Service canadien du renseignement de sécurité

Canadian Security Intelligence Service

Évolution de la désinformation :
UN AVENIR « HYPERTRUQUÉ »

Canada

Publication nº 2023-10-01 de la série *Regards sur le monde : avis d'experts*

Also available in English under the title: *The Evolution of Disinformation: A Deepfake Future*

Le présent rapport est fondé sur les opinions exprimées et les courts articles offerts par les conférenciers à l'occasion d'un atelier organisé par le Service canadien du renseignement de sécurité dans le cadre de son programme de liaison-recherche et de la collaboration avec les intervenants (LRCI) et la Direction de l'analyse et de l'exploitation des sources d'information (AXSI). Le présent rapport est diffusé pour nourrir les discussions. **Il ne s'agit pas d'un document analytique et il ne représente la position officielle d'aucun des organismes participants.** L'atelier s'est déroulé conformément à la règle de Chatham House; les intervenants ne sont donc pas cités, et les noms des conférenciers et des participants ne sont pas révélés.

www.canada.ca

Publié en octobre 2023
Imprimé au Canada

Nº de cat. PS74-19/2023F-PDF
ISBN : 978-0-660-49646-7

Évolution de la désinformation
Un avenir « hypertruqué »

Points saillants d'un atelier non classifié de la Direction de la liaison-recherche et de la collaboration avec les intervenants (LRCI) et la Direction de l'analyse et de l'exploitation des sources d'information (AXSI)

Le 24 mai 2023, Ottawa

Table des matières

L'atelier et ses objectifs

Le 24 mai 2023, la Direction de la liaison-recherche et de la collaboration avec les intervenants (LRCI) et la Direction de l'analyse et de l'exploitation des sources d'information (AXSI) du Service canadien du renseignement de sécurité (SCRS) ont tenu un atelier lors duquel les participants se pencheront sur l'ensemble complexe de menaces que représentent les technologies d'hypertrucage dans les campagnes de désinformation.

L'atelier, qui s'est déroulé selon la règle de Chatham House, était axé sur les travaux de huit éminents spécialistes issus de centres de recherche dans les sources ouvertes, ainsi que sur les observations de professionnels de la sécurité ayant acquis tout un éventail d'expériences au pays comme à l'étranger. Les exposés présentés à l'atelier composent l'essentiel du présent rapport. **Les opinions qui y sont exprimées appartiennent à ces experts indépendants et ne sont pas celles du SCRS.**

La Liaison-recherche et de la collaboration avec les intervenants a pour objectif de favoriser un dialogue entre des professionnels du renseignement et d'éminents experts de différentes disciplines au sein d'universités, de groupes de réflexion, d'entreprises privées ou d'autres établissements de recherche. Il se peut que certains spécialistes invités défendent des idées ou tirent des conclusions qui ne concordent pas avec les points de vue et les analyses du SCRS, et c'est précisément ce qui rend utile la tenue d'un tel dialogue.

Sommaire

Le présent rapport est fondé sur les opinions exprimées et les courts articles offerts par les conférenciers à l'occasion d'un atelier organisé par le Service canadien du renseignement de sécurité dans le cadre de son programme de liaison-recherche et de la collaboration avec les intervenants (LRCI) et la Direction de l'analyse et de l'exploitation des sources d'information (AXSI). Le présent rapport est diffusé pour nourrir les discussions. ***Il ne s'agit pas d'un document analytique et il ne représente la position officielle d'aucun des organismes participants.*** *L'atelier s'est déroulé conformément à la règle de Chatham House; les intervenants ne sont donc pas cités, et les noms des conférenciers et des participants ne sont pas révélés.*

Les menaces pour la sécurité et la démocratie qui sont associées à la désinformation sont jugées importantes et d'actualité, voire permanentes. Nés des progrès de l'intelligence artificielle (IA), les hypertrucages sont considérés comme une version moderne de la désinformation qui pose de nouvelles difficultés aux gouvernements, aux personnes et aux sociétés. Protéger l'intégrité de l'écosystème informationnel est primordial pour la démocratie, mais aussi pour l'ensemble de la société.

Progrès technologiques et applications utiles à la société

Le terme « hypertrucages » (traduction de l'anglais « deepfake », un mot-valise formé des équivalents de « apprentissage profond » et de « faux ») est maintenant utilisé plus largement pour désigner tout contenu mettant en scène des personnes qui a été produit ou modifié au moyen d'algorithmes d'apprentissage profond. La manipulation de vidéos, d'images, d'enregistrements sonores ou vocaux et la création de textes à l'aide de techniques d'IA générative ont rapidement évolué pour gagner en accessibilité et donner des résultats de plus en plus réalistes. À bien des égards, ces progrès ouvrent des possibilités excitantes :

- Les avancées de l'IA générative ont été considérablement favorisées par la disponibilité de grands modèles linguistiques et de création d'images. Elles visaient à augmenter la puissance, les capacités et l'accessibilité de ces modèles et à donner aux utilisateurs davantage de contrôle sur le style et le contenu des productions grâce à des invites textuelles détaillées.

- Les hypertrucages peuvent servir à créer des contenus divertissants, comme des remplacements de visages réalistes ou le doublage de films, d'émissions de télévision ou de jeux vidéo, ce qui ouvre de nouvelles perspectives créatives et permet de donner vie à des personnages de fiction. Ils peuvent aussi servir à réaliser des effets spéciaux et à restaurer des œuvres, par exemple pour recréer ou améliorer des scènes qu'il serait difficile ou coûteux de tourner. Ainsi, ils peuvent être utilisés pour vieillir ou rajeunir des acteurs ou pour ramener à la vie des artistes décédés, le temps d'un film ou d'une publicité.

- À titre d'outils de marketing et d'enseignement (dans le cadre de séances de formation et de simulations), les hypertrucages peuvent mettre en scène des situations réalistes dans des domaines comme la médecine, l'entraînement militaire ou l'intervention en cas d'urgence : les professionnels peuvent ainsi mettre leurs compétences à l'épreuve dans un environnement sûr et contrôlé.

Menaces pour la société et pour la sécurité

Plus la capacité de synthétiser des contenus gagne en accessibilité et en précision, plus la probabilité de détournement augmente. L'une des grandes préoccupations associées aux hypertrucages est la possibilité de répandre de la désinformation et de manipuler le débat politique, ce qui est facteur de confusion, de méfiance et d'instabilité sociale dans les sociétés démocratiques.

- Les hypertrucages posent des questions sérieuses de respect de la vie privée, car ils peuvent être utilisés pour créer des contenus sexuels non consensuels grâce au collage d'un visage sur des vidéos osées. C'est une atteinte à la vie privée et à la réputation de la cible, qui peut aussi souffrir de détresse émotionnelle. De plus, les hypertrucages font naître de nombreux problèmes juridiques et éthiques. Ils peuvent violer le principe de propriété intellectuelle et les lois sur la vie privée.

- En outre, les hypertrucages peuvent ébranler la confiance dans les contenus visuels. Plus la technologie se complexifie, plus il devient difficile de distinguer le contenu original du contenu manipulé, ce qui rend difficile de se fier à des preuves audiovisuelles et exacerbe le problème de la désinformation.

- Parce qu'ils sont faciles d'accès, les hypertrucages peuvent avoir des répercussions négatives sur le plan social, notamment par la cyberintimidation, le harcèlement et d'éventuels troubles sociaux. Ils peuvent être instrumentalisés pour exploiter ou manipuler des personnes, donc entraîner des atteintes à la réputation des victimes, des blessures psychologiques ou des divisions sociales.

Il est plus probable que les hypertrucages fassent évoluer des activités liées à la menace pour la sécurité plutôt qu'ils en engendrent de nouvelles, mais il importe de comprendre les risques qui y sont associés et d'élaborer des solutions technologiques fiables, des lignes directrices sur le plan éthique et des mécanismes juridiques pour remédier aux problèmes qu'ils posent et en atténuer les conséquences.

- Utilisée de tout temps par des acteurs étatiques et non étatiques, la désinformation discrédite les institutions démocratiques et minimise leur rôle, amplifie les théories du complot et la radicalisation, et attise la méfiance envers les autorités. Les hypertrucages en accélèrent la propagation et en augmentent l'efficacité, tout en facilitant le ciblage du personnel gouvernemental et militaire, l'hameçonnage, la fraude psychologique et la reproduction de données biométriques.

- Les hypertrucages peuvent servir à générer du bruit pour inonder un espace dans lequel des renseignements sont collectés, ce qui a pour effet de distraire des renseignements véridiques. En outre, des conversations, des vidéos et des textes factices peuvent déformer la perception des sources humaines. Le fait que l'on compte aujourd'hui davantage sur les renseignements fondés sur des sources ouvertes (OSINT) renforce le poids des informations « hypertruquées » dans l'écosystème informationnel.

- Les hypertrucages peuvent aussi servir à empoisonner les données utilisées pour entraîner des systèmes d'apprentissage profond, afin de polluer délibérément ces derniers au moyen d'information dommageable. Par exemple, les algorithmes visant à détecter les cyberattaques pourraient être mis en échec par l'injection de données empoisonnées dans les grands ensembles de données employés pour les entraîner.

- Du point de vue de la sécurité publique, les hypertrucages peuvent servir à commettre des fraudes, à exercer des formes de coercition et d'extorsion, à créer de fausses preuves d'actes criminels, à se faire passer pour quelqu'un d'autre en vue de

mener des activités illégales ou à imputer de telles activités à tort à des tiers.

Perspectives

Les hypertrucages sont conçus pour la tromperie et le cerveau humain n'est pas toujours capable de repérer les produits de technologies complexes. Bien que les géants technologiques aient commencé à signaler le contenu « hypertruqué » comme étant de la désinformation, les systèmes de détection qui font appel à la fois à des êtres humains et à des modèles de prédiction sont plus performants. Les gouvernements ont un rôle à jouer pour favoriser les utilisations des technologies d'hypertrucage qui profitent à la population et à la démocratie et qui les protègent, et les citoyennes et les citoyens ont la possibilité de se prémunir contre les hypertrucages et d'en préserver leur communauté.

- Les hypertrucages bousculent la législation en vigueur, notamment dans les domaines de la diffamation, de la propriété intellectuelle et du droit à la vie privée. À l'heure actuelle, les médias sociaux sont peu imputables à titre de diffuseurs de contenu résultant d'hypertrucages. Il faut prioriser l'adaptation et la mise à jour des lois, car celles-ci doivent tenir compte des problèmes posés par les hypertrucages et clarifier les questions relatives à l'obligation de rendre des comptes, à la responsabilité et aux droits des personnes visées par la manipulation de ces technologies.
- Il est important que les politiques favorisent la recherche-développement visant à élaborer des technologies capables de détecter les hypertrucages et de les contrer. Il est primordial pour les autorités de resserrer leur collaboration avec les experts compétents, afin d'établir des normes et des lignes directrices sur l'utilisation responsable de la technologie d'hypertrucage, et de trouver l'équilibre entre innovation propice aux avancées technologiques et réglementation nécessaire face aux risques.
- L'authentification des contenus est une approche prometteuse : au lieu d'essayer de détecter le contenu généré par l'IA, elle consiste à incorporer l'authentification de ce dernier dans

l'infrastructure même d'Internet grâce à un marqueur cryptographique incrusté dans l'ADN de chaque élément.

- Les normes et le discours sociétaux sur les hypertrucages devraient instaurer un climat dans lequel les gens sont non seulement sceptiques face à ce qu'ils voient, mais encouragés à remettre en question les informations provenant de leurs pairs. L'éducation numérique, surtout si elle est destinée aux leaders de la pensée et aux influenceurs, aide à sensibiliser aux risques et renforce la confiance dans les médias.

Il ne fait aucun doute que la technologie d'hypertrucage continuera d'évoluer rapidement pour produire des contenus de plus en plus réalistes, de façon de plus en plus efficace et de plus en plus économique. Le fait d'envisager les hypertrucages dans leur globalité permet d'adopter des approches complètes, en vue d'optimiser les avantages de ces nouvelles technologies tout en atténuant les risques pour la sécurité nationale et individuelle qui y sont associés, sans compromettre le droit à la vie privée, ni la confiance du public dans les médias et dans les sources d'information.

Les hypertrucages, une vraie menace pour l'avenir du Canada

La création ou la diffusion d'informations fausses, trompeuses ou sensationnalistes est loin d'être une nouveauté : les premiers cas remontent au 15e siècle[1, 2]. Cependant, en raison notamment d'une récente résurgence du nationalisme de droite, les « fausses nouvelles », en particulier la désinformation, sont jugées préoccupantes par le milieu universitaire comme par la population[3].

La désinformation est la diffusion d'informations fausses délibérément trompeuses[4] : en plus d'être inexacte, elle vise à leurrer l'auditoire et à faire de sérieux dégâts[5]. Internet et les médias sociaux augmentent la rapidité avec laquelle elle se propage et l'énorme influence qu'elle peut avoir. En outre, elle a actuellement un autre avantage sur ses formes plus anciennes : l'existence des hypertrucages.

Les hypertrucages consistent à manipuler des contenus à l'aide d'outils de pointe faisant appel à l'intelligence artificielle (IA) pour modifier ou créer de toutes pièces des images, des voix, des vidéos ou des textes[6]. Ils peuvent être employés pour placer n'importe qui ou n'importe quoi dans une situation à laquelle il ou elle n'a jamais participé (une conversation, une activité, un lieu ou autre)[7, 8]. L'IA permet de produire des textes (articles, billets de blogue et commentaires) qui peuvent être publiés rapidement en ligne parmi le contenu authentique, peu importe leur degré de véracité[9].

Si la technologie d'hypertrucage est aussi utilisée pour créer des contenus sains, divertissants et satiriques, les gouvernements doivent envisager les maux potentiels ou les menaces pour la sécurité publique qui y sont associés. Comme le démontre le présent article, les hypertrucages nécessitent l'attention et l'action des États démocratiques et de celles et ceux qui tiennent aux libertés et à la sécurité que procure la vie dans un tel État.

Progrès de l'IA

Le Conseil du Trésor du Canada (CT) définit l'IA comme une technologie informatique qui effectue des tâches nécessitant habituellement l'intervention d'un cerveau biologique, comme comprendre la langue parlée, apprendre des comportements ou

résoudre des problèmes[10]. Autrement dit, l'IA, c'est le fait pour un ordinateur d'accomplir des tâches actuellement réservées aux êtres humains, comme la reconnaissance vocale, la prise de décision, la reconnaissance d'objets ou la traduction d'une langue à une autre[11, 12].

Au cours des dernières années, l'IA a considérablement progressé dans sa capacité à réaliser ces « tâches humaines ». ChatGPT (openai.com) est un exemple éloquent de ces progrès[13]. Le sigle « GPT » renvoie à « generative pre-trained transformer », qui signifie « transformateur génératif préentraîné », une locution qui désigne le type de modèle linguistique qui sert de moteur à ChatGPT. À la différence des robots conversationnels traditionnels, ChatGPT dialogue d'une façon semblable à celle des êtres humains : il produit des réponses adaptées au contexte et au ton du « dialogue » qu'il a avec les personnes qui l'utilisent. ChatGPT fournit aussi des réponses (habituellement) exactes aux questions posées par ces personnes, puisqu'il a été formé à partir de données tirées de l'Internet[14].

La synthèse d'images à partir d'invites textuelles est une autre avancée importante des contenus de synthèse : une personne entre des mots-clés que le modèle d'IA utilise pour créer une image unique[15]. Avec suffisamment de données d'entraînement et de nouvelles avancées, ces modèles pourront sans doute bientôt produire des images montrant des personnes existantes dans des scènes réalistes sur le plan visuel, mais complètement factices.

L'IA a aussi réduit la vitesse à laquelle les tâches humaines sont généralement accomplies. Par exemple, des évaluations ont déterminé que le programme AlphaFold avait prédit avec exactitude les structures de plus de 200 millions de protéines[16], ce qui a fait faire un bond considérable à la recherche sur la probabilité et le traitement des maladies. Sans l'IA, il aurait fallu des années ou des décennies à la communauté scientifique pour parvenir au même résultat[17].

Menace pour l'avenir du Canada

En raison des progrès de l'IA, les hypertrucages deviennent rapidement de plus en plus réalistes, ce qui les rend plus difficiles à

détecter et à signaler[18]. Les applications permettant de les créer sont aussi plus accessibles et moins techniques[19]. Malgré cela, ils semblent méconnus ou mal compris et il y a une incapacité à les reconnaître ou à les détecter[20].

Compte tenu de ces facteurs, la question de savoir si la population canadienne devrait s'inquiéter des hypertrucages se pose. En réponse à cela, on peut réfléchir à une déclaration faite Maria Ressa, lauréate du prix Nobel de la paix :

> *Sans les faits, il n'y a pas de vérité. Sans vérité, il n'y a pas de confiance. Sans confiance, il n'y a pas de réalité commune ni de démocratie, et il devient impossible de gérer les problèmes existentiels de notre monde*[21, 22].

Au bout du compte, il s'agit d'une question de faits. Si une société démocratique n'est pas capable de distinguer la réalité de la fiction, comment peut-elle survivre? Comment le Canada va-t-il fonctionner s'il y a des clivages dans la population en fonction des différents ensembles de faits non vérifiables auxquels elle croit? Si la désinformation est ingérable ou indétectable, comment le pays va-t-il élaborer des solutions à ces problèmes bien réels? Qu'est-ce que cela signifie pour les valeurs et pour le mode de vie canadiens? En outre, que se produit-il quand les hypertrucages sont utilisés à des fins malveillantes ou en vue de nuire au Canada et à ses alliés?

Conséquences délétères de l'utilisation des hypertrucages

Les effets délétères des hypertrucages sont illustrés par des exemples récents. En janvier 2023, une jeune femme, Blaire, mieux connue sous le pseudonyme de QTCinderella pour ses vidéos diffusées en direct sur Twitch et son travail de youtubeuse, a découvert qu'un site d'hypertrucages pornographiques utilisait son visage et son apparence dans des vidéos pornographiques, avec ceux d'autres diffuseuses sur Twitch[23]. Un camarade diffuseur sur Twitch, Brandon Ewing, avait payé le site pour visionner ces films de Blaire et de ses comparses[24, 25].

En 2019, Rana Ayyub, journaliste d'enquête pour le *Washington Post,* a critiqué un parti politique indien parce qu'il protégeait le violeur d'une fillette de huit ans[26]. En représailles, une vidéo pornographique « hypertruquée » la mettant en scène a été produite et est devenue virale en 48 heures. Après la publication de cet hypertrucage, la jeune femme a reçu des menaces de mort, ainsi que des messages racistes et misogynes. Sans surprise, Mme Ayyub a complètement disparu des réseaux sociaux et a cessé ses reportages pendant un certain temps[27].

Ces cas d'hypertrucages pornographiques ne sont pas rares. Plus de 90 % des hypertrucages visibles en ligne sont des vidéos pornographiques non consensuelles de femmes. En octobre 2022, 57 millions de résultats étaient répertoriés pour la recherche « deepfake porn » (hypertrucages pornographiques) sur Google seulement[28]. Les femmes sont presque toujours les cibles ou les objets non consentantes de ces vidéos et la législation actuelle n'offre pas aux victimes beaucoup de protection ni de possibilités d'obtenir justice[29, 30].

Il convient de noter que tous les dommages imputables aux hypertrucages n'ont pas un caractère pornographique. Par exemple, des criminels ont utilisé à maintes reprises des hypertrucages d'Elon Musk pour organiser de prétendues distributions de cryptomonnaie, causant ainsi des pertes financières équivalant à des millions de dollars[31, 32]. Dans certains cas, des fraudeurs ont aussi recouru à des clonages de la voix de hauts dirigeants de banques et d'autres sociétés dont la valeur nette est élevée[33]. Ils appelaient alors les bureaux de l'entreprise visée, se faisaient passer pour leur président-directeur général ou leur gestionnaire, et ordonnaient au personnel de transférer de l'argent à leur compte bancaire[34, 35].

À cause de la pandémie de COVID-19, le monde s'est habitué à utiliser des services en ligne pour organiser des réunions, des entrevues ou des cours virtuels, par exemple[36]. Comme la technologie d'hypertrucage est de plus en plus largement disponible et accessible, il devient de plus en plus compliqué de vérifier la véritable identité des personnes qui apparaissent à l'écran. En 2022, le FBI a constaté que des criminels se servaient des hypertrucages au cours d'entretiens

d'embauche virtuels pour des emplois à distance, quand il était probable que la société n'interagirait qu'en ligne avec la personne sélectionnée[37, 38].

Autres considérations liées à l'IA

Il est évident que l'IA est un outil puissant, qui peut apporter des solutions à des problèmes et favoriser leur résolution. Cependant, elle peut aussi procurer à une personne ou à une organisation le pouvoir d'infliger des dommages considérables :

1. Atteintes à la confidentialité : Les systèmes d'IA peuvent collecter, traiter, analyser et entreposer des volumes colossaux de données. Si un de ces systèmes est piraté ou si une violation de sécurité se produit, ces données (qui peuvent comprendre les antécédents médicaux et les informations biographiques ou bancaires d'une personne) peuvent être aisément volées, manipulées ou utilisées à des fins d'extorsion au service d'objectifs malveillants[39, 40, 41].

2. Manipulation sociale : L'IA peut servir à suivre, à analyser et à prévoir les activités en ligne d'une personne, qui peut donc prêter le flanc à la manipulation ou être forcée de se livrer à des activités dont elle s'abstiendrait autrement[42, 43].

3. Préjugés : Les systèmes d'IA sont conçus, fabriqués, entraînés et testés par des êtres humains, qui ont des préjugés implicites et explicites. Les décisions prises par ces systèmes peuvent donc refléter ces préjugés, au détriment de certains groupes[44, 45].

Prochaines étapes pour les gouvernements

On estime que 5 milliards de personnes dans le monde utilisent Internet ou y ont accès[46] et que 36 millions d'entre elles sont canadiennes[47]. Cela signifie qu'une part considérable de l'auditoire canadien et mondial a accès à la technologie reposant sur l'IA et à la désinformation qui y est associée, donc peut être influencée par celles-ci.

Les hypertrucages et les autres technologies de pointe faisant appel à l'IA menacent la démocratie, car certains acteurs cherchent à exploiter l'incertitude ou à maintenir en vie des versions des « faits » fondées sur des informations falsifiées ou créées de toutes pièces. Ce problème sera exacerbé si les gouvernements ne sont pas capables de prouver que les contenus officiels sont réels et factuels.

Les hypertrucages et les contenus de synthèse peuvent aussi faciliter des activités lourdes de conséquences sur les plans psychologique, économique ou de la réputation[48, 49]. Comme indiqué plus haut, l'utilisation ou l'exploitation de systèmes d'IA peut favoriser des atteintes à la vie privée, des manipulations sociales ou des dommages liés aux préjugés inhérents à ces technologies. Il incombe aux gouvernements d'intervenir pour atténuer ces menaces qui pèsent sur leurs citoyennes et leurs citoyens.

En effet, quelques-uns des experts les plus reconnus dans le domaine de l'IA (Yoshua Bengio, Elon Musk et Geoffrey Hinton) ont souligné les dangers que l'IA présente pour la population[50, 51, 52, 53]. M. Hinton a même démissionné de son poste à Google pour pouvoir communiquer plus librement sur le sujet[54, 55]. Le degré de préoccupation exprimé par ces experts, associé aux impacts potentiels pour leurs citoyens, devrait pousser les gouvernements à s'attaquer à l'IA et à ces impacts.

Le Secrétariat du CT du Canada (SCT) a signifié que le gouvernement du Canada s'engageait à utiliser l'IA à l'appui de certains services fournis à la population ou pour les améliorer, de façon compatible avec les principes fondamentaux du droit administratif. Cet engagement est à la base de sa Directive sur la prise de décisions automatisée, qui « a pour objet de veiller à ce que les systèmes décisionnels automatisés soient déployés d'une manière qui permet de réduire les risques pour les clients, les institutions fédérales et la société canadienne, et qui donne lieu à une prise de décisions plus efficace, exacte et conforme, qui peut être interprétée en vertu du droit canadien »[56].

L'Initiative de citoyenneté numérique lancée par Patrimoine Canada est une autre mesure positive. Cette stratégie à plusieurs axes « [...] soutient la démocratie et l'inclusion sociale au Canada [en renforçant] la résilience des citoyens face à la désinformation [et en établissant des partenariats qui favorisent un écosystème informationnel sain] ». Elle vise à aider la population et le gouvernement du Canada à mieux comprendre la désinformation et ses répercussions, afin de déterminer les mesures à prendre et de faciliter l'élaboration des futures politiques[57].

Les capacités de l'IA continueront de progresser et d'évoluer. Le réalisme des contenus « hypertruqués » ou synthétiques va s'améliorer et le contenu généré par l'IA va devenir plus répandu. Cela signifie que les politiques, les directives et les initiatives gouvernementales (actuelles et futures) doivent progresser et évoluer au même rythme, notamment dans leur efficacité à cerner les contenus malveillants générés à l'aide de l'IA et à les distinguer de ceux qui sont positifs et utiles à la société.

Peut-être par nécessité, les gouvernements démocratiques sont notoirement lents à rédiger des politiques, des procédures ou des lois et à les faire appliquer[58, 59, 60]. À l'opposé, l'IA progresse et évolue vite. Si les gouvernements évaluent et combattent l'IA chacun de leur côté et à leur rythme habituel, leurs interventions seront rapidement obsolètes. Il est essentiel que les gouvernements partenaires et alliés, les universitaires et les experts de l'industrie collaborent pour garantir l'intégrité de l'information diffusée partout dans le monde et remédier aux utilisations néfastes de l'IA en pleine mutation.

Réaction humaine à la désinformation et aux hypertrucages

Comme ils sont réalistes et semblent « criants de vérité », les hypertrucages sont exceptionnellement efficaces pour mettre en scène des personnes et des faits créés de toutes pièces. Les gens qui les relaient ne le font pas nécessairement parce qu'ils les croient vrais, mais pour consolider leur identité et leur statut social. Les hypertrucages changent rarement les opinions, mais peuvent entraîner une radicalisation par le chaos et la confusion qu'ils sèment. Le présent article passe en revue les incidences des hypertrucages dans les sphères sociale et privée, et propose des interventions qui pourraient permettre d'atténuer ces effets délétères[61].

Un hypertrucage est une supercherie ou la falsification ultraréaliste d'un contenu numérique (image, vidéo ou enregistrement sonore) créée au moyen de réseaux de neurones faisant appel à des modèles d'apprentissage automatique nommés « réseaux antagonistes génératifs » (GAN). Les hypertrucages sont employés dans une grande variété de contextes, des arts au divertissement, en passant par la publicité et l'enseignement. Leur utilisation la plus fréquente est toutefois la pornographie : en date d'octobre 2019, 96 % des hypertrucages sur Internet étaient pornographiques[62]. Cependant, c'est leur contribution avérée et potentielle à l'épidémie de fausses nouvelles qui retient l'attention des universitaires et des médias, même s'ils font aussi peser un certain nombre de menaces sur la société[63]. Les images et les vidéos ont du poids, car elles paraissent représenter la vie réelle directement. De nombreux hypertrucages qui mettent en scène des personnalités politiques faisant des déclarations contraires à leurs prises de position officielles sont apparus, comme celui qui a été téléversé par piratage sur un site de nouvelles ukrainien, où l'on voyait le président de l'Ukraine, Volodymyr Zelensky, ordonner à ses soldats de déposer les armes[64] ou cet autre de Barack Obama proférant des insultes sur Donald Trump[65]. Si ces tromperies devenaient virales, elles pourraient avoir un effet irréversible sur les affaires mondiales.

Les êtres humains traitent les données visuelles naturellement, donc de façon très fluide[66], et les gens croient ce qu'ils voient[67]. En outre, les images détaillées qui caractérisent les hypertrucages pourraient favoriser la proximité psychologique. Toute mésinformation concrète

(y compris toute désinformation) prédispose ses destinataires à penser que les événements décrits sont plus proches et plus probables qu'ils ne le sont réellement, ce qui exacerbe le sentiment de menace[68] et la probabilité que les nouvelles sur cette menace soient relayées[69].

Quelle est l'efficacité des hypertrucages?

Les études sur la capacité des gens à distinguer les hypertrucages des images authentiques ne sont pas unanimes. Certaines concluent que les hypertrucages ne parviennent pas mieux que les sources traditionnelles de fausses nouvelles, comme les textes et les enregistrements audio, à générer de faux souvenirs, car ils ne sont pas plus crédibles ou plus puissants[70, 71]. Cependant, même si ces études sont récentes, la technologie évolue si vite qu'elles ne portent pas sur les dernières techniques de création d'images à l'aide de l'intelligence artificielle (IA) disponible en ligne. Pire, certaines ne sont fondées que sur une seule vidéo[72].

Selon Lago et coll. (2022), les plus récentes images créées par l'IA sont perçues comme véritables[73]. En effet, les participants à cette étude ont trouvé les visages artificiels générés par les GAN de pointe plus vrais que des images authentiques, ce qui souligne le potentiel qu'ont les hypertrucages de simuler le réel et d'éviter l'écueil du sentiment étrange et perturbant qui naît quand des robots humanoïdes et des images créées par ordinateur sont trop proches de la réalité (l'effet « vallée de l'improbable »)[74]. De plus, d'après Köbis et coll. (2021), les personnes ne sont pas capables de détecter les hypertrucages avec exactitude, et ni la sensibilisation, ni les incitatifs financiers n'améliorent leurs résultats à cet égard[75]. Une étude encore plus récente a conclu que les vidéos « hypertruquées » étaient à la fois plus crédibles que les supercheries sous forme d'images et de textes et qu'il y avait plus de probabilités que les personnes interagissent avec[76]. L'évolution rapide des GAN rendra bientôt les hypertrucages impossibles à distinguer des contenus authentiques, si ce n'est pas déjà fait.

L'exactitude importe-t-elle?

La véracité de l'information n'est pas un facteur prépondérant quand il s'agit de diffuser du contenu en ligne[77]. Même quand ils et elles les repèrent, les internautes peuvent relayer des hypertrucages dans leur cercle social. En effet, l'étude Vosoughi et coll. (2018) a découvert que les fausses nouvelles se propageaient plus vite et plus loin en ligne que les informations factuelles[78].

Leur usage le plus commun donne un indice sur la raison pour laquelle les personnes propagent et regardent des hypertrucages : le terme « deepfake » (l'équivalent anglais d'hypertrucage) a été inventé sur un forum Reddit créé pour diffuser des vidéos pornographiques tournées par des actrices dont les visages ont été remplacés artificiellement par ceux d'autres femmes, pour la plupart des célébrités[79]. Il est peu probable que les consommateurs d'hypertrucages pornographiques soient trompés par ces images, étant donné qu'ils les visionnent sur des sites ou sous des titres comportant généralement la mention « fake » (faux). Il n'y a aucune prétention à la vérité. Ainsi, le fait de savoir qu'une vidéo est fausse n'empêche pas les amateurs d'hypertrucages pornographiques de trouver la satisfaction qu'ils recherchent — il y contribue même peut-être. Ce constat pourrait s'appliquer à beaucoup d'hypertrucages politiques. D'autres usages des hypertrucages ne dépendent pas non plus de leur véracité : c'est le cas de ceux qui ont des visées artistiques ou éducatives.

La plupart des hypertrucages véhiculés sur les médias sociaux le sont pour les mêmes raisons que les fausses nouvelles, à savoir générer des clics. Pour y parvenir, les fausses nouvelles misent sur deux caractéristiques qui suscitent l'attention : la nouveauté et la négativité[80]. La diffusion de faits inédits a une valeur sociale, car elle donne à croire que celui ou celle qui le fait à des informations exclusives[81]. Par ailleurs, l'attrait qu'exercent les mauvaises nouvelles est bien établi. Les gens prêtent davantage l'oreille aux pertes potentielles qu'aux gains[82]. Une certaine noblesse est associée à la communication d'informations négatives : l'émetteur met en garde ses relations contre une menace. Ainsi, les professionnels de la santé sont plus enclins à

propager les fausses rumeurs visant à éviter des effets indésirables (p. ex. une cause de cancer) que celles censées avoir un effet favorable (p. ex. guérir le cancer)[83, 84].

Les gens échangent des nouvelles avec leur communauté idéologique parce que cela répond à une motivation profondément humaine consistant à renforcer ses liens sociaux[85]. Cela confirme aussi l'identité d'une personne comme membre d'un groupe idéologique[86]. La défense de l'identité prend donc le pas sur la vérification de l'exactitude[87]. En effet, l'identité d'un individu (p. ex. sa nationalité, sa religion, sa race ou son parti politique) joue sur ce qu'elle croit vrai[88]. Cependant, parce que les gens vivent de plus en plus dans des bulles d'information, les croyances partisanes conduisent généralement davantage à fermer les yeux sur des vérités dérangeantes qu'à croire en des contre-vérités. Par conséquent, les personnes peuvent voir et relayer une vidéo « hypertruquée » conforme à leurs convictions sans jamais avoir de raison de croire qu'il s'agit d'un hypertrucage. À titre d'exemple, une étude consistant à montrer de fausses photographies à des sujets a révélé que les personnes d'orientation conservatrice étaient plus susceptibles de se « souvenir » de la poignée de main entre Barack Obama et le président de l'Iran et les personnes d'obédience libérale, des vacances de George W. Bush avec une célébrité pendant l'ouragan Katrina (aucun de ces événements n'a eu lieu)[89]. Il a été démontré que les hypertrucages radicalisent un auditoire contre l'opposition[90]. Par conséquent, comme les informations de toute autre source, il est plus probable que les hypertrucages durcissent des opinions existantes plutôt qu'ils les changent.

L'âge augmente la probabilité d'être trompé par les hypertrucages et l'idéologie politique influence la façon dont les nouvelles « hypertruquées » sont perçues[91]. Si les républicains et les démocrates aux États-Unis ont la même propension à diffuser des fausses nouvelles[92], les conservateurs les moins scrupuleux (c'est-à-dire celles et ceux qui ont le moins tendance à respecter les normes sociales en matière de contrôle des pulsions) ont plus de chances de propager de la mésinformation, car ils souhaitent le chaos[93].

Conséquences

Les hypertrucages génèrent des risques pour les sociétés, les entreprises et les consommateurs. L'étude de Caldwell et coll. (2020) classe les supercheries audio et vidéo comme la plus importante menace posée par l'usage de l'intelligence artificielle à des fins criminelles et terroristes[94]. Europol (2022) a signalé que les hypertrucages pouvaient être utilisés pour harceler et humilier les gens en ligne, pratiquer l'extorsion et la fraude, falsifier des identités en ligne et tromper les mécanismes d'identification des clients, exploiter sexuellement des enfants en ligne, contrefaire ou manipuler les éléments de preuve électroniques dans le cadre d'enquêtes criminelles et perturber les marchés financiers[95].

Les hypertrucages menacent aussi les structures de gouvernance. L'incertitude qu'ils provoquent permet aux gens de vivre dans leur propre réalité subjective, ce qui accentue les clivages sociaux et entrave la bonne marche de la démocratie[96]. Ce phénomène est particulièrement dangereux en période électorale, où des puissances nationales et étrangères essaieront probablement de manipuler l'issue du scrutin[97]. Les parties rivales peuvent être tentées de soumettre l'électorat à des documents « hypertruqués » longtemps avant un vote pour les prédisposer à avoir certaines opinions ultérieurement[98].

Pour les entreprises, les menaces sont notamment les faux commentaires sur leurs articles, la diffamation et le sabotage, et les atteintes à l'image, à la réputation et à la crédibilité[99]. Par exemple, en 2019, des criminels ont réussi à se faire passer pour le directeur de la société mère d'une compagnie grâce à un logiciel d'imitation de la voix. Trompant ainsi le président-directeur général d'une entreprise britannique d'énergie, ils l'ont convaincu de leur transférer un montant de 243 000 $ US.

Outre ce « hameçonnage par hypertrucage », l'intelligence artificielle rendra certaines technologies obsolètes. Les menaces pour les consommateurs comprennent de nouvelles possibilités de chantage, d'intimidation, de sabotage, de harcèlement, de diffamation, de pornographie vengeresse et de vol d'identité.

Le caractère personnalisé des hypertrucages pornographiques ajoute à la détresse des victimes et à la menace qui pèse sur elles[100]. La plupart de ces hypertrucages présentent des célébrités, à qui leur réputation peut éviter dans une certaine mesure d'être vues comme les véritables actrices dans ces vidéos. Ces personnalités ont aussi des tribunes pour s'exprimer publiquement et les moyens juridiques et financiers de contester la véracité de ces vidéos. En revanche, le commun des mortels ne jouit même pas de ces minces protections[101].

Même si les citoyens ne croient pas la mésinformation qui leur est présentée ou ne se préoccupent pas de la vérité, les hypertrucages peuvent aggraver l'incertitude qui entoure l'information et amoindrir la confiance dans les médias[102]. Aux États-Unis, les fausses nouvelles ont conduit 50 % des républicains et 38 % des démocrates à réduire leur consommation de nouvelles[103]. Comme l'a montré la COVID-19, la population peut prêter le flanc à la mésinformation en période de crise, à cause du climat de complot et d'incertitude qui règne alors[104]. Les hypertrucages exacerbent ce problème.

Le recours aux hypertrucages contre des personnalités publiques génère ce qu'on appelle « le dividende du menteur » : les personnes qui font face à des accusations peuvent soutenir que les preuves factuelles contre elles sont des hypertrucages[105]. L'omniprésence des hypertrucages peut donc prédisposer les gens à douter de l'authenticité de toute information. Ainsi, pour discréditer les preuves d'une incartade sexuelle, le ministre malaisien des Affaires économiques a prétendu qu'il s'agissait d'un hypertrucage, sans que rien n'étaye son affirmation[106]. Plus récemment, les avocats d'Elon Musk ont utilisé le même procédé pour contrer des poursuites[107].

Les hypertrucages ont beaucoup d'avantages potentiels. Ils sont à l'origine de nouvelles formes d'art intrigantes, font d'excellents outils pédagogiques et sont une source de plaisir et d'amusement sans conséquence. Ils peuvent aussi offrir des occasions d'affaires[108]. Le métavers de Facebook sera en grande partie composé d'objets « hypertruqués ». Les hypertrucages permettent des campagnes de marketing d'un genre nouveau (notamment grâce à l'élimination de la barrière de la langue), le recours à des ambassadeurs et ambassadrices

de marque virtuels (Lil Miquela est une influenceuse factice qui a plus de trois millions d'abonnés) et tout un éventail d'innovations techniques. Ainsi, il existe aujourd'hui trois présentateurs ou présentatrices de nouvelles virtuels créés à partir de véritables êtres humains. Les hypertrucages peuvent aussi servir à améliorer la mémoire, par exemple en faisant paraître vivante une personne décédée.

Entre de mauvaises mains, les hypertrucages sont toutefois une forme nouvelle de virus social. Comme pour tous les virus, il est difficile d'en prédire l'évolution et les répercussions. La plus importante menace qu'ils présentent pour la société est leur capacité à orienter le débat public. Lorsque la désinformation pèse sur ce dernier, sa dangerosité est à la mesure de l'altération de la compréhension et de la mémoire collective qu'elle entraîne. La présence croissante des hypertrucages pourrait aussi amener les gens à douter d'une grande partie de ce qu'ils voient.

Solutions

Les hypertrucages sont conçus pour nous tromper : l'esprit humain n'est pas prêt à repérer constamment et avec exactitude les produits de technologies complexes. Certains géants des technologies ont commencé à signaler certains contenus comme relevant de la désinformation, mais ce n'est pas la panacée. Les fausses nouvelles sont moins relayées quand elles sont accompagnées d'avertissements[109]. Cependant, il n'est pas évident que ces derniers empêchent d'y croire[110]. En effet, le fait d'avoir été exposé à de la mésinformation renforce l'impression que les fausses nouvelles sont vraies, ce qui pourrait annihiler le poids de ces avertissements. Par ailleurs, les systèmes de détection réunissant êtres humains et modèles prédictifs sont plus efficaces que ceux qui reposent sur les personnes d'une part et sur les méthodes automatiques d'autre part[111].

Afin d'atténuer le problème, il est notamment possible d'afficher des avertissements en amont, pour que les gens sachent que l'information qu'ils sont sur le point de consulter pourrait être fausse. Ces avertissements doivent être précis, car les mises en garde générales sur l'éventuelle présence de mésinformation sont inefficaces[112]. En

outre, ils doivent être accompagnés d'un récit causal, qui explique à la fois les faits et l'objet de la désinformation. Les sociétés peuvent éduquer leurs clients à leurs produits, à leurs marques et à leurs services pour les aider à reconnaître les sources d'information crédibles qu'elles cautionnent[113].

D'un point de vue juridique, la responsabilité des distributeurs est peu engagée en cas de circulation d'hypertrucages sur les médias sociaux[114]. Aux États-Unis, le débat juridique est centré sur l'article 230 du Communications Decency Act [loi sur la décence dans les communications], qui évite aux entreprises d'être tenues comptables du contenu véhiculé sur leurs systèmes[115]. L'appareil judiciaire pourrait préciser la responsabilité civile des personnes et des organisations qui créent et qui diffusent des hypertrucages, tout en améliorant la protection juridique accordée aux victimes de diffamation.

Individuellement, nous avons peu d'outils pour empêcher les attaques faisant appel à des hypertrucages. Quand ces derniers menacent des réputations, les personnes visées peuvent enregistrer leurs activités pour pouvoir davantage nier les actions décrites dans ces productions, mais cela pose des problèmes de confidentialité[116]. Collectivement, des méthodes de diffusion des faits peuvent contribuer à protéger les communautés, mais elles doivent s'appuyer sur une compréhension approfondie de l'écosystème informationnel dans lequel baigne le public en question et des moyens d'y évoluer.

Une personne est plus facilement convaincue et corrigée par quelqu'un qu'elle connaît. Par conséquent, il faut amener les normes et le discours sociétaux sur les hypertrucages à changer pour instaurer un climat dans lequel les gens sont non seulement sceptiques face à ce qu'ils voient, mais sont encouragés à remettre en question les informations provenant de leurs pairs.

Pour modifier les normes en vigueur dans une société, il faut miser sur les leaders de la pensée et sur les gens qui jouent un rôle central sur les réseaux sociaux. Les ressources éducatives, comme la formation en informatique, constituent des outils utiles, en particulier lorsqu'elles sont offertes aux personnes d'influence. Il a été constaté que des

vidéos expliquant des hypertrucages à caractère politique avaient réduit l'incertitude et, ce faisant, rehaussé la confiance dans les nouvelles diffusées par les médias[117], mais l'évolution des normes ne peut être le fruit que d'une action collective.

Quand de vraies personnes ont recours à des faux : utilisation des hypertrucages par la population

En traitement des signaux, en infographie et en vision informatique, il a toujours été essentiel de synthétiser des sons, des images et des vidéos réalistes. Avec des outils datant d'avant l'avènement de l'intelligence artificielle (IA), le processus est généralement long, coûteux et exigeant sur le plan technique pour le commun des mortels. Cependant, les progrès rapides de l'IA ces dernières années ont beaucoup abaissé le seuil des ressources, du temps et de l'expertise nécessaires pour créer des faux convaincants. Ces avancées ont frappé le public à la fin de l'année 2017, quand un compte Reddit appelé « DeepFake », un fourre-tout sur lequel étaient publiés des produits de l'apprentissage profond et des contenus falsifiés, a commencé à diffuser des vidéos pornographiques sur lesquelles avaient été transplantés les visages de célébrités à l'aide d'un algorithme faisant appel à un réseau de neurones profond (RNP). Depuis, des algorithmes plus élaborés synthétisant des sons, des images et des vidéos réalistes sont apparus, ainsi qu'une pléthore d'outils logiciels de source ouverte et de services commerciaux. Par ailleurs, le terme « hypertrucage » (« deepfake » en anglais) est aujourd'hui exploité plus largement pour désigner toute supercherie créée ou éditée au moyen d'algorithmes d'apprentissage profond.

Les hypertrucages ne sont que la partie émergée de cette inquiétante tendance. Parce qu'ils donnent l'illusion de la présence et des activités d'une personne, ils peuvent causer des dommages bien réels quand ils sont utilisés à des fins offensives. Par exemple, une fausse vidéo mettant en scène une personnalité politique agissant de façon inappropriée pourrait suffire à faire pencher la balance en sa défaveur si elle était diffusée peu avant un scrutin. Un enregistrement audio d'une haute dirigeante d'entreprise commentant la situation financière de sa société pourrait faire chuter l'action de cette dernière. L'utilisation d'un visage humain de synthèse réaliste comme photo de profil d'un faux compte sur les médias sociaux peut renforcer considérablement le poids d'une tromperie. Un prédateur en ligne peut se faire passer pour un membre de la famille ou du cercle d'amis de sa victime au cours d'une conversation vidéo, afin de l'attirer. S'ils ne sont pas contrôlés, les hypertrucages peuvent amplifier la désinformation en ligne et le danger qui y est associé, donc fondamentalement ébranler la confiance d'une société dans les contenus numériques.

Les dernières avancées de la production de contenus à l'aide de l'IA (ou « IA générative ») ont été considérablement favorisées par la disponibilité de modèles de création de textes et d'images à grande échelle, notamment ceux de la famille Generative Pre-trained Transformer (GPT) d'OpenAI, DALL-E et Midjourney. Ces progrès, qui ont stimulé l'imagination du public quant aux superpouvoirs de l'IA et qui ont ouvert la perspective de l'intelligence artificielle générale (IAG), apportent aussi de nouvelles occasions et de nouveaux défis pour la fabrication d'hypertrucages. Ils se concentrent sur trois grands axes : i) augmenter la puissance et les capacités des modèles; ii) rendre les modèles plus accessibles; iii) donner aux utilisateurs davantage de contrôle sur le style et le contenu des productions grâce à des commandes textuelles détaillées.

L'un des progrès les plus importants de l'IA générative est la puissance et la capacité accrues des modèles. Parce que des volumes considérables de données sont disponibles pour leur entraînement, ces modèles peuvent apprendre des combinaisons compliquées et générer des produits de grande qualité. Ils peuvent produire des images, des vidéos et des sons réalistes et complexes, qu'il est presque impossible à distinguer de ceux qui sont créés par des personnes. Ils ont un large éventail d'applications, allant de la synthèse d'images réalistes pour les environnements virtuels à celle de voix réalistes pour les assistants virtuels.

Un deuxième progrès important des modèles d'IA générative est leur accessibilité. De nombreux outils ont maintenant des interfaces Web nécessitant peu ou pas de codage ou d'effort à l'installation. Cela facilite leur utilisation par les non-initiés, qui peuvent en tirer parti.

Enfin, l'élaboration d'outils fonctionnant grâce à l'IA qui offrent davantage de contrôle sur le style et le contenu des produits au moyen d'invites textuelles détaillées est un dernier progrès important. Cela permet aux usagers de préciser le style et le contenu de la production qu'ils souhaitent par des commandes textuelles entrées dans le modèle d'IA. Cela peut être utile pour créer des contenus sur mesure destinés à des campagnes de marketing, des contenus personnalisés pour les médias sociaux ou des simulations réalistes aux fins de formation.

Actuellement, la production d'hypertrucages peut prendre trois formes différentes : les images, les vidéos et les sons ou les voix.

Images

Les images ultraréalistes créées par les modèles faisant appel à des réseaux antagonistes génératifs (GAN) sont un exemple parlant d'hypertrucages. Ces modèles sont composés de deux RNP qui sont entraînés en tandem. L'un, le « générateur », synthétise des images et l'autre, le « discriminateur », différencie les images de synthèse des vraies. Au cours de cet entraînement, les deux RNP sont en concurrence : le générateur essaie de produire des images de plus en plus réalistes pour tromper le discriminateur, qui tente d'améliorer l'exactitude de son tri. Une fois que les deux réseaux atteignent l'équilibre, l'entraînement est terminé. Le générateur est ensuite utilisé pour créer des images réalistes à partir de bruit blanc à l'entrée.

De récents travaux, appelés StyleGAN, ont montré la supériorité des modèles faisant appel aux GAN pour ce qui est de la capacité à produire des visages humains réalistes en haute résolution. Ces modèles peuvent aussi servir à modifier ou à transférer les attributs et les expressions de visages. Un modèle encore plus récent de création d'images est le modèle de diffusion. Comme celui qui fait appel aux GAN, il crée des images réalistes à partir de bruit à l'entrée. Par contre, le mécanisme d'entraînement du modèle de diffusion est différent : il utilise un RNP pour simuler le processus physique de la diffusion, dans lequel un signal structuré est lentement dissous jusqu'à ce que l'équilibre thermodynamique soit atteint, au terme d'un processus de diffusion aléatoire (imaginez une goutte d'encre qui se dissout dans une tasse d'eau). Le RNP est ensuite utilisé pour, à l'inverse, transformer le bruit obtenu à partir de l'entrée en image structurée. Les modèles de diffusion ont permis de créer des visages humains à la pointe du réalisme et les systèmes logiciels comme Stable Diffusion sont largement utilisés.

Vidéos

Le terme « deepfake » (équivalent anglais de « hypertrucage ») provient de vidéos dans lesquelles des visages étaient remplacés par d'autres à l'aide d'un dispositif de transposition d'images. Plus précisément, les visages d'une cible sont remplacés ceux d'une source à l'aide d'un modèle de type « auto-encodeur ». Cet auto-encodeur est formé de deux RNP, soit un encodeur et un décodeur, entraînés à l'aide des visages de la cible et de la source. L'encodeur conserve les expressions du visage et le port de tête de la cible*c*, que le décodeur combine avec l'identité de la source*c*. Cette méthode d'échange de visages sur des vidéos est passée dans le grand public grâce à des applications logicielles de source ouverte sur GitHub (github.com).

Il existe aussi des techniques de création de vidéos qui reproduisent les mouvements du haut du corps et ceux de l'ensemble du corps. Des variantes de cette méthode permettent d'animer une seule image d'un visage à partir de la vidéo source d'une autre personne. Ces méthodes sont notamment appelées « Reenact GAN » (GAN de réinterprétation) et « First Order Motion » (mouvement de premier ordre). Elles font appel à des modèles fonctionnant grâce à des RNP pour transférer le mouvement d'un visage tiré de la vidéo source à l'image envoyée à l'entrée afin de créer une séquence vidéo du sujet figurant sur cette dernière image avec les mêmes mouvements faciaux que ceux de la personne figurant dans la vidéo source. Plusieurs jeunes pousses ont commercialisé des outils de production de vidéos avec remplacement de visage ou de réinterprétation d'une scène par une autre personne (p. ex. Synthesia et Canny AI).

Sons et voix

Les modèles faisant appel aux RNP sont aussi employés pour générer des voix humaines synthétiques réalistes. Il existe deux types d'hypertrucages audio, qui diffèrent par les modalités d'entrée. Les modèles permettant de passer du texte à la voix (comme Parrotron et Spectron) transforment un texte écrit à l'entrée en texte dit avec la voix de la cible à la sortie, tandis que les modèles de conversion vocale utilisent la voix d'une personne source à l'entrée. Le système

neuronal de synthèse de la parole avec adaptation au locuteur sur lequel reposent ces modèles comprend généralement : i) des composantes de modélisation acoustique, qui vont des simples spectrogrammes à la vectorisation neuronale des caractéristiques du locuteur et du style, plus complexe (comme Tacotron et ses variantes); ii) des vocodeurs, comme WaveNet ou WaveRNN, pour la génération de formes d'ondes vocales; iii) des algorithmes de conversion à base de modèles faisant appel à des auto-encodeurs ou à des GAN. Plusieurs entreprises, comme Lyrebird, Respeecher, Murf.ai, ElevenLabs et Dessa, offrent des services d'imitation vocale à la demande.

Génération multimodale

La synthèse d'images à partir d'invites textuelles s'est considérablement améliorée au cours des deux dernières années et des progrès récents ont été réalisés dans les modèles de diffusion avec transformateur faisant appel à un mécanisme d'attention. Plusieurs modèles de passage de la langue à l'image à grande échelle ont été élaborés, dont DALL-E, proposé par OpenAI en 2021, qui emploie un transformateur autorégressif pour générer des images de haute qualité à partir de l'ensemble de données MS-COCO sans étiquettes d'entraînement. D'autres modèles, comme CogView, Parti, Make-A-Scene et, dernièrement, MidJourney, utilisent aussi des modèles à transformateur autorégressif pour créer des images à partir de textes. En 2022, DALL-E2, une version mise à jour de DALL-E, a été publiée à l'aide d'un modèle de diffusion avec vectorisation d'images CLIP, ce qui lui permet de produire des échantillons plus variés, de qualité supérieure, de façon plus efficace. D'autres modèles, comme GLIDE, Stable-Diffusion et Imagen, font aussi appel à des modèles de diffusion pour améliorer la synthèse d'images à partir d'éléments de texte.

Ces modèles puissants de synthèse d'images à partir d'invites textuelles ont inspiré plusieurs études axées sur l'élaboration des prochains modèles d'édition d'images dirigée par des invites textuelles, notamment DiffEdit, Prompt-to-prompt, Null-text Inversion, Imagic et Muse. Ces modèles effectuent l'édition locale d'une image à partir d'une entrée textuelle (on parle d'édition sémantique) : ils permettent d'y apporter la modification souhaitée et une disposition scénique

facultative (grâce à une carte de segmentation). Cependant, leur optimisation permet souvent de garder au maximum les caractéristiques de l'image originale, tout en apportant des modifications significatives à des zones locales. Ce type de synthèse complète est facile à repérer dans les données d'entraînement.

Ces dernières années, l'utilisation de méthodes faisant appel à l'IA pour produire des doublages de n'importe quelle vidéo à partir d'autres voix est devenue de plus en plus populaire. Ces méthodes visent à créer des mouvements de bouche réalistes, synchronisés avec l'enregistrement sonore d'une personne qui parle, dans une vidéo donnée, ce qui permet de doubler la vidéo, ou d'en refaire la trame vocale dans une autre langue. Pour y parvenir, on emploie souvent des modèles basés sur l'apprentissage automatique capables d'apprendre les liens entre le son et les mouvements de la bouche. Ces modèles impliquent généralement un entraînement sur de vastes ensembles de données (paires audiovisuelles) visant à apprendre le mappage entre domaines visuel et sonore. D'autres approches consistent à utiliser des techniques de détection d'un repère facial pour prédire les mouvements des lèvres à partir de l'entrée audio. Les derniers progrès comprennent le recours à des techniques de traduction automatique neuronale pour générer des doublages dans différentes langues et l'intégration de techniques de traitement du langage naturel pour des doublages plus précis et mieux adaptés au contexte.

Conclusion

Bien qu'il soit difficile de prédire l'avenir des hypertrucages, une chose est sûre : la technologie continuera d'évoluer rapidement pour produire des contenus de plus en plus réalistes, de façon de plus en plus efficace et de plus en plus économique. Les différents intervenants devront agir pour contrôler l'éventuel détournement de ces outils à des fins de désinformation. La mesure la plus urgente consiste pour les fournisseurs de services et d'outils à réguler l'usage de ces derniers et à apposer des filigranes sur les produits ainsi créés, afin qu'il soit possible d'en retrouver l'origine et de les repérer une fois qu'ils circulent sur les réseaux sociaux. Les entreprises qui gèrent ces

derniers doivent aussi filtrer et limiter la propagation virale de contenu artificiel et des campagnes de désinformation bien orchestrées qui y font appel. Pour aider les utilisateurs à dénoncer la désinformation, les médias grand public peuvent effectuer rapidement des vérifications des faits et de l'authenticité des contenus. La population doit également renforcer sa connaissance des productions synthétiques et sa sensibilité à ces dernières, et il faut l'encourager à éviter de relayer des informations non fiables. Enfin, les organismes gouvernementaux ont un rôle essentiel à jouer pour orienter les stratégies nationales de recherche, afin que davantage de ressources soient investies dans l'étude de mesures visant à lutter contre les hypertrucages, mais aussi pour axer les efforts législatifs sur le contrôle du problème.

Commercialisation de l'intelligence artificielle : usage en technologie, dans l'industrie et dans le monde des affaires

Les avancées de l'intelligence artificielle (IA) ne semblent pas devoir ralentir et de nouvelles capacités technologiques apparaissent à un rythme qui rend difficile d'imaginer ce qui pourrait bientôt transformer en profondeur la vie quotidienne et la gestion des affaires et ce qui pourrait devenir obsolète. Les entreprises technologiques comme OpenAI, Google, Microsoft, Meta, NVIDIA, Apple et Adobe investissent fortement dans la recherche-développement et intègrent les fonctions novatrices faisant appel à l'IA à leurs produits aussi vite que possible, pour ne manquer aucune occasion de créer une rupture avec le passé.

Les débats actuels sur l'IA concernent principalement l'IA générative et les produits faisant appel à un transformateur, comme ChatGPT ou GPT-4. Ces technologies sont intégrées à des moteurs de recherche comme Bing (bing.com), tandis que des quantités phénoménales d'images synthétisées par l'IA grâce à des modèles de diffusion circulent sur les médias sociaux. L'IA générative, quand elle est utilisée pour générer des contenus, englobe la création et la manipulation d'images, de vidéos, d'enregistrements sonores et même de contenus en trois dimensions (3D). Tous ces produits peuvent être générés à partir de données d'entrée (textes, images, sons, etc.) ou non (bruit aléatoire).

Les activités et les occasions commerciales dérivées de l'IA sont sans limite et concernent presque tous les secteurs d'activités (télécommunications, santé, transport, éducation, énergie, divertissement, etc.). Le présent article met l'accent sur ce qui suit :

1. la commercialisation récente des principales technologies d'IA générative permettant de synthétiser des contenus (ce qui est réel et ce qui ne l'est pas);
2. les nouvelles capacités technologiques et les produits qui pourraient créer une rupture à l'avenir.

Les hypertrucages, des contenus synthétiques encore inoffensifs

Depuis l'apparition de ce qu'on appelle les technologies d'hypertrucage, le public se préoccupe des risques qu'elles présentent sur le plan de

la désinformation, de la fraude et du harcèlement, surtout qu'elles ont été largement adoptées pour créer de la pornographie non consensuelle. Ces technologies permettent de synthétiser des contenus, souvent à l'aide de l'IA générative, pour manipuler des sujets humains dans une vidéo (par exemple par le remplacement ou la manipulation de visages, ou le doublage). Le risque qu'elles soient utilisées à des fins malveillantes est particulièrement palpable, parce qu'elles peuvent générer des contenus extrêmement convaincants (résultats que seuls des studios professionnels d'effets spéciaux pouvaient obtenir auparavant), donc faire faire et dire n'importe quoi à des gens sur des vidéos. De plus, elles sont accessibles à tout le monde, car il est facile de les apprivoiser et elles ne nécessitent aucune compétence technique.

Bien que les technologies employées pour manipuler des contenus aient constamment progressé, les hypertrucages en tant que tels semblent relativement anodins, donc pas aussi catastrophiques que de nombreux experts l'avaient prévu. Ils ont été utilisés dans des cas de fraude et de harcèlement en ligne, dans le contexte d'élections politiques ou par des activistes russes au cours de la guerre en Ukraine, mais jusqu'ici, ils n'ont été un outil de désinformation efficace (surtout par rapport aux fausses nouvelles en général). Bon nombre des hypertrucages (audio et vidéo) qui circulent sur Internet sont créés par des personnes qui ne sont pas expertes, donc même s'ils sont impressionnants, ils paraissent toujours artificiels et il n'y a pas besoin d'algorithmes de pointe pour les détecter.

Élaboration de produits à partir de l'IA générative

Applications mobiles, filtres de réalité augmentée et vidéos sur les médias sociaux

Outre certaines applications mobiles et certains filtres de réalité augmentée tenant du gimmick et du divertissement (Snap, TikTok, etc.), les technologies d'hypertrucage peuvent ne pas sembler avoir d'utilisation plus profonde que la manipulation de vidéos et l'ajout de nouveaux effets visuels aux publications sur les réseaux sociaux. Les outils les plus communs servent à incorporer le visage d'un usager

dans un extrait vidéo (par exemple Zao, Reface) et à remplacer le visage d'un utilisateur par celui d'une célébrité (Impressions.ai). Pour utiliser le premier, il suffit de téléverser une seule photo et de choisir une vidéo parmi une sélection. Pour le deuxième, il faut téléverser une vidéo et sélectionner le modèle préentraîné d'une célébrité. La réinterprétation d'une scène avec un autre visage grâce à une technique appelée « first order motion » (mouvement de premier ordre) gagne également en popularité : le portrait choisi arbitrairement d'une personne est téléversé et la scène immédiatement rejouée. Ainsi, les internautes peuvent créer des vidéos virales de personnalités politiques qui chantent ou animer les photos de personnes disparues (par exemple, DeepNostalgia, MyHeritage).

La demande de nouveaux outils d'expression personnelle plus impressionnants incite la recherche universitaire et commerciale à repousser les limites en matière de synthèse de contenus (par exemple pour augmenter la résolution des produits, offrir des services en temps réel, procurer plus de contrôle aux usagers, éliminer des artéfacts ou améliorer l'accessibilité des outils). Cela a permis d'élaborer de nouveaux filtres plus perfectionnés (qui permettent de rajeunir les sujets, de les faire changer de genre ou de leur donner l'apparence de personnages de dessins animés) et de créer des algorithmes offrant des modalités d'entrée novatrices (réinterprétation d'une scène à partir d'une seule image, entrée de consignes par texte, avatars photoréalistes tirés de vidéos, etc.).

Assistants virtuels, vidéos de marketing et outils de traduction universelle

Bien que les personnages numériques soient essentiels à de nombreuses applications de divertissement, d'autres secteurs commerciaux ont étudié la possibilité de les employer pour améliorer, automatiser et élargir leurs services à l'aide de l'IA générative. Plusieurs sociétés ont élaboré des assistants virtuels humanoïdes (par exemple Soul Machines ou Uneeq), mais les consommateurs n'en sont pas friands en raison de leur apparence, qui les place dans la « vallée de l'improbable » (nom donné à l'inconfort ressenti face à des visages créés par ordinateur imparfaits)[118, 119]. Malgré des avancées

technologiques dans l'utilisation de processeurs graphiques (par exemple, Epic Games/MetalHumans, unrealengine.com) ou de l'IA générative pour améliorer le photoréalisme de ces avatars (par exemple, Samsung Neon, Pinscreen, etc.), les assistants virtuels peinent toujours à remplacer les êtres humains. Leurs réactions sont encore trop simplistes et leur voix et les expressions de leur visage manquent souvent d'émotion et d'empathie.

Cependant, compte tenu des récents progrès accomplis en matière de grands modèles linguistiques (LLM) comme ChatGPT et des dernières études sur la synthèse de mouvements, l'adoption massive d'agents humanoïdes faisant appel à l'IA très convaincants et réalistes pourrait être plus proche que jamais (deux ou trois ans), surtout si ces agents peuvent interagir en temps réel. Pour le moment, plusieurs jeunes pousses (comme Synthesia ou Colossyan) se penchent sur l'utilisation de vidéos créées de toutes pièces à partir d'enregistrements de véritables êtres humains, sans interaction, pour produire des films de marketing ou de formation à grande échelle pour des entreprises. Il est possible de choisir un acteur ou une actrice et une voix sur une interface Web pour produire du contenu vidéo automatiquement sur un serveur à partir d'un texte. Ces solutions font généralement appel à une application permettant de passer du texte écrit au texte dit (application fournie par un tiers ou « maison », qui permet de personnaliser les voix) et d'un générateur de vidéos à partir d'entrées vocales, entraîné à l'aide d'un extrait sonore et d'images d'une vidéo (par exemple, pour Synthesia, il faut fournir 10 minutes de vidéo d'un acteur lisant son texte face à la caméra, dans un studio bien éclairé).

Ces méthodes sont plus perfectionnées que le populaire algorithme wav2lip et donnent des résultats de meilleure qualité, avec une plus haute résolution. Des technologies semblables ont aussi été adoptées par la société chinoise Tencent et par des entreprises coréennes comme DeepBrain pour créer des lecteurs et lectrices de nouvelles et des supports de marketing à grande échelle. Tencent, par exemple, ne facture que 145 $ US pour chaque sujet (moitié du corps ou corps complet) et prend en charge à la fois l'anglais et le chinois. Malgré leur grande fidélité, les résultats obtenus à partir de la voix manquent toujours de fluidité pendant les conversations, donc leur adoption reste limitée.

Google a récemment annoncé à sa conférence I/O le lancement d'un service de traduction universelle pour entreprises appelé « Universal Translator », qui permet aux créateurs de contenu éducatif de traduire leurs vidéos dans de nombreuses langues. À partir de la traduction de ce qui est dit, cette application crée les mouvements des lèvres correspondants dans les vidéos. Cette traduction vocale (interprétation) est également produite à l'aide d'un modèle d'interprétation générative, qui imite la voix et le ton d'un orateur dans une autre langue. Pour l'instant, cette solution n'est offerte qu'à un nombre restreint de créateurs de contenus habilités (comme l'Université d'État de l'Arizona), ce qui peut en limiter les utilisations malveillantes.

Réduction des coûts et des délais pour les effets spéciaux et le doublage à Hollywood

Que ce soit pour générer des cascades numériques, ramener à la vie des vedettes décédées ou rajeunir un acteur ou une actrice, les effets spéciaux créés par ordinateur ont été largement utilisés dans certaines des superproductions les plus mémorables (par exemple *Star Wars, Dangereux 7, Terminator : Sombre destin* et *L'étrange histoire de Benjamin Button*). Cependant, ces effets sont généralement réalisés par des studios spécialisés de pointe (Industrial Light & Magic, Weta Digital, MPC, Framestore, etc.), coûtent des millions de dollars et nécessitent des mois de travail pour quelques secondes de film. Les effets spéciaux sur les visages humains sont particulièrement onéreux et difficiles à réaliser à cause de la « vallée de l'improbable ».

Quand des applications libres d'accès permettant de créer des hypertrucages (comme les GAN utilisés pour remplacer les visages et Deep Face Lab) ont été rendues disponibles sur Internet, les amateurs et les artistes travaillant à partir d'hypertrucages ont commencé à créer de courtes vidéos divertissantes, dans lesquelles ils intervertissaient des célébrités. Même s'il était possible de générer des hypertrucages très convaincants, leur résolution était toujours trop basse pour les productions cinématographiques. Cependant, ces méthodes ont rapidement attiré l'attention de producteurs d'effets spéciaux, qui y ont vu un outil pouvant permettre de perfectionner leurs méthodes conventionnelles pour leur faire économiser et

améliorer la narration. Certains, comme Industrial Light & Magic (ILM), ont étudié l'utilisation de technologies d'hypertrucage pour rajeunir des acteurs (comme Mark Hamill dans *Star Wars* ou Harrison Ford, dans *Indiana Jones 5*). Pour ce faire, ils ont remplacé les visages des acteurs âgés ou de leurs cascadeurs par des images neuronales produites à partir de vidéos de ces mêmes acteurs plus jeunes et les ont combinées à des modèles 3D et à des techniques de compositing vidéo.

Toutes les jeunes pousses dans le domaine de l'IA, comme Pinscreen et Metaphysic, offrent des solutions complètes de création d'effets spéciaux à l'aide de l'IA qui permettent de substituer des visages dans les productions cinématographiques. Metaphysic est connue pour ses hypertrucages mettant en scène Tom Cruise, qui circulent sur TikTok, et pour avoir implanté le visage d'Elvis dans un extrait d'*America's Got Talent*.

Pinscreen a innové en élaborant un certain nombre de techniques d'animation neuronale des visages reposant sur des GAN (notamment PaGAN, pour « photoreal avatar GAN », soit un GAN permettant de générer des avatars photoréalistes), à l'origine conçues pour améliorer le réalisme d'avatars en 3D destinés à des interactions en 3D ou à des métavers. En 2022, la société a commencé à se tourner vers les effets spéciaux grâce à un partenariat avec Netflix et Amazon Studios. Elle s'est ainsi mise à travailler sur plusieurs séries télévisées à grand retentissement (comme *Manifest*), des superproductions (comme *La petite Nemo* et *Le monde des rêves*) et des publicités (pour Nike, Balenciaga, etc.) faisant appel à l'IA générative. Les services d'effets spéciaux faisant appel à l'IA comprennent le traitement de bout en bout du remplacement de visages, l'animation des visages, le vieillissement et le rajeunissement, ainsi que le doublage. Le principal avantage de Pinscreen est de pouvoir travailler sur de très courtes scènes de films et de produire des contenus haute-fidélité, 4K à grande gamme dynamique, ce qui permet de traiter des prises de vue en gros plan, des points de vue extrêmement latéraux et des éclairages spectaculaires et changeants. Ce processus nécessite une amplification de données au moyen de GAN spécialisés et des procédures d'amélioration faisant appel à l'IA pour générer des données inédites

à partir des quelques images collectées dans les films, ainsi qu'un renforcement de l'architecture en vue de créer des vidéos haute résolution cohérentes sur le plan temporel.

Malgré la demande croissante de services d'effets spéciaux faisant appel à l'IA, comme le remplacement de visages, le vieillissement et le rajeunissement, leur utilisation reste relativement marginale et varient beaucoup d'une émission à l'autre. Elle pourrait devenir importante dans le marché du doublage des films et des émissions de télévision, car elle permettrait de regarder ces productions dans n'importe quelle langue avec les mouvements des lèvres des acteurs parfaitement synchronisés avec ce qu'ils ou elles disent. Les longs métrages sont bien plus compliqués à traiter que les vidéos qui sont tournées dans un cadre contrôlé (comme la lecture de nouvelles ou le tournage de supports de marketing ou de formation), en raison de la complexité des scènes, du manque de données d'entraînement et des critères de qualité extrêmement exigeants du cinéma (image 4K à grande gamme dynamique).

En 2022, Pinscreen est devenue la première société au monde à doubler intégralement un long métrage complet à l'aide de son système exclusif reposant sur l'IA générative, soit le film *The Champion* (traduit de l'allemand et du polonais vers l'anglais). Pour ce faire, elle a combiné l'IA générative de pointe à un processus intégré d'effets spéciaux pour doubler ce film de 90 minutes en moins de trois mois. Son approche permet de traiter un film déjà tourné et ne nécessite que des enregistrements vidéo supplémentaires des acteurs jouant le doublage. D'autres intervenants sur le marché des effets spéciaux faisant appel à l'IA, comme Flawless.ai, essaient d'entrer sur le marché du doublage, mais disposent de moyens techniques limités : ils ne peuvent produire que de nouvelles animations des visages à partir d'enregistrements vocaux, au lieu de travailler à partir de vidéos. Ils ont effectué le doublage de certains extraits vidéo, mais pas de films complets.

Production d'images à partir d'invites textuelles grâce aux modèles de diffusion

Grâce à des percées comme Dall-E d'OpenAI et aux dernières avancées des modèles de diffusion avec transformateur, comme Stable Diffusion, il est maintenant possible de produire des images plus efficacement et de façon plus poussée qu'avec les méthodes habituelles faisant appel aux GAN pour ce qui est de la qualité, de la résolution et de la diversité. Cette dernière propriété est particulièrement importante, car elle est très efficace pour synthétiser des images à partir de textes : les utilisateurs entrent les invites textuelles de leur choix, et le modèle synthétise une image correspondant précisément à cette invite. La fonction d'invite textuelle repose généralement sur l'utilisation d'un encodeur de type « CLIP » capable de faire correspondre l'invite à une incrustation textuelle, qui est ensuite utilisée comme condition pour créer une image par un processus d'élimination progressive du bruit (le générateur), habituellement grâce à l'utilisation répétée d'un réseau de neurones profond (RNP) qui fonctionne à l'aide d'une architecture U-Net permettant de générer des images à partir d'images.

Bien qu'il soit plus simple et plus fiable que celui des GAN, l'entraînement des modèles de diffusion nécessite des ressources considérables : généralement des semaines d'entraînement et des centaines de processeurs graphiques hautement performants (de type A100). Par conséquent, ces modèles sont souvent entraînés par des sociétés disposant de grandes quantités de ces processeurs (comme OpenAI, Stability.ai ou Google). Les laboratoires universitaires et les sociétés de plus petite envergure se contentent de modèles préentraînés qu'ils adaptent à leurs besoins. Les dernières applications commerciales, et les plus populaires, sont notamment Dall-E2, Midjourney (robot sur Discord), la solution offerte par Stability.ai (Dream Studio, une interface Web) et les interfaces de protocoles d'application. Bien que ces outils permettent de produire des images incroyablement réalistes, celles-ci ont toujours tendance à comporter des artéfacts visibles et il n'est pas encore possible de contrôler les menus détails des images synthétisées. Un certain degré de contrôle a récemment été obtenu au moyen d'esquisses ou de grandes lignes abstraites des images d'origine (comme ControlNet), mais les images ainsi engendrées

comportent toujours des détails ou des aspects imprévisibles. En conséquence, les méthodes reposant sur la diffusion ne peuvent pas encore générer de vidéos ayant la qualité requise pour des productions professionnelles, car il n'est pas facile de les contrôler et d'en garantir la cohérence sur le plan temporel.

Résumé et possibilités futures

Les capacités de l'IA générative à synthétiser des contenus (images, vidéos et sons) sont en constante évolution. Les images ainsi produites sont de meilleure qualité (plus haute résolution, moins d'artéfacts et résultats plus réalistes sur le plan sémantique) et plus variées, ce qui permet d'utiliser des invites textuelles en langage naturel à l'entrée. Comme à l'avènement des GAN, les professionnels de la recherche s'efforcent d'offrir un meilleur contrôle, des résultats plus prévisibles et des images cohérentes sur le plan temporel pour les vidéos, ainsi que la possibilité de gérer d'autres modalités, comme le contenu neuronal en 3D. En raison de l'accessibilité de cette technologie et de sa capacité à produire du contenu convaincant, la population s'est inquiétée de ce qu'elle puisse être utilisée à des fins malveillantes. Pour l'instant, ces technologies de synthèse de contenus et les hypertrucages n'ont pas été exploités largement à des fins offensives, même s'il s'agit d'une menace.

Ces prochaines années, la société devrait connaître d'autres percées technologiques dans le domaine de l'IA générative. Ces percées ouvriront de nouvelles perspectives commerciales, notamment plusieurs services généraux de création de vidéos en ligne (par exemple, un YouTube pouvant produire la vidéo souhaitée immédiatement à partir de n'importe quelle invite textuelle), des vidéos pleinement interactives en temps réel (comme des publicités capables d'interagir avec l'auditoire en temps réel), ainsi que des environnements totalement immersifs et photoréalistes générés par l'IA pour des métavers. Compte tenu des nouveaux casques de réalité augmentée ou de réalité virtuelle dont la sortie a été annoncée, comme le Vision Pro d'Apple et le MetaQuest 3 de Meta, la demande de contenu 3D sophistiqué devrait croître et l'IA générative jouer un rôle clé dans la création de contenu.

Répercussions des hypertrucages sur la sécurité nationale

La société est à l'aube d'une ère où la réalité peut être manipulée, la vérité déformée et la confiance brisée. Les technologies d'hypertrucage font peser une menace grave et imminente sur la sécurité nationale. Il existe aujourd'hui divers moyens technologiques de créer des hypertrucages, dont la facilité d'emploi varie. Certains des outils les plus utilisés permettent de remplacer des visages[120], de faire du doublage[121] ou de cloner des voix[122], ou encore de produire des hypertrucages grâce à des réseaux antagonistes génératifs (GAN)[123].

Les adversaires, qu'il s'agisse d'acteurs étatiques, de criminels ou d'organisations clandestines, s'emploient constamment à perfectionner leurs compétences dans l'art de la manipulation publique. Les hypertrucages peuvent leur faciliter la tâche pour ce qui est d'exploiter des failles, de semer la discorde dans la population ou d'ébranler les fondements mêmes de la démocratie.

Le présent document donne un aperçu de la capacité, de l'impact et des dangers sous-estimés des hypertrucages, et souligne l'urgence de se doter de stratégies efficaces et complètes pour lutter contre cette menace très complexe qui évolue rapidement.

Incidences, cas de figure et conséquences[124]

La capacité des hypertrucages à démolir des réputations en un instant illustre leur nature insidieuse. La dissémination stratégique d'une seule vidéo de synthèse peut causer l'indignation publique, des pertes financières et des dommages irréparables à des personnes et à des organisations. Imaginez une vidéo « hypertruquée » dans laquelle un PDG fait des commentaires racistes. Les répercussions d'une telle attaque intentionnelle pourraient entraîner une chute catastrophique du prix de l'action de sa société, ce qui ébranlerait non seulement la stabilité financière de cette dernière, mais aussi sa crédibilité aux yeux des investisseurs et des autres parties prenantes. Les hypertrucages pornographiques ont déjà fait des dégâts. Une étude datant de 2019 a conclu que 96 % des 14 000 vidéos « hypertruquées » mises en ligne étaient pornographiques. Or, ce nombre est infime comparé au volume actuel de pornographie non consensuelle mettant en scène des personnalités, comme des journalistes ou des célébrités[125].

La capacité de nuire des hypertrucages ne se limite pas aux menaces externes. Des « initiés » peuvent procurer un avantage imbattable à des adversaires et à des acteurs malveillants. Un membre du personnel ayant accès à des informations sensibles peut se servir des hypertrucages pour orchestrer la fuite de données classifiées ou mener d'autres activités illicites, et ainsi porter atteinte à la sécurité nationale, compromettre l'intégrité de son organisation, mettre en danger des collègues ou causer d'importantes pertes financières. Des protocoles rigoureux de filtrage et de surveillance du personnel doivent donc être mis en place pour repérer toute menace interne potentielle et y remédier. En outre, il est impératif d'élaborer des technologies poussées de détection des hypertrucages conçues spécifiquement pour prévenir les menaces internes. Enfin, celles et ceux qui abusent de leur position de confiance doivent être sévèrement punis, afin de décourager d'autres tentatives.

Les hypertrucages peuvent être incorporés à des campagnes de fraude psychologique visant à manipuler des personnes et des organisations à des fins malveillantes. Par exemple, une vidéo « hypertruquée » mettant en scène un être cher qui demande une forte somme d'argent, surtout s'il semble en danger, peut entraîner d'importantes pertes financières pour la cible. Afin de lutter contre ce fléau, il est primordial de mieux sensibiliser la population à la fraude psychologique faisant appel aux hypertrucages, d'élaborer des protocoles efficaces de détection et d'intervention, et de favoriser la coopération entre corps policiers.

Les hypertrucages ont des incidences économiques qui sont loin de se limiter aux atteintes à la réputation, car ils se sont aussi avérés utiles pour l'espionnage économique. Les personnes et les organisations adverses peuvent exploiter des hypertrucages pour attaquer des sociétés ou des industries, pour tenter d'obtenir un avantage concurrentiel ou pour perturber des secteurs essentiels de l'économie. Afin d'y remédier, il faut notamment appliquer des technologies de détection efficaces aux communications d'entreprise, améliorer la culture médiatique des chefs d'entreprise et établir des protocoles clairs de lutte contre l'espionnage économique faisant appel à des hypertrucages. En outre, ces derniers peuvent faciliter la fraude

financière, si des individus s'en servent pour se faire passer pour des personnes ayant accès à des informations financières sensibles. Afin d'éviter cela, il convient de mettre en place une authentification multifactorielle pour les transactions financières sensibles.

Les hypertrucages ont le potentiel de pénétrer les systèmes les plus cruciaux et les plus sécurisés d'une nation, ce qui constitue un grave danger pour la cybersécurité. Par exemple, des criminels peuvent les exploiter pour contourner les systèmes d'authentification biométrique, ce qui leur permettrait d'obtenir un accès non autorisé à des installations sécurisées et à des informations personnelles sensibles. Les conséquences de tels piratages sont redoutables : non seulement elles compromettent la vie privée des gens, mais elles menacent la sécurité nationale. Pour les contrer, il faut instaurer des mesures anti-tromperie pour les systèmes d'authentification biométrique, afin que les forteresses numériques restent inviolables.

Les hypertrucages se sont aussi avérés des armes furtives, permettant de mener des opérations clandestines à l'abri de tout soupçon. De fausses preuves peuvent être fabriquées, ou des vidéos de surveillance manipulées, ce qui ébranle la confiance dans les éléments de preuve visuels et entrave les opérations de renseignement. Pour préserver l'intégrité de ces opérations et renforcer la sécurité nationale, il est indispensable de se doter de techniques d'analyse judiciaire permettant de détecter les hypertrucages dans les preuves audiovisuelles.

Incidences pour la sécurité nationale et le renseignement

Pour prévenir les piratages reposant sur les hypertrucages ou en atténuer l'effet, il est fondamental de renforcer la sensibilisation aux menaces associées aux hypertrucages et aux stratégies de lutte contre ces dernières chez les personnes et les organisations qui risquent grandement de faire l'objet de tentatives d'extorsion ou de coercition, mais aussi d'élaborer des protocoles de cybersécurité efficaces.

Comme les hypertrucages peuvent permettre de fabriquer de toutes pièces des preuves frauduleuses ou de manipuler la perception de la population, ils peuvent entraîner une distorsion de la vérité. Pour y

remédier, il est primordial de renforcer la transparence et la responsabilisation dans l'utilisation d'éléments de preuve liés aux hypertrucages, d'élaborer des systèmes de détection et de vérification, et de promouvoir la vérification des faits et des contenus par les médias lorsqu'ils font des reportages.

La perturbation des infrastructures essentielles ou des organismes gouvernementaux peut avoir de graves conséquences. Pour éviter cela, il faut miser sur l'application de protocoles efficaces de cybersécurité, sur la sensibilisation de la population aux menaces associées aux hypertrucages et sur la conduite régulière d'exercices visant à former les organismes gouvernementaux au repérage et à la neutralisation des hypertrucages.

Dans le cadre de campagnes de cyberguerre, les hypertrucages peuvent être utilisés contre des infrastructures essentielles, pour perturber le fonctionnement des gouvernements et pour semer la panique dans les marchés financiers. Afin de prévenir les cyberattaques faisant appel à des hypertrucages, il est crucial de sensibiliser les organismes gouvernementaux et les fournisseurs d'infrastructures essentielles aux menaces de cyberguerre découlant des hypertrucages, d'élaborer des protocoles efficaces de détection et d'intervention, et de favoriser la coopération internationale. Par ailleurs, les hypertrucages peuvent affecter les relations diplomatiques ou retarder la négociation de traités. La prise en charge de ce risque passe notamment par la conception d'une technologie de détection efficace des hypertrucages dans les communications diplomatiques, l'amélioration de la culture numérique des diplomates et des représentants du gouvernement, et l'établissement de voies de communication claires en prévision d'éventuels incidents diplomatiques impliquant des hypertrucages.

Les organisations terroristes ont certainement compris le potentiel des hypertrucages pour la diffusion de propagande et la coordination d'attaques. Même en l'absence d'hypertrucages, le terrorisme met en péril la sécurité d'innocents et la stabilité de l'infrastructure essentielle. Par conséquent, il est capital d'améliorer la conscience qu'ont les corps policiers et les services de renseignement des menaces

terroristes associées aux hypertrucages, de mettre en place des protocoles de détection et d'intervention efficaces, et de s'engager résolument en faveur de la coopération internationale pour lutter contre cette utilisation délétère des hypertrucages.

Les hypertrucages peuvent aussi répandre de la désinformation sur les moyens et les mouvements militaires, ce qui pourrait causer des conflits armés. Pour éviter ce type de conflit, il est essentiel de sensibiliser les armées et les services de renseignement aux menaces de désinformation associées aux hypertrucages, de mettre au point des protocoles de détection et d'intervention efficaces et de promouvoir la coopération internationale.

Incidences pour la démocratie

L'utilisation des hypertrucages pour répandre de fausses informations ou de la propagande peut semer la confusion et susciter la méfiance de la population. Afin de combattre ce fléau, il est crucial d'élaborer des technologies poussées de détection des hypertrucages, d'améliorer la culture médiatique et de sanctionner la propagation d'hypertrucages nuisibles.

Parce qu'ils permettent de manipuler l'opinion publique, ce qui peut jouer sur les résultats d'élections, voire les faire basculer, les hypertrucages peuvent entraver le bon fonctionnement des processus démocratiques. Afin de sensibiliser le public aux hypertrucages et à la désinformation en ligne et d'inciter les partis politiques à collaborer pour éviter que l'issue de l'élection britannique de 2019 soit influencée par ces hypertrucages, Future Advocacy, un groupe défendant diverses causes, a créé et diffusé une vidéo qui mettait en scène les candidats Boris Johnson et Jeremy Corbyn se concédant mutuellement la victoire[126]. Pour remédier à cette utilisation des hypertrucages, il faut élaborer des technologies permettant de les repérer dans le cadre des campagnes politiques, développer la culture médiatique de l'électorat et sanctionner la propagation d'hypertrucages nuisibles pendant les élections.

Les hypertrucages peuvent aisément répandre de la désinformation contre certains représentants ou certains ministères et organismes

du gouvernement, ce qui peut susciter la méfiance de la population et entraîner des poursuites. Face à ce danger, il est donc essentiel d'adopter des technologies permettant de détecter les hypertrucages dans les communications gouvernementales, d'améliorer la culture médiatique des représentants du gouvernement et d'établir des protocoles clairs de lutte contre les campagnes de désinformation impliquant des hypertrucages.

Les hypertrucages peuvent être utilisés pour créer des vidéos de propagande convaincantes visant à influencer ou à faire basculer l'opinion publique. Face à cette menace, il est crucial de renforcer la transparence et de financer les commissions électorales afin qu'elles puissent enquêter pour prévenir la propagation des hypertrucages. De la même façon, dans le contexte des opérations d'ingérence, il faudra élaborer des processus rigoureux de vérification des faits, éduquer la population aux risques associés aux hypertrucages et promouvoir l'esprit critique.

En outre, les hypertrucages permettent de manipuler (ou de créer de toutes pièces) des preuves dans le cadre d'enquêtes criminelles ou de poursuites juridiques, ce qui peut se traduire par des condamnations erronées. Par conséquent, il faut établir des directives sur l'emploi de preuves numériques ou des procédures sur la vérification des vidéos utilisées comme éléments de preuve en cour.

Stratégies d'endiguement et d'atténuation

Les stratégies d'endiguement et d'atténuation peuvent être d'ordre technologique, juridique ou social et nécessitent une collaboration et une gouvernance responsables.

Il faut concevoir et développer des technologies de détection des hypertrucages. Les algorithmes d'apprentissage automatique peuvent servir à déceler des incohérences dans les vidéos, les sons ou les images. La technologie de la chaîne de blocs pourrait aussi être utilisée pour créer des sauvegardes inaltérables des contenus, afin de les rendre difficiles à falsifier.

Il faut également envisager des stratégies juridiques rendant illégales la production et la diffusion des hypertrucages (au-delà des répercussions juridiques associées à la « fraude »). Les lois devraient protéger les personnes dont la réputation est entachée par les hypertrucages.

Il faut encourager et promouvoir la culture médiatique et l'esprit critique chez la population. Il serait utile d'éduquer le public afin qu'il puisse repérer les hypertrucages et connaisse les risques et les conséquences potentielles qui y sont associés. Par ailleurs, il faut aussi soutenir les journalistes et les organes de presse qui vérifient les faits et les contenus avant leur publication.

Il faut instituer des partenariats et une collaboration internationaux pour combattre les hypertrucages. De telles ententes peuvent faciliter l'échange de connaissances, de ressources et d'expertise, et favoriser une lutte coordonnée contre les menaces associées aux hypertrucages.

Enfin, les sociétés technologiques doivent assumer la responsabilité du contenu échangé sur leurs plateformes. Par exemple, elles peuvent investir dans l'élaboration et la mise en service d'outils de détection et simplifier le signalement des hypertrucages à leur clientèle, afin d'en faciliter le retrait.

Conclusion

Le spectre des hypertrucages fait peser une menace sans précédent sur la sécurité nationale. L'évolution rapide et la prolifération de cette technologie n'exigent rien de moins qu'une action résolue et complète.

Pour protéger les nations démocratiques, les gouvernements doivent investir dans des technologies poussées de détection des hypertrucages capables de démasquer des imposteurs numériques et de révéler au grand jour des intentions malveillantes. En même temps, les cadres juridiques doivent être renforcés pour rendre illégales la création et la diffusion des hypertrucages, mais aussi pour offrir de solides protections aux personnes dont la réputation est fragile.

Toutefois, il sera impossible de remporter la bataille contre les hypertrucages à l'aide de la technologie et de la loi seulement. Les

citoyens doivent être munis des armes que sont l'esprit critique et la culture médiatique, afin de pouvoir distinguer le vrai du faux. En cultivant une société sceptique, informée et résiliente, les gouvernements peuvent opposer un bouclier aux effets corrosifs des hypertrucages.

Ce combat dépasse les frontières et nécessite une solidarité mondiale. La collaboration entre nations, qui permettra d'échanger connaissances, ressources et expertise, sera la pierre angulaire de la défense contre cette menace. Ensemble, il est possible de nouer une alliance internationale contre les hypertrucages qui soit résiliente et inébranlable afin de préserver l'intégrité de nos sociétés.

Par ailleurs, les sociétés technologiques doivent être tenues de rendre des comptes. Elles exercent un pouvoir et une influence énormes, donc elles devraient prioriser l'élaboration et l'emploi de technologies visant à détecter les hypertrucages, afin de créer des mécanismes de signalement conviviaux et de retirer rapidement les hypertrucages de leurs plateformes. Ainsi, elles deviendraient des défenseuses de la sécurité nationale.

Les hypertrucages représentent un défi colossal, mais les gouvernements démocratiques et leurs alliés doivent rester unis et vigilants contre lui, en plus de tenir fermement leurs engagements à protéger la sécurité démocratique, diplomatique et économique.

Extraire du sens des contenus artificiels : le renseignement à l'ère des hypertrucages

Il est bien établi que les vidéos « hypertruquées » auront probablement un effet négatif sur notre environnement informationnel. Cependant, l'impact des hypertrucages sur les services de renseignement et les organismes chargés de protéger la sécurité nationale est moins étudié, notamment pour ce qui est du contexte de la menace dans lequel ceux-ci opèrent, des méthodes de collecte de renseignements, de l'utilisation de procédés automatisés de détection de la menace, de la réception des produits de renseignement et des dilemmes éthiques.

Même si les hypertrucages sont impressionnants sur le plan technologique, il y a plus de chances qu'ils fassent évoluer les menaces pour la sécurité nationale et le renseignement plutôt qu'ils en engendrent de nouvelles. En revanche, les dilemmes éthiques qu'ils suscitent risquent d'engendrer un ensemble de difficultés plus graves et nécessiteront un excellent jugement pour faire tout un éventail de choix difficiles.

Évolution des menaces

En règle générale, la littérature sur les hypertrucages met l'accent sur les menaces pour les sociétés démocratiques qui y sont associées. Il importe de souligner que la plupart de ces menaces ne seront probablement pas inédites : il s'agira de progrès et du renforcement des activités liées à la menace que les services de renseignement et de sécurité nationale gèrent déjà. Cela comprend la désinformation, le ciblage du personnel gouvernemental et militaire par des forces adverses, l'hameçonnage et la fraude psychologique, ainsi que l'imitation des données biométriques.

Désinformation

Dans le présent chapitre, la désinformation est définie comme suit : « toute fausse information visant à manipuler des personnes, des organisations et des pays, à leur causer des préjudices ou à les orienter dans la mauvaise direction ». Dans le même ordre d'idées, la malinformation est de l'information dérivée de la vérité, mais souvent exagérée de façon trompeuse, ce qui peut avoir des effets néfastes[127]. La malinformation découle habituellement d'informations volées ou

piratées, dont une partie peut être altérée pour donner de la crédibilité à un récit faussé sur lequel un adversaire souhaite insister, puis publiée sur Internet afin d'y être distribuée.

Les environnements médiatiques en ligne actuels regorgent de désinformation et de malinformation, qui servent un certain nombre d'objectifs. Du point de vue de la sécurité nationale, cela implique globalement des activités d'ingérence étrangère et de radicalisation. Il est bien établi que des États comme la Russie, la Chine et l'Iran mènent des campagnes de désinformation et de malinformation au service de leurs objectifs politiques. De la même façon, les groupes extrémistes violents répandent des discours sur l'effondrement de la société, la corruption des institutions, les théories du complot et la nécessité de commettre des actes violents et révolutionnaires pour redonner à l'humanité le statut qui lui revient (cela peut toutefois aussi se faire par l'ironie et des mèmes)[128]. Même si leurs objectifs politiques diffèrent, il importe de noter que ces deux groupes sont souvent unis dans leur volonté de discréditer les institutions démocratiques et d'en minimiser le rôle, d'amplifier les théories conspirationnistes et d'attiser la méfiance envers ce qu'ils considèrent en général comme « le système ».

Pour ces acteurs, les hypertrucages présentent l'avantage d'abaisser le coût des campagnes de désinformation. Auparavant, il fallait du temps, des efforts et des compétences pour produire des supercheries et de fausses informations. Aujourd'hui, grâce aux hypertrucages, on peut générer sans délai des documents utilisables rapidement et diffusables dans le monde entier encore plus vite. En fonction du degré d'accessibilité et de propagation des outils servant à créer des hypertrucages, ces derniers pourraient aussi permettre de prendre part à des guerres d'information, ce qui rendrait plus trouble un environnement informationnel déjà complexe.

Ciblage du personnel gouvernemental et militaire

Des acteurs adverses emploieront probablement des hypertrucages pour s'attaquer à du personnel gouvernemental, militaire et chargé

de la sécurité nationale, habituellement en vue d'en faire des cibles ou de perturber leur travail.

Déjà, des campagnes de désinformation ont lieu aux endroits où sont postés des contingents occidentaux, afin d'attiser la méfiance et de dégrader les relations avec les populations. Par exemple, quand des unités canadiennes ont été déployées en Lettonie dans le cadre de la présence avancée rehaussée de l'Organisation du Traité de l'Atlantique Nord (OTAN), une opération de désinformation et de malinformation les ciblant a commencé[129]. Ces tentatives de semer la défiance n'ont pas cessé depuis 2017[130]. Des hypertrucages pourraient être utilisés dans le cadre de ces campagnes. Par ailleurs, les familles des militaires pourraient être prises pour cible par des hypertrucages concernant leurs proches en mission à l'étranger, l'objectif étant de causer de la détresse et d'infliger des blessures mentales.

Fraude psychologique

Les menaces liées à l'hameçonnage et à la fraude psychologique constituent une troisième source de préoccupation. La fraude psychologique est la manipulation des utilisateurs légitimes d'un outil visant à obtenir des informations confidentielles. En général, les auteurs utilisent le téléphone ou Internet pour se faire passer pour une personne ayant autorité sur leur victime, pour une collègue, pour un membre de la famille, voire pour du personnel de soutien technique, afin de l'amener à révéler des informations sensibles. Cela comprend aussi l'hameçonnage qui consiste, pour un acteur malveillant, à mystifier une victime ou à lui envoyer un courriel imitant une marque habituellement bien connue pour la convaincre de fournir des informations confidentielles[131].

Les hypertrucages, qui permettent de reproduire le visage, l'image et la voix des personnes, pourraient servir à tromper les gens de façon plus poussée. Cela fera certainement les choux gras des criminels, mais les services de renseignement adverses pourraient aussi les employer pour s'attaquer à des personnalités politiques, à des agents de renseignement et à d'autres détenteurs d'information classifiée, afin de gagner leur confiance en vue d'accéder à des données sensibles.

De plus, lors d'un conflit, des hypertrucages peuvent être utilisés comme ruses de guerre : des enregistrements vidéo ou sonores falsifiés pourraient être utilisés pour envoyer de faux ordres aux soldats, ou encore de fausses informations visant à perturber les opérations militaires.

Piratage des données biométriques

Outre la fraude sociale visant à obtenir des informations sensibles ou classifiées, des adversaires pourraient employer les hypertrucages pour imiter des données biométriques, afin d'avoir accès directement à ces informations. La recherche indique que les hypertrucages pourraient déjà avoir la capacité de tromper les lecteurs biométriques, comme les systèmes de reconnaissance faciale[132]. Étant donné qu'un nombre croissant d'applications collectent et utilisent les données biométriques, il est probable qu'un grand nombre d'institutions détenant ces données puissent les vendre ou être victimes de piratages[133]. En outre, ces données pourraient servir à créer des hypertrucages encore plus réalistes.

Collecte de renseignements

Afin de faire face aux menaces actuelles et à leur évolution, ainsi qu'aux futures, les services de renseignement et ceux qui sont chargés de la sécurité nationale doivent collecter les renseignements nécessaires à l'accomplissement de leur mandat. Malheureusement, il est probable que les hypertrucages affecteront ce travail au moins de deux façons : en générant du bruit et en affectant la fiabilité des informations de sources ouvertes.

Génération de bruit

Les États adverses peuvent recourir aux hypertrucages pour perturber la collecte de renseignements, y compris électromagnétiques, si des hypertrucages envahissent un espace informationnel, créant ainsi beaucoup de bruit ou de distractions. Les hypertrucages pourraient également être utilisés tactiquement contre une initiative précise, ou soupçonnée, de collecte.

Des hypertrucages pourraient aussi déformer la perception des sources humaines en leur faisant croire à la réalité d'une conversation, d'une vidéo ou d'un texte synthétiques. Ces sources pourraient ensuite transmettre cette information à des agents de renseignement en toute bonne foi. L'incapacité d'une telle source à faire la distinguer le vrai du faux pourrait affecter la collecte et l'analyse des renseignements.

Informations de sources ouvertes

Un deuxième problème concerne l'utilisation des informations de sources ouvertes par des organismes gouvernementaux et non gouvernementaux. L'invasion de l'Ukraine par la Russie en 2022 est le dernier événement mondial en date à avoir démontré l'importance et la valeur de l'OSINT et de son analyse[134]. Du moissonnage sur les médias sociaux à l'analyse d'images satellite publiques, les techniques d'OSINT servent à découvrir les mouvements de troupes, à localiser les fortifications défensives, à évaluer le moral des combattants, à confirmer des attaques ou des frappes militaires et les pertes qu'elles ont causées et à enquêter sur des crimes de guerre. Bien que leur qualité soit variable, les services de sécurité nationale tout comme les journalistes et les organisations humanitaires ont développé leurs propres méthodes ou trouvé des sources fiables pour alimenter leurs investigations. Par conséquent, les informations de sources ouvertes sont une cible de choix pour les hypertrucages. Les acteurs adverses qui souhaitent diviser les alliés, affaiblir leur détermination, nier des crimes de guerre ou falsifier des informations emploieront probablement des hypertrucages contre les producteurs d'OSINT. Cela pourrait amener ces derniers à élaborer des reportages ou des rapports inexacts, et ces inexactitudes pourraient ensuite être utilisées pour discréditer leur travail. Dans le meilleur des cas, les hypertrucages risquent de compliquer considérablement la tâche déjà prenante de vérification des OSINT.

Automatisation des procédés

Les hypertrucages pourraient également avoir un impact sur les procédés automatisés destinés à déjouer les activités adverses. L'empoisonnement des données consiste à contaminer intentionnellement

les données amassées pour entraîner les systèmes d'apprentissage automatique au moyen d'informations pernicieuses[135]. Les algorithmes servant à détecter les cyberattaques ou les campagnes de désinformation ou de malinformation pourraient aussi être mis en échec par l'injection de données empoisonnées dans les grands ensembles d'information qui servent à les entraîner. De plus, des chercheurs ont découvert que les systèmes conçus pour repérer les hypertrucages sont vulnérables à l'empoisonnement des données, ce qui les rend moins efficaces[136].

Réception des informations

Comme indiqué ci-dessus, une préoccupation majeure associée aux hypertrucages est le rôle que ces derniers pourraient jouer dans la détérioration d'un environnement informationnel déjà mal en point. De ce fait, alors qu'ils évoluent dans un espace où la vérité est de plus en plus contestée, les représentants du gouvernement, ainsi que les cadres de l'appareil du renseignement et de la sécurité nationale et leurs analystes, devraient s'attendre à ce que la population, voire certaines personnalités politiques, aient du mal à accepter leurs conclusions.

Il est sain que la population se pose des questions sur les évaluations de renseignement ou les interroge. Cependant, si ces doutes reposent sur des accusations ou sur des soupçons découlant de théories du complot, de mésinformation, de désinformation ou de malinformation plutôt que sur l'intérêt pour la bonne gouvernance, cela placera les organismes de l'appareil du renseignement et de la sécurité nationale dans une posture très délicate. L'acceptation sociale dont ces organismes ont besoin pour faire leur travail sera notamment compromise si des pans importants de la population refusent d'entrée de jeu les résultats de leurs enquêtes, ou en fait fi parce qu'ils considèrent trop difficile de distinguer la vérité. Ce souci pourrait être exacerbé chez les organismes et les services qui ont des problèmes de transparence depuis longtemps.

Pour compliquer davantage ces considérations, les mises en garde au sujet des hypertrucages pourraient au contraire aggraver le problème

dans certains environnements informationnels. Deux chercheurs, M. Chesney et Mme Citron, avancent que les efforts visant prévenir la population des effets pervers des hypertrucages pourraient avoir la conséquence inverse, ce qu'ils appellent « le dividende du menteur ». Ainsi, les personnes, les sociétés et les gouvernements accusés d'actes répréhensibles pourront prétendre que les éléments de preuve à charge (en particulier les images, les sons ou les vidéos) sont des hypertrucages, afin d'échapper à leurs responsabilités[137].

Questions éthiques associées aux hypertrucages

Compte tenu de ce qui précède, les chercheurs et les universitaires, particulièrement celles et ceux qui ont un point de vue scientifique, juridique ou technique, s'efforcent de trouver des solutions techniques et réglementaires. Peu d'articles ont étudié les dilemmes éthiques posés par les hypertrucages, en particulier pour les organismes et les services gouvernementaux. Le présent chapitre traitera brièvement de trois de ces dilemmes : l'utilisation des hypertrucages et le respect des normes démocratiques; les dilemmes propres au secteur privé; le risque d'exagération du problème.

Les démocraties doivent-elles utiliser des hypertrucages?

La première difficulté est que, si les techniques d'hypertrucages sont à la fois efficaces et économiques, les pays démocratiques seront tentés de les utiliser dans le cadre de leurs opérations de défense, de sécurité et de renseignement. D'un côté, ces États pourraient vouloir utiliser ces techniques parce qu'elles ne coûtent pas cher et sont plus faciles d'emploi que d'autres formes de collecte de renseignement ou d'activités clandestines, plus dangereuses.

Les services souhaitant utiliser les hypertrucages peuvent arguer que les ruses de guerre existent depuis des siècles. En outre, un objectif clé des opérations d'information actuelles est la diffusion de propagande visant à obtenir un avantage concurrentiel sur un opposant[138]. Cela comprend des tentatives de susciter un sentiment d'impuissance chez l'armée ou la population d'un État adverse, pour qu'elles abandonnent toute velléité de combat[139]. Par conséquent, il

ne serait pas surprenant que des États fabriquent des hypertrucages dans le cadre de ces campagnes, afin d'atteindre leurs buts rapidement, facilement et avec le moins de sang versé possible. De la même façon, de nombreux services de renseignement mènent des opérations perturbatrices visant à éviter que des activités malveillantes se produisent sur leur territoire de compétence ou contre leurs intérêts. Les hypertrucages, notamment des bandes sonores ou des vidéos frauduleuses, pourraient être utilisés pour égarer ou tromper des adversaires.

Il y a toutefois un inconvénient majeur à le faire. Il est attendu que les États autoritaires se livrent activement à de la propagande et il est très probable qu'ils se tournent vers les hypertrucages pour atteindre leurs objectifs politiques. Cependant, comme les États démocratiques sont fondés sur l'état de droit (même s'il est imparfaitement appliqué), la désinformation ne leur apportera pas forcément les mêmes avantages (il n'est pas non plus évident qu'ils soient très bons dans la conduite d'opérations d'information[140]). Si l'on apprend ou si l'on croit que les démocraties, leur armée ou leurs services de renseignement emploient activement des hypertrucages, le dividende du menteur sera certainement utilisé dans des cas qui importeront ultérieurement, en particulier si l'Occident essaie de convaincre des auditoires nouveaux ou sceptiques.

En outre, étant donné que la désinformation est largement reconnue comme un problème touchant les États démocratiques, il est discutable d'utiliser les hypertrucages pour en générer davantage. Après tout, les services de renseignement occidentaux semblent avoir eu plus de chance avec le désamorçage de la désinformation pendant l'invasion de l'Ukraine par la Russie en 2022 qu'avec la création d'un autre tissu de mensonges[141].

Dilemmes pour le secteur privé

La deuxième catégorie de problèmes éthiques concerne le rôle du secteur privé dans la création et la détection des hypertrucages. L'intelligence artificielle et les outils d'hypertrucage pourraient permettre à un grand nombre d'acteurs indépendants et d'intermédiaires

de mener des campagnes de désinformation, mais il est possible que les véritables bénéficiaires de ces campagnes seront un petit nombre d'importantes sociétés technologiques dont la valeur nette est considérable. Ces sociétés ont les moyens d'amasser de grands ensembles de données, de les exploiter et de les traiter pour alimenter des systèmes d'apprentissage automatique qui peuvent servir à la fois à créer et à détecter les hypertrucages. Il faudra apporter un soin et une réflexion particuliers pour déterminer les modalités de collaboration des États avec ces sociétés et d'utilisation de leurs produits. De nombreux ensembles de données destinés à l'apprentissage automatique comprennent des images obtenues par des moyens discutables[142]. En outre, la façon dont les biais intégrés à l'IA peuvent exacerber le racisme systémique a suscité des préoccupations et des chercheurs ont montré que les images « hypertruquées » pouvaient aviver les préjugés raciaux propres aux interfaces de programmation de reconnaissance faciale sur le Web[143]. Les algorithmes d'hypertrucages pourraient véhiculer des préjugés raciaux et autres qui en affecteront l'efficacité.

Les lois et la réglementation sur le respect de la vie privée fourniront des indications sur ce que les États démocratiques seront autorisés à faire. Les États devront toutefois faire preuve de finesse pour déterminer le type d'entreprises avec lesquelles collaborer, la façon d'évaluer les pratiques de ces dernières et les modalités de gestion des questions de responsabilité.

La menace est-elle exagérée?

Enfin, pour tous les enjeux abordés dans ce chapitre, il y a aussi un risque d'exagération. La désinformation est un problème grave, et même les hypertrucages rudimentaires peuvent y contribuer. Cependant, de nombreuses affirmations sur les perturbations que pourrait créer la propagande reposant sur l'IA sont pure spéculation et n'ont pas été prouvées[144]. L'IA engendrera des difficultés, mais le bruit qui l'entoure à l'heure actuelle n'est pas la réalité. Les hypertrucages sont vraiment impressionnants, mais cela ne rend pas forcément leur utilisation pratique. Par exemple, une vidéo « hypertruquée » d'un dirigeant mondial déclarant une guerre peut

être vite vérifiée et démentie par un simple examen des événements sur le terrain.

En outre, il n'est pas encore évident actuellement que la propagande faisant appel aux hypertrucages sera plus efficace pour véhiculer des opinions que la diffusion d'images et de mèmes rudimentaires, qui sont déjà relayés largement et rapidement. Des études ont montré que les fausses nouvelles se répandent non pas parce qu'elles sont logiques ou réalistes, mais parce qu'elles font vibrer une corde sensible chez la personne qui les republie[145]. En ce sens, les États devraient se préoccuper davantage de certains propos, plutôt que de l'apparence des contenus qui les servent. Ils doivent prendre les hypertrucages au sérieux, mais globalement, ces productions vont faire évoluer le contexte actuel de la menace, pas le bouleverser. Une réaction excessive aux hypertrucages pourrait donc affecter l'analyse de la menace et les politiques d'intervention.

Conclusion

Dans ce chapitre, il a été question de certaines des difficultés et des occasions que les services chargés de la sécurité nationale et les services de renseignement rencontreront ces prochaines années. Cependant, bien que la technologie soit impressionnante, les hypertrucages ont plus de chances de faire évoluer les activités liées à la menace existantes que d'en créer de nouvelles. S'il y a un avantage aux hypertrucages, c'est que la plupart des États démocratiques ne partent pas de zéro, mais disposent déjà de politiques et de procédures leur permettant de les gérer (même si celles-ci devront aussi évoluer). Par exemple, lors de la collecte de supports numériques, il importera de créer des chaînes de responsabilité, afin d'en favoriser la protection et la vérification futures, notamment au cours de procédures visant à faire appliquer la loi.

Bon nombre des problèmes liés aux hypertrucages les plus difficiles ne seront pas réglés par des outils technologiques ou juridiques, mais par des pratiques éthiques, qui nécessiteront du discernement. Cela passe par une réflexion sur les modalités de collaboration des États avec le secteur privé, en particulier avec les sociétés qui contrôlent

déjà d'importants services technologiques, et sur ce que cela implique en matière de surveillance et d'examen. De plus, s'il y a beaucoup de bonnes raisons pour les États démocratiques d'envisager d'utiliser les hypertrucages dans le cadre de leurs propres opérations de sécurité nationale et de renseignement, cette approche pourrait présenter davantage d'inconvénients que d'avantages.

Élaborer des interventions solides et axées sur les droits de la personne face aux hypertrucages, aux contenus synthétiques et à l'intelligence artificielle générative audiovisuelle

Il y a d'importantes leçons à tirer d'un examen des risques, des menaces et des solutions possibles en matière d'hypertrucages, de contenus synthétiques et d'intelligence artificielle (IA) générative audiovisuelle sous l'angle des préoccupations civiles et citoyennes (notamment celles des défenseurs des droits de la personne et des journalistes). Dans le monde entier, ces personnes et ces communautés affrontent déjà des maux semblables à ceux qu'engendrent les contenus artificiels. Cependant, bien qu'elles soient les plus à risque, elles sont tenues à l'écart des décisions concernant ces nouvelles technologies.

Lancée en 2018, l'initiative « Prepare, Don't Panic: Deepfakes and Synthetic Media » (Face aux hypertrucages et aux contenus synthétiques, ne paniquez pas, préparez-vous) de WITNESS vise à intervenir rapidement dans l'univers des contenus artificiels et met l'accent sur l'infrastructure technique, les nouveaux outils, la culture numérique et les aspects politiques et législatifs. Elle découle d'études approfondies, de consultations des membres du secteur concerné (au cours d'ateliers menés en Europe, en Afrique du Sud[146], au Brésil[147], en Asie du Sud-Est[148] et aux États-Unis[149]) et de nombreux ateliers et consultations en ligne dans toutes les régions du monde[150].

Menaces et risques du point de vue de la société civile

Lors des consultations menées par WITNESS durant les cinq dernières années, les acteurs de la société civile mentionnent inlassablement un certain nombre de maux tangibles et de risques associés aux contenus de synthèse. Ils soulignent invariablement que les femmes sont particulièrement sensibles aux menaces posées par les contenus artificiels parce que la technologie permet de nouvelles formes de violence sexiste.

Les contenus synthétiques, ou leur simple existence, permettent de nier toute implication et de discréditer des preuves pourtant solides en prétendant qu'elles sont fausses (on appelle cela « le dividende du menteur ») ou qu'aucun contenu n'est fiable (souvent pour décrédibiliser les journalistes, les militants et les organisations de la société civile dans leur ensemble, ainsi que le contenu digne de confiance qu'ils publient). Les participants à l'étude de WITNESS sont préoccupés à

l'idée que ces affirmations (ou ces craintes) servent à justifier l'adoption de lois restreignant plus largement la liberté d'expression.

Les personnes interrogées soulignent constamment le risque que les contenus synthétiques alimentent la désinformation et incitent à la violence, notamment par les canaux actuels de propagation rapide, comme les applications de messagerie. Elles mentionnent aussi que ces contenus peuvent être exploités par des acteurs étrangers et nationaux contre des groupes et des communautés dont les membres sont déjà fragiles en raison de leur origine ethnique, de leur religion, de leur identité politique, de leur métier ou d'autres caractéristiques.

Les participants font généralement le lien entre ces dangers et les difficultés actuelles liées aux lacunes de la culture médiatique, au manque de ressources des journalistes et à l'accès restreint qu'ont les rouages essentiels de la société civile et les citoyens aux outils de détection et d'authentification.

Les participants aux ateliers organisés par WITNESS affirment que ces menaces s'ajoutent à celles que leurs propres gouvernements font actuellement peser sur la société civile et les citoyens en vue de restreindre l'espace alloué à la société civile, par exemple la diffusion de désinformation visant cette dernière, la surveillance et le harcèlement des journalistes et des défenseurs des droits de la personne, et les tentatives de rendre leurs activités illégales.

Au cours de l'année écoulée, étant donné l'accessibilité, la facilité d'utilisation et la capacité de personnalisation grandissantes des outils d'IA générative, davantage de personnes ont pu utiliser ces derniers. Elles ont ainsi pu imaginer (ou expérimenter) la façon dont ils pourraient les affecter. Ce changement a entraîné une réévaluation des risques et des méfaits potentiels de cette technologie.

Plus les participants ont testé les outils de production de contenu et pris conscience de leur facilité d'emploi pour créer des contenus individualisés, générer des variantes de ce contenu et synthétiser des images d'événements de la vie réelle (avec peu de données d'entrée), plus ils et elles évoquent régulièrement les difficultés découlant de

l'invasion de l'écosystème informationnel par le contenu synthétique, étant donné le volume insuffisant de la vérification des faits, par exemple. Ils et elles placent ces constatations dans le contexte de situations critiques, comme les élections et les crises de santé publique[151].

Principes de renforcement de la résilience civile

Pour ce qui est de renforcer la résilience civile face à la manipulation et à la synthèse à l'aide de l'IA, un certain nombre de principes essentiels peuvent être dégagés de cette étude.

Il faut prioriser : 1) les personnes du monde entier qui rencontrent des obstacles semblables; 2) les journalistes et les membres de la société civile qui œuvrent en faveur d'un écosystème informationnel fiable

Pour intervenir face aux hypertrucages et aux contenus synthétiques, ainsi que face à la capacité de multiplication et à la facilité de création associées à l'IA générative audiovisuelle et à la technologie d'hypertrucage, il faut prêter attention à celles et ceux qui risquent le plus de faire l'objet d'attaques ciblées comme d'un travail plus vaste visant à saper la confiance dans le contenu essentiel. La plupart des communautés et des groupes concernés auront déjà été la cible de technologies plus anciennes. Par exemple, partout dans le monde, les femmes qui sont journalistes et qui font partie de la sphère publique sont agressées au moyen d'images de synthèse sexuelles non consensuelles, tandis que les défenseurs des droits de la personne et les journalistes d'investigation sont régulièrement accusés d'avoir falsifié leur documentation et leurs enquêtes.

De la même façon, les groupes en question sont déjà limités dans leur capacité de défense par des obstacles tout à fait concrets. Par exemple, les journalistes et le personnel électoral locaux ainsi que les personnalités politiques communautaires qui sont des femmes ou des membres déclarés de la communauté LGBTQI sont souvent pris pour cible, sous-financées et surchargés.

Éviter de sortir les contenus synthétiques de leur contexte, notamment historique.

Bien que les contenus synthétiques soient le fruit d'une nouvelle technologie, les menaces qu'ils représentent ne sont pas nouvelles. Comme indiqué ci-dessus, les mesures visant à lutter contre les menaces associées aux contenus artificiels et à tirer parti des possibilités qu'ils offrent devraient reposer sur l'expérience, surtout celle qu'ont acquise des populations vulnérables, des membres essentiels de la société civile et des intermédiaires au sein des médias. Les populations marginalisées savent comment la désinformation est utilisée contre elles, et les responsables des stratégies d'intervention communautaire et de la vérification des faits sont versés dans la lutte contre des formes plus traditionnelles de manipulation audio et vidéo (appelés « trucages simples »). Quant aux médias sociaux, ils sont déjà aux prises avec la gestion de la satire (pour laquelle les hypertrucages sont couramment employés) à l'échelle mondiale.

Responsabiliser résolument tous les maillons de la chaîne, y compris les acteurs qui conçoivent et qui mettent en service la technologie et ceux qui créent et distribuent du contenu synthétique (médias et médias sociaux).

Toute solution nécessitera la plus grande attention de tous les maillons de la chaîne de fabrication des contenus synthétiques, c'est-à-dire des acteurs à l'origine du modèle de base de l'IA générative et de la technologie connexe à ceux qui la mettent en service et qui distribuent des contenus synthétiques. Les interventions ne doivent pas imposer aux utilisateurs finaux de reconnaître les contenus synthétiques ou de révéler qu'ils emploient ces outils si les maillons de la chaîne de production de ces contenus ne sont pas tous imputables.

Par exemple, il n'est pas viable de tout miser sur une stratégie axée sur la culture médiatique qui vise à permettre à la population d'analyser une vidéo pour en déterminer la provenance. Ainsi, il est vain d'encourager les personnes qui voient une image sur leur fil d'actualité à essayer de repérer un défaut, comme une main tordue résultant du processus de synthèse, et de promouvoir de tels indicateurs. Ces

derniers découlent de défaillances de l'algorithme auxquelles les progrès techniques remédieront rapidement.

Les gouvernements, les médias sociaux, les entreprises de technologie et les organes de presse ont tous un rôle à jouer dans l'élaboration de mécanismes (réglementation, politiques, fonctions, processus, etc.) qui permettent de lutter activement contre ces menaces sans faire porter la responsabilité du contenu uniquement à celles et ceux qui le créent ou qui le consomment, mais qui imputent une partie de la responsabilité (le cas échéant) aux intervenants en amont. Il est crucial d'établir une infrastructure technique, des normes, des politiques applicables aux plateformes du monde entier et des lois et des règlements qui soient axés sur les droits de la personne.

Les gouvernements et les législateurs peuvent soutenir un large éventail de solutions techniques et politiques qui aident à mettre en place des garde-fous clairs axés sur les droits de la personne, qui imposent le respect des droits et qui portent une attention particulière aux droits fondamentaux que sont la protection de la vie privée et la liberté d'expression.

Pour des citoyens numériques bien informés

Les solutions proposées ici sont dérivées des résultats de l'étude réalisée par WITNESS, et les considérations qui précèdent doivent être prises en compte dans leur application.

Bien qu'insuffisante, la culture médiatique et numérique est plus nécessaire que jamais.

Il est essentiel que la société et les gouvernements favorisent plus largement la culture médiatique dans leurs interventions. C'est d'autant plus vrai que l'utilisation de contenus de synthèse et d'hypertrucages en est encore à ses balbutiements et que les trucages simples (qui consistent à sortir un contenu de son contexte, à le modifier légèrement ou à y apposer une légende trompeuse) sont actuellement bien plus répandus que les contenus synthétiques.

Les techniques et les approches propres aux contenus synthétiques ne doivent pas être élaborées isolément de celles qui relèvent de la culture médiatique ou qui visent les trucages simples. Par exemple, il est possible d'appliquer à la culture médiatique dans son ensemble, aux trucages simples et aux hypertrucages l'approche SIFT[152], dont les principes essentiels sont les suivants : **S**top (Réfléchissez), **I**nvestigate the source (Renseignez-vous sur la source), **F**ind alternative coverage (Trouvez d'autres informations sur le même sujet) et **T**race the original context (Retrouvez le contexte d'origine). Les campagnes visant à renforcer la culture médiatique doivent s'inscrire dans le cadre plus général de la mésinformation et de la désinformation, afin de promouvoir un regard critique sur le contenu consommé en ligne. Elles ne doivent pas mettre l'accent sur les défauts techniques actuels, en particulier ceux de l'IA générative (comme le conseil bien connu — et maintenant caduc — selon lequel les visages remplacés par hypertrucage ne cligneraient pas des yeux), mais sur des principes plus généraux et sur l'utilisation d'outils de détection et d'authentification mieux adaptés au contexte, à mesure qu'ils seront disponibles.

Une des principales stratégies à envisager dans le cadre des campagnes axées sur la culture médiatique consiste à éviter d'ajouter à l'effervescence suscitée par l'IA générative et les contenus synthétiques, et à atténuer la capacité de ces derniers à ébranler la confiance dans le contenu numérique. Les dénonciations du genre : « C'est un hypertrucage! » et l'affirmation plus générale : « On ne peut plus se fier à rien, car tout peut être falsifié » sont de plus en plus répandues, surtout lorsqu'il est question des communautés vulnérables et des voix essentielles de la société civile. Ces campagnes devraient donc être soigneusement adaptées à l'envergure du problème et éviter les discours alarmistes qui l'aggravent.

Parallèlement aux stratégies de culture médiatique destinées au grand public, il est important de se concentrer sur la culture des médias eux-mêmes et sur l'apport des journalistes, afin que ceux-ci ne donnent pas de conseils simplistes axés sur des indicateurs visuels, ni ne contribuent au cycle de la panique instrumentalisé dans certains contextes contre des voix essentielles de la société.

Les outils de détection devraient être mis à la disposition de celles et ceux qui en ont le plus besoin.

Les outils de détection font partie de la solution. En règle générale, ceux qui existent actuellement ne sont pas fiables à grande échelle et il faut faire appel à un expert pour en évaluer les résultats. Dans un certain nombre de cas observés un peu partout dans le monde, l'emploi par le grand public des outils de détection disponibles en ligne a alimenté la confusion et renforcé le doute entourant des vidéos authentiques, au lieu de les dissiper[153]. À l'heure actuelle, cependant, il y a un manque criant d'initiatives visant à doter de moyens supplémentaires et d'outils utiles les journalistes et les responsables de la vérification des faits qui s'efforcent de discréditer les tromperies réalistes ou de démentir les allégations prétendant que les contenus multimédias journalistiques sont faux.

Dans ce contexte, l'accessibilité aux outils de détection ne se limite pas aux aspects techniques : il s'agit d'aider à comprendre comment *utiliser efficacement* les outils de détection et comment *communiquer efficacement* les résultats aux divers intervenants. Si les plateformes peuvent jouer un rôle dans la modération du contenu de synthèse, il ne faudrait pas qu'elles se contentent d'un processus automatisé de repérage sans nuances, à cause du manque de fiabilité des modèles actuels de détection. Il faut admettre non seulement que la plupart des contenus synthétiques servent la communication personnelle et n'ont pas de conséquence néfaste, mais aussi que les contenus combineront de plus en plus des éléments de synthèse et des éléments authentiques, ce dont les tentatives de détection devront tenir compte. Cependant, les plateformes pourraient fournir des marques qui aideraient tant les membres de la société civile que les médias qui font des analyses. Cette mesure viendrait en complément de la culture médiatique au sens large évoquée plus haut.

Même si la provenance vérifiable et les filigranes peuvent servir de marques favorisant une participation numérique éclairée, l'infrastructure et les outils nécessaires devraient être élaborés en tenant compte des droits de la personne et de l'accessibilité.

Un large éventail d'initiatives visent à étudier comment fournir des marques d'authenticité et de provenance aux consommateurs de contenus (par exemple, celles de la Coalition for Content Provenance and Authenticity, C2PA.org, et de la Content Authenticity Initiative, contentauthenticity.org). Plusieurs solutions ont été proposées pour incorporer des filigranes aux produits de l'IA, mais elles reposent toutes sur la participation et l'intégration de tous les maillons de la chaîne de responsabilité (comme indiqué plus haut), ce qui comprend les personnes et les organismes qui conçoivent le modèle de base et les outils, ainsi que les médias sociaux et les principaux organes de presse. Tous partagent la responsabilité primordiale de la transparence quant à la façon dont les contenus que voient les citoyens sont créés. L'adhésion des médias sociaux, des grands médias et des développeurs de modèles d'IA et d'outils est essentielle, car ces intervenants ont un rôle capital à jouer pour garantir que la provenance vérifiable ou les filigranes font partie du contenu audiovisuel ou y sont liés dès le début de son cycle de vie, mais aussi que leur clientèle et leur auditoire sont bien au courant de la nature du contenu qu'ils consomment.

Dans ce cadre, une des responsabilités fondamentales d'un gouvernement démocratique consiste à veiller à ce que ces technologies soient mises en service avec le souci de l'accessibilité et du respect de la vie privée, et réglementées quand c'est nécessaire. Le travail de WITNESS a fait ressortir un certain nombre de préoccupations pour les droits de la personne en lien avec les méthodes permettant de certifier l'authenticité et la provenance, d'insérer des filigranes et d'accroître la transparence. Il est notamment primordial que la provenance du contenu ne soit pas intrinsèquement ou systématiquement liée à l'identité d'une personne et il est également essentiel de garantir la participation des intervenants du monde entier[154]. Comme les développeurs de contenus synthétiques, les entreprises, les organisations et les gouvernements à l'origine des initiatives qui font la promotion de l'utilisation de marques de provenance et de filigranes ont la responsabilité d'évaluer la capacité de nuire de ces technologies[155].

Il manque encore de politiques adaptées au contexte et d'expertise locale dans la modération des contenus de synthèse à grande échelle,

afin de lutter contre les menaces associées à ces contenus et de protéger la liberté d'expression.

Avec les contenus de synthèse, la quantité de contenus créés et diffusés en ligne augmentera. L'automatisation de la modération, notamment à l'aide de l'IA (outils de détection, ou marques de provenance et filigranes), sera un volet des interventions. Cependant, ces méthodes devraient être conçues avec les groupes visés et fondées sur les principes de respect des droits de la personne. En outre, il faudrait instaurer des processus clairs permettant d'intégrer les expériences et les experts locaux à la boucle de modération.

Il est capital de préserver la possibilité d'employer les contenus de synthèse à des fins satiriques et parodiques, comme c'est souvent le cas[156], tout en sachant bien qu'il s'agit d'une zone grise, dans laquelle il est facile d'enfumer l'auditoire en véhiculant des contenus néfastes ou malveillants sous le couvert de l'humour.

Un nouveau défi pour la démocratie : trouver des repères à l'ère de l'intelligence artificielle générative

L'humanité est à l'aube d'une nouvelle étape de l'évolution humaine, qui aura un effet profond sur la société et sur la démocratie. On pourrait l'appeler « l'ère de l'intelligence artificielle (IA) générative » : une époque où les relations humaines avec les machines modifieront la trame même de la société. Trouver des repères en cette période de changement considérable, avec toutes les possibilités qu'elle ouvre et tous les risques qu'elle engendre, sera l'un des plus grands défis du siècle pour la démocratie et pour la société.

Cela fait dix ans que les impacts d'une nouvelle forme d'IA, appelée « IA générative », sont étudiés. Le nom de cette nouvelle technologie donne un indice sur ce qui la rend si exceptionnelle : il s'agit d'un domaine nouveau de l'apprentissage automatique qui permet aux machines de générer, c'est-à-dire de créer de toutes pièces, des données, ou des choses.

Ces nouvelles données peuvent avoir pour support n'importe quel format numérique. L'IA générative peut synthétiser n'importe quoi : des sons, des images, des textes et des vidéos. En pratique, elle s'apparente à un moteur surpuissant pour toutes les informations et toutes les connaissances. Elle sera de plus en plus utilisée non seulement pour générer tous les contenus numériques, mais aussi comme outil d'automatisation dans la production de toute activité humaine intelligente et créative.

Imaginez un partenaire créatif capable d'écrire des histoires captivantes, de composer des musiques enchanteresses ou de concevoir des œuvres d'art visuel à couper le souffle. Maintenant, imaginez que ce partenaire est un modèle d'IA, c'est-à-dire un outil qui puise dans les vastes répertoires de connaissances humaines numérisées pour apprendre et qui affine constamment ses capacités à donner vie à nos rêves les plus ambitieux.

C'est ça, l'IA générative : une virtuose numérique qui appréhende les nuances de l'intelligence humaine et les utilise pour inventer quelque chose d'inédit et de saisissant, ou d'inédit et de terrifiant. Parce qu'elle exploite le potentiel des techniques d'apprentissage profond et des réseaux neuronaux, elle transcende la programmation

traditionnelle. Elle permet donc aux machines de réfléchir, d'apprendre et de s'adapter comme jamais auparavant.

La révolution de l'IA est déjà en train de devenir un rouage essentiel de l'environnement numérique, mis en service harmonieusement dans l'infrastructure physique et numérique de l'Internet, des médias sociaux et des téléphones intelligents. Néanmoins, si l'IA générative fait partie du champ des possibles depuis moins d'une décennie, elle n'a basculé dans le domaine grand public qu'en novembre dernier, avec la sortie de l'application ChatGPT, un important modèle linguistique (c'est-à-dire un système d'IA capable d'interpréter et de générer du texte).

ChatGPT est maintenant l'application la plus populaire de tous les temps. Elle a atteint les 100 millions d'utilisateurs en deux mois et aujourd'hui, elle en compte 100 millions par mois en moyenne. Presque tout le monde a une histoire à raconter sur ChatGPT, des étudiants qui l'emploient pour rédiger leurs travaux aux médecins qui s'en servent pour résumer les notes sur leurs patients. Si l'IA générative suscite un très fort engouement, elle est en même temps source de préoccupations cruciales pour l'intégrité de l'information et elle pose la question de notre capacité collective à suivre la cadence du changement.

Des hypertrucages à l'IA générative

L'environnement numérique développé au cours des 30 dernières années (qui repose sur Internet, les médias sociaux et les téléphones intelligents) est devenu essentiel aux affaires, à la communication, à la géopolitique et à la vie quotidienne. Si le rêve utopique de l'ère de l'information s'est concrétisé, son sombre revers est aussi de plus en plus évident.

Cet environnement permet à des acteurs malveillants de se livrer à des opérations criminelles et politiques bien plus efficacement qu'avant et en toute impunité. En 2023, les coûts mondiaux associés à la cybercriminalité devraient atteindre 10 000 milliards de dollars canadiens. Si l'on comparait son poids à celui des pays les plus

performants sur le plan économique, la cybercriminalité se classerait au troisième rang, après les États-Unis et la Chine.

Cependant, les acteurs malveillants ne sont pas les seuls à causer du tort à cet environnement. À eux seuls, le volume d'information à traiter et l'incapacité des êtres humains à l'interpréter peuvent aussi avoir des effets délétères. Ce phénomène est appelé « censure par le bruit » : il se passe tellement de choses qu'il est impossible de repérer ou de choisir les messages auxquels prêter attention.

Tout cela était bien présent à l'esprit des observateurs quand les premiers contenus produits par l'IA sont apparus, en 2017. Au fur et à mesure que la possibilité d'utiliser l'IA pour générer des données inédites a gagné en viabilité, ses adeptes les plus enthousiastes ont commencé à employer cette technologie pour créer des « hypertrucages ». Ce terme désigne aujourd'hui un contenu synthétisé au moyen de l'IA qui met en scène une personne disant ou faisant quelque chose qu'elle n'a jamais dit ou fait. Même si la scène a été créée de toutes pièces, elle paraît réelle à l'œil et à l'oreille.

La capacité pour l'IA de cloner l'identité des gens, mais surtout de générer des contenus artificiels sur toutes formes de contenu numérique (vidéo, audio, texte et image) est révolutionnaire, et pas seulement parce que l'IA est employée pour concevoir de faux contenus : les ramifications de cette innovation vont bien plus loin. Selon ce nouveau paradigme, l'IA sera utilisée dans la production de toutes les formes d'informations.

Intégrité de l'information et risque existentiel

Dans l'ouvrage *Deepfakes: The Coming Infocalypse*, paru en 2020, l'idée était avancée que l'arrivée de contenu généré par l'IA allait présenter des risques graves et existentiels, non seulement pour les personnes et les entreprises, mais pour la démocratie même. En effet, depuis la publication de ce livre il y a trois ans, les gens ont commencé à être exposés « dans leur milieu naturel » à d'importants volumes de contenu produit par l'IA.

Au début de l'année 2022, au commencement de l'invasion de l'Ukraine par la Russie, une vidéo « hypertruquée » du président ukrainien, M. Zelensky, pressant son armée de se rendre est apparue sur les médias sociaux. Si ce message avait été publié à un moment crucial de la résistance ukrainienne, il aurait pu avoir des effets dévastateurs. Bien qu'il ait été rapidement démenti, cet exemple d'utilisation offensive de contenus artificiels est annonciateur de ce qui nous attend.

En 2021, d'après la Commission fédérale du commerce des États-Unis, les fraudes par usurpation d'identité faisant appel à des hypertrucages, comme celle qui a permis à des voleurs dans le domaine des cryptomonnaies de se faire passer pour le président-directeur général de Tesla, Elon Musk, ont rapporté plus de 2,3 millions de dollars canadiens en six mois. Pendant ce temps, un nouveau type d'arnaque (appelé « fraude fantôme »), par laquelle des fraudeurs utilisent des identités « hypertruquées » pour accumuler des dettes et blanchir de l'argent, a déjà causé pour environ 4,5 milliards de dollars canadiens de pertes.

Le problème n'est pas juste que n'importe qui peut être victime d'une attaque par hypertrucage. Globalement, la prolifération de contenus générés par l'IA a un impact profond sur la confiance numérique. La santé de l'écosystème informationnel était préoccupante avant l'arrivée de l'IA, mais si tout ce que les gens consommaient en ligne (soit la majeure partie de la nourriture intellectuelle humaine) pouvait être produit par l'IA, quelles incidences cela aurait-il sur la démocratie, et sur la société dans son ensemble? Comment savoir en quoi avoir confiance? Comment faire la différence entre contenu authentique et contenu synthétique?

Protéger l'intégrité de l'écosystème informationnel est une priorité cruciale pour la démocratie, mais aussi pour l'ensemble de la société. Non seulement l'IA peut tout imiter, mais elle peut désormais générer n'importe quel contenu numérique, ce qui signifie que tout contenu authentique (par exemple, une vidéo révélant des violations des droits de la personne ou montrant une personnalité politique accepter un pot-de-vin) peut être dénoncé comme un faux, ou un produit de l'IA : ce phénomène a été baptisé « le dividende du menteur ».

Le risque principal pour la démocratie est un avenir dans lequel l'IA serait utilisée pour générer toutes les informations et toutes les connaissances, ce qui saperait la foi dans l'information numérique. Cependant, la démocratie (comme la société) ne peut pas fonctionner sans un médium d'information et de communication en qui tout le monde ait confiance. Alors que l'IA devient partie intégrante de cet écosystème, il est donc vital pour la société de défendre avec sérieux l'intégrité de l'information.

Solutions : authentification de l'information

Il existe des moyens techniques et sociaux d'y parvenir. L'une des approches les plus prometteuses est l'authentification des contenus. Au lieu d'essayer de détecter tout ce qui est produit par l'IA (ce qui sera vain si l'IA est utilisée pour générer toutes les informations à l'avenir), le dispositif d'authentification est enchâssé dans la trame même d'Internet. Cela pourrait être fait avec un marqueur cryptographique, afin que l'origine et le mode création d'un contenu (fabrication par l'IA ou non) puissent toujours être vérifiés. Ce type de cryptographie est incrusté dans l'ADN du contenu, alors c'est plus qu'un filigrane : il en fait partie intégrante, il ne peut être ni retiré, ni imité.

Cependant, le simple fait de « signer » ainsi le contenu ne suffit pas. La société doit aussi adopter une norme ouverte permettant à cet ADN ou à cette marque d'authentification d'être vue chaque fois que les gens échangent des contenus pour dialoguer sur Internet, que ce soit par courriel, sur YouTube ou sur les réseaux sociaux. Une telle norme ouverte est déjà en cours d'élaboration par la Coalition for Content Provenance and Authenticity (C2PA, voir C2PA.org), une organisation sans but lucratif qui s'intéresse à la provenance et à l'authentification du contenu, dont sont membres la BBC, Microsoft, Adobe et Intel.

En somme, cette approche mise sur la transparence fondamentale de l'information. Au lieu d'essayer de déterminer la vérité (une quête insensée), elle vise à permettre à tout le monde de prendre ses décisions sur la fiabilité des informations en fonction du contexte.

Tout comme nous avons besoin de la liste des ingrédients composant nos aliments, la société a besoin d'une infrastructure numérique qui permette à ses membres d'évaluer l'information en ligne dont découlent quasiment toutes les décisions qu'ils prennent et d'en jauger la fiabilité.

Résilience de la société

Néanmoins, il est impossible de relever ce défi à l'aide de la seule technologie. La collectivité peut se doter d'outils pour signer les contenus et d'une norme ouverte pour vérifier l'information partout sur Internet, mais le plus difficile est de comprendre que l'humanité est à l'aube d'un monde très différent, où des progrès technologiques exponentiels vont modifier la trame même de la société. Cela signifie qu'il faut revoir les anciennes façons de réfléchir. Nos systèmes analogiques sont maintenant obsolètes. Il convient de repenser ce qu'implique le fait d'être citoyen dans une démocratie dynamique et pour cela, il faut d'abord appréhender cette nécessité. Au fond, ce n'est pas une question de technologie, mais d'humanité.

Bien que les dernières avancées de l'IA aient engendré des débats sur l'avènement de l'intelligence artificielle générale (le stade auquel les machines prendront le pouvoir parce qu'elles seront plus intelligentes que les personnes), la société n'en est pas (encore) là. Les êtres humains ont toujours la possibilité de décider comment intégrer l'IA à la société et c'est leur responsabilité. À titre de démocratie, ce défi est l'un des plus importants de notre époque, alors ne gâchons pas notre chance de le relever.

Évolution de la désinformation : un avenir « hypertruqué »

Atelier non classifié organisé par la Direction de la liaison-recherche et de la collaboration avec les intervenants (LRCI) et la Direction de l'analyse et de l'exploitation des sources d'information (AXSI) du Service canadien du renseignement de sécurité (SCRS)

24 mai 2023, Ottawa

ORDRE DU JOUR

8 h 30 – 8 h 45	Accueil et prière d'ouverture
8 h 45 – 9 h 00	Mot d'ouverture
9 h 00 – 10 h 15	**Module 1 – Les hypertrucages en contexte**
10 h 15 – 10 h 30	Pause
10 h 30 – 12 h 00	**Module 2 – Capacités actuelles et possibilités futures**
12 h 00 – 13 h 00	Dîner
13 h 00 – 14 h 30	**Module 3 – Incidences pour le renseignement et pour la sécurité nationale**
14 h 30 – 14 h 45	Pause
14 h 45 – 16 h 15	**Module 4 – Incidences pour la démocratie**
16 h 15	Mot de la fin

Notes

1. *A brief history of fake news*, Center for Information Technology and Society, Université de Californie à Santa Barbara, 2022, consultable à l'adresse : https://www.cits.ucsb.edu/fake-news/brief-history.

2. Soll, J., *The Long and Brutal History of Fake News*, 2016, consultable à l'adresse : https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/.

3. *Critical Disinformation Studies: History, Power, and Politics*, 2021, disponible à l'adresse : https://doi.org/10.37016/mr-2020-76.

4. Désinformation en ligne, 2023, consultable à l'adresse : https://www.canada.ca/fr/campagne/desinformation-enligne.html.

5. United Nations (no date) Countering Disinformation | United Nations, consultable à l'adresse : https://www.un.org/en/countering-disinformation.

6. Public-Private Analysis Exchange Program, *Increasing Threat of Deepfake Identities*, 2021, consultable à l'adresse : https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

7. Johnson, D. et Johnson, A., « What are deepfakes? How fake AI-powered audio and video warps our perception of reality », *Business Insider*, 15 juin 2023, consultable à l'adresse : https://www.businessinsider.com/guides/tech/what-is-deepfake.

8. Sample, I., « What are deepfakes — and how can you spot them? », *The Guardian*, 13 janvier 2020, consultable à l'adresse : https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them.

9. Public-Private Analysis Exchange Program, Increasing *Threat of Deepfake Identities*, 2021, consultable à l'adresse : https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf, 2021.

10. Secrétariat du Conseil du Trésor du Canada, ***Directive sur la prise de décisions automatisée***, 2019, consultable à l'adresse : https://www.tbs-sct.canada.ca/pol/doc-fra.aspx?id=32745.

11. Copeland, B.J., *Artificial intelligence (AI) | Definition, Examples, Types, Applications, Companies, & Facts*, 2023, consultable à l'adresse : https://www.britannica.com/technology/artificial-intelligence.

12. Burns, E., Laskowski, N. et Tucci, L., « Artificial intelligence (AI) », *Enterprise AI*, 2023, consultable à l'adresse : https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.

13. *Introducing ChatGPT* (sans date), consultable à l'adresse : https://openai.com/blog/chatgpt.

14. Pocock, K., « What is ChatGPT and what is it used for? », *PC Guide*, 2023, consultable à l'adresse : https://www.pcguide.com/apps/what-is-chat-gpt/.

15. Allyn, B., « Surreal or too real? Breathtaking AI tool DALL-E takes its images to a bigger stage », *NPR*, 20 juillet 2022, consultable à l'adresse : https://www.npr.org/2022/07/20/1112331013/dall-e-ai-art-beta-test.

16. Database, A.P.S. (pas de date), *AlphaFold Protein Structure Database*, consultable à l'adresse : https://alphafold.ebi.ac.uk/.

17. *AlphaFold* (pas de date), consultable à l'adresse : https://www.deepmind.com/research/highlighted-research/alphafold.

18. Bond, S., « As tech evolves, deepfakes will become even harder to spot », *NPR*, 3 juillet 2022, consultable à l'adresse : https://www.npr.org/2022/07/03/1109607618/as-tech-evolves-deepfakes-will-become-even-harder-to-spot.

19. Tucker, P., « Deepfakes Are Getting Better, Easier to Make, and Cheaper », *Defense One*, 2021, consultable à l'adresse : https://www.defenseone.com/technology/2020/08/deepfakes-are-getting-better-easier-make-and-cheaper/167536/.

20. iProov, « Deepfake Statistics & Solutions | Protect Against Deepfakes », *iProov*, 26 août 2022, consultable à l'adresse : https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection.

21. CBC, « Disinformation, Dictators & Democracy: A discussion with Maria Ressa and Ron Deibert », 10 mai 2023, consultable à l'adresse : https://www.cbc.ca/radio/ideas/disinformation-democracy-ressa-deibert-1.6837181.

22. Prix Nobel de la paix 2021, 10 décembre 2021, consultable à l'adresse : https://www.nobelprize.org/prizes/peace/2021/ressa/lecture/.

23. Cole, S., « 'You Feel So Violated': Streamer QTCinderella Is Speaking Out Against Deepfake Porn Harassment », *Vice News*, 13 février 2023, consultable à l'adresse : https://www.vice.com/en/article/z34pq3/deepfake-qtcinderella-atrioc.

24. Farokhmanesh, M., « The Debate on Deepfake Porn Misses the Point », *WIRED,* 1er mars 2023, consultable à l'adresse : https://www.wired.com/story/deepfakes-twitch-streamers-qtcinderella-atrioc-pokimane/.

25. Gan, J., « Atrioc returns to Twitch six weeks after deepfake controversy, working on DMCA takedowns », *Dexerto*, 15 mars 2023, consultable à l'adresse : https://www.dexerto.com/twitch/atrioc-returns-to-twitch-six-weeks-after-deepfake-controversy-working-on-dmca-takedowns-2086445/.

26. Desk, I.T.W., « I was vomiting: Journalist Rana Ayyub reveals horrifying account of deepfake porn plot », *India Today*, 21 novembre 2018, consultable à l'adresse : https://www.indiatoday.in/trending-news/story/journalist-rana-ayyub-deepfake-porn-1393423-2018-11-21.

27. Hany Farid, Robert Chesney et Danielle Citron, « All's Clear for Deepfakes? Think Again », 11 mai 2020, consultable à l'adresse : https://www.ischool.berkeley.edu/news/2020/alls-clear-deepfakes-think-again.

28. Savin, J., « Deepfake porn is on the rise — and everyday women are the target », *Cosmopolitan*, 25 novembre 2022, consultable à l'adresse : https://www.cosmopolitan.com/uk/reports/a41534567/what-are-deepfakes/.

29. Dunn, S., *Women, Not Politicians, Are Targeted Most Often by Deepfake Videos,* 2021, consultable à l'adresse : https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/.

30. Reuters, « Fake Elon Musk giveaway featured in cryptocurrency scams-U.S. FTC », *Reuters*, 27 mai 2021, consultable à l'adresse : https://www.reuters.com/technology/fake-elon-musk-giveaway-featured-cryptocurrency-scams-us-ftc-2021-05-17/.

31. Serrano, J., « Please Don't Invest in This Crypto Scam Because Deepfake Elon Musk Told You To », *Gizmodo*, 27 mai 2022, consultable à l'adresse : https://gizmodo.com/elon-musk-deepfake-invest-bitcoin-scam-bitvex-1848982652.

32. Jenkinson, G., « 'Yikes!' Elon Musk warns users against latest deepfake crypto scam », *Cointelegraph*, 26 mai 2022, consultable à l'adresse : https://cointelegraph.com/news/yikes-elon-musk-warns-users-against-latest-deepfake-crypto-scam.

33. Kohli, A., « From Scams to Music, AI Voice Cloning Is on the Rise », *Time,* 29 avril 2023, consultable à l'adresse : https://time.com/6275794/ai-voice-cloning-scams-music/.

34. Damiani, J., « A Voice Deepfake Was Used To Scam A CEO Out Of $243,000 », *Forbes,* 3 septembre 2019, consultable à l'adresse : https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=120b6b6b2241.

35. Ropek, L., « Bank Robbers in the Middle East Reportedly 'Cloned' Someone's Voice to Assist with $35 Million Heist », Gizmodo, 30 octobre 2021, consultable à l'adresse : https://gizmodo.com/bank-robbers-in-the-middle-east-reportedly-cloned-someo-1847863805.

36. Mileva, G., « The Ultimate Virtual Events Statistics You Need To Know in 2023 », *Influencer Marketing Hub* [prépublication], consultable à l'adresse : https://influencermarketinghub.com/virtual-event-statistics/.

37. Thubron, R., « FBI warns of more criminals using deepfakes in remote interviews for tech jobs », *TechSpot*, 29 juin 2022, consultable à l'adresse : https://www.techspot.com/news/95119-fbi-warns-more-criminals-using-deepfakes-remote-job.html.

38. Muncaster, P., *FBI: Beware Deepfakes Used to Apply for Remote Jobs*, 2022, consultable à l'adresse : https://www.infosecurity-magazine.com/news/fbi-beware-deepfakes-remote-jobs/.

39. Crews, J., « AI's Dark Side: The Potential Misuses of Artificial Intelligence & Why You Should Be Concerned », www.linkedin.com [prépublication], 2023, consultable à l'adresse : https://www.linkedin.com/pulse/ais-dark-side-potential-misuses-artificial-why-you-should-crews.

40. « AI and Privacy: The privacy concerns surrounding AI, its potential impact on personal data », *The Economic Times*, 25 avril 2023, consultable à l'adresse : https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms?from=mdr.

41. Hern, A. et Milmo, D., « 'I didn't give permission': Do AI's backers care about data law breaches? », *The Guardian*, 10 avril 2023, consultable à l'adresse : https://www.theguardian.com/technology/2023/apr/10/i-didnt-give-permission-do-ais-backers-care-about-data-law-breaches.

42. Hassan. S., « How AI Can Be Used to Manipulate People », *Psychology Today*, 6 avril 2023, consultable à l'adresse : https://www.psychologytoday.com/ca/blog/freedom-of-mind/202304/how-ai-can-be-used-to-manipulate-people#:~:text=By%20analyzing%20patterns%20in%20people%27s,punishments%20based%20on%20predicted%20behavior.

43. Rathenau Instituut, « AI and manipulation on social and digital media », 2022, consultable à l'adresse : https://www.rathenau.nl/en/digitalisering/ai-and-manipulation-social-and-digital-media.

44. Marr, B., « The Problem With Biased AIs (and How To Make AI Better) », *Forbes,* 30 septembre 2022, consultable à l'adresse : https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=8035ac547700.

45. Larkin, Zoe, « AI Bias—What Is It and How to Avoid It? », 2022, consultable à l'adresse : https://levity.ai/blog/ai-bias-how-to-avoid.

46. Ruby, D., « Internet User Statistics In 2023 — (Global Data & Demographics) », *DemandSage* [prépublication], 2023, consultable à l'adresse : https://www.demandsage.com/internet-user-statistics/#:~:text=5.07%20billion%20people%20around%20the,the%20internet%20as%20of%202023.

47. Statista, *Canada: number of internet users 2013-2023*, 2023, consultable à l'adresse : https://www.statista.com/statistics/243808/number-of-internet-users-in-canada/#:~:text=Canada%3A%20number%20of%20internet%20users%202013%2D2023&text=As%20of%20January%202023%2C%20Canada,percent%20of%20the%20country%27s%20population.

48. Hancock, Jeffrey T. et Jeremy N. Bailenson, « The Social Impact of Deepfakes », *Cyberpsychology, Behavior, and Social Networking*, vol. 24, n° 3, 2021, doi: 10.1089/cyber.2021.29208.jth.

49. Cook, J., « Deepfake Technology: Assessing Security Risk », 2022, consultable à l'adresse : https://www.american.edu/sis/centers/security-technology/deepfake_technology_assessing_security_risk.cfm#:~:text=Often%2C%20they%20inflict%20psychological%20harm,technology%20to%20conduct%20online%20fraud.

50. Pringle, E., « One of A.I.'s 3 'godfathers' says he has regrets over his life's work. 'You could say I feel lost' », Fortune, 31 mai 2023, consultable à l'adresse : https://

fortune.com/2023/05/31/godfather-of-ai-yoshua-bengio-feels-lost-regulation-calls/.

51. Metz, C. et Schmidt, G., « Elon Musk and Others Call for Pause on A.I., Citing 'Risks to Society' », *The New York Times*, 29 mars 2023, consultable à l'adresse : https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

52. Vallance, B.C., « Elon Musk among experts urging a halt to AI training », *BBC News*, 30 mars 2023, consultable à l'adresse : https://www.bbc.com/news/technology-65110030.

53. Brown, S., « Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI », *MIT Sloan*, 2023, consultable à l'adresse : https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai#:~:text=AI%20concerns%3A%20Manipulating%20humans%2C%20or%20even%20replacing%20them&text=And%20it%20seems%20very%20hard,own%20subgoals%2C%E2%80%9D%20Hinton%20said.

54. Goodyear, S., « The 'godfather of AI' says he's worried about 'the end of people' », *CBC,* 4 mai 2023, consultable à l'adresse : https://www.cbc.ca/radio/asithappens/geoffrey-hinton-artificial-intelligence-advancement-concerns-1.6830857.

55. Taylor, J. et Hern, A., "'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation," *The Guardian*, 30 mai 2023, consultable à l'adresse : https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning.

56. Secrétariat du Conseil du Trésor du Canada, ***Directive sur la prise de décisions automatisée***, 2023, consultable à l'adresse : https://www.tbs-sct.canada.ca/pol/doc-fra.aspx?id=32592.

57. Patrimoine Canada, Initiative de citoyenneté numérique — *la désinformation en ligne et les autres préjudices et menaces en ligne,* 2023, consultable à l'adresse : https://www.canada.ca/fr/patrimoine-canadien/services/desinformation-en-ligne.html. (La traduction accessible en ligne le 28 juin 2023 ne correspond pas à l'anglais, d'où les crochets dans la traduction du paragraphe cité dans le présent article.)

58. Reynolds, C., « AI pioneer Yoshua Bengio says regulation in Canada is too slow, warns of 'existential' threats », *The Globe and Mail*, 24 mai 2023, consultable à l'adresse : https://www.theglobeandmail.com/business/technology/article-ai-pioneer-yoshua-bengio-says-regulation-in-canada-is-too-slow-warns/.

59. Datta, B., « Can Government Keep Up with Artificial Intelligence? », *NOVA | PBS,* 10 août 2017, consultable à l'adresse : https://www.pbs.org/wgbh/nova/article/ai-government-policy/.

60. Paas-Lang, C., « AI is having a moment. What should the government do about it? », *CBC*, 24 avril 2023, consultable à l'adresse : https://www.cbc.ca/news/politics/ai-regulation-mps-canada-1.6818095.

61. Le présent chapitre est tiré de l'article de Nieweglowska, M., Stellato, C. et Sloman, S. A., « Deepfakes : Vehicles for Radicalization, Not Persuasion. », *Current Directions in Psychological Science*, 2023, 09637214231161321.

62. Ajder, H., Patrini, G., Cavalli, F. et Cullen, L., *The State of Deepfakes: Landscape, Threats, and Impact*, septembre 2019.

63. Vaccari, C. et Chadwick, A., « Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news », *Social Media+ Society*, vol. 6, n° 1, 2020, 2056305120903408.

64. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

65. https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed.

66. Prior, M., « Visual Political Knowledge: A Different Road to Competence? » *The Journal of Politics*, vol. 76, n° 1, 2013, pp. 41 à 57.

67. Messaris, P. et Limus, A., « The Role of Images in Framing News Stories », dans l'ouvrage *Framing Public Life : Perspectives on Media and Our Understanding of the Social World*, édité par Reese, Stephen D., Gandy, Oscar H., Grant, August E., 2001, pp. 215 à 226 (Mahwah, Lawrence Erlbaum).

68. Rim, S., Amit, E., Fujita, K., Trope, Y., Halbeisen, G. et Algom, D., « How words transcend and pictures immerse: On the association between medium and level of construal », *Social Psychological and Personality Science*, vol. 6, n° 2, 2015, pp. 123 à 130.

69. Kirkpatrick, A., « The spread of fake science: Lexical concreteness, proximity, misinformation sharing, and the moderating role of subjective knowledge », *Public Understanding of Science* vol. 30, n° 1, 2020, pp. 55 à 74.

70. Barari, S., Lucas, C. et Munger, K., « Political deepfake videos misinform the public, but no more than other fake media », *OSF Preprints*, 2021.

71. Vaccari, C. et Chadwick, A., « Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news », *Social Media+ Society*, vol. 6, n° 1, 2020, 2056305120903408.

72. Vaccari, C. et Chadwick, A., « Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news », *Social Media+ Society*, vol. 6, n° 1, 2020, 2056305120903408.

73. Lago F., Pasquini C., Böhme R., Dumont H., Goffaux V. et Boato G., « More Real Than Real: A Study on Human Visual Perception of Synthetic Faces », rubrique Applications Corner, *IEEE Signal Processing Magazine*, vol. 39, n° 1, janvier 2022, pp. 109 à 116.

74. Mori, M., K. F. MacDorman et N. Kageki, « The Uncanny Valley », *IEEE Robotics and Automation Magazine*, vol. 19, n° 2, 2012, pp. 98 à 100.

75. Köbis, N. C., Doležalová, B. et Soraperra, I., « Fooled twice: People cannot detect deepfakes but think they can », *IScience*, vol. 24, nº 11, 2021, pp. 1 à 17.

76. Lee, J., et Shin, S. Y., « Something that they never said: multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news », *Media Psychology*, vol. 25, nº 4, 2022, pp. 531 à 546.

77. Pennycook, G., Epstein, Z., Mosleh, M. et coll., « Shifting attention to accuracy can reduce misinformation online », *Nature*, vol. 592, nº 7855, 2021, pp. 590 à 595.

78. Vosoughi, S., Roy, D. et Aral, S., « The spread of true and false news online », *Science*, vol. 359, no 6380, 2018, pp. 1146 à 1151.

79. Ajder, H., Patrini, G., Cavalli, F. et Cullen, L., *The State of Deepfakes: Landscape, Threats, and Impact*, septembre 2019.

80. Chesney, R. et Citron, D.K., « Deep fakes: a looming challenge for privacy, democracy, and national security », *California Law Review*, vol. 107, nº 6, 2019, pp. 1753 à 1820.

81. Groh, M., Epstein, Z., Obradovich, N., Cebrian, M. et Rahwan, I., « Human detection of machine-manipulated media », *Communications of the ACM*, vol. 64, nº 10 (octobre), 2021, pp. 40 à 47.

82. Yechiam, E. et Hochman, G., « Loss attention in a dual task setting », *Psychological Science,* vol. 25, nº 2, 2014, pp. 494 à 502.

83. Kirkpatrick, A., « The spread of fake science: Lexical concreteness, proximity, misinformation sharing, and the moderating role of subjective knowledge », *Public Understanding of Science* vol. 30, nº 1, 2020, pp. 55 à 74.

84. Bebbington, Jan, Russell, Shona et Thomson, I., « Accounting and sustainable development: Reflections and propositions », *Critical Perspectives on Accounting*, vol. 48, 2017, 10.1016/j.cpa.2017.06.002.

85. Baumeister, R. F. et Leary, M. R., « The need to belong: Desire for interpersonal attachments as a fundamental human motivation », *Psychological Bulletin*, vol. 117, nº 3, 1995, pp. 497 à 529.

86. Tajfel, H., « Social identity and intergroup behaviour », *Social science information*, vol. 13, nº 2, 1974, pp. 65 à 93.

87. Zhou, Y. et Shen, L., « Confirmation Bias and the Persistence of Misinformation on Climate Change », *Communication Research*, 2021.

88. Liv., N. et Greenbaum, D., « Deep Fakes and Memory Malleability: False Memories in the Service of Fake News », *AJOB Neuroscience,* vol. 11, nº 2, 2020, pp. 96 à 104.

89. Frenda, S. J., Knowles, E. D., Saletan, W. et Loftus, E. F., « False memories of fabricated political events », *Journal of Experimental Social Psychology*, vol. 49, nº 2, 2013, pp. 280 à 286, doi:10.1016/j.jesp.2012.10.013.

90. Dobber, T., Metoui, N., Trilling, D., Helberger, N. et de Vreese, C., « Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? », *The International Journal of Press/Politics*, vol. 26, n° 1, 2021, pp. 69 à 91.

91. Caramancion, K. M., « The demographic profile most at risk of being disinformed », *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS),* avril 2021, pp. 1 à 7 (IEEE).

92. Harper, C. A. et Baguley, T., « "You are fake news" : Ideological (a)symmetries in perceptions of media legitimacy », *OSF Preprints*, 2019.

93. Lawson, M. A. et Kakkar, H., « Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news », *Journal of Experimental Psychology: General*, vol. 151, n° 5, 2022, p. 1154.

94. Caldwell, M., Andrews, J. T. A., Tanay, T. et Griffin, L. D., « AI-enabled future crime », *Crime Science*, vol. 9, n° 1, 2020, p. 14.

95. Europol, *Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab,* Office des publications de l'Union européenne, Luxembourg, 2022, consultable à l'adresse : https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

96. Chesney, R. et Citron, D.K., « Deep fakes: a looming challenge for privacy, democracy, and national security », *California Law Review*, vol. 107, n° 6, 2019, pp. 1753 à 1820.

97. Liv., N. et Greenbaum, D., « Deep Fakes and Memory Malleability: False Memories in the Service of Fake News », *AJOB Neuroscience*, vol. 11, n° 2, 2020, pp. 96 à 104.

98. Cialdini, R. B., Pre-suasion: A revolutionary way to influence and persuade, 2018 (New York, Simon & Schuster Paperbacks).

99. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. et Dwivedi, Y. K., « Deepfakes: Deceptions, mitigations, and opportunities », *Journal of Business Research*, vol. 154, article n° 113368, 2023.

100. Harris, D., « Deepfakes: False pornography is here and the law cannot protect You », *Duke Law & Technology Review*, 5 janvier 2019.

101. Gieseke, A. P., « "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography », *Vanderbilt Law Review*, vol. 73, n° 5, 2020, pp. 1479 à 1515.

102. Vaccari, C. et Chadwick, A., « Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news », *Social Media+ Society*, vol. 6, n° 1, 2020, 2056305120903408.

103. Mitchell, Amy, *Many Americans SAY made-up news is a critical problem that needs to be fixed*, 17 août 2020, téléchargé le 23 février 2021 à l'adresse : https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

104. Ognyanova, K., Lazer, D., Robertson, R. E. et Wilson, C., « Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power », *Harvard Kennedy School (HKS) Misinformation Review*, 2020.

105. Chesney, R. et Citron, D.K., « Deep fakes: a looming challenge for privacy, democracy, and national security », *California Law Review*, vol. 107, nº 6, 2019, pp. 1753 à 1820.

106. Barari, S., Lucas, C. et Munger, K., « Political deepfake videos misinform the public, but no more than other fake media », *OSF Preprints,* 2021.

107. https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused#:~:text=People%20are%20arguing%20in%20court%20that%20real%20images%20are%20deepfakes%20%3A%20NPR&text=Press-,People%20are%20arguing%20in%20court%20that%20real%20images%20are%20deepfakes,claim%20that%20anything%20is%20fake.

108. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. et Dwivedi, Y. K., « Deepfakes: Deceptions, mitigations, and opportunities », *Journal of Business Research*, vol. 154, article nº 113368, 2023.

109. Pennycook, G., Cannon, T. D. et Rand, D. G., « Prior exposure increases perceived accuracy of fake news », *Journal of Experimental Psychology, General,* vol. 147, nº 12, 2018, pp. 1865 à 1880.

110. Ternovski, J., Kalla, J. et Aronow, P. M., « Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments », *OSF Preprints*, 2021.

111. Groh, M., Epstein, Z., Obradovich, N., Cebrian, M. et Rahwan, I., « Human detection of machine-manipulated media », *Communications of the ACM*, vol 64, nº 10 (octobre), 2021, pp. 40 à 47.

112. Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N. et Cook, J., « Misinformation and its correction: Continued influence and successful debiasing », *Psychological Science in the Public Interest*, vol. 13, nº 3, 2012, pp. 106 à 131.

113. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. et Dwivedi, Y. K., « Deepfakes: Deceptions, mitigations, and opportunities », *Journal of Business Research*, vol. 154, article nº 113368, 2023.

114. Kietzmann, J., Lee, L. W., McCarthy, I. P. et Kietzmann, T. C., « Deepfakes: Trick or treat? », *Business Horizons*, vol. 63, nº 2, 2020, pp. 135 à 146.

115. Chesney, R. et Citron, D.K., « Deep fakes: a looming challenge for privacy, democracy, and national security », *California Law Review*, vol. 107, nº 6, 2019, pp. 1753 à 1820.

116. Chesney, R. et Citron, D.K., « Deep fakes: a looming challenge for privacy, democracy, and national security », *California Law Review*, vol. 107, nº 6, 2019, pp. 1753 à 1820.

117. Vaccari, C. et Chadwick, A. , « Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news », *Social Media+ Society*, vol. 6, nº 1, 2020, 2056305120903408.

118. Mori, Masahiro, « The Uncanny Valley » (1970), traduit par Karl F. MacDorman et Norri Kageki, 6 juin 2012, consultable à l'adresse : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6213238&tag=1.

119. Cherry, K., « What Is the Uncanny Valley? », *Verywell Mind*, 2022, consultable à l'adresse : https://www.verywellmind.com/what-is-the-uncanny-valley-4846247.

120. Par exemple, FakeApp ou DeepFace Lab.

121. Par exemple, Lyrebird ou Descript.

122. Par exemple, https://voice.ai/.

123. Hypertrucages produits à partir des GAN : Il s'agit des hypertrucages les plus perfectionnés et les plus complexes. Des réseaux antagonistes génératifs (GAN) sont utilisés pour créer des images, des vidéos ou des enregistrements sonores réalistes. Cela nécessite une puissance de calcul et une expertise technique importantes, donc ce type d'hypertrucages est généralement produit par des chercheurs en intelligence artificielle ou par des équipes professionnelles. En d'autres termes, c'est le type d'hypertrucages qui peut vraiment porter atteinte aux intérêts du Canada en matière de sécurité nationale.

124. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

125. https://abcnews.go.com/Politics/sharing-deepfake-pornography-illegal-america/story?id=99084399#:~:text=In%202019%2C%20synthetic%20media%20expert,%2C%E2%80%9D%20Ajder%20told%20ABC%20News.

126. *BBC News*, « The fake video where Johnson and Corbyn endorse each other », 2019, consulté le 10 mars 2022 à l'adresse : https://www.bbc.com/news/av/technology-50381728.

127. Centre canadien pour la cybersécurité, *Repérer les cas de mésinformation, désinformation et malinformation (ITSAP.00.300)*, février 2022, consultable à l'adresse : https://www.cyber.gc.ca/fr/orientation/reperer-les-cas-de-mesinformation-desinformation-et-malinformation-itsap00300.

128. Voir, par exemple, *You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories and Our Polluted Media Landscape*, Cambridge (Massachusetts), MIT Press, 2021.

129. Tom Blackwell, « Russian fake-news campaign against Canadian troops in Latvia includes propaganda about litter, luxury apartments », National Post, 17 novembre 2017, consultable à l'adresse : https://nationalpost.com/news/canada/russian-fake-news-campaign-against-canadian-troops-in-latvia-includes-propaganda-about-litter-luxury-apartments.

130. Murray Brewster, « Canadian-led NATO battlegroup in Latvia targeted by pandemic disinformation campaign », 25 mai 2020, consultable à l'adresse : https://www.cbc.ca/news/politics/nato-latvia-battle-group-pandemic-covid-coronavirus-disinformation-russia-1.5581248.

131. Centre canadien pour la cybersécurité, *Glossaire*, consulté le 12 mai 2023 à l'adresse : https://www.cyber.gc.ca/fr/glossaire.

132. Hannah Smith et Katherine Mansted, *Weaponised deep fakes; National security and democracy,* Canberra, Australian Strategic Policy Institute, 2020.

133. Ash Carter et Laura Manley, *Tech Factsheets for Policymakers: Deepfakes,* Cambridge (Massachusetts), President and Fellows of Harvard College, 2020; Commissariat à la protection de la vie privée du Canada, « Des données au bout des doigts : La biométrie et les défis qu'elle pose à la protection de la vie privée », février 2011. Il faut aussi songer que les États autoritaires collectent et centralisent des données biométriques dans le cadre de leur programme national de surveillance. Voir *Riddle Russia*, « Biometrics as a Kremlin tool », 5 mai 2023, consultable à l'adresse : https://ridl.io/biometrics-as-a-kremlin-tool/.

134. *The Economist*, « Open-source intelligence is piercing the fog of war in Ukraine », 13 janvier 2023, consultable à l'adresse : https://www.economist.com/interactive/international/2023/01/13/open-source-intelligence-is-piercing-the-fog-of-war-in-ukraine.

135. Payal Dhar, « Protecting AI Models from "Data Poisoning" New ways to thwart backdoor control of deep learning systems », *IEEE Spectrum*, 24 mars 2023, consultable à l'adresse : https://spectrum.ieee.org/ai-cybersecurity-data-poisoning.

136. Xiaoyu Cao et Neil Zhenqiang Gong, « Understanding the Security of Deepfake Detection », dans Pavel Gladyshev, et coll. (éditeurs), *Digital Forensics and Cyber Crime, ICDF2C 2021* dans la série, *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 441, 2022 (Springer, Cham). https://doi.org/10.1007/978-3-031-06365-7_22.

137. Chesney et Citron, « Deep Fakes: A Looming Challenge ».

138. RAND Corporation, Information Operations, consulté le 17 mai 2023 à l'adresse : https://www.rand.org/topics/information-operations.html.

139. Clint Watts, *Messing with the Enemy: Surviving in a Social Media World of Hackers, Terrorists, Russians and Fake News*, New York, Harper Collins Publishers, 2018.

140. Brett Boudreau, *The Rise and Fall of Military Strategic Communications at National Defence 2015-2021: A Cautionary Tale for Canada and NATO, and a Roadmap for Reform*, Institut canadien des affaires mondiales, mai 2022, consultable à l'adresse : https://www.cgai.ca/the_rise_and_fall_of_military_strategic_communications_at_national_defence_2015_2021.

141. Stephanie Carvin, « Deterrence, Disruption and Declassification: Intelligence in the Ukraine Conflict », *CIGI Online*, 2 mai 2022, consultable à l'adresse : https://www.cigionline.org/articles/deterrence-disruption-and-declassification-

intelligence-in-the-ukraine-conflict/; David Klepper, « "Pre-bunking" shows promise in fight against misinformation », Associated Press, 24 août 2022, consultable à l'adresse : https://apnews.com/article/technology-misinformation-eastern-europe-902f436e3a6507e8b2a223e09a22e969.

142. Chloe Xiang, « AI Is Probably Using Your Images and It's Not Easy to Opt Out », *Vice*, 26 septembre 2022, consultable à l'adresse : https://www.vice.com/en/article/3ad58k/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out.

143. Shahroz Tariq, Sowon Jeon et Simon S. Woo, « Evaluating Trustworthiness and Racial Bias in Face Recognition APIs Using Deepfakes », *Computer,* vol. 56, n° 5, mai 2023, doi : 10.1109/MC.2023.3234978.

144. James R. Ostrowski, « Shallowfakes: The danger of exaggerating the AI disinfo threat », *The New Atlantis*, printemps 2023, consultable à l'adresse : https://www.thenewatlantis.com/publications/shallowfakes.

145. Cameron Martel, Gordon Pennycook and David G. Rand, « Reliance on emotion promotes belief in fake news », *Cognitive Research: Principles and Implications,* vol. 5, article n° 47, 2020, consultable à l'adresse : https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-020-00252-3.

146. *What We Learned from the Pretoria Deepfakes Workshop* (rapport complet), blogue de WITNESS, consultable à l'adresse : https://blog.witness.org/2020/02/report-pretoria-deepfakes-workshop/.

147. *WITNESS Media Lab | Deepfakes: Prepare Now (Perspectives from Brazil),* WITNESS Media Lab, consultable à l'adresse : https://lab.witness.org/brazil-deepfakes-prepare-now/.

148. *WITNESS Media Lab | Deepfakes: Prepare Now (Perspectives from South and Southeast Asia),* WITNESS Media Lab, consultable à l'adresse : https://lab.witness.org/asia-deepfakes-prepare-now/.

149. *Deepfakes and Disinformation,* MediaJustice, consultable à l'adresse : https://mediajustice.org/news/deepfakes-and-disinformation/.

150. Rapports consultables à l'adresse : wit.to/Synthetic-Media-Deepfakes.

151. *WITNESS, Fortifying the Truth in the Age of Synthetic Media and Generative AI*, mai 2023, https://blog.witness.org/2023/05/generative-ai-africa/.

152. Voir https://hapgood.us/2019/06/19/sift-the-four-moves/ et https://ctrl-f.ca/.

153. https://blog.witness.org/2021/07/deepfake-detection-skills-tools-access/.

154. Ticks or It Didn't Happen: Confronting Key Dilemmas in Authenticity Infrastructure for Multimedia, consultable à l'adresse : https://lab.witness.org/ticks-or-it-didnt-happen/ et *Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism*, consultable à l'adresse : https://journals.sagepub.com/doi/10.1177/14648849211060644.

155. https://blog.witness.org/2021/12/witness-and-the-c2pa-harms-and-misuse-assessment-process/.

156. *Just Joking: Deepfakes, Satire, and the Politics of Synthetic Media*, consultable à l'adresse : https://cocreationstudio.mit.edu/just-joking/.