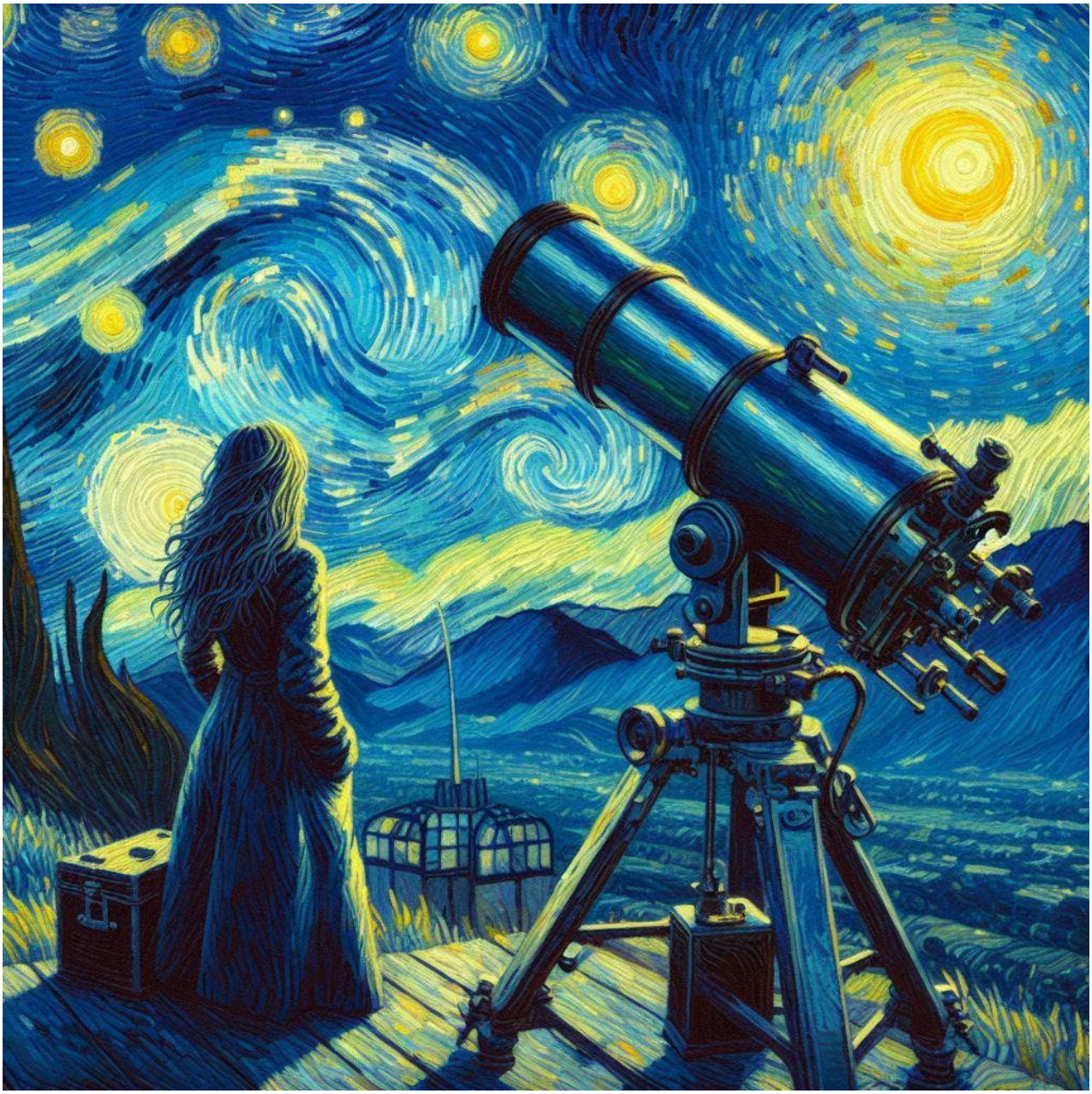
Combattre les hypertrucages dans l'IA générative



(Image générée par DALL-E)

Jeremy Marks et Carolina Cohoon

Introduction

Les Canadiens vivent dans un paysage numérique où la frontière entre la vérité et la fiction est souvent floue. Des programmes comme ChatGPT, DALL-E, développé par OpenAI, et Bing AI développé par Microsoft avec une version personnalisée de ChatGPT d'Open AI peuvent être utilisés en conjonction avec d'autres outils pour améliorer le réalisme des « hypertrucages », ou une « copie numérique hyperréaliste d'une personne qui peut être manipulée pour faire ou dire n'importe quoi ». Ces images et vidéos frauduleuses mettent les spectateurs au défi de distinguer la réalité de l'illusion; elles sont également de plus en plus courantes (« Deepfake Crime by the Numbers, » 2023; Fallis, 2020).

Les hypertrucages posent des défis importants aux formatrices et formateurs et aux personnes apprenantes. Même si les falsifications et les tromperies ne sont pas nouvelles, la sophistication des hypertrucages est surprenante, et les informaticiens mettent en garde contre les dangers que l'IA générative fait peser sur la vérification des faits et sur la confiance du public dans la couverture médiatique des événements mondiaux *Deepfakes, Explained|MIT Sloan*, 2020. Les hypertrucages exposent également les gens à des escroqueries et à des vols d'identité. Des acteurs malveillants utilisent des images et des voix empruntées à des membres de la famille et à des amis pour inciter les utilisateurs à partager des informations personnelles et privées. (Takruri, 2023; U.S. Department of Homeland Security, n.d.).

Dans ce monde où la tromperie numérique pilotée par l'IA est de plus en plus fréquente, nous nous tournons vers les formatrices et formateurs pour qu'ils apprennent aux personnes apprenantes à utiliser des techniques de vérification des faits numériques afin de leur donner les moyens de distinguer les faits des falsifications. Les formatrices et formateurs ayant d'excellentes compétences en matière de vérification des faits peuvent disséquer les hypertrucages pour les personnes apprenantes, en leur montrant comment appliquer des techniques de recherche efficaces pour révéler la falsification et les trucages. L'enseignement de la vérification des faits est devenu un devoir éducatif de la plus haute importance, car « les apparences peuvent être trompeuses » en ce qui concerne les contenus générés par l'IA. Chaque utilisateur du monde numérique doit renforcer sa capacité à penser de manière critique et à remettre en question ce qu'il voit.

Les formateurs et formatrices qui enseignent aux personnes apprenantes à vérifier les hypertrucages remplissent également un devoir démocratique. Le Canada, comme d'autres sociétés ouvertes, ne peut survivre que si les citoyens savent distinguer la vérité et les faits de la tromperie et du mensonge. Les Canadiens doivent apprendre à distinguer les images factuelles des fausses impressions et éviter d'être trompés et de faire de mauvais choix civiques. Étant donné que la plupart des Canadiens apprennent le lien entre les faits, l'intégrité et la démocratie à l'école, il est logique que les formateurs et formatrices soient en première ligne dans la lutte contre les hypertrucages. Et plus nos vies se déplacent en ligne, plus cette responsabilité éducative et civique s'accroît.

Argument

Dans une société libre comme le Canada, les formatrices et formateurs doivent enseigner aux personnes apprenantes à distinguer les faits des mensonges. Instiller l'engagement à vérifier les faits est directement lié à l'intégrité académique et à l'honnêteté intellectuelle (Eaton et Hughes, 2022). Lorsque les personnes apprenantes acceptent l'obligation de vérifier l'exactitude des informations, elles s'engagent non seulement à dire la vérité dans leur travail, mais aussi à se prémunir contre la tromperie et la déception.

Le philosophe allemand Karl Popper affirmait que « tout est critiquable ». Il estimait que, dans l'enseignement et la recherche, « chaque source - tradition, raison, imagination, observation... est admissible et peut être utilisée, mais aucune n'a d'autorité ». Pour qu'une source ait de l'autorité, les mensonges doivent être réfutés (Popper & Gombrich, 2013, 493-494). Le point de vue de Popper est pertinent pour les hypertrucages, car il part du principe qu'il faut d'abord douter pour croire. Étant donné que des personnes conçoivent des images et des vidéos hypertruquées pour manipuler les spectateurs afin qu'ils croient à des mensonges ou qu'ils perdent leur capacité à distinguer le vrai du faux, les spectateurs doivent apprendre des techniques pour se protéger de la tromperie.

Une bonne technique de vérification des faits exige des personnes apprenantes qu'elles soumettent ce qu'elles voient à un questionnement et à une recherche rigoureuse. Bien que nous puissions prendre plaisir à regarder une vidéo hypertruquée d'un politicien qui en aurait traité un autre d'un nom peu flatteur, ou à voir une célébrité se ridiculiser, nous ne devrions pas croire ce que nous

voyons simplement parce que cela nous fait plaisir (Sample, 2023; Petkauskas, 2021; BuzzFeed Video, 2018). Les personnes apprenantes doivent aborder les images et les vidéos avec scepticisme et développer une stratégie de dissection pour détecter les signes de falsification. Cette approche consistant à tenter de réfuter quelque chose avant de l'accepter est liée à ce que Popper appelait la « falsifiabilité », ou la croyance selon laquelle « nous devrions essayer autant que possible de renverser notre solution, plutôt que de la défendre » (Popper, 2005).

Si travailler à réfuter un hypertrucage est un exercice nécessaire à la vérification des sources, il s'agit également d'une pratique épistémologique. Lorsque nous vérifions les faits, nous cherchons à déterminer ce qui est réel (Amazeen, 2015). Les acteurs malveillants utilisent des textes, des images et des vidéos falsifiés à des fins de propagande pour provoquer l'instabilité, la peur, la panique et la violence. Nous en avons été témoins avec le déni des élections aux États-Unis et l'utilisation d'émissions de propagande pour inciter au meurtre de masse pendant le génocide rwandais de 1994 (Institute of Strategic Dialogue et al., 2023; Maximino, 2014). Les formatrices et formateurs et les personnes apprenantes font leur devoir démocratique lorsqu'ils maîtrisent la vérification des faits pour dénoncer les hypertrucages comme de la propagande et de la manipulation (Parkin, 2019).

Les hypertrucages sont également dangereux sur le plan intellectuel et politique parce qu'ils sont visuels. Comme l'affirment les chercheurs, « les médias visuels se voient accorder une grande confiance de la part du grand public » depuis l'arrivée de la photographie au XIX^{e} siècle (Langguth et al., 2021). En 1961, l'historien américain Daniel J. Boorstin s'est attaqué à la capacité des images à tromper en décrivant ce qu'il a appelé les « pseudo-événements », ou l'idée que si un journaliste ne peut pas « trouver une histoire, alors il doit en fabriquer une »). Les pseudo-événements étaient conçus pour inciter les lecteurs à acheter les journaux, ce qui a amené Boorstin à faire la remarque suivante : « Nous nous sommes tellement habitués à nos illusions que nous les prenons pour la réalité ». Il a également reconnu le pouvoir émotionnel de l'image photographique : « Chacun d'entre nous fournit le marché et la demande pour les illusions qui inondent notre expérience [...]. Nous voulons et croyons à ces illusions parce que nous souffrons d'attentes extravagantes. Nous exigeons trop du monde » (Boorstin, 2012).

Les formatrices et formateurs doivent aborder le lien cognitif et émotionnel entre les « pseudo-événements » de Boorstin et l'appétit humain pour l'illusion sous la forme de l'hypertrucage. Les gens génèrent des hypertrucages pour alimenter les préjugés, attirer l'attention à des fins de profit personnel et créer la controverse (Rasser, 2019). La sophistication croissante de la technologie de l'IA rend aussi plus difficile pour les formatrices et formateurs et les personnes apprenantes (ainsi que pour les gouvernements) de distinguer la vérité du mensonge (Nishimura, A., 2023; Goethe, 2019). Un exemple récent est la controverse qui a entouré une photo d'hypertrucage d'un diplomate américain qui aurait abattu un markhor, l'animal national du Pakistan. Cette fausse photo avait pour but de susciter une controverse diplomatique et a dû être publiquement démystifiée (Collins, 2023).

Le danger est d'autant plus grand que nous vivons à une époque où tout le monde passe de plus en plus de temps en ligne, s'exposant ainsi facilement aux hypertrucages. Les gouvernements, les entreprises et les établissements d'enseignement attendent des formateurs et formatrices et des personnes apprenantes qu'ils deviennent des utilisateurs fluides des technologies numériques les plus récentes (et émergentes), y compris l'IA générative, afin de se préparer à l'emploi sur un marché mondial (Fan, 2021). De plus, si l'Organisation des Nations Unies pour l'éducation, la science et la culture (UNESCO) a encouragé les gouvernements à « maîtriser l'IA dans les écoles afin de protéger les élèves et les enseignantes et enseignants », on ne sait pas très bien ce que cela signifie dans la pratique (Morrison, 2023).

Les formatrices et formateurs d'adultes sont en mesure d'aider les personnes apprenantes à devenir des utilisateurs critiques et perspicaces de l'IA générative. Mais pour ce faire, ils doivent apprendre l'art de la vérification des faits dans le domaine de l'intelligence artificielle. Les formatrices et formateurs doivent maîtriser une technique de vérification des faits pour enseigner un processus d'application de l'honnêteté, de la factualité et de l'empirisme aux textes, aux images, aux vidéos et aux fichiers audio. Certaines entreprises technologiques, comme Intel, suggèrent que les écoles publiques (et élémentaires) peuvent entamer ce processus en introduisant l'IA pour encourager la préparation numérique des enfants (Teaching AI: Artificial Intelligence in the Classroom, n.d.). Mais la préparation n'est qu'un premier pas vers le développement d'une approche critique de l'information, conforme à l'exigence de Popper selon laquelle la vérité doit posséder une « correspondance avec les faits » (Popper et Gombrich, 2013).

Bien qu'une familiarité générale avec l'IA puisse familiariser les personnes apprenantes avec les falsifications, ce n'est qu'en enseignant et en pratiquant les techniques de vérification des faits que les personnes apprenantes les maitriseront. La vérification des faits est un travail de détection qui nécessite une pratique fréquente. Bien qu'il n'existe pas de méthode unique pour apprendre à vérifier les hypertrucages, voici quelques étapes simples :

1. **Identifier un hypertrucage** : Définir pour les personnes apprenantes ce qu'est un hypertrucage, quelles sont ses caractéristiques reconnaissables et comment les utilisateurs les créent. Discuter de la façon dont ils menacent potentiellement les institutions démocratiques, le sens de la communauté et la sécurité personnelle.
2. **Repérer un hypertrucage** : Apprendre aux personnes apprenantes à rechercher les incohérences de couleur et de lumière dans les photos et les vidéos, y compris les mouvements inhabituels des yeux, les gestes physiques étranges et les sons de mauvaise qualité ou non synchronisés. Donnez-leur l'occasion de s'exercer à utiliser ces techniques pour repérer les hypertrucages.
3. **Faire une démonstration** : Démontrer comment vérifier les faits d'une photo ou d'une vidéo suspecte. Comparez-les avec des photos réelles et des fichiers vidéos/audios vérifiés pour faire des comparaisons éclairées. Illustrer votre démarche et montrer comment utiliser les outils numériques tels que les moteurs de recherche pour faciliter votre investigation.
4. **Pratiquer :** Demandez aux personnes apprenantes de vérifier les faits d'un hypertrucage présumé. Lorsqu'elles ont terminé la tâche, demandez-leur de partager leurs résultats et d'expliquer leur processus d'investigation.
5. **Discuter des conséquences de l'hypertrucage :** Animez des discussions avec les personnes apprenantes sur comment et pourquoi les gens créent des hypertrucages et sur les dangers de l'utilisation de la technologie des hypertrucages à des fins de divertissement et de marketing. Discutez avec elles de ce qu'elles ont trouvé de persuasif et d'émotionnellement puissant dans l'hypertrucage qu'elles ont démystifié.

(Microsoft & OpenAI, 2023; The Social Institute, 2023; *Be MediaWise Lesson 12: How to Detect Deepfakes and Avoid Disinformation*, 2023)

En fin de compte, les meilleurs outils dont disposent les formatrices et formateurs et les personnes apprenantes consistent à développer, à cultiver et à appliquer leur esprit critique et leurs compétences en matière de vérification des faits, et ce dans un environnement de collaboration en personne, hybride ou en classe virtuelle.

Conclusion

Pour le moment, il n'existe pas de solution logicielle au problème des hypertrucages. Mais comme par le passé, les êtres humains doivent s'appuyer sur leur engagement et leur capacité de raisonnement et cultiver leur aptitude à remettre en question ce qu'ils voient. Cela me fait penser à la célèbre question du philosophe français Michel Montaine : « Qu'est-ce que je sais? » (Vázquez, 2022). Voici ce que les formatrices et formateurs et les personnes apprenantes doivent se demander alors que nous sommes confrontés aux hypertrucages et aux capacités de l'IA générative en général.

Tout comme les formatrices et formateurs ont besoin de savoir que leurs personnes apprenantes produisent un travail factuel et honnête, les gouvernements doivent déterminer que les courriels, les lettres et les sondages d'opinion qu'ils reçoivent de leurs électeurs sont authentiques (Kreps et Kriner, 2023; Israel, 2023). Comme les écoles, les gouvernements doivent embaucher des personnes qui partagent l'engagement de Karl Popper en faveur de la falsifiabilité et qui abordent l'information avec un œil critique. Il n'est pas exagéré de dire que le sort de notre société démocratique libre repose sur ce type d'autodiscipline intellectuelle dans les écoles, les entreprises et l'État (Harloe, 2013).

Le théoricien de l'éducation Paolo Freire a écrit un jour : « En tant qu'enseignant, je ne peux pas aider les étudiants à surmonter leur ignorance si je ne suis pas engagé en permanence à essayer de surmonter la mienne » (Freire, 1998). La tâche de Freire est celle à laquelle les formatrices et formateurs sont confrontés aujourd'hui. Nous devons montrer à nos personnes apprenantes comment distinguer les informations numériques exactes de celles qui sont trompeuses. Mais pour ce faire, nous devons d'abord devenir nous-mêmes des vérificateurs de faits compétents.

Sources

Amazeen, M. A. (2015). Revisiting the Epistemology of Fact-Checking. *Critical Review*, *27*(1).

Be MediaWise lesson 12: How to detect deepfakes and avoid disinformation. (2023, 24 octobre). PBS Newshour Classroom. Consulté le 6 novembre 2023, sur le site : https://www.pbs.org/newshour/classroom/lesson-plans/2022/12/lesson-plan-how-to-detect-deepfakes-to-ensure-you-dont-fall-for-disinformation

Bhuiyan, J. (2023, 31 octobre). 'Is this an appropriate use of AI or not?': teachers say classrooms are now AI testing labs. *The Guardian*. Consulté le 4 novembre 2023, sur le site : https://www.theguardian.com/technology/2023/oct/31/educators-teachers-ai-learning-classrooms-misuse

BuzzFeed Video. (2018). *You Won't Believe What Obama Says In This Video!* [Vidéo]. YouTube. Consulté le 4 novembre 2023, sur le site : https://www.youtube.com/watch?v=cQ54GDm1eL0

Collins, D. (2023). Deepfake Trophy Hunting Photo of U.S. Diplomat Foreshadows a Troubling Future. *Outdoor Life*. Consulté le 6 novembre 2023, sur le site : https://www.outdoorlife.com/hunting/deepfake-hunting-photo-us-diplomat/

Deepfake Crime By The Numbers. (2023, 30 juin). *EDsmart*. Consulté le 6 novembre 2023, sur le site : https://www.edsmart.org/deepfake-statistics/

Dickson, B. (2020, 4 mars). What Is a Deepfake? *PCMAG*. Consulté le 5 novembre 2023, sur le site : https://www.pcmag.com/news/what-is-a-deepfake

Eaton, S. E., & Hughes, J. C. (2022). Academic Integrity in Canada: Historical Perspectives and Current Trends. In *Academic Integrity in Canada: An Enduring and Essential Challenge* (Vol. 1). Springer. https://link.springer.com/chapter/10.1007/978-3-030-83255-1_1

Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy and Technology*, 623–643. https://doi.org/10.1007/s13347-020-00419-2

Fan, Z. (2021, 27 octobre). How To Prepare Students For The Jobs Of 2030. *Forbes*. Consulté le 4 novembre 2023, sur le site : https://www.forbes.com/sites/forbestechcouncil/2021/10/27/how-to-prepare-students-for-the-jobs-of-2030/?sh=570c4d414ad3

Farmer, B. M. (2021, octobre). The impact of deepfakes: How do you know when a video is real? - 60 Minutes. *CBS News*. Consulté le 4 novembre 2023, du site : https://www.cbsnews.com/news/deepfakes-real-fake-videos-60-minutes-2021-10-10/

Freire, P. (1998). *Pedagogy of Freedom: Ethics, Democracy, and Civic Courage*. Rowman & Littlefield.

Goethe, T. S. (2019, 26 avril). *War, propaganda and Misinformation: the evolution of Fake news*. R. Consulté le 4 novembre 2023, sur le site : https://reporter.rit.edu/features/war-propaganda-and-misinformation-evolution-fake-news

Hagan, E. (2021, 8 octobre). Deepfakes Can Be Used to Hack the Human Mind: Digital copies of persons are easy to manipulate and can easily impact viewers. *Psychology Today*. Consulté le 5 novembre 2023, sur le site : https://www.psychologytoday.com/us/blog/spontaneous-thoughts/202110/deepfakes-can-be-used-hack-the-human-mind

Harloe, K. (2013). Questioning the Democratic, and Democratic Questioning [Digital]. In *Classics in the Modern World: A Democratic Turn?* (pp. 3–14). Oxford University Press. https://shc.stanford.edu/arcade/interventions/questioning-democratic-and-democratic-questioning

Helmus, T. C., & Marcellino, W. (2023, 31 octobre). Lies, Misinformation Play Key Role in Israel-Hamas Fight. *The RAND Blog*. Consulté le 4 novembre 2023, sur le site : https://www.rand.org/blog/2023/10/lies-misinformation-play-key-role-in-israel-hamas-fight.html

Institute of Strategic Dialogue, Craig, J., Simmons, C., & Bhatnagar, R. (2023, janvier). *How January 6 inspired election disinformation around the world - ISD*. ISD. Consulté le 6 novembre 2023, sur le site : https://www.isdglobal.org/digital_dispatches/how-january-6-inspired-election-disinformation-around-the-world/

Israel, S. (2023, 30 mars). Here's What Happened When ChatGPT Wrote to Elected Politicians. *The New Republic*. Consulté le 8 novembre 2023, sur le site : https://newrepublic.com/article/171459/chatgpt-ai-cornell-experiment-politicians-constituents-emails

Kreps, S., & Kriner, D. (2023). How AI Threatens Democracy | Journal of Democracy. *Journal of Democracy*, *34*(4), 122–131. https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication*, *6*, 1–12. https://doi.org/10.3389/fcomm.2021.632317

Mahadevan, A. (2023, 23 octobre). *As misinformation surges during the Israel-Hamas war, where is AI?* Poynter. Consulté le 4 novembre 2023, sur le site : https://www.poynter.org/fact-checking/2023/israel-hamas-war-artificial-intelligence-misinformation-fake-images/

Maximino, M. (2014, 3 décembre). *Propaganda, media effects and conflict: Evidence from the Rwandan genocide - The Journalist's Resource*. The Journalist's Resource. Consulté le 4 novembre 2023, sur le site : https://journalistsresource.org/politics-and-government/propaganda-conflict-evidence-rwandan-genocide/

Milmo, D. (2023, 25 octobre). AI-created child sexual abuse images 'threaten to overwhelm internet.' *The Guardian*. Consulté le 4 novembre 2023, sur le site : https://www.theguardian.com/technology/2023/oct/25/ai-created-child-sexual-abuse-images-threaten-overwhelm-internet

Morrison, N. (2023, 7 septembre). Governments Urged To Get A Grip On AI In Schools. *Forbes*. Consulté le 4 novembre 2023, sur le site : https://www.forbes.com/sites/nickmorrison/2023/09/07/governments-urged-to-get-a-grip-on-ai-in-schools/?sh=165dfd67302b

Nishimura, A. (2023, 2 novembre). *Human subjects protection in the era of Deepfakes*. Lawfare. Consulté le 4 novembre 2023 sur le site : https://www.lawfaremedia.org/article/human-subjects-protection-in-the-era-of-deepfakes

OpenAI (2023). Bing Chat Enterprise (version du 6 novembre) [Grand modèle de langage]. https://www.bing.com/?PC=ER02

Parkin, S. (2019, 22 juin). The rise of the deepfake and the threat to democracy: The rise of the deepfake and the threat to democracy. *The Guardian*. Consulté le 6 novembre 2023 sur le site : https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy

Petkauskas, V. (2021, 28 septembre). Report: number of expert-crafted video deepfakes double every six months. *Cyber News*. Consulté le 4 novembre 2023 sur le site : https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months/

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Popper, K. R., & Gombrich, E. H. (2013). *The Open Society and Its Enemies*. Princeton University Press.

Rasser, M. (2019, 14 août). Why Are Deepfakes So Effective?: It's because we often want them to be true. *Scientific American*. Consulté le 5 novembre 2023, sur le site : https://blogs.scientificamerican.com/observations/why-are-deepfakes-so-effective/

Saenz, A., & Liptak, K. (2023, 30 octobre). White House tackles artificial intelligence with new executive order. *CNN.com*. Consulté le 4 novembre 2023 sur le site : https://www.cnn.com/2023/10/30/politics/white-house-tackles-artificial-intelligence-with-new-executive-order/index.html

Sample, I. (2023, 28 octobre). What are deepfakes – and how can you spot them? *The Guardian*. Consulté le 4 novembre 2023 sur le site : https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

Somers, M. (2020, 21 juillet). *Deepfakes, explained | MIT Sloan*. MIT Sloan. Consulté le 6 novembre 2023 sur le site : https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Takruri, L. (2023, 26 octobre). What are deepfakes and how do fraudsters use them? *Onfido*. Consulté le 6 novembre 2023 sur le site : https://onfido.com/fr//

Teaching AI: Artificial Intelligence in the Classroom: Learn strategies for teaching AI in K–12 classrooms, and prepare your students for the future. (n.d.). Intel.com.

The Dangers of Manipulated Media and Video: Deepfakes and More, ADL. (2023, 6 juin). *The Dangers of Manipulated Media and Video: Deepfakes and More | ADL*. ADL. Consulté le 4 novembre 2023 sur le site : https://www.adl.org/resources/blog/dangers-manipulated-media-and-video-deepfakes-and-more The Social Institute. (2023, 17 mars). *Understanding Deepfakes: Equip students to spot misinformation on social media | The Social Institute*. Consulté le 6 novembre 2023 sur le site : https://thesocialinstitute.com/blog/understanding-deepfakes-equip-students-to-spot-misinformation-on-social-media/

U.S. Department of Homeland Security. (n.d.). *The Increasing Threat of Deepfake Identities*. Consulté le 6 novembre 2023 sur le site : https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

COALITION ONTARIENNE DE FORMATION DES ADULTES

L'adaptation en français a été effectuée grâce au Fonds de traduction
de la Coalition ontarienne de formation des adultes (COFA) qui reçoit un financement du ministère
du Travail, de l'Immigration, de la Formation et du Développement des compétences.