Comprendre les menaces et les défis – Désinformation visuelle et multimodale (DVM)

RÉSUMÉ D'UNE ÉTUDE RÉALISÉE PAR LE CENTRE DE RECHERCHE INFORMATIQUE DE MONTRÉAL EN COLLABORATION AVEC LE LABORATOIRE SUR L'INTÉGRITÉ DE L'INFORMATION DE L'UNIVERSITÉ D'OTTAWA

uOttawa | Laboratoire sur l'intégrité de l'information
Information Integrity Lab

CRIM

Image de couverture : image générée par DALL-E en utilisant les mots-clés : Janus, un visage réel, un visage numérique.

INTRODUCTION

La désinformation représente une menace importante pour la société en raison de ses répercussions dévastatrices sur les institutions et les personnes. Ses effets se font sentir dans de nombreux secteurs, notamment l'exécution des processus démocratiques, l'éducation, la science, la sécurité nationale et la défense. En interférant avec le discours public, elle entrave les efforts déployés pour régler des problèmes urgents et existentiels tels que ceux liés au changement climatique et à la santé publique. La propension à la désinformation ainsi que l'augmentation et l'intensité de celle-ci sont alarmantes, les campagnes à grande échelle devenant de plus en plus fréquentes. Les campagnes peuvent se propager plus rapidement que la diffusion de faits vérifiables, et ainsi fausser et influencer l'opinion publique et le discours sociétal.

De nombreux accélérants et facteurs exacerbent le problème que constitue la désinformation : celle-ci se diffuse rapidement sur les réseaux sociaux et échappe aux efforts de réglementation, ce qui rend difficile l'élaboration de processus fiables de vérification. En outre, des technologies avancées qui rendent possible la falsification réaliste de textes, d'images, de vidéos et d'enregistrements audio viennent brouiller la distinction entre le contenu inventé et le contenu réel, et ces outils deviennent de plus en plus conviviaux et largement accessibles. Or, les méthodes de détection de la désinformation accusent du retard par rapport à ces outils de création et de falsification, particulièrement lorsqu'il s'agit de désinformation visuelle et multimodale (DVM), ce qui nuit considérablement à la lutte efficace contre la désinformation.

Face à ce problème qui prend de plus en plus d'importance, le Laboratoire sur l'intégrité de l'information (LabInfo) de l'Université d'Ottawa et le Centre de recherche informatique de Montréal (CRIM) ont formé un partenariat stratégique. Cette collaboration vise à réaliser des études ainsi qu'à acquérir et à diffuser des connaissances sur de nouveaux outils et de nouvelles techniques expressément adaptés pour s'attaquer au domaine de la désinformation visuelle et multimodale en évolution. Compte tenu des progrès rapides dans ce secteur, le partenariat est déterminé à effectuer de la recherche et du développement continus en suivant le rythme de l'évolution des technologies et des stratégies de désinformation.

Le présent rapport résume une étude approfondie réalisée par le CRIM en partenariat avec le Laboratoire sur l'intégrité de l'information de l'Université d'Ottawa, intitulée Désinformation visuelle et multimodale : analyse, défis et solutions, qui offre une vue d'ensemble de la DVM et explore les méthodes permettant de la contenir et de lutter contre celle-ci. Il porte sur les aspects sociaux, scientifiques et technologiques de la DVM et fournit un point de référence aux personnes qui cherchent à lutter contre ce phénomène, des chercheurs universitaires aux développeurs de solutions techniques, en passant par les décideurs politiques, les professionnels des médias, les éducateurs et les groupes de sensibilisation du public.

Nous commençons par définir la DVM et discuter de ses répercussions sur la société, en insistant sur la nécessité de solutions adaptées. Nous examinons diverses initiatives actuelles visant à étudier et à contrer la désinformation, en tenant compte de la nature multidisciplinaire et évolutive du domaine. Nous passons ensuite en revue les méthodes et outils actuels de production et de détection de la DVM. Cela comprend les approches universitaires, les outils commerciaux et les bibliothèques à code source ouvert, ainsi qu'un accent mis sur les nouvelles méthodes d'IA générative qui modifient les façons dont la désinformation est créée en premier lieu. Le présent rapport se termine par un examen des solutions technologiques potentielles. Cependant, il reconnaît que les outils actuels de lutte contre la DVM ne sont pas encore tout à fait en mesure de résoudre le problème, soulignant la nécessité d'un développement et d'une innovation continus dans ce domaine.

La désinformation est l'une des principales menaces de notre époque, et les intervenants de tous les domaines doivent posséder une compréhension approfondie des aspects technologiques qui sous-tendent la production et la détection de la DVM. Cette compréhension est essentielle non seulement pour mettre au point des méthodes de lutte efficaces, mais aussi pour favoriser de meilleures pratiques d'« hygiène numérique » et une compréhension critique parmi nos publics. Pour s'assurer que notre compréhension de la DVM et nos interventions relatives à celle-ci demeurent actuelles et efficaces, le Laboratoire sur l'intégrité de l'information et le CRIM s'engagent à fournir régulièrement des mises à jour sur les travaux en cours et les faits nouveaux dans ce domaine dynamique et critique.

Jennifer Irish
Directrice, Laboratoire sur l'intégrité de l'information,
Institut de développement professionnel de
l'Université d'Ottawa

Françoys Labonté
Directeur général
Centre de recherche informatique
de Montréal

Voici un rapport sommaire d'une étude exhaustive réalisée par le CRIM en partenariat avec le laboratoire, intitulée Désinformation visuelle et multimodale : analyse, défis et solutions et rédigée par Marc Lalonde, Houman Zolfaghari, Ph. D., Mohamed Dahmane, Ph. D., Hamed Ghodrati, Ph. D., Gilles Boulianne, Ph. D., Nicolas Rutherford et Richard Pinet.

Édité par Nicolas Rutherford et Marc Lalonde

Les défis en évolution de la désinformation numérique

La manipulation de documents visuels dans l'intention d'influencer l'opinion publique existe depuis longtemps, certains exemples notables datant de la guerre de Sécession[1]. Depuis le début des années 1990 et l'arrivée de Photoshop, la retouche d'images est devenue plus accessible que jamais. La propagation rapide de la désinformation est de plus en plus comparée à une avalanche écrasante de contenu inexact et trompeur, et est souvent qualifiée d'« infodémie » en raison de son incidence étendue et généralisée (Brennen et coll., 2021). Ce problème a été considérablement intensifié par la prolifération des médias sociaux, où les contenus, y compris la désinformation, se propagent rapidement sur diverses plateformes et, souvent, ne font pas l'objet de vérifications suffisantes. Les entreprises de médias sociaux, dont les modèles de revenus reposent en grande partie sur une mobilisation importante des utilisateurs et le partage de contenu, ont fait face à des critiques pour ne pas avoir efficacement freiné cette tendance. En outre, les outils avancés d'IA comme ChatGPT, même s'ils sont bénéfiques à de nombreux égards, font en sorte qu'il est beaucoup plus difficile de vérifier l'authenticité de l'information et rendent le problème encore plus complexe. Cela découle du fait que, si ces outils facilitent la création de messages malveillants, ChatGPT peut également générer un nombre infini de variantes du même message, et dans plusieurs langues. Ces technologies donnent ainsi accès à de nouveaux moyens de production et de diffusion de la désinformation, ce qui fait en sorte qu'il est de plus en plus difficile de distinguer les contenus authentiques des contenus faux.

La désinformation visuelle et multimodale diffère de la désinformation textuelle traditionnelle, car elle intègre au texte des images, des vidéos et de l'audio. Cette intégration de plusieurs types de médias entraîne des difficultés uniques en leur genre, car elle peut rendre l'information plus persuasive et plus difficile à vérifier que le texte seul.

Le degré de complexité de la production de la DVM peut aller des trucages simples *(shallowfakes)*, dans lesquels du texte est associé à des images ou des vidéos hors contexte, à des trucages un peu plus perfectionnés *(cheapfakes)*, qui reposent sur des outils de manipulation de base d'images et de vidéos. Ces manipulations comprennent le recadrage simple d'une image, l'insertion de parties d'autres images, la modification de la luminosité ou de la couleur, le flou, la modification de la vitesse d'élocution dans les vidéos ou la distorsion des données graphiques. À ceux-ci s'ajoutent les hypertrucages *(deepfakes)*, qui représentent un niveau de manipulation plus avancé. Les hypertrucages utilisent l'intelligence artificielle et des techniques d'apprentissage profond pour fabriquer ou modifier considérablement du contenu audiovisuel et faire en sorte qu'il semble authentique et réel.

[1] Un rapport du Centre for Data Ethics and Innovation (CDEI) au Royaume-Uni mentionne l'existence d'une image produite au plus fort de la guerre de Sécession, dans laquelle il y a eu une « substitution de visage » concernant le président Lincoln : https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation#about-this-cdei-snapshot-paper (en anglais seulement).

Incidence sociale de la DVM

Les répercussions de la désinformation visuelle et multimodale sont vastes et touchent des domaines comme la politique, l'économie et la cohésion sociale. Bien que les effets négatifs généraux de la désinformation soient largement reconnus et s'appliquent à diverses formes, l'incidence particulière des éléments visuels et multimodaux, comme la DVM, est moins bien documentée. Bien que les enjeux mentionnés liés au fait d'influencer le discours politique et la prise de décision, de travestir la stabilité économique et de nuire aux relations sociales en diffusant de fausses informations soient vrais pour la désinformation en général, la DVM amplifie ces effets en raison de son pouvoir de persuasion accru, ce qui en fait une forme de désinformation particulièrement puissante. Une compréhension approfondie de ces répercussions est essentielle à l'élaboration de stratégies efficaces pour résoudre les problèmes uniques en leur genre posés par la DVM.

Répercussions politiques de la DVM

Le DVM touche de façon significative les institutions politiques. Le département de la Sécurité intérieure des États-Unis et l'Australian Strategic Policy Institute ont signalé des risques associés aux hypertrucages, y compris la propagande en ligne et son influence sur les élections et les processus législatifs[2]. Ces rapports soulignent également l'érosion de la confiance du public envers les institutions en raison de la DVM.

La communauté du renseignement des États-Unis reconnaît les implications géopolitiques de la DVM, des adversaires potentiels utilisant des hypertrucages pour déstabiliser les États-Unis et leurs alliés (Appel et Prietzel, 2022). Les publicités politiques qui utilisent de nouvelles technologies d'imagerie et de vidéo peuvent, malgré les avertissements, avoir des effets psychologiques profonds.

Répercussions sur le secteur de la santé

Dans le secteur des soins de santé, la mésinformation et la désinformation peuvent entraîner la méfiance du public à l'égard des recommandations des experts en matière de santé et la croyance en des traitements non prouvés ou frauduleux. Le rôle des éléments visuels dans la mésinformation et la désinformation pendant la pandémie de COVID19, particulièrement dans le cadre des discussions conflictuelles sur la vaccination, a notamment été signalé par Brennen et coll. (2020). Ils ont observé qu'une grande partie de l'information douteuse qui circulait à l'époque était accompagnée de visuels alléchants et ont mis l'accent sur l'incidence de la DVM dans de tels débats sur la santé publique.

Répercussions sur les secteurs financier et commercial

Les secteurs financier et commercial sont également vulnérables à la DVM. Un rapport de 2022[3] de la Federal Trade Commission (FTC) met en évidence l'éventail des menaces en ligne pour les consommateurs, notamment les hypertrucages utilisés pour nuire à la réputation des entreprises et des organismes gouvernementaux, comme ce fut le cas pour Tesla[4] ou pour l'autorité fiscale de la Chine[5]. Bien qu'il n'y ait pas encore de preuve directe que la DVM touche les systèmes financiers mondiaux, elle peut avoir de graves répercussions sur les organisations pendant les crises de relations publiques et déstabiliser les économies dont les institutions financières sont faibles.

Répercussions sur les communautés marginalisées

La DVM peut nuire de façon disproportionnée aux communautés marginalisées, en ciblant les femmes, les personnes LGBTQ2+, les personnes de couleur et même, dans certains cas, les chercheurs qui étudient la haine et le racisme en ligne (Paris et Donovan, 2019). Elle peut exacerber les vulnérabilités et élargir les inégalités sociales, ce qui met en évidence la nécessité d'interventions ciblées dans ces communautés.

[2] https://www.aspi.org.au/report/weaponised-deep-fakes (en anglais seulement).

[3] https://www.ftc.gov/reports/combatting-online-harms-through-innovation (en anglais seulement).

[4] https://core.ac.uk/download/pdf/276953172.pdf (en anglais seulement).

[5] https://findbiometrics.com/fraudsters-use-deepfake-biometrics-hack-chinas-taxation-system-040103/ (en anglais seulement).

S'y retrouver dans les complexités de la DVM

La désinformation visuelle et multimodale peut avoir sur le public une incidence plus importante que la désinformation textuelle. Les recherches montrent que le contenu comportant des éléments visuels, en particulier sur les médias sociaux, suscite une mobilisation accrue, comme en témoigne l'augmentation des clics, des mentions J'aime et des partages pour les gazouillis contenant des images (Cao et coll., 2020). L'une des principales raisons de cela est que l'interprétation d'images nécessite habituellement moins d'effort cognitif par rapport au texte, l'interprétation d'éléments textuels s'accompagnant souvent de barrières linguistiques ou relatives à l'alphabétisation. En outre, l'expérience sensorielle associée à la vue d'une image a une incidence durable sur les perceptions de crédibilité et de réalité du message et d'engagement envers celui-ci. On estime que cette incidence augmente avec la richesse de la représentation du message, comme on le constate lors du passage d'un texte à une image, ou d'une image à une vidéo.

Des études ont montré que les messages multimodaux sont généralement perçus comme plus crédibles que les messages textuels (Dan et coll., 2021). Ils suscitent également des réactions émotionnelles plus fortes et sont considérés comme ayant une plus grande valeur probante. L'impact émotionnel accru des éléments visuels peut influencer considérablement le comportement, surtout dans un monde dominé par l'information visuelle où le potentiel de sélection et de manipulation des images est intrinsèquement lié à la manipulation des perceptions et des opinions.

La détection de la DVM, cependant, est difficile tant pour les algorithmes que pour les humains. Le grand public a souvent de la difficulté à déceler la désinformation, une partie importante de celui-ci surestimant sa capacité à distinguer l'information légitime de l'information fausse (Lyons et coll., 2021; Languein, 2022). La complexité de la détection découle non seulement du contenu lui-même, mais aussi du contexte de sa production, y compris de l'intention du créateur. La croyance intuitive et commune selon laquelle « les images ne mentent pas » fait en sorte que les citoyens sont mal outillés pour lutter eux-mêmes contre la DVM, d'autant plus que les avancées en matière d'IA rendent les productions de DVM de plus en plus perfectionnées et réalistes.

Parallèlement, la DVM se diffuse plus rapidement, plus loin et plus profondément sur les réseaux sociaux que les informations véridiques (Morrow et coll., 2020; Svahn et Perfumi, 2022). Des études sur les hypertrucages ont montré que ceux-ci sont perçus comme plus convaincants et crédibles que les faux articles de nouvelles, ce qui donne lieu à une plus grande probabilité de partage sur les médias sociaux. La diffusion rapide et presque sans effort de ces nouvelles a une incidence négative sur les opinions politiques et la prise de décisions (Dan et coll., 2021).

Au cours des dernières années, des progrès importants ont été réalisés dans les modèles d'IA en ce qui concerne le traitement du langage, suivis de développements rapides pour rendre ces techniques plus pratiques et accessibles. Les exemples comprennent Transformers (Google, 2017), Bert (Google, 2018), GPT2 (OpenAI, 2019), GPT3 (OpenAI, 2020) et ChatGPT (OpenAI, 2022). En 2023, bon nombre de ces modèles étaient accessibles au public. La combinaison de ces grands modèles de langage avec d'autres modalités de production et de manipulation d'images et de vidéos, comme DALL-E 2 (OpenAI, 2022), GPT4 (OpenAI, 2023) et Gen2 (RunwayAI, 2023), entraîne une transformation rapide de la technologie et de la société qui s'accompagne de conséquences importantes concernant la prolifération et le raffinement de la désinformation.

Méthodes de production de la DVM

Les outils et les techniques peuvent être divisés en deux catégories : ceux qui transforment des documents existants et ceux qui génèrent des documents.

1 - Outils de transformation de documents

Divers outils sont disponibles pour transformer des documents existants :

- Logiciel de substitution de visage : Des programmes comme FaceSwap et DeepFaceLab, tous deux hébergés sur GitHub et classés parmi les 250 principaux dépôts sur 28 millions, permettent aux utilisateurs de remplacer le visage d'une personne par celui d'une autre dans une vidéo.
- Logiciel de conversion de voix et de conversion de texte en parole : Ces outils peuvent modifier l'enregistrement vocal d'une personne ou générer du nouveau contenu parlé avec une voix particulière à partir de texte.
- Logiciel de synchronisation labiale : Cette technologie modifie des vidéos existantes pour faire en sorte qu'une personne semble dire quelque chose de différent, en synchronisant les mouvements de la bouche avec les phonèmes supposément prononcés.
- Logiciel de manipulation de l'apparence : Ces programmes peuvent modifier l'apparence du visage d'une personne, par exemple pour le vieillir ou le rajeunir.
- Logiciel de synthèse de performance virtuelle : Les mouvements corporels entiers d'une personne peuvent être modifiés ou créés en transposant les mouvements d'une autre personne.

2 - Production et manipulation de documents

Le domaine de la génération de contenu a connu une croissance importante ces dernières années. Des outils qui génèrent des images de personnes inexistantes sont utilisés pour créer de faux profils de médias sociaux, évitant ainsi la nécessité de réutiliser les photos de personnes réelles, qui pourraient être retracées au moyen de recherches d'images inversées. Les grands modèles de langage (GML) comme GPT sont également utilisés pour créer de faux profils de médias sociaux grâce à la combinaison de photos de personnes inexistantes avec des biographies, des intérêts et des passe-temps inventés. Ces GML peuvent également produire du texte d'accompagnement pour des images ou des vidéos dans des publications sur les médias sociaux. Enfin, il convient de noter que les GML peuvent aider à générer des mèmes Internet nuisibles, un autre type populaire de désinformation visuelle et multimodale (Pramanick et coll., 2021).

Les progrès récents en matière d'apprentissage profond[6] ont amélioré la technologie de synthèse d'images. Désormais, de nouvelles techniques permettent la création de contenus entièrement synthétiques, sans rapport avec la réalité, basés sur des descriptions de texte ou des invites textuelles. Voici quelques modèles de texte à image bien connus :

- Réseaux antagonistes génératifs (RAG) : Bien que complexes et moins populaires récemment, les RAG ont été parmi les premiers modèles génératifs développés, capables de synthétiser des images faciales réalistes.
- Modèles génératifs fondés sur les GML : Ces modèles génèrent, au moyen de millions de paires image-texte, des images hyperréalistes à partir de phrases simples. Des entreprises comme OpenAI (avec ChatGPT et DALL-E) et Stability AI (avec Stable Diffusion) sont des chefs de file dans ce domaine. Le modèle Parti de Google, qui produit des images très réalistes, est un autre acteur important dans ce secteur.
- Modèles de diffusion : GLIDE et DALL-E 2 d'OpenAI ainsi qu'Imagen de Google, qui sont basés sur des techniques de diffusion, font progresser la génération d'images réalistes et améliorent le rendu du texte dans les images.

[6] L'apprentissage profond est un sous-ensemble de l'apprentissage automatique où des réseaux neuronaux artificiels, des algorithmes inspirés par le cerveau humain, apprennent à partir de grandes quantités de données. Cette approche permet au système d'apprendre automatiquement des structures complexes et de prendre des décisions fondées sur son apprentissage.

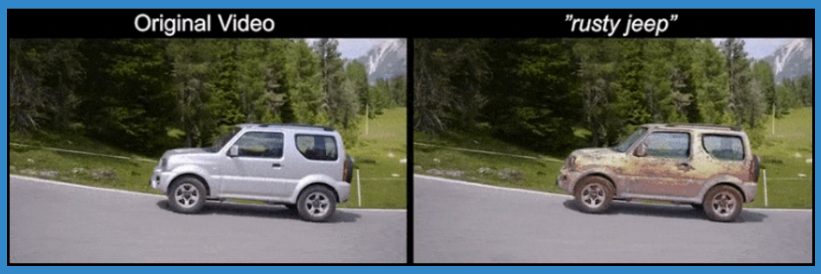
L'IMAGE CI-DESSUS EST UNE REPRÉSENTATION HYPERRÉALISTE ENTIÈREMENT CRÉÉE PAR LE MODÈLE MIDJOURNEY. LE CARACTÈRE ARTIFICIEL DE L'IMAGE EST MIS EN ÉVIDENCE PAR LE TEXTE ININTELLIGIBLE FIGURANT SUR LA COUVERTURE DU LIVRE.

Source: https://twitter.com/EliotHiggins/status/1638187198162821127?s=20

Ces modèles, en plus de générer des images, permettent également aux utilisateurs de modifier facilement celles-ci au moyen d'invites textuelles, rendant possible l'apport de changements comme la transformation de l'arrière-plan, l'ajout d'objets et la modification du style sans manipulation directe des pixels. Les entreprises sont de plus en plus conscientes des utilisations malveillantes potentielles de ces technologies, comme en témoigne le refus de DALL-E 3.0 de traiter les demandes concernant des personnalités publiques[7].

Les progrès rapides des modèles de langage pour la génération d'images ont également stimulé la recherche dans la génération de vidéos. Ici, la complexité du maintien de la cohérence temporelle du contenu et du rendu visuel a représenté un défi, mais les outils les plus récents peuvent maintenant créer des vidéos à haute résolution d'une durée considérable.

Le montage vidéo basé sur des invites textuelles évolue également et permet aux utilisateurs de décrire les changements à apporter à un objet, qui sont mis en œuvre par le modèle génératif. Ce domaine progresse rapidement et s'accompagne d'avancées importantes en matière de qualité d'image et de durée de la trame. Déjà, des organisations politiques ont utilisé des images générées par l'IA aux fins de publicités partisanes. La facilité d'utilisation et la disponibilité en ligne de ces outils, qui ne nécessitent aucune expérience en matière de programmation, accélèrent la production de contenus douteux. Compte tenu du fait que ces outils progressent sur les plans de la qualité et de la convivialité, la possibilité de production de contenu discutable augmente.



Source: https://www.unite.ai/consistent-ai-video-content-editing-with-text-guided-input/

DANS L'EXEMPLE CI-DESSUS, L'UTILISATION DE L'INVITE TEXTUELLE « JEEP ROUILLÉE » AMÈNE UN MODÈLE À MODIFIER AUTOMATIQUEMENT L'APPARENCE DU VÉHICULE DANS LA VIDÉO D'ORIGINE, SANS AUCUNE AUTRE INTERVENTION HUMAINE.

[7] https://www.theverge.com/2023/9/20/23881241/openai-dalle-third-version-generative-ai (en anglais seulement).

État actuel de la lutte contre la DVM

Lutter contre la désinformation visuelle et multimodale, et la désinformation en général, est une tâche indéniablement complexe et exigeante. Cette complexité se reflète dans les listes exhaustives de recommandations contenues dans des rapports comme celui du Forum des politiques publiques[8] et les 35 recommandations de Wardle[9]. Les recherches s'accordent sur le fait qu'aucune solution ou aucun intervenant ne peut en soi régler entièrement le problème et que des stratégies efficaces nécessitent une combinaison de nouvelles technologies, de pratiques organisationnelles et de changements sociétaux, selon Bateman (2020). Cela requiert une approche multidimensionnelle qui intègre divers types de mesures élaborées et déployées conjointement par de multiples acteurs.

Par exemple, Helmus (2022) insiste sur la nécessité de se concentrer sur cinq domaines clés : l'élaboration d'outils de détection, la mise en œuvre de normes de certification pour l'authenticité des documents audiovisuels, la prise en compte d'approches réglementaires, la promotion d'approches fondées sur le renseignement comme le renseignement de sources ouvertes (OSINT) en journalisme, et l'amélioration des compétences médiatiques.

L'IA joue un rôle essentiel dans le filtrage de la grande quantité de messages publiés quotidiennement sur les réseaux sociaux. Son utilisation comme outil pour repérer la désinformation en ligne est en hausse, malgré les difficultés liées aux biais, au rendement et aux faux positifs, selon Svahn et Perfumi (2022).

Parmi ces stratégies variées, les médias sociaux occupent une place cruciale. Leurs systèmes internes jouent un rôle de premier plan pour ce qui est de détecter et de gérer le contenu constituant de la DVM. Les approches relatives aux médias sociaux en ce qui concerne la DVM comprennent ce qui suit :

- **Modération du contenu** : Ce processus consiste à examiner et à surveiller le contenu des plateformes en ligne afin d'assurer le respect des règles de la plateforme. Il réunit souvent des outils algorithmiques, des modérateurs humains et des rapports d'utilisateurs, leur combinaison précise dépendant de la façon dont la plateforme est gérée (Morrow et coll., 2020). Le contenu inapproprié peut être supprimé, être rendu moins visible ou perdre sa capacité à générer des revenus. Toutefois, cette approche présente des difficultés. Une modération efficace nécessite que les examinateurs humains soient adéquatement rémunérés, protégés par la loi, bien formés et soutenus, ce qui comprend du soutien en santé mentale. En outre, même s'ils s'efforcent d'être impartiaux, les modérateurs humains peuvent avoir des préjugés, et il existe un risque que l'on mise de manière trop importante sur les systèmes automatisés pour la prise de décisions, une préoccupation soulignée par la FTC (2022).
- **Étiquetage** : Cette méthode consiste à ajouter des annotations visuelles ou textuelles au contenu généré par les utilisateurs pour fournir un contexte supplémentaire. La notion d'amélioration des publications sur les médias sociaux au moyen d'informations dont la véracité a été confirmée est attrayante, mais soulève des préoccupations éthiques comme la censure et l'incidence sur la liberté d'expression. Les réactions des utilisateurs aux étiquettes sont mitigées. Certains utilisateurs les considèrent comme essentielles pour que les plateformes fournissent une vérification des faits, tandis que d'autres les perçoivent comme excessives et susceptibles de porter atteinte à la liberté d'expression. L'efficacité des étiquettes dépend fortement de leur conception, et notamment de facteurs comme leur taille et leur libellé. Des lignes directrices détaillées sont souvent utilisées pour aider à créer des étiquettes claires et percutantes tout en veillant à ce qu'elles transmettent le message voulu sans être intrusives ou trop directives[10].
- **Contextualisation** : Il s'agit d'une forme d'étiquetage où des renseignements supplémentaires, qui ne figurent pas dans la publication d'origine, sont ajoutés pour fournir plus de contexte. Il peut s'agir de détails sur la source d'une image ou de renseignements sur l'auteur de la publication. La contextualisation vise à permettre aux utilisateurs de mieux comprendre le contenu, ce qui les aidera à mieux évaluer sa crédibilité. Par exemple, le fait de connaître l'origine d'une image peut donner des précisions sur son authenticité, tandis que l'information sur l'auteur peut aider à en savoir plus sur les préjugés qu'il peut avoir ou sur sa crédibilité. Cette approche vise à soutenir une base d'utilisateurs mieux informés, capables d'interagir de façon critique avec le contenu auquel ils sont exposés en ligne.

[8] https://ppforum.ca/wp-content/uploads/2022/01/DEMX-R2.pdf (en anglais seulement).

[9] https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c (en anglais seulement).

[10] https://firstdraftnews.org/articles/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow/ (en anglais seulement).

D'autres approches de lutte contre la DVM et d'atténuation de celle-ci peuvent être classées comme suit :

Approches sociétales

L'amélioration de la littératie numérique et des compétences médiatiques dans l'ensemble de la population, en mettant l'accent sur les jeunes et les journalistes, est considérée comme essentielle pour la consommation contemporaine de l'information.

Les journalistes, en particulier, font face à un flux constant d'informations qui doivent être rapidement validées, mais ils n'ont souvent pas les compétences nécessaires en informatique, en édition d'images et en montage vidéo pour effectuer de telles vérifications. La vérification de la DVM comme les hypertrucages, pour lesquels les détails qui permettent de déterminer qu'une image est un hypertrucage sont souvent difficilement perceptibles, constitue une difficulté de taille.

Parallèlement, des programmes comme le News Literacy Project visent à améliorer la littératie numérique du primaire au niveau universitaire, en particulier dans les écoles de journalisme et de communication. Toutefois, comme le soulignent Vaccari et Chadwick (2020), les modèles traditionnels de vérification des faits peuvent être insuffisants lorsqu'il s'agit de formes plus perfectionnées de désinformation, comme les hypertrucages.

Outre ces programmes éducatifs, on reconnaît de plus en plus le besoin de ressources ciblant d'autres groupes démographiques, comme les personnes âgées, qui sont particulièrement vulnérables à la désinformation (Brashier et Shacter, 2020). Les plateformes en ligne et les organisations mettent au point de plus en plus de ressources et d'outils pour aider ces segments de la population à naviguer dans le paysage numérique de façon plus sécuritaire et en faisant preuve d'un meilleur discernement[11].

Solutions politiques et juridiques

La lutte contre la DVM au moyen de mesures juridiques est souvent considérée comme une approche moins prometteuse, principalement en raison des solides protections juridiques entourant la liberté d'expression et de la nature intrinsèquement lente des procédures judiciaires, comme l'a souligné Bateman (2020). On craint également que des cadres juridiques rapidement mis en œuvre encouragent par inadvertance la censure sur les plateformes de médias sociaux, car celles-ci peuvent réglementer le contenu de façon excessive pour éviter des répercussions juridiques (Langguth et coll., 2021).

En outre, bien que les lois existantes puissent offrir des cadres pour s'attaquer au problème, leur portée et leur efficacité sont souvent limitées en raison de questions de compétence. Ces questions existent à la fois à l'échelle internationale, où les normes juridiques et les capacités d'exécution diffèrent d'un pays à l'autre, et à l'échelle nationale, où les lois peuvent varier d'un État ou d'une région à l'autre. Cette mosaïque de règlements crée un paysage juridique complexe pour ce qui est de lutter efficacement contre la propagation et les répercussions de la DVM.

Outre les questions de compétence, la nature nuancée de contenus comme les hypertrucages – qui peuvent chevaucher la ligne qui sépare l'expression légitime et la désinformation malveillante – fait en sorte qu'il est difficile de mettre en œuvre des lois qui visent tous les aspects de manière exhaustive sans empiéter sur le principe de la liberté d'expression. Par conséquent, les mesures juridiques ont un rôle à jouer, mais leur efficacité est limitée et elles doivent être assorties d'autres mesures.

Solutions technologiques

Des solutions algorithmiques pour la détection de la DVM sont de plus en plus proposées, l'IA jouant un rôle important à cet égard. Des approches d'apprentissage automatique, par exemple, analysent les incohérences du signal d'image pour détecter les trucages simples, qui sont créés en insérant dans l'image originale des pixels provenant d'autres images. Cependant, les outils perfectionnés utilisés pour créer des hypertrucages et générer des images et des vidéos à partir d'invites textuelles posent des difficultés pour ces stratégies.

[11] https://www.buffalo.edu/cii/projects/DART.html (en anglais seulement).

Survol des stratégies de détection de la DVM

Les approches d'intelligence artificielle conçues pour détecter la DVM deviennent de plus en plus cruciales en raison de leur capacité à traiter le volume croissant de contenus échangés sur le Web. Ces approches peuvent être classées en quatre grandes catégories : analyse de l'intégrité sémantique, recherche d'images inversée, analyse des artefacts, détection d'images et de vidéos générées.

Analyse de l'intégrité sémantique

Ce type d'analyse consiste à valider la cohérence sémantique d'une image. Par exemple, la confirmation de l'identité des personnes dans une image peut aider à détecter les substitutions de visages[12] ou les inexactitudes historiques. Une tâche plus complexe est l'analyse de l'intégrité sémantique textuelle-visuelle, un nouveau domaine tirant parti des progrès réalisés dans les réseaux neuronaux multimodaux capables de traiter à la fois les images et le texte. L'objectif est de s'assurer que le message textuel s'aligne sémantiquement avec le contenu visuel de l'image, en détectant les scénarios où une image réelle et non altérée est prise hors contexte et jumelée à un texte trompeur. Des technologies comme PROVES[13] sont en cours de développement pour certifier et vérifier l'information sémantique d'une image tout en permettant de simples opérations d'édition comme le recadrage et les ajustements de couleur. Le directeur du programme SemaFor de la Defense Advanced Research Projects Agency (DARPA) reconnaît qu'il est difficile d'aligner correctement tous les éléments sémantiques, par exemple le texte d'une nouvelle, l'image qui l'accompagne et les éléments que celle-ci contient[14]. Le projet ELSA met également l'accent sur l'analyse sémantique et explore de nouvelles façons de comprendre et de détecter les fausses données au moyen d'approches d'apprentissage automatique qui combinent l'analyse syntaxique et l'analyse perceptive.

Recherche d'images inversée

Les méthodes de recherche d'images inversée sont conçues pour trouver des images similaires en fonction du contenu visuel. Elles sont plus efficaces que de simples recherches par mots-clés dans les métadonnées des images, qui sont souvent supprimées par les plateformes de médias sociaux pendant le téléversement. L'idée de base est de construire une empreinte digitale numérique à partir des caractéristiques visuelles « stables » de l'image recherchée, puis de comparer cette empreinte digitale à celles d'une base de données d'images. Ce processus consiste généralement à identifier automatiquement les points clés perceptuellement importants dans une image et à rattacher une signature numérique à chaque point en fonction des pixels environnants. Des services comme Google Image Search et Tineye balaient continuellement le Web pour collecter de nouvelles images, calculer leurs empreintes digitales et les stocker pour les comparer à des milliards d'empreintes digitales stockées.

Source: https://www.unite.ai/consistent-ai-video-content-editing-with-text-guided-input/

[12] https://par.nsf.gov/servlets/purl/10346314 (en anglais seulement).

[13] https://openaccess.thecvf.com/content/WACV2022/papers/Xie_PROVES_Establishing_Image_Provenance_Using_Semantic_Signatures_WACV_2022_paper.pdf (en anglais seulement).

[14] https://www.darpa.mil/news-events/2021-03-02 (en anglais seulement).

Analyse des artefacts

La manipulation des images numériques laisse des traces qui peuvent être utilisées pour détecter les images falsifiées. L'analyse des artefacts, une technique de vérification de copies images, tient compte des artefacts générés par les opérations post capture et ceux induits lors de l'acquisition par le capteur lui-même. Ces artefacts fournissent des informations utiles pour établir l'authenticité d'une image. L'analyse englobe les artefacts causés pendant l'acquisition d'images (par la lentille et le capteur électronique), l'intégrité physique de la scène (incohérences comme les ombres incompatibles), les artefacts causés par les modèles génératifs et les artefacts visuels (incohérences visuelles non naturelles). La recherche sur la détection de ces artefacts est nécessaire, car ces solutions ont un potentiel de détection important étant donné qu'elles ne reposent pas sur une analyse ou des hypothèses explicites au sujet du contenu de l'image.

Détection d'images et de vidéos générées

Des modèles génératifs comme les RAG et les modèles de texte à image peuvent créer des images et des vidéos à partir d'invites textuelles. Pour différencier les images synthétiques des images réelles, les méthodes d'analyse tentent de détecter les artefacts produits lors de la création d'images par ces modèles génératifs. Il est également possible de retracer la classe de modèle génératif utilisé pour créer une image synthétique à partir d'une analyse détaillée des artefacts. Cependant, il est nécessaire de pouvoir compter sur des approches universelles capables de détecter les images générées par n'importe quel modèle. Des progrès sont réalisés en ce qui concerne les RAG, mais des travaux supplémentaires sont requis pour inclure d'autres modèles génératifs. Les ensembles de données publics comme CIFAKE et COCOFake, inspirés d'ensembles de données populaires en vision par ordinateur, sont essentiels au développement de ces méthodes.

La lutte contre la désinformation visuelle et multimodale englobe également d'autres stratégies, allant de l'authentification et de la provenance au tatouage numérique, entre autres choses.

Authentification/provenance

La Content Authenticity Initiative (CAI) développe une infrastructure pour assurer la traçabilité des documents visuels (images, vidéos), de la capture à la visualisation, et ainsi fournir aux utilisateurs un accès aux métadonnées et une liste des modifications appliquées. Ce « système sécurisé de bout en bout » pourrait s'étendre aux modèles génératifs et permettre aux créateurs de divulguer l'utilisation de l'IA générative dans la création de contenu. La Coalition for Content Provenance and Authenticity (C2PA) dirige également la création de normes techniques pour la certification de la source et de la provenance des documents médias. Toutefois, le succès de cette initiative dépend de son adoption généralisée par l'industrie, y compris les fabricants de capteurs, les développeurs de logiciels et les utilisateurs importants comme les médias. Des acteurs clés de l'industrie comme Adobe, Microsoft et Intel participent activement à cette initiative.

Tatouage numérique

Le tatouage numérique intègre dans une image une signature visuellement imperceptible et détectable au moyen d'un logiciel. Longtemps utilisé dans la gestion de la propriété intellectuelle, il devient un élément essentiel de l'infrastructure de provenance; il s'avère efficace lorsqu'il s'agit de manipulations bénignes comme la compression ou l'ajustement de contraste, mais moins utile lorsqu'il s'agit de manipulations comme les hypertrucages. L'IA contribue également aux solutions de tatouage numérique, comme Glaze, de l'Université de Chicago, qui protège les images contre les copies de style, et PhotoGuard, du MIT, qui injecte des perturbations imperceptibles pour immuniser les images contre les manipulations d'édition basées sur l'IA. Toutefois, l'efficacité du tatouage numérique est limitée par la grande quantité de matériel non tatoué sur Internet et son inefficacité lorsqu'il s'agit d'images utilisées hors contexte.

Responsabilité des chercheurs/développeurs

Étant donné la pression sociale croissante, les chercheurs et les développeurs en IA sont priés de tenir compte de l'utilisation malveillante potentielle de modèles partagés publiquement sur des plateformes comme GitHub. Les grands acteurs de l'industrie peuvent héberger leurs modèles derrière des API à accès contrôlé, contrairement aux entités de petite taille. Les interventions possibles comprennent l'injection de signatures dans des modèles préentraînés pour détecter leur implication dans la production de DVM et le fait de rendre des ensembles de données « radioactifs » afin de vérifier s'ils ont été utilisés pour entraîner des modèles soupçonnés de produire de la désinformation. Le succès de ces mesures repose sur la participation généralisée de la communauté des développeurs et sur des solutions de suivi robustes pour les modèles et les ensembles de données.

Désinformation et technologies de la parole

Les progrès des technologies de la parole ont considérablement accru l'incidence de la désinformation audio. Les technologies de conversion vocale et de synthèse texte-parole permettent la création de vidéos réalistes comportant des discours fabriqués. Les avancées récentes ont éliminé de nombreuses contraintes techniques, telles que la quantité de données, ce qui rend possible une synthèse vocale de haute qualité à partir de seulement quelques secondes d'enregistrement vocal. Des outils gratuits et faciles à utiliser rendent désormais ces technologies accessibles à tous. Les recherches sur le vol d'identité vocale, notamment dans le cadre des campagnes biennales ASVSpoof, mettent l'accent sur le fait de détecter si un échantillon de voix est authentique. Le défi consiste à effectuer cette détection à l'aveugle, sans connaître les techniques de synthèse utilisées ou sans être exposés à celles-ci pendant la formation, et de demeurer fiable dans des environnements bruyants. Bien qu'offrant des performances impressionnantes pour la vérification du locuteur, les systèmes de détection sont actuellement moins efficaces lorsqu'il s'agit d'hypertrucages. Les technologies de désinformation fondées sur la parole devraient produire des résultats encore plus réalistes à l'avenir grâce à l'intégration d'émotions subtiles et d'éléments non verbaux, ce qui rendra la détection plus difficile, tandis que la production sera facilitée.

Le défi en évolution et le besoin d'interventions multidisciplinaires

Le paysage de la désinformation visuelle et multimodale (DVM) est vaste et en constante évolution. La DVM pose d'importants défis dans divers secteurs, de la politique à la santé publique en passant par le secteur financier et la cohésion sociale. Le perfectionnement croissant des outils de création et de manipulation de contenu numérique, en particulier avec les progrès de l'IA, fait en sorte qu'il est de plus en plus difficile de distinguer un contenu authentique et d'un contenu falsifié. Cette complexité est aggravée par la diffusion rapide de l'information sur les médias sociaux et d'autres plateformes numériques, les capacités de détection et de contremesures actuelles étant souvent dépassées.

La lutte contre la VMD et ses effets nécessite une approche multidimensionnelle englobant des stratégies technologiques, juridiques, éducatives et sociétales. Bien que l'IA et l'apprentissage automatique fournissent des moyens prometteurs de détecter la DVM et de lutter contre celle-ci, ils soulèvent également des préoccupations relatives aux préjugés et à l'utilisation éthique. Les cadres juridiques doivent pouvoir concilier la dissuasion efficace contre l'utilisation malveillante et la protection de la liberté d'expression. Les initiatives éducatives, en particulier dans le domaine des connaissances médiatiques, sont essentielles pour outiller le public afin qu'il puisse faire une évaluation critique de l'information à laquelle il est exposé.

La collaboration entre divers intervenants – notamment les gouvernements, les entreprises technologiques, le milieu universitaire et la société civile – est essentielle pour élaborer des solutions complètes et adaptables à ce problème complexe. Tandis que les technologies continuent de progresser, il faudra, pour garder une longueur d'avance dans la lutte contre la désinformation, faire preuve d'une innovation, d'une vigilance et d'un engagement continus afin de préserver l'intégrité de l'information en cette ère du numérique.

Dans ce contexte, le CRIM et le Laboratoire sur l'intégrité de l'information de l'Université d'Ottawa continueront leur collaboration dans l'étude de la désinformation visuelle et multimodale, tout en fournissant des mises à jour régulières. De plus, nous nous engageons à promouvoir le développement d'outils et de méthodes visant à détecter et à contrer les menaces que représentent la désinformation et la mésinformation pour notre discours politique, notre cohésion sociale, ainsi que notre sécurité économique et nationale.

Bibliographie

Appel, Markus et Fabian Prietzel. « The detection of political deepfakes ». *Journal of Computer-Mediated Communication*, vol. 27, no 4 (juillet 2022) : zmac008. https://academic.oup.com/jcmc/article/27/4/zmac008/6650406.

Bateman, Jon. « Deepfakes and synthetic media in the financial system: Assessing threat scenarios ». *Carnegie Endowment for International Peace* (2020).

Brashier, Nadia M, et Schacter, Daniel L. « Aging in an Era of Fake News ». *Current directions in psychological science*, vol. 29,3 (2020) : 316-323. https://pubmed.ncbi.nlm.nih.gov/32968336/

Brennen, J. Scott, Felix M. Simon et Rasmus Kleis Nielsen. « Beyond (Mis)Representation: Visuals in COVID-19 Misinformation ». *The International Journal of Press/Politics*, vol. 26, no 1 (janvier 2021) : 277-299. https://journals.sagepub.com/doi/10.1177/1940161220964780.

Cao, Juan, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo et Jintao Li. « Exploring the Role of Visual Content in Fake News Detection ». Dans *Disinformation, Misinformation, and Fake News in Social Media*, édité par Kai Shu, Suhang Wang, Dongwon Lee et Huan Liu, 141-161. Lecture Notes in Social Networks. Cham : Springer International Publishing (2020). https://link.springer.com/chapter/10.1007/978-3-030-42699-6_8.

Dan, Viorela, Britt Paris, Joan Donovan, Michael Hameleers, Jon Roozenbeek, Sander van der Linden et Christian von Sikorski. « Visual Mis- and Disinformation, Social Media, and Democracy ». *Journalism & Mass Communication Quarterly*, vol. 98, no3 (septembre 2021) : 641-664. https://journals.sagepub.com/doi/10.1177/10776990211035395.

FTC. *Combatting Online Harms Through Innovation; Federal Trade Commission Report to Congress*. s.d. https://www.ftc.gov/reports/combatting-online-harms-through-innovation.

Helmus, Todd. *Artificial Intelligence, Deepfakes, and Disinformation: A Primer.* RAND Corporation, (2022). https://www.rand.org/pubs/perspectives/PEA1043-1.html.

Langguth, Johannes, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková et Daniel Thilo Schroeder. « Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes ». *Frontiers in Communication*, vol. 6 (24 mai 2021) : 632317. https://www.frontiersin.org/articles/10.3389/fcomm.2021.632317/full.

Languein, Adela. *Combatting Visual Misinformation on Social Media: A Review of Strategies and Concepts*. Mémoire de maîtrise, Université Concordia, (2022). https://spectrum.library.concordia.ca/id/eprint/990735/1/Languein_MA_S2022.pdf.

Lyons, Benjamin A., Jacob M. Montgomery, Andrew M. Guess, B. Nyhan, J. Reifler. « Overconfidence in news judgments is associated with false news susceptibility ». *Proceedings of the National Academy of Sciences of the United States of America*, 118, (2021) Article e2019527118. https://www.pnas.org/doi/full/10.1073/pnas.2019527118

Morrow, Garrett, Briony SwireThompson, Jessica Polny, Matthew Kopec et John Wihbey. « The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation ». *SSRN Electronic Journal*, (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3742120.

Paris, Brit et Joan Donovan. « Deepfakes and cheap fakes: the manipulation of audio and visual evidence ». *Data & Society*. (2019). https://datasociety.net/library/deepfakes-and-cheap-fakes/

Pramanick, Shraman, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov et Tanmoy Chakraborty. « Detecting Harmful Memes and Their Targets ». Dans *Findings of the Association for Computational Linguistics* : ACL-IJCNLP 2021, 278396. En ligne : Association for Computational Linguistics, (2021). https://aclanthology.org/2021.findings-acl.246/.

Svahn, Mattias et Serena Coppolino Perfumi. « Towards a Positioning Model for Evaluating the Use and Design of Anti-Disinformation Tools ». *JeDEM – EJournal of EDemocracy and Open Government*, vol. 14, no 2 (23 décembre 2022) : 109-129. https://jedem.org/index.php/jedem/article/view/746.

Vaccari, Cristian et Andrew Chadwick. « Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News ». *Social Media + Society*, vol. 6, no 1 (janvier 2020) : 205630512090340. https://journals.sagepub.com/doi/10.1177/2056305120903408.

uOttawa

Laboratoire sur l'intégrité de l'information

Information Integrity Lab

CRIM