protection déployées tout au long du cycle de vie n'ont pas été supprimées en catimini ou compromises.

Sous-thème potentiel : Améliorer les mesures de protection dans des contextes multimodaux

Les modèles d'IA générative sont de plus en plus capables d'accepter des entrées multimodales et de générer des sorties multimodales (texte, image, vidéo, etc.). Ainsi, l'amélioration des mesures de protection pour les contextes multimodaux devient de plus en plus complexe. Par exemple, l'association d'une image apparemment anodine et avec un message-guide tout aussi anodin peut conduire à la création de contenu préjudiciable si les mesures de protection préexistantes ne sont pas en mesure de maintenir leur efficacité face à des combinaisons de texte, d'images et d'autres types d'entrées. Des recherches supplémentaires sont nécessaires pour améliorer les mesures de protection tant au sein des modalités individuelles qu'à travers les combinaisons de modalités, à mesure que les modèles industriels multimodaux deviennent plus largement adoptés.

---

[13] S Hong, J Lee, et 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*. Vol. 38, nº 19, 2024.

[14] P Henderson, X Qi, Y Zeng, T Xie, P Chen, R Jia et P Mittal, "Safety Risks from Customizing Foundation Models via Fine-tuning", Stanford Human-Centered AI, 11 janvier 2024.

2. Comment évaluer les techniques actuelles de transparence du contenu numérique et leur mise en œuvre?

Il est essentiel que la mise en œuvre d'une technique de transparence du contenu numérique soit sécurisée, fiable et accessible[15] et qu'elle garantisse la protection des renseignements personnels. Cela constitue une base essentielle pour atténuer les risques liés au contenu synthétique qui pourraient autrement être pris pour du contenu non synthétique.

L'utilité d'une technique est limitée si sa mise en œuvre peut être exploitée par des acteurs malveillants. Par exemple, un auteur de menaces pourrait affirmer une provenance fausse ou usurpée. De plus, il est facile de perdre la provenance après que des modifications mineures ou apparemment bénignes ont été apportées au contenu en raison de problèmes liés à la robustesse, ou la technique elle-même pourrait compromettre la protection des renseignements personnels du créateur ou de l'utilisateur. Si un filigrane numérique est robuste, il sera également difficile de le retirer, et les renseignements qu'il contient pourraient être suivis par différentes entités sans le consentement des personnes dont les renseignements personnels y figurent[16]. Il peut également s'agir d'un problème plus important en matière de protection des renseignements personnels avec les métadonnées, car celles-ci contiennent habituellement plus de renseignements sur leur créateur et le contenu, y compris les modifications.

Sous-thème potentiel : Cartographie de l'écosystème des normes et protocoles actuels d'authentification du contenu, y compris les interactions entre les outils, les systèmes, les plateformes, les administrations et les vulnérabilités potentielles en matière de sécurité et de protection des renseignements personnels et de sécurité au niveau de l'écosystème

Les problèmes de sécurité, de robustesse et de protection des renseignements personnels liés à la mise en œuvre des méthodes actuelles d'authentification du contenu se manifestent souvent au niveau plus général de l'écosystème, en raison de la création et de la diffusion du contenu numérique sur différents matériel informatique, logiciels, plateformes et services en ligne. Parmi les nouveaux protocoles d'authentification et d'étiquetage du contenu, on retrouve :

- l'étiquette d'attribution de preuve sécurisée,
- une spécification bien connue de la norme C2PA et entièrement ouverte autorisée sous licence,

[15] Dans les cas d'accessibilité, il faut également tenir compte des contextes culturels et multilingues.
[16] Center for Democracy and Technology, "Privacy Principles for Digital Watermarking", 2 juin 2008.

- le protocole des numéros,
- une solution de chaîne de blocs décentralisée qui utilise également la norme C2PA,
- la norme du Code international normalisé de contenu sur les identificateurs de contenu (une empreinte digitale),
- l'utilisation d'autres techniques comme les filigranes numériques.

L'adoption de ces normes est encore naissante, mais elle ne cesse de croître.

L'évaluation de la mise en œuvre des normes émergentes sur les filigranes numériques, qu'ils soient visibles ou invisibles, et des zones de vulnérabilité en matière de sécurité et de protection des renseignements personnels au sein de l'écosystème peut aider les organisations à corriger le tir, à améliorer leur mise en œuvre et à coordonner la résolution de ces problèmes. Elle peut également éclairer la conception de protocoles améliorés, ce qui constitue en soi une voie de recherche prometteuse.

Sous-thème potentiel : Améliorer les critères de référence pour mettre à l'essai la suppression, l'altération et la falsification des filigranes numériques par des acteurs malveillants selon différentes modalités de contenu

L'un des principaux problèmes liés aux filigranes numériques est leur robustesse et leur sécurité face à des modifications d'apparence bénigne ou à des tentatives malveillantes de suppression. Même les meilleurs filigranes numériques conçus pour résister à des catégories précises de perturbations peuvent présenter des vulnérabilités face à des attaques malveillantes. Les filigranes numériques, même s'ils sont appliqués de manière apparente sur de petites portions de contenu synthétique ou non synthétique, peuvent être facilement supprimés, parfois même par accident. Il est souvent tout aussi simple de supprimer délibérément des filigranes numériques dissimulés si leur application n'est pas rigoureuse. D'autres recherches comparatives dans ce domaine s'imposent afin de renforcer la sécurité des filigranes numériques sur différents types de contenu (image, audio, vidéo, texte).

Une autre piste à explorer serait de déterminer s'il est possible de mettre en place un suivi des modifications apportées au contenu à l'aide de filigranes numériques. Autrement dit, en cas de modification d'un contenu marqué d'un filigrane numérique, comment cela doit-il se refléter dans le tatouage numérique?

Sous-thème potentiel : Améliorer l'évaluation des techniques de transparence du contenu numérique

À l'heure actuelle, on peut noter un manque d'évaluations et de points de repère systématiques, réalistes et normalisés pour diverses techniques de transparence du

contenu numérique, y compris le suivi des données de provenance ainsi que les méthodes de détection. Ces lacunes soulignent la nécessité d'améliorer la compréhension du rendement des mises en œuvre actuelles en matière de sécurité, de confidentialité, de fiabilité, d'interopérabilité et d'accessibilité. Des critères publics pour des techniques précises peuvent être développés pour faciliter l'alignement des mises en œuvre avec des normes particulières et reconnues.

Comment pouvons-nous développer des ensembles de données de référence réalistes pour mieux refléter la manière dont le contenu généré par l'IA apparaît dans le monde réel? En plus des mesures de rendement normalisées, quelles caractéristiques un cadre d'évaluation de grande qualité devrait-il mettre à l'essai (p. ex. la fiabilité, utilité, transparence, etc.)? Comment pouvons-nous accroître le nombre d'évaluations interlinguistiques?

3. Quels sont les mécanismes de propagation du contenu synthétique dans l'environnement d'information et les effets systémiques connexes (p. ex. en ce qui concerne la confiance du public)? Comment les techniques de transparence du contenu numérique sont-elles adoptées et utilisées dans le monde réel et quels sont leurs effets systémiques potentiels?

Le milieu de la recherche doit se concentrer sur les enquêtes portant sur les répercussions systémiques plus larges des systèmes d'IA et de leurs résultats[17]. Compte tenu de l'adoption massive et transfrontalière des systèmes d'IA, qui ont des répercussions sur la vie de milliards d'utilisateurs, nous devons être attentifs aux effets secondaires et aux externalités négatives, notamment les effets sur la confiance publique découlant de la production et de la distribution généralisées de contenu synthétique. Par ailleurs, il est crucial d'analyser les effets de l'adoption de diverses techniques de transparence du contenu afin de garantir que tous les risques involontaires introduits par des ensembles particuliers de mesures d'atténuation sont également pris en compte.

Sous-thème potentiel : L'effet du contenu synthétique sur les écosystèmes d'information mondiaux et la confiance du public, et ce qui peut être abordé par des techniques de transparence du contenu

Le contenu synthétique se propage à grande vitesse sur Internet, et nous ne mesurons pas encore très bien les répercussions que cela peut avoir sur la confiance du public à l'égard de l'information. Il serait donc pertinent d'élaborer un cadre de classement des différents types de contenu synthétique et de leurs diverses répercussions sur la confiance du public, et de déterminer les modèles de menace qui peuvent être abordés à l'aide de techniques de transparence du contenu et de leurs intervenants respectifs. Cela pourrait également éclairer les types de nouvelles mesures d'atténuation et les interventions qui permettraient de renforcer la confiance du public.

Sous-thème potentiel : Adoption et convivialité des techniques de transparence du contenu

Les techniques de transparence du contenu numérique ne peuvent être efficaces que si elles sont adoptées par une grande partie de la population. Il est donc essentiel de comprendre comment elles sont actuellement utilisées « dans le monde réel » et les facteurs qui pourraient freiner leur adoption : Comment les utilisateurs perçoivent-ils la divulgation de contenu et y réagissent-ils? Peuvent-ils raisonner en fonction des répercussions? Quels renseignements sur la provenance sont jugés utiles et à qui doivent-ils être visibles? En quoi cela influence-t-il le comportement et le niveau de confiance des

---

[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 et W Isaac, "Sociotechnical Safety Evaluations of Generative AI systems", Google Deepmind, 18 octobre 2023.

utilisateurs? Dans quelle mesure l'efficacité des outils dépend-elle du contexte, de la culture et d'autres facteurs? Pouvons-nous anticiper leur fonctionnement avant leur déploiement à grande échelle? À quel moment devrait-il être possible de se retirer?

Sous-thème potentiel : Répercussions de l'adoption de techniques et de mesures d'atténuation pour le contenu synthétique sur les écosystèmes mondiaux de l'information et des médias, notamment dans la majorité du globe

Les répercussions des techniques de transparence du contenu restent encore à déterminer. Par exemple, la mise en œuvre de certaines mesures d'atténuation pourrait creuser les fossés numériques préexistants entre les populations et certains groupes démographiques. Des études préliminaires menées par des organisations de la société civile, comme WITNESS, ont permis de cerner des cas où la mise en œuvre de l'authentification des contenus pourrait enfreindre les droits de la personne ou avoir des répercussions négatives sur ces droits[18]. Il est essentiel de mener des recherches empiriques approfondies pour comprendre les répercussions des techniques et des mesures d'atténuation, tant positives que négatives, sur les écosystèmes médiatiques qui fonctionnent à l'extérieur de l'hémisphère Nord.

Sous-thème potentiel : Validation du contenu réel et gestion du spectre entre le contenu original et celui généré par l'IA

Les techniques de transparence du contenu peuvent être utilisées pour authentifier le contenu réel plutôt que pour signaler le contenu généré par l'IA; il serait donc utile de comprendre les répercussions de ce paradigme d'étiquetage et les faux positifs potentiels. Ces systèmes peuvent-ils être utilisés à des fins antagonistes, p. ex. pour discréditer du contenu humain réel? Quel niveau de modification de l'IA fait passer l'étiquette de « réel » à « généré par l'IA »?

4. Quelles approches techniques ou non techniques peuvent être utilisées pour appuyer ou faire progresser les techniques de transparence du contenu numérique?

Il existe des approches émergentes qui peuvent contribuer à l'amélioration des efforts d'atténuation des risques, mais elles sont encore à un stade préliminaire par rapport aux méthodes actuelles d'authentification et de vérification de l'origine du contenu. Le développement de ces méthodes est crucial, puisque les capacités d'IA se perfectionnent,

[18] J Castellanos et S Gregory, "WITNESS and the C2PA Harms and Misuse Assessment Process", *WITNESS Blog*, 2 decembre 2021.

que l'utilisation antagoniste des modèles d'IA devient plus sophistiquée et que le contenu synthétique lui-même devient plus complexe.

Sous-thème potentiel : Concevoir des filigranes numériques textuels qui résistent à la traduction

Les filigranes numériques textuels insérés par les méthodes existantes ne résistent pas à la traduction d'une langue à l'autre[19]. Ce manque de « cohérence interlinguistique » pose un défi sur le plan de la robustesse dans un contexte aussi bien bénin qu'antagoniste. D'autres recherches dans ce domaine sont nécessaires pour assurer la robustesse et la fiabilité de ces filigranes pendant la traduction.

Sous-thème potentiel : Élaborer un indice et améliorer les techniques de détection basées sur le contenu

Les premières recherches suggèrent que les essais rédigés en anglais par des apprenants de langue seconde sont plus susceptibles d'être signalés comme générés par l'IA par les outils de détection[20]. Des recherches ont également mis en évidence des performances hétérogènes des IA génératives à travers plusieurs langues et cultures. Des recherches supplémentaires sont nécessaires pour mieux comprendre, mesurer et améliorer la performances des techniques de détection des contenus synthétiques, y compris l'atténuation de l'occurrence systématique d'incidents de faux positifs.

Sous-thème potentiel : Améliorer la compréhension sémantique et l'attribution pour la détection multimodale de contenu synthétique

Les méthodes actuelles de détection du contenu synthétique sont souvent insuffisantes dans les cas d'utilisation pratique. Il est donc nécessaire de déployer des efforts avancés intégrant d'autres types de renseignements, comme la prise en compte de renseignements sur les réseaux de comptes, les activités des utilisateurs et des modèles apparaissant dans différentes sources de contenu. Cette approche doit aller au-delà d'une méthode unique fondée sur le contenu pour détecter le contenu synthétique. Un autre défi consiste à trouver une solution aux réseaux sophistiqués d'acteurs antagonistes en utilisant à la fois la génération de contenu synthétique et non synthétique. Il sera essentiel de mettre au point des méthodes de détection qui intègrent ces signaux d'attribution plus larges pour améliorer la robustesse par rapport aux tactiques malveillantes avancées.

---

[19] Z He, B Zhou, H Hao, A Liu, X Wang, Z Tu, Z Zhang et R Wang, "Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models", 4 juin 2024.

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

Sous-thème potentiel : Collaboration humaine avec les outils de détection de l'IA, leurs résultats et l'interprétation des résultats

L'interaction entre les humains et les outils de détection de l'IA dans certains contextes à risque élevé, ainsi que l'interprétation subséquente des résultats pour prendre d'autres mesures ont été des domaines sous-étudiés. Ce travail est effectué en partie au sein de la société civile, incluant les droits de la personne et des organisations journalistiques. Par exemple, le groupe de travail d'intervention rapide sur les hypertrucages de WITNESS fait appel à des vérificateurs de faits et des journalistes dans le processus de détection du contenu généré par l'IA. Il faudra approfondir la recherche sur les approches hybrides qui évaluent les interactions humaines avec des outils de détection pour la vérification des faits, le suivi de la provenance et l'atténuation des identités préjudiciables, entre autres applications, afin d'examiner la diffusion et les répercussions de ces outils dans la société.

Sous-thème potentiel : Améliorer et développer des techniques DCT novatrices qui renforcent la performance et la sécurité des implémentations et des normes actuelles

Les problèmes de sécurité, de confidentialité, de robustesse et d'efficacité liés à la mise en œuvre des techniques DCT actuelles sont principalement dus à des faiblesses des algorithmes sous-jacents, tels que les algorithmes de hachage perceptuel, de tatouage numérique et d'authentification de contenu. L'identification des faiblesses fondamentales en matière de sécurité, de confidentialité et d'efficacité peut contribuer à améliorer les techniques DCT existantes et à développer des techniques novatrices pour surmonter ces problèmes.

Glossaire

Techniques de transparence du contenu numérique : méthodes visant à faciliter l'accès et l'exposition à l'information sur l'origine ou l'historique du contenu numérique.

Filigrane numérique : technique permettant d'intégrer de l'information dans un contenu (image, texte, audio, vidéo) tout en la rendant difficile à supprimer. Ce filigrane peut aider à vérifier l'authenticité du contenu ou les caractéristiques de sa provenance, de ses modifications ou de son moyen de transmission[21].

Métadonnées : informations décrivant les caractéristiques des données[22].

Authentification du contenu : utilisation de méthodes de suivi des données de provenance (méthodes techniques permettant de suivre l'origine ou l'historique du contenu, y compris le filigrane, les métadonnées et les empreintes numériques) afin de déterminer la nature du contenu[23].

Suivi des données de provenance : techniques permettant de consigner l'origine et l'historique du contenu[24].

Sécurité **:** protéger les techniques de transparence du contenu numérique et les systèmes utilisés pour exécuter ces techniques contre l'accès, l'utilisation, la divulgation, la falsification, l'usurpation, la perturbation ou la destruction non autorisés[25].

Fiabilité : veiller à ce que les techniques de transparence du contenu numérique soient fiables, résistent aux modifications et manipulations bénignes, puissent être utilisées efficacement pour identifier l'origine du contenu numérique, et puissent être appliquées et améliorées en fonction de l'état actuel des connaissances techniques et en préservant l'intégrité du contenu.

Divulgation : fournir aux utilisateurs des informations sur la façon dont le contenu a été créé, modifié ou publié, ainsi que des renseignements sur la façon dont les techniques et les mesures d'atténuation pour le contenu synthétique sont appliquées.

---

[21] Chandra, Dunietz et Roberts, "Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency", *NIST AI*, 20 novembre 2024 .

[22] C Johnson, M Badger, D Waltermire, J Snyder et C Skorupka, "Guide to Cyber Threat Information Sharing", NIST Computer Security Resource Center, 4 octobre 2016.

[23] Chandra, Dunietz et Roberts, "Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency", NIST AI, 20 novembre 2024 .

[24] Ibid.

[25] *NIST Computer Security Resource Centre*, "Infosec Glossary", n.d.

Accessibilité : offrir aux personnes, aux organisations et aux populations, en particulier à celles qui disposent de moins de ressources, la possibilité d'obtenir les avantages de la transparence du contenu numérique. Des techniques appliquées de façon inégale dans le monde pourraient renforcer les disparités numériques existantes ou en produire de nouvelles.