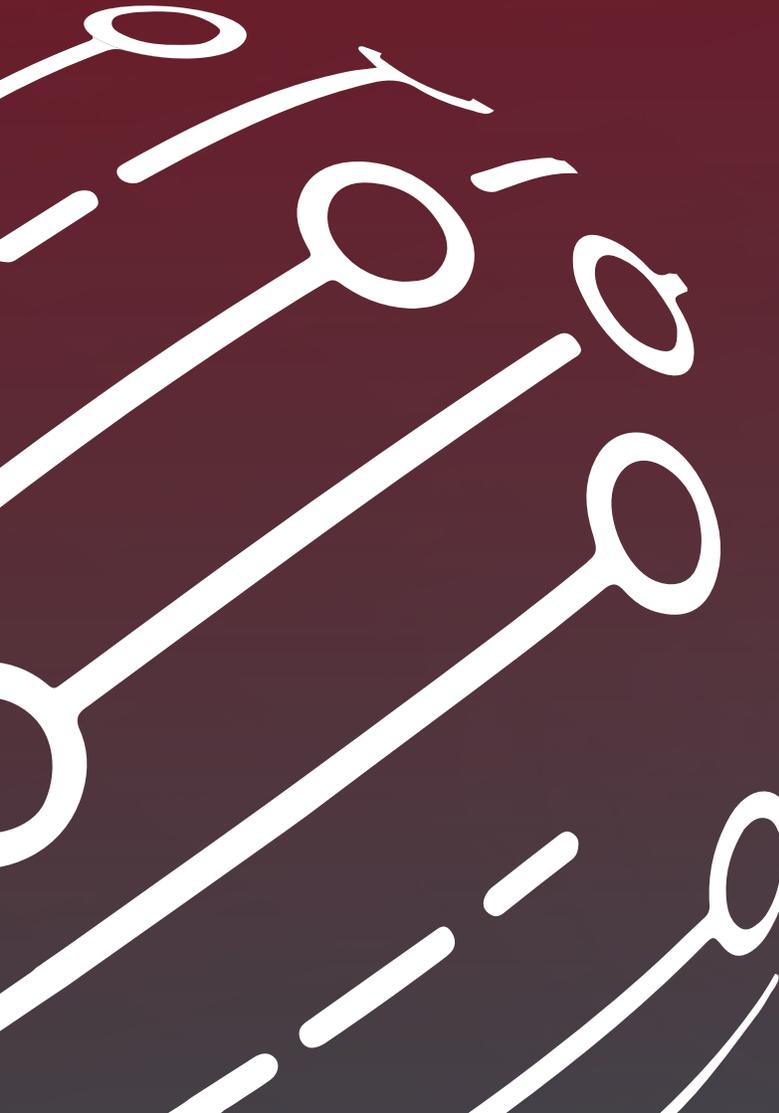


Désinformation Visuelle et Multimodale:

Analyse, enjeux, solutions

MARS 2025



 **CRIM**
CENTRE DE
RECHERCHE INFORMATIQUE
DE MONTRÉAL



uOttawa

Laboratoire sur l'intégrité
de l'information

Information Integrity Lab

Désinformation Visuelle et Multimodale:

Analyse, enjeux, solutions

Auteurs:

Marc Lalonde

Spécialiste en vision par ordinateur, CRIM

Gilles Boulianne, Ph. D.

Chercheur sénior en traitement automatique de la parole, CRIM

Nicolas Rutherford

Analyste sénior, Laboratoire sur l'intégrité de l'information, Université d'Ottawa

Mario Beaulieu

Spécialiste en vision par ordinateur, CRIM

Hamed Ghodrati, Ph. D.

Chercheur industriel, CRIM

Mohamed Dahmane, Ph.D.

Chercheur en vision par ordinateur, CRIM





À propos du CRIM

Le Centre de recherche informatique de Montréal (CRIM) est un organisme à but non lucratif spécialisé en recherche appliquée dans les domaines de l'intelligence artificielle, de la science des données et de l'architecture logicielle. Le CRIM collabore avec des partenaires industriels et publics pour concevoir des solutions en cybersécurité, intelligence artificielle et traitement des données, en mettant la recherche au service des défis concrets.



uOttawa

Laboratoire sur l'intégrité
de l'information

Information Integrity Lab

À propos de l'InfoLab

Le Laboratoire sur l'intégrité de l'information (InfoLab) de l'Université d'Ottawa se consacre à l'analyse et à la lutte contre la désinformation et la désinformation dans des domaines clés comme la démocratie, les changements climatiques et la santé publique. Il rassemble analystes, chercheurs et praticiens pour développer des analyses, outils et stratégies visant à renforcer l'intégrité de l'information et la résilience du public dans un environnement numérique en constante évolution.

Remerciements

Les auteurs de ce rapport tiennent à remercier les collaborateurs suivants :

Anastasia Menneghier, stagiaire (2023)

Yasmina Kheddache, stagiaire (2024-2025)

Célia Djilali, stagiaire (2024)

Dereck Piché, stagiaire (2024)

Houman Zolfaghari, directeur scientifique (2023-2024)

License

Ce travail est sous licence Creative Commons Attribution – NonCommercial – ShareAlike 4.0 International. Vous pouvez le partager et le redistribuer avec attribution, mais son usage commercial est interdit. Toute modification ou adaptation doit être diffusée sous la même licence, avec mention des changements, sans implication d'approbation par l'auteur ou l'éditeur. Les images sont utilisées avec autorisation et ne peuvent être reproduites ou partagées sans l'accord des détenteurs des droits.

Table Des Matières

Résumé Exécutif	1
Contexte et Enjeux	1
Une Nouvelle Ère de Désinformation.....	2
Objectifs Et Structure Du Document	3
1. Qu'est-ce que la désinformation visuelle et multimodale	4
1.1 Contexte	4
1.2 Définition et taxonomie.....	4
1.3 Comment est produite la désinformation visuelle et multimodale.....	6
1.3.1 Recontextualisation.....	6
1.3.2 Manipulation de contenu	6
1.3.3 Fabrication de contenu	7
1.4 Pourquoi est produite la désinformation visuelle et multimodale.....	7
2. Impacts de la DVM	9
2.1 L'information et les médias	9
2.2 La politique et les institutions.....	10
2.3 La santé et ses institutions	11
2.4 Économie, finance, commerce.....	15
2.5 Droits et sécurité des personnes	17
3. DVM et crise climatique	18
3.1 Un changement climatique sans précédent	18
3.2 L'émergence et l'expansion de la désinformation climatique.....	19
3.3 Cas d'étude	20
3.4 Discussion	21

4. Urgence	22
4.1 Impact plus grand du visuel par rapport au texte	22
4.2 Propagation intensifiée	22
4.3 Détection difficile	23
4.4 Problème peu étudié	23
4.5 Nouvelle vague d'outils pour la création de contenus	24
5. Acteurs dans la lutte contre la DVM	25
5.1 Au niveau politique	25
5.2 ONG	26
5.3 Groupes de recherche/instituts	26
5.4 Industrie	29
5.5 Grands projets	29
6. État de la lutte contre la DVM : solutions possibles	31
6.1 Solutions par les plateformes	31
6.1.1 Modération	32
6.1.2 Labelling	32
6.1.3 Contextualisation	33
6.2 Approches sociétales	33
6.3 Solutions politiques/légales	34
6.3.1 Cadres juridiques et réglementaires actuels	34
6.3.2 Lacunes et défis des cadres juridiques	35
6.4 Solutions technologiques	35
7. Lutte à la DVM : le paysage des solutions technologiques	36
7.1 Méthodes de production de DVM	36
7.2 Stratégies de détection impliquant l'image	37
7.2.1 Analyse d'intégrité sémantique	38
7.2.2 Recherche d'images inversées	39
7.2.3 Analyse des artéfacts	40
7.2.4 Détection d'images générées	41

7.3	Désinformation et technologies de la parole	42
7.3.1	État des lieux	42
7.3.2	Le développement de la technologie	42
7.3.3	La disponibilité des outils et leur facilité d'utilisation	42
7.3.4	Les moyens de lutte	44
7.3.5	Le futur de la désinformation en parole	44
7.4	Stratégies additionnelles ou complémentaires	45
7.4.1	Authentification/provenance	45
7.4.2	Tatouage numérique	45
7.4.3	Responsabilité des chercheurs/développeurs	46
7.4.4	Dépôts d'images/vidéos signalées	47
7.5	Vérification de faits	47
7.6	Jeux de données disponibles	48
7.7	État des outils de détection	50
7.8	Lutte à la DVM : Discussion	53
8.	Lutte à la DVM : l'influence de l'ia générative	54
8.1.1	Génération d'images	54
8.1.2	Manipulation d'images	56
8.1.3	Modèles de génération et d'édition de vidéos	56
8.2	Détection de désinformation par modèles multimodaux	57
8.2.1	Détection d'images générées	57
8.2.2	Analyse sémantique	58
	Annexe I - Lexique	60
	Annexe II - Outils De Génération Vidéo	62
	Annexe III - Méthodes D'analyse Sémantique	63
	Annexe IV - Analyse D'artéfacts	65
	Annexe V - Désinformation Et Audio : Références Scientifiques Et Techniques, Démonstrations	69
	Annexe VI - Jeux De Données Pour La Lutte À La DVM	70
	Bibliographie et Lectures Supplémentaires	72



RÉSUMÉ EXÉCUTIF

Contexte et Enjeux

La désinformation visuelle et multimodale (DVM) représente une menace croissante pour la société, avec des implications majeures pour les institutions, la confiance du public et les processus de prise de décision. Contrairement aux formes plus traditionnelles de désinformation, souvent centrées sur du contenu textuel, la DVM exploite la puissance combinée du texte, des images, des vidéos et de l'audio pour produire des contenus persuasifs et hautement trompeurs. En jouant sur plusieurs sens et en renforçant des récits biaisés à travers divers formats médiatiques, la DVM capte davantage l'attention du public, rendant la désinformation plus convaincante et plus difficile à réfuter. Ces fausses informations ne se limitent pas à induire en erreur : elles façonnent activement l'opinion publique, influencent les décisions politiques et sapent la confiance dans les institutions démocratiques, le journalisme et la science.

L'impact de la DVM est particulièrement marqué dans des domaines critiques comme la politique, la santé publique, la sécurité, la défense et l'action climatique. En politique, les deepfakes et images manipulées ont été utilisés pour déformer la réalité, influencer des élections et renforcer la méfiance envers les dirigeants et les institutions. En matière de santé publique, des campagnes de désinformation reposant sur des visuels trafiqués et des infographies trompeuses ont diffusé des fausses informations sur les vaccins, les pandémies et les traitements médicaux. La désinformation climatique repose quant à elle sur des images fabriquées, des montages vidéo manipulés et du contenu généré par l'IA pour brouiller la compréhension des faits scientifiques, minimiser l'urgence climatique et amplifier le scepticisme à l'égard des politiques environnementales.

L'essor rapide de la DVM repose sur plusieurs facteurs clés :

- » Propagation accélérée via les plateformes numériques : Les réseaux sociaux et les plateformes de partage de contenu sont devenus les principaux vecteurs d'information et de désinformation. L'énorme quantité de contenu numérique, combinée aux algorithmes qui favorisent l'engagement, permet aux récits faux ou trompeurs de se diffuser à grande vitesse, avant même que les efforts de vérification puissent intervenir. L'interactivité de ces plateformes accentue le phénomène en incitant les utilisateurs à partager involontairement du contenu trompeur.
- » Progrès technologiques en matière d'IA et de manipulation des médias : Les outils d'intelligence artificielle, comme la génération de deepfakes, la synthèse vocale et les modèles de transformation texte-image, sont devenus de plus en plus performants. Autrefois réservées à des experts, ces technologies sont désormais accessibles au grand public via des applications simplifiées, abaissant ainsi la barrière à la production de fausses informations sophistiquées. À mesure que ces outils progressent, il devient de plus en plus difficile de distinguer un contenu authentique d'un contenu généré ou modifié par l'IA.
- » Failles dans la détection et les contre-mesures : Bien que des initiatives de fact-checking et des outils de détection basés sur l'IA existent, ils peinent à suivre le rythme rapide des innovations en matière de DVM. La plupart des contre-mesures restent réactives, identifiant et démentant les fausses informations une fois qu'elles ont déjà largement circulé. De plus, l'analyse des contenus multimodaux reste un défi : détecter une image manipulée ne permet pas toujours de comprendre comment elle s'intègre dans un récit trompeur combinant texte, audio et vidéo. Par ailleurs, des techniques d'IA adversariale sont développées pour contourner les outils de détection, compliquant encore davantage la lutte contre la désinformation manipulée.

Une Nouvelle Ère de Désinformation

La désinformation n'est pas un phénomène nouveau, mais elle a franchi un cap décisif, transformant en profondeur la manière dont les individus perçoivent la réalité et interagissent avec l'information. Historiquement, la diffusion de récits trompeurs reposait principalement sur l'oral et l'écrit, nécessitant des efforts délibérés pour élaborer des messages mensongers et un appui sur les médias traditionnels pour les propager. La manipulation de preuves visuelles et sonores demandait auparavant un savoir-faire spécialisé, des équipements coûteux et un travail minutieux. Pendant des décennies, la crédibilité des photographies, vidéos et enregistrements audio reposait sur leur statut de preuves factuelles. Aujourd'hui, cette confiance est ébranlée.

Les avancées de l'intelligence artificielle et des outils d'édition numérique ont ouvert une ère où des falsifications peuvent être créées et diffusées en temps réel avec un réalisme saisissant. La technologie des deepfakes permet de générer des vidéos hyperréalistes, clonant l'apparence et la voix d'une personne à partir de très peu de données. Les modèles d'IA dédiés à la génération d'images fabriquent des événements fictifs, créant des visuels intégrés à des campagnes de désinformation. L'IA appliquée au texte peut produire des articles trompeurs ou imiter des experts, tandis que la synthèse vocale rend possible la reproduction fidèle d'une voix, renforçant ainsi la crédibilité de contenus fallacieux.

Cette mutation introduit des risques sans précédent :

- » **Érosion de la crédibilité des preuves** : Les documents visuels et sonores, autrefois perçus comme des éléments de preuve incontestables, peuvent désormais être falsifiés avec une facilité déconcertante. Cette situation alimente la méfiance envers le journalisme, la justice et l'espace public, rendant plus difficile la distinction entre information et manipulation.
- » **Démocratisation de la désinformation sophistiquée** : La création de faux contenus de haute qualité nécessitait autrefois des ressources et des compétences avancées. Désormais, les outils basés sur l'IA ont abaissé cette barrière, rendant la tromperie accessible à tout internaute disposant d'un smartphone ou d'un ordinateur.
- » **Propagation virale sur les réseaux sociaux** : Le contenu manipulé se diffuse de manière exponentielle, atteignant des millions d'utilisateurs avant même que des vérifications puissent être effectuées. Les algorithmes des plateformes privilégient l'engagement, ce qui favorise involontairement la désinformation par rapport aux contenus factuels.
- » **Arme d'influence stratégique** : Les technologies de manipulation de l'IA sont utilisées par des acteurs étatiques et non étatiques à des fins politiques, économiques et idéologiques. Ces outils servent à manipuler des élections, discréditer des journalistes, provoquer des troubles sociaux et fausser la perception des enjeux mondiaux comme le changement climatique.

Face à ces défis, une réaction coordonnée et urgente est nécessaire. Cela implique des avancées technologiques en matière de détection, des mesures réglementaires adaptées, des initiatives éducatives renforçant la culture numérique, ainsi qu'une coopération accrue entre les industries technologiques, les gouvernements et la société civile.

Objectifs et Structure du Document

Ce document a pour but de dresser un état des lieux de la désinformation visuelle et multimodale (DVM) et des moyens pour la circonscrire et la combattre. L'étude porte à la fois sur les aspects sociaux, scientifiques et technologiques. Ce rapport vise à proposer une référence commune aux acteurs de la lutte contre la DVM, qu'ils se situent du côté académique ou technique, agissant au niveau de l'analyse et de l'action stratégique, ou bien au niveau de solutions logicielles, afin de servir de support pour la réflexion et l'établissement de plans d'action.

Les premières sections du rapport définissent la DVM, décrivent son impact social et présentent un argumentaire sur la nécessité de solutions majeures pour y faire face. Une section dédiée traite spécifiquement de la désinformation sur la crise climatique et les enjeux qui s'y rattachent.

On y trouve aussi un survol de plusieurs initiatives actuelles qui visent à étudier et combattre la désinformation. Évidemment, la scène est en pleine évolution, et ce survol ne cherche pas à être exhaustif.

Les sections suivantes sont plus techniques. Elles dressent l'état actuel des méthodes et outils de production et de détection de DVM, en incluant tant les approches académiques que les outils, les logiciels et les librairies commerciaux et open-source. On y trouve aussi des discussions sur les nouvelles méthodes basées sur l'IA générative qui sont en train de transformer en profondeur les moyens utilisés pour la désinformation. Chaque modalité, c.-à-d. image, audio et vidéo, sera aussi traitée séparément pour en dégager les spécificités.

Plusieurs pistes de solutions technologiques sont envisagées, même si à l'heure actuelle, les outils de lutte contre la DVM ne sont généralement pas encore à la hauteur du problème.

Une série d'annexes complètent les détails des survols techniques.



1. QU'EST-CE QUE LA DÉSINFORMATION VISUELLE ET MULTIMODALE

1.1 Contexte

Il est indéniable que la désinformation sous toutes ses formes est un phénomène en grande croissance. Autant dans la littérature scientifique que dans les publications grand public, on parle de « tsunami » de mésinformation, d'« infodémie » (Jia, 2020; Brennen et al., 2021), d'« infocalypse » (Pawelec, 2022), etc.

Étant donné que la désinformation sévit en grande partie sur les plateformes de réseaux sociaux, son impact est aggravé par la multiplication de ces plateformes et la plus grande facilité avec laquelle du contenu peut être partagé entre plateformes. Celles-ci sont critiquées pour leur relative inaction à combattre le phénomène, possiblement parce qu'elles en tirent un profit financier à cause du volume « d'échanges ».

Plus généralement, on observe deux phénomènes :

- » **La militarisation de l'information ou l'IA** (AI weaponization)¹ est au centre du champ de bataille de la désinformation, aussi bien en tant que facilitateur de diffusion que d'outil de mitigation (Duberry, 2022).
- » **L'ère post-vérité** (post-truth era) où l'opinion personnelle, l'idéologie, l'émotion et la croyance l'emportent sur la réalité des faits.²

Tous ces éléments mettent en évidence la complexité de la lutte à la désinformation. Socialement, il importe de trouver un équilibre entre la modération de contenu et le droit de parole. Au niveau technique, il faut s'attaquer au problème du volume grandissant de contenu douteux, aggravé par la facilité d'utilisation des nouveaux outils d'intelligence artificielle comme ChatGPT4. Pis encore, les outils de détection actuels (aussi basés sur l'IA) ne sont pas toujours adéquats.

Dans ce contexte, il existe dans la littérature scientifique un consensus à l'effet que la lutte à la désinformation est rendue encore plus compliquée quand celle-ci

comporte du contenu visuel (image, vidéo) ou multimodal (visuel+texte, visuel+audio, etc.). Il est donc important de faire un état des lieux visant la désinformation visuelle et multimodale : cartographier les initiatives, recenser les acteurs, identifier les tendances à surveiller, et surtout étudier les nouveaux outils de production et de détection.

1.2 Définition et taxonomie

Le terme « **désinformation** » et ses variantes (**mésinformation**, **malinformation**) sont utilisés dans plusieurs sens différents (Jia 2020). Par exemple, certains chercheurs définissent la mésinformation comme de l'information qui est fautive, indépendamment de l'intention de l'auteur. D'autres font la différence entre mésinformation propagée intentionnellement et mésinformation propagée accidentellement, ou même voient la désinformation comme une sous-catégorie de mésinformation. Il existe même des concepts de « définition interdisciplinaire » pour réconcilier toutes ces définitions.

Dans le contexte de ce rapport, on retient les définitions proposées par le Centre canadien pour la cybersécurité:³

La mésinformation désigne le fait de diffuser de la fautive information sans avoir nécessairement de mauvaises intentions.

Par désinformation, on entend le fait de diffuser de la fautive information dans le but de manipuler ou de tromper des personnes, des organisations et des États ou bien de leur faire du tort.

Pour ce qui est de **la malinformation**, il s'agit du fait de diffuser de l'information qui repose sur un fait, mais qui est souvent exagérée de façon à tromper ou même à causer des préjudices.

¹ <https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html>

² Certains font aussi référence à l'ère du « grand doute » (deep doubt era) : <https://arstechnica.com/information-technology/2024/09/du-to-ai-fakes-the-deep-doubt-era-is-here/>

³ <https://www.cyber.gc.ca/fr/orientation/reperer-les-cas-de-mesinformation-desinformation-et-malinformation-itsap00300>

À l’instar de plusieurs auteurs (ex. Languein, 2022), le terme « fake news » n’est pas utilisé dans ce document parce qu’il est jugé trop vague et qu’il peut englober à la fois mésinformation, désinformation, malinformation et contenu véridique utilisé hors contexte.⁴

On associe souvent la désinformation à un contenu textuel qui rapporterait des faits inexacts. Ce document se penche plutôt sur la désinformation visuelle et multimodale, qui utilise comme support l’image, la vidéo et l’audio, qui peuvent aussi être accompagnés de texte. Le tableau ci-dessous résume les différents types de désinformation:

Type	Description	Exemples
Texte	<ul style="list-style-type: none"> Création de contenu textuel 	<ul style="list-style-type: none"> Génération d’un texte à l’aide de ChatGPT
Image	<ul style="list-style-type: none"> Modification d’une image Génération d’image 	<ul style="list-style-type: none"> Collage d’une portion d’image dans une autre Hypertrucages (Deepfakes)
Vidéo	<ul style="list-style-type: none"> Modification d’une vidéo Génération de vidéo 	<ul style="list-style-type: none"> Modification de la vitesse de la vidéo Modèles génératifs
Audio / audio + vidéo	<ul style="list-style-type: none"> Modification d’une bande audio 	<ul style="list-style-type: none"> Voice spoofing Ralentissement du débit de parole d’une personne
Texte + image/vidéo	<ul style="list-style-type: none"> Ajout d’une image ou d’une vidéo accompagnant un texte 	<ul style="list-style-type: none"> Recontextualisation dans publication sur réseaux sociaux
Texte dans image	<ul style="list-style-type: none"> Insertion de texte dans une image 	<ul style="list-style-type: none"> Mèmes internet

Comme ce rapport fait référence à plusieurs technologies d’intelligence artificielle, il convient de donner quelques définitions de bases.

Terme français	Terme anglais	Définition
Modèle génératif	Generative model	Modèle issu de l’apprentissage automatique qui permet à la fois de générer de nouveaux exemples à partir des données d’entraînement et d’évaluer la probabilité qu’un nouvel exemple provienne ou ait été généré à partir des données d’entraînement. ⁵
Mème internet	Internet meme	Contenu multimédia créé par la superposition de contenu textuel sur une séquence vidéo ou sur une image, afin d’être diffusé rapidement sur Internet, notamment sur les réseaux sociaux.
Fausse nouvelle	Fake news	Publication qui imite la structure d’un article de presse, qui comprend à la fois des renseignements véridiques et des renseignements erronés. ⁶
Génération texte-à-image	Text-to-image generation	La synthèse texte-à-image utilise un modèle génératif pour produire des images inspirées d’un texte ou invite fourni en entrée.
Hypertrucage	Deepfake	Procédé de manipulation audiovisuelle qui recourt aux algorithmes d’apprentissage profond pour créer des trucages ultra réalistes.
Grand modèle de langue	Large language Model (LLM)	Modèle de langue entraîné sur un très vaste corpus, le plus souvent multilingue et comportant un très grand nombre de paramètres. Exemples : Bert, GPT.
Stable Diffusion DALL-E, Midjourney, etc.		Modèles qui offrent une approche prometteuse pour générer des images détaillées et réalistes à partir de descriptions textuelles.
Remplacement de visage	Face swapping	Technique d’édition d’image qui permet de remplacer le visage d’une personne par celui d’une autre.

⁴ À noter également, « fake news » est souvent utilisé de manière abusive pour discréditer des points de vue opposés et/ou pour miner la crédibilité des médias/journalistes professionnels à travers le monde.

⁵ <https://datafranca.org/wiki>

⁶ <https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/26542775/fausse-nouvelle>

Terme français	Terme anglais	Définition
Usurpation d'identité par la voix, Hameçonnage par la voix	Voice spoofing, Vishing	Consiste à usurper l'identité d'une personne en clonant sa voix par une technique d'intelligence artificielle.
Faux bon marché	Cheapfake	Fabrication de faux contenus qui consiste à manipuler une image/vidéo avec un logiciel bon marché pour en changer le sens : recadrage, floutage, insertion d'une portion d'image dans une autre, etc.
Faux superficiel	Shallowfake	Fabrication de faux contenus qui consiste à associer à un texte une image ou une vidéo prise hors contexte.
Multimodal		Relatif au traitement et à l'analyse de plusieurs types de données différentes : textuelles, visuelles, sonores et même les combinaisons de celles-ci.-

1.3 Comment est produite la désinformation visuelle et multimodale

Comme nous l'avons vu brièvement à la section précédente, il existe plusieurs manières de produire de la désinformation sous forme multimodale, et ce, avec un degré de sophistication variable. Selon Heley et al.⁷, on trouve trois classes de DVM : recontextualisation, manipulation de contenu et fabrication de contenu. Le besoin de formaliser/standardiser cette taxonomie de DVM, en particulier pour la communauté de la vérification de faits, a mené à la création d'une autre proposition⁸ de codification (schéma) appelée MediaReview.⁹ Il est important de préciser ceci d'entrée de jeu : ce n'est pas parce qu'une image a été modifiée qu'elle est automatiquement un véhicule de désinformation; de la même façon, une image ou un texte généré par l'IA n'est pas nécessairement problématique, et à l'inverse des documents pourraient être véridiques et intacts, mais

leur association pourrait poser problème. En bout de ligne, la question de fond, d'une grande complexité, est la suivante : quelle est l'intention de l'auteur?

1.3.1 Recontextualisation

L'approche la plus simple est connue sous le terme shallowfake (faux superficiel) (Ajder & Glick, 2021) et elle consiste à associer à une image ou une vidéo réelle, soit un texte réel factuellement exact mais pris hors contexte, p. ex. la légende d'une autre image/vidéo, soit un texte fabriqué de toutes pièces à caractère tendancieux ou contenant de fausses affirmations.

1.3.2 Manipulation de contenu

La manipulation de contenu se présente sous plusieurs formes selon le type de modalité (texte, image, vidéo, audio) en cause. Pour les images, les techniques de manipulation sont souvent connues sous le vocable 'cheapfakes' (faux bon marché) qui englobent toutes les manipulations qui peuvent être réalisées avec un logiciel bon marché : celles-ci vont du simple rognage/ recadrage à l'insertion de portions d'images provenant d'autres documents, en passant par les changements de luminosité/couleur (e.g. assombrir une image, rehausser certaines couleurs, jouer avec le contraste), le floutage de certaines régions de l'image, etc., tout ceci dans le but de masquer ou mettre en évidence des informations spécifiques, rehausser l'émotion véhiculée, etc. Ces techniques incluent également la manipulation des axes d'un graphique pour en déformer l'interprétation. Du côté de l'audio/vidéo, le ralentissement ou l'accélération du débit vocal d'une vidéo¹⁰ est plutôt simple à réaliser, mais potentiellement efficace. Côté textuel, on peut penser à des changements apportés aux entités nommées associées à un texte qui accompagne une image ou une vidéo. Par exemple, dans la légende d'une photo prise en un lieu X à une date Y lors d'un événement qui rassemble certaines personnes, on pourrait remplacer le lieu X par un autre, modifier la date pour laisser croire que l'événement est plus récent, changer le nom des personnes, etc.

⁷ <https://pmc.ncbi.nlm.nih.gov/articles/PMC9451169/>

⁸ <https://docs.google.com/document/d/1jRbX2lesVQrWvKpehb8ntSMKeOD88bZp3nK8ZAjq6E4/edit>

⁹ <https://www.niemanlab.org/2020/01/is-this-video-missing-context-transformed-or-edited-this-effort-wants-to-standardize-how-we-categorize-visual-misinformation/>

¹⁰ <https://www.samsungsds.com/en/insights/what-are-cheapfakes.html>

1.3.3 Fabrication de contenu

Plus loin dans l'échelle de complexité, on retrouve les techniques de fabrication audiovisuelle où le contenu, modifié ou créé de toutes pièces, apparaît réel et légitime. Pour la modification de contenu, il existe une grande panoplie de techniques, souvent regroupées sous le terme « deepfakes » : ce sont des méthodes sophistiquées capables de créer des visages hyperréalistes, faire des substitutions de visages (face swaps) où le visage d'une personne est greffé au corps d'une autre, substituer des vêtements (clothing-swap), manipuler l'arrière-plan d'une image, faire de la synthèse vocale possiblement combinée à du lip sync afin de « faire parler » un individu artificiellement, etc. Pour la création de contenu, la vague des modèles génératifs est naturellement à l'épicentre du phénomène : il est maintenant facile de créer de toutes pièces des images et même des vidéos hyperréalistes, comme nous le verrons plus en détails à la section 8.

1.4 Pourquoi est produite la désinformation visuelle et multimodale

À retenir :

- » La désinformation est une arme qui existe depuis longtemps du côté du texte et du discours.
- » Jusqu'à récemment, la manipulation d'images était soit assez primitive, soit coûteuse à produire de façon convaincante.
- » En 2022, on a vu l'apparition de techniques très puissantes, avec des outils de plus en plus abordables, pour la création et la manipulation réalistes et rapides d'images et de vidéos, basée sur des commandes et descriptions textuelles.
- » Les images et vidéos produites par ces dernières vont devenir de plus en plus indistinguables des documents factuels.

La manipulation de documents audiovisuels existe depuis longtemps. Un rapport du CDEI britannique¹¹ mentionne l'existence d'une image où un « face swap » impliquant le président Lincoln a été réalisé au plus fort de la guerre de sécession américaine. Depuis l'arrivée de Photoshop, au début des années 1990, les retouches d'images sont devenues plus faciles que jamais à réaliser.

Dans plusieurs cas, les motifs sont artistiques (satire, parodie) ou pratiques (ex. recadrage de photos pour mise en page journalistique). La vraie menace provient des acteurs malveillants qui trouvent intérêt à trafiquer des images pour leur propre bénéfice. Weikmann & Lecheler (2022) les classent en trois groupes : acteurs politiques (gouvernements, partis politiques, élus/candidats), médiatiques (disséminateurs accidentels de mésinformation) et privés (plateformes, groupes d'individus, etc.). Un rapport récent du GAO américain¹² mentionne que la menace provenant d'acteurs gouvernementaux étrangers est plus sérieuse que celle d'acteurs privés. Vaccari & Chadwick (2020) rapportent que certains acteurs politiques sont animés d'une volonté de semer le chaos et de créer de l'incertitude systémique par la diffusion de messages contradictoires, absurdes ou déroutants. Même si les motivations pour la production de DVM sont les mêmes que pour celles de désinformation textuelle, les particularités de la DVM (complexité de production, grande dissémination, difficulté à faire la vérification de faits) impliquent des profils d'acteurs potentiellement différents.

Le déploiement de la désinformation suit des schémas souvent similaires : on mène des campagnes de désinformation pour distraire, embrouiller (confuse), manipuler, semer la discorde/incertitude. Le but n'est pas toujours de tromper, mais parfois de simplement semer le doute, ou même capter l'attention des groupes visés pour leur faire « gaspiller » leur temps d'écran limité.

De façon générale, des motivations politiques ou économiques sont en cause, par exemple pour renforcer le positionnement du producteur ou pour discréditer un adversaire ou un concurrent. Ce faisant, les acteurs malveillants sont pragmatiques et cherchent le meilleur

¹¹ <https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deep-fakes-and-audiovisual-disinformation#about-this-cdei-snapshot-paper>

¹² <https://www.gao.gov/assets/gao-24-107600.pdf>

rapport qualité-prix : si la production d'un deepfake est jugée comme étant trop coûteuse ou ayant une rentabilité incertaine, un cheapfake sera privilégié, et le réflexe sera de le produire avec un outil facilement accessible (ex. solution commerciale¹³). L'étude de Hameleers (2024) vient d'ailleurs crédibiliser ce point de vue après avoir étudié la perception d'une population de sujets exposés à un message politique modifié par deepfake ou cheapfake : « this study offers tentative evidence that highly sophisticated modes of visual disinformation are not clearly more effective than less sophisticated and more accessible forms of deception ».¹⁴

Langguth et al. (2021) listent les conditions qui vont affecter la rentabilité d'un deepfake :

- » Doit impliquer des personnes connues.
- » Doit être compris avec un minimum de contexte.

- » Doit induire une émotion suffisamment forte pour entraîner une action ou un changement d'opinion.

Cependant, les avancées en IA sont fulgurantes, plus spécifiquement en ce qui concerne les modèles génératifs (p. ex. texte-à-image) qui sont peu coûteux et faciles à utiliser. On en a pour preuve l'augmentation croissante de codes source de modèles génératifs présents en libre-service sur des plateformes d'hébergement de codes telles que Github. On peut observer notamment que, entre 2019 et 2022, le nombre de dépôts Github créés par année autour de la thématique des modèles texte-à-image a plus que doublé (diagramme ci-dessous). Ainsi, il est permis de penser que la production de médias synthétiques deviendra de plus en plus rentable avec la multiplication de ces modèles et, surtout, leur plus grande accessibilité. Soulignons finalement l'inquiétude de Goldstein et al. (2023) qui jugent qu'il est difficile d'imaginer les possibilités à venir quand on considère la hausse rapide de joueurs et d'innovations dans le domaine.

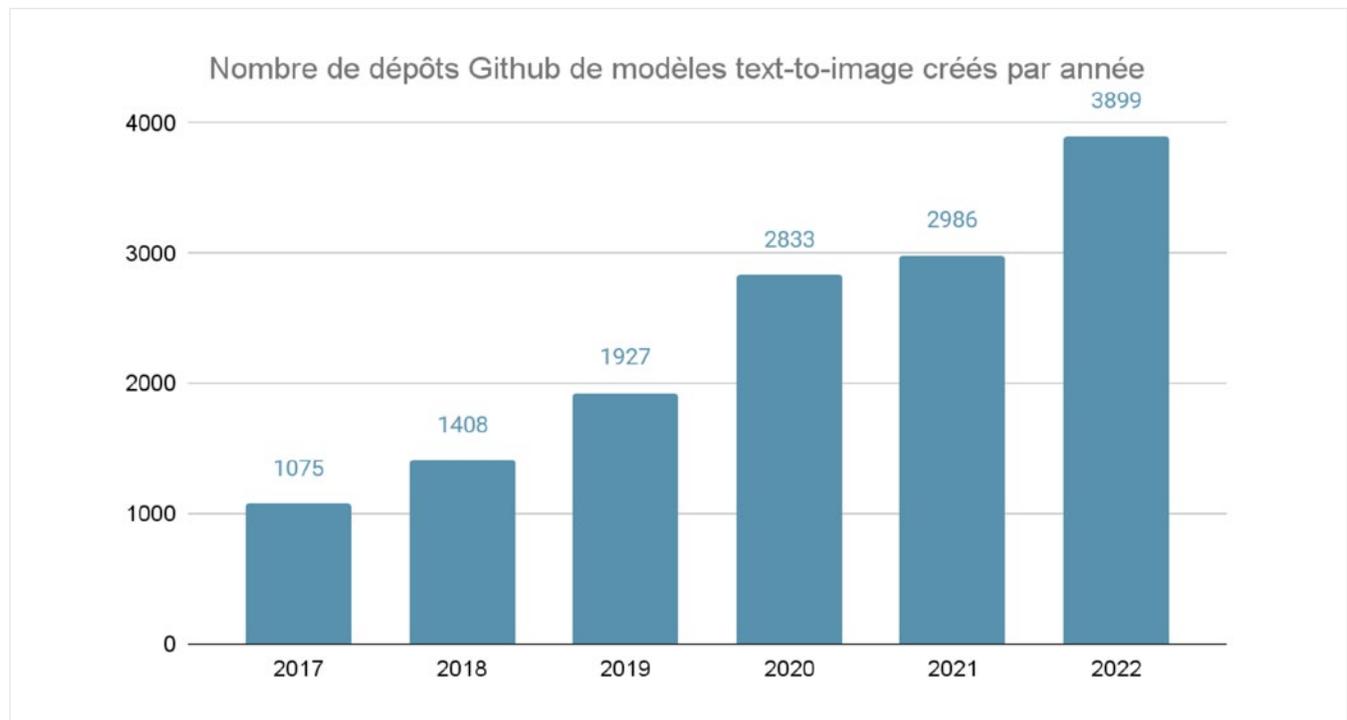


Figure 1 - Histogramme¹⁵ du nombre de dépôts Github texte-à-image créés par année

¹³ <https://public-assets.graphika.com/reports/graphika-report-deepfake-it-till-you-make-it.pdf>

¹⁴ <https://academic.oup.com/ijpor/article/36/1/edae004/7617425>

¹⁵ Analyse effectuée via les mots clés « text-to-image », « text2image », « image-generation », « txt2img » pour les années 2017 à 2022 via l'API Github Rest.

2. IMPACTS DE LA DVM

À retenir :

- » Tous les aspects de la société et de ses institutions sont vulnérables à la menace de DVM.
- » Les médias traditionnels, mis sous pression par les nouveaux acteurs médiatiques/non traditionnels aux principes plus élastiques, sont mal équipés pour se défendre contre le risque de diffuser la DVM par mégarde : faible littératie technique des journalistes en général, outils de vérification de faits mal adaptés et peu performants.
- » Le monde politique s'inquiète de la montée en puissance de la DVM à cause de l'érosion de la confiance des citoyens envers les institutions, les médias, la science.
- » Le monde de la santé n'est pas en reste avec la DVM, notamment autour de la COVID-19 et de la vaccination. Même le secteur financier/commercial est impacté par la fausse représentation véhiculée sur Internet, sans oublier les citoyens eux-mêmes.

On sait que de grands pans de la société sont affectés à divers degrés par la désinformation et la mésinformation¹⁶, mais l'influence de leur aspect visuel est moins bien documentée. Cette section fait un tour d'horizon de l'impact de la désinformation visuelle et multimodale sur la politique, l'économie et le tissu social. Par ailleurs, on sait également que la désinformation est très présente dans les échanges publics à propos de la crise climatique, à tel point qu'elle représente une sérieuse menace (et peut-être même la menace la plus sérieuse) à la mise en application des plans concrets de lutte aux changements climatiques.¹⁷ Pour cette raison, l'impact de la DVM sur la crise climatique fera l'objet d'une analyse séparée plus poussée à la section 3.

2.1 L'information et les médias

Les médias sont fortement impactés par la désinformation qui circule sur Internet, amplifiée par les réseaux sociaux et les algorithmes favorisant la viralité des contenus sensationnalistes. Parallèlement, les groupes médiatiques occidentaux traversent une période de fragilité économique, marquée par une baisse drastique des revenus publicitaires au cours de la dernière décennie. Cette situation a conduit à des réductions de personnel, limitant leur capacité à produire un journalisme d'investigation approfondi et à vérifier efficacement les faits. Combinée à la pression constante de publier ou diffuser l'information rapidement, cette situation compromet l'efficacité des processus de vérification des faits. Couplée à ceci, Thomson et al. (2022) soulignent la faible adoption des outils de vérification par les journalistes, indiquant que « seulement 11 % des journalistes dans le monde utilisent des outils de vérification sur les réseaux sociaux ».

Le rapport de l'UNESCO¹⁸ corrobore ces observations et conclut que plusieurs tendances convergent pour transformer le paysage médiatique mondial. On observe en parallèle :

1. L'effondrement des modèles économiques traditionnels, qui fragilise la viabilité financière des médias et réduit leur capacité à investir dans le journalisme d'enquête et la production d'informations de qualité.
2. La transformation numérique des salles de rédaction, qui accélère le rythme de production de l'information, parfois au détriment de la rigueur journalistique.
3. L'émergence de nouveaux écosystèmes médiatiques, alimentés par la montée en puissance des plateformes numériques et des réseaux sociaux, redistribuant les dynamiques de pouvoir et exposant les publics à davantage de désinformation et de contenus non vérifiés.

¹⁶ <https://blogs.lse.ac.uk/impactofsocialsciences/2024/10/08/misinformation-is-a-threat-to-society-lets-not-pretend-otherwise/>

¹⁷ <https://www.un.org/en/climatechange/rampant-disinformation-delaying-climate-action>

¹⁸ <https://www.mediadefence.org/ereader/publications/modules-de-synthese-sur-les-litiges-relatifs-aux-droits-numeriques-et-a-la-liberte-d-expression-en-ligne/fausses-nouvelles-mesinformation-et-propagande/la-mesinformation-la-désinformation-et-la-malinformation/?lang=fr>

Ces difficultés sont exacerbées par une crise de confiance croissante envers les médias traditionnels : « Seulement 29 % des citoyens américains disent faire confiance aux informations, et seulement 13 % disent faire confiance aux informations sur les réseaux sociaux » (Languein, 2022). Plus largement, l'environnement médiatique a connu une transformation radicale. Les médias traditionnels ne détiennent plus le monopole de la production et de la diffusion de l'information. L'essor des technologies numériques a permis une diversification sans précédent des sources d'information, accompagnée de nouvelles pratiques de partage et de contribution. Cette évolution intensifie la concurrence, désormais alimentée par des individus et des groupes qui disposent également des moyens de se faire entendre et de toucher de vastes audiences.

La désinformation visuelle (DVM) rend la vérification des faits encore plus complexe. Les journalistes ne disposent pas des outils nécessaires pour effectuer une vérification efficace des contenus visuels. On observe également des disparités en termes de littératie numérique, de compétences dans l'utilisation d'Internet et de maîtrise des bases de l'édition photo parmi ces professionnels. De plus, comme le montre ce rapport, il n'existe actuellement aucun outil véritablement adapté pour les soutenir dans cette tâche. Parmi les besoins identifiés, Dan et al. (2021) soulignent les suivants :

- » **Développement d'outils spécifiques** : Les journalistes ont besoin de logiciels capables d'analyser et de vérifier rapidement l'authenticité des contenus visuels, notamment pour détecter les manipulations dans les images ou les vidéos.
- » **Formation spécialisée** : Il est essentiel de proposer des programmes de formation axés sur l'identification et la correction de la désinformation visuelle, incluant l'utilisation des outils technologiques et des techniques de démystification (debunking).
- » **Accès à des bases de données fiables** : Des ressources centralisées contenant des informations vérifiées sur les images, les vidéos et leurs métadonnées pourraient faciliter la vérification des faits.
- » **Normes et protocoles de vérification adaptés** : L'élaboration de lignes directrices spécifiques pour la vérification des contenus visuels permettrait d'har-

moniser les pratiques et d'améliorer l'efficacité des journalistes face à la désinformation.

- » **Soutien institutionnel et collaboratif** : Une meilleure collaboration entre les rédactions, les plateformes numériques et les organisations de vérification pourrait renforcer la lutte contre la désinformation visuelle.

2.2 La politique et les institutions

Les institutions politiques sont aussi grandement affectées par la désinformation visuelle et médiatique (DVM). Les rapports du Department of Homeland Security américain et de l'Australian Strategic Policy Institute (ASPI) ont tiré la sonnette d'alarme sur l'émergence des deepfakes, soulignant leur capacité à saper la confiance publique et à perturber les processus démocratiques. Ces technologies, désormais accessibles à un large éventail d'acteurs, permettent de manipuler des images, vidéos ou enregistrements audio de manière quasi indétectable, exposant ainsi les institutions à des campagnes de désinformation massives et ciblées.

En matière de propagande, les deepfakes permettent de produire rapidement et à moindre coût un volume important de contenus manipulateurs, souvent « assez bons » pour tromper une partie significative de la population. Ces outils amplifient les opérations d'influence, en propageant des récits divisifs et en exploitant les biais cognitifs du public pour polariser les sociétés.

Par ailleurs, les deepfakes contribuent à une érosion généralisée de la confiance, non seulement dans les institutions démocratiques, mais aussi dans les médias traditionnels et les figures d'autorité.

Le « dividende du menteur » aggrave cette dynamique en permettant à des acteurs malveillants de discréditer des preuves réelles simplement en invoquant l'existence potentielle de deepfakes. Face à ces défis, les rapports appellent à l'urgence de développer des technologies de détection robustes, de renforcer les normes d'authentification numérique et de sensibiliser le public pour limiter les impacts destructeurs des deepfakes sur la gouvernance et la stabilité institutionnelles.

Selon le U.S. Intelligence Community (Appel & Prietzel, 2022), il est admis que l'incidence de DVM aura

des impacts significatifs sur la géopolitique et voici quelques extraits pertinents de l'article :

- » **Déstabilisation des démocraties** : « Les deepfakes pourraient devenir un facteur perturbateur dans de nombreux domaines, notamment dans le domaine politique, où les enregistrements audio et vidéo ont souvent été décisifs pour le destin des politiciens et des gouvernements ».
- » **Érosion de la confiance** : « L'avènement de la technologie des deepfakes a marqué un tournant dans la capacité à déformer la réalité [...]. Alors que les vidéos étaient autrefois considérées comme des preuves irréfutables, nous sommes désormais confrontés à de nouvelles générations de logiciels utilisant des algorithmes d'apprentissage profond pour manipuler le contenu audiovisuel ».
- » **Arme dans les conflits internationaux** : « L'émergence du deepfake de Zelenski en 2022 a démontré l'importance actuelle de ce sujet ».
- » **Menaces sur les relations diplomatiques** : « Les deepfakes pourraient contribuer au problème plus large des fausses informations sur les réseaux sociaux ('des informations fabriquées qui imitent le contenu des médias d'information en termes de forme, mais pas en termes de processus organisationnel ou d'intention') ».

Défis en matière de responsabilité et de preuves : « Les deepfakes peuvent passer inaperçus, même si des informations pertinentes (contexte, contenu ou anomalies techniques) sont disponibles pour leur détection ». Un exemple marquant est la publicité¹⁹ diffusée en avril 2023 par le Comité national républicain (RNC), entièrement générée à l'aide de l'intelligence artificielle (IA), en réaction à l'annonce de la candidature de Joe Biden pour l'élection présidentielle de 2024. Cette vidéo projette une vision dystopique des États-Unis sous une éventuelle réélection de Biden, illustrant des explosions fictives à Taïwan, des opérations militaires dans les rues américaines, ainsi que des immeubles désertés à Wall Street.

Cette utilisation de l'IA dans les publicités politiques a suscité des préoccupations quant à la diffusion de contenus trompeurs et à l'impact potentiel sur les électeurs. En réponse, des discussions ont émergé aux États-Unis concernant la nécessité de réglementer ces pratiques. En mai 2024, le président de la Commission fédérale des communications (FCC) a proposé d'obliger les annonceurs politiques à signaler clairement l'utilisation de contenus générés par l'IA dans les publicités télévisées et radiophoniques, afin d'assurer une transparence accrue envers les électeurs.

2.3 La santé et ses institutions

La mésinformation est abondante dans le domaine de la santé²⁰, avec de sérieuses conséquences sur la santé des citoyens. Elle peut entraîner une baisse de la couverture vaccinale, favorisant ainsi la réémergence de maladies évitables comme la rougeole. De plus, les messages erronés sur les traitements ou la gravité de certaines maladies peuvent provoquer des comportements à risque, tels que le refus de traitements éprouvés ou l'adhésion à des remèdes non validés.

Psychologiquement, la mésinformation engendre peur, anxiété et méfiance envers les institutions de santé publique. Cette méfiance compromet les campagnes d'information et complique la mise en place de mesures sanitaires efficaces. Sur le plan social, elle alimente les chambres d'écho en ligne et polarise davantage les communautés, rendant le consensus sur les politiques de santé difficile à atteindre.

Enfin, les conséquences de la mésinformation s'étendent à l'érosion de la confiance dans les experts et les autorités scientifiques, affaiblissant ainsi la capacité des gouvernements à gérer des crises sanitaires.

La pandémie de COVID-19 et sa gestion par les différents ordres de gouvernements ont entraîné l'apparition d'un volume important de matériel informationnel d'origine douteuse, abondamment partagé sur les réseaux sociaux. Inévitablement, une partie de cette désinformation comporte un aspect visuel, comme l'ont rapporté Brennen et al. (2020). Ces auteurs ont mis en évidence six cadres principaux utilisés dans la désinformation visuelle :

¹⁹ <https://youtu.be/kLMMxgtxQ1Y>

²⁰ <https://www.sciencedirect.com/science/article/pii/S0277953619305465>

- » **Agence autoritaire** : Ce cadre concerne les affirmations sur les actions ou inactions des autorités publiques, telles que les gouvernements ou les organisations internationales. Ces affirmations peuvent exagérer ou minimiser le rôle et les efforts de ces entités. Les visuels associés imitent souvent des documents officiels ou des logos pour renforcer leur crédibilité.
- » **Virulence** : Ce cadre déforme la propagation ou la gravité du virus, soit en exagérant le danger, soit en suggérant que le virus n'existe pas. Les visuels servent souvent de preuves prétendues montrant une large contamination ou banalisant les risques réels.
- » **Efficacité médicale** : Ce cadre met en avant l'existence supposée de traitements, de cures ou de mesures préventives contre la COVID-19. Il peut promouvoir faussement des thérapies ou des produits comme solutions efficaces. Les visuels utilisés représentent fréquemment des percées médicales ou des équipements pour appuyer ces affirmations.
- » **Intolérance** : Ce cadre véhicule des messages xénophobes, racistes ou partisans, en ciblant souvent des groupes ou des individus spécifiques comme responsables de la pandémie. Les visuels renforcent ces affirmations, soit par des images explicites, soit par des indices subtils.
- » **Prophétie** : Les affirmations dans ce cadre suggèrent que la pandémie avait été prédite auparavant, la présentant comme un événement prophétisé. Les visuels dans ce contexte font souvent référence à d'anciens documents, livres ou captures d'écran manipulées pour donner une impression de prévision.
- » **Satire** : Ce cadre comprend des contenus satiriques ou humoristiques qui peuvent néanmoins être pris au sérieux par certains publics. Les visuels sont souvent exagérés ou intentionnellement absurdes, mais peuvent induire en erreur si le contexte satirique n'est pas clair.

Les visuels dans la désinformation liée à la COVID-19 jouent trois rôles principaux:

- » **Illustrer et mettre sélectivement en avant** : Les visuels mettent en évidence certains éléments de la désinformation pour rendre certains aspects plus saillants tout en minimisant d'autres. Cela peut inclure

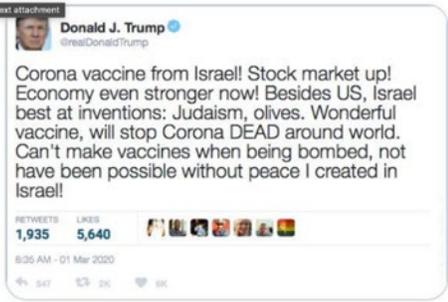
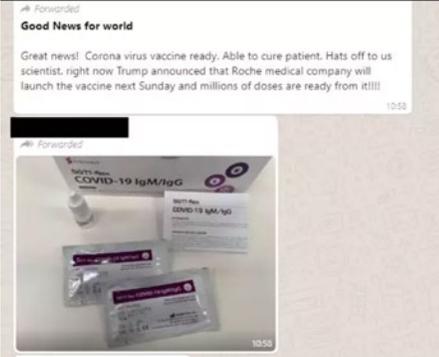
l'utilisation d'images pour attirer l'attention sur des points clés ou rendre la désinformation plus mémorable et engageante.

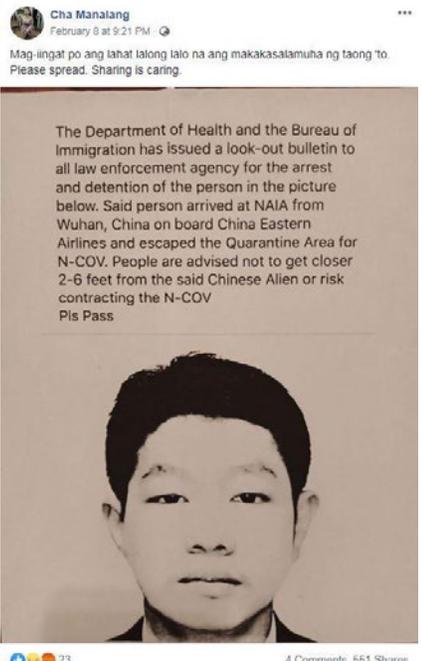
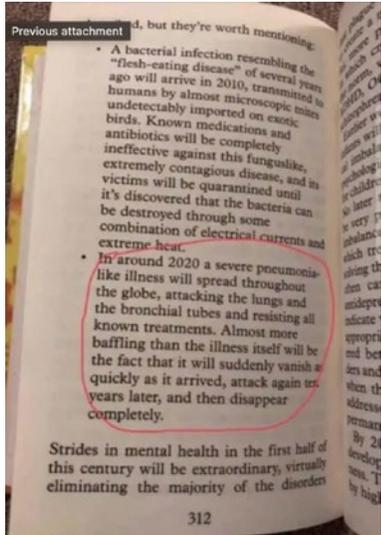
- » **Servir de preuve** : Les visuels sont souvent présentés comme des preuves pour étayer des affirmations fausses. Cela comprend des images ou des vidéos étiquetées pour correspondre au récit, même lorsqu'elles sont sorties de leur contexte ou mal identifiées, donnant ainsi une apparence d'authenticité et de crédibilité.
- » **Imiter les autorités** : Les visuels imitent des documents officiels, des logos ou des formats pour créer une illusion de crédibilité. En apparaissant comme provenant de sources autorisées, ces visuels visent à légitimer la désinformation, augmentant sa fiabilité perçue.

La majorité des visuels ne sont pas manipulés, mais mal étiquetés, ce qui leur confère une apparence d'authenticité trompeuse. Ces observations soulignent l'importance des visuels dans la diffusion de fausses informations et la nécessité d'une vigilance accrue face à leur rôle dans l'amplification des messages trompeurs.



Le tableau suivant donne quelques exemples identifiants les cadres les plus courants présents dans le corpus de la recherche des auteurs, et indique la prévalence de chacun dans leur ensemble de données.

Cadre	Description	Prévalence (%)	
Agence autoritaire	Affirmations valorisées sur les actions des autorités publiques	40	 <p>8:41 AM · Mar 1, 2020 · Twitter Web App</p>
Virulence	Exagère ou minimise la propagation du virus, ou prétend que la maladie n'est pas réelle	33	 <p>31 comments 268 shares</p>
Efficacité médicale	Fournit des informations médicales, met en avant tests, vaccins, traitements ou préventifs	29	

Cadre	Description	Prévalence (%)	
Intolérance	Exprime le racisme, la xénophobie ou le sexisme	15	
Prophétie	Suggère que le virus a été prédit auparavant	10	
Satire	Contenu satirique ou humoristique	6	

2.4 Économie, finance, commerce

Le secteur financier et commercial peut également être impacté par la désinformation. Un solide rapport de la Federal Trade Commission américaine (FTC, 2022)²¹ fait état des divers types de menaces en ligne qui planent sur les consommateurs de ce pays et des stratégies possibles pour les combattre. Les principales catégories de menaces en ligne identifiées dans le rapport sont :

- » **Contenu trompeur et frauduleux** : Inclut les escroqueries ou les contenus frauduleux destinés à tromper ou à nuire aux individus.
- » **Contenu manipulé** : Tels que les vidéos deepfake et les faux avis individuels visant à induire les personnes en erreur.
- » **Interfaces trompeuses et exploitantes** : Interfaces de sites web ou d'applications mobiles conçues pour tromper ou exploiter intentionnellement les utilisateurs.
- » **Contenu illégal** : Inclut la vente illégale d'opioïdes, l'exploitation et l'abus sexuels des enfants, la pornographie vengeresse, le harcèlement, la cyberintimidation, les crimes haineux, la glorification de la violence et l'incitation à la violence.
- » **Abus par les terroristes et extrémistes violents** : Utilisation par les terroristes des plateformes numériques pour se promouvoir, diffuser de la propagande et glorifier des actes de violence.
- » **Campagnes de désinformation** : Coordonnées par des comptes inauthentiques ou des individus pour influencer les élections ou l'opinion publique.
- » **Vente de produits contrefaits** : Y compris l'utilisation de plateformes numériques pour la distribution de produits contrefaits

Les fraudeurs peuvent créer du contenu visuel manipulé afin de tromper les utilisateurs de plusieurs manières :

- » **Deepfakes** : Ils utilisent les nombreuses techniques d'intelligence artificielle pour créer des vidéos, images, ou fichiers audio qui semblent authentiques, mais

qui ont été manipulés pour se faire passer pour une autre personne. Cela permet d'impliquer faussement des individus dans des actions ou discours qu'ils n'ont pas commis.

- » **Faux avis et recommandations** : Des images et vidéos peuvent être créées et manipulées pour donner l'impression qu'un produit ou un service a été approuvé par des utilisateurs légitimes. Ces faux avis sont souvent destinés à attirer de nouveaux clients de manière frauduleuse.
- » **Attaques à la réputation** : Il existe plusieurs cas documentés de vidéos créées pour nuire à des entreprises. Par exemple, un incident impliquant Tesla²², où une fausse information alimentant des débats sur la sécurité des voitures autonomes a temporairement affecté le cours de l'action de l'entreprise. Un cas majeur de fraude fiscale en Chine²³ a exposé des failles dans les systèmes de sécurité des données et de reconnaissance faciale : des identités synthétiques créées à partir d'images faciales acquises sur le marché noir ont permis de contourner les contrôles biométriques du système fiscal chinois et de créer une société écran, laquelle a émis des factures frauduleuses pour un montant total de 500 millions de yuans (environ 76,2 millions USD).

L'analyse de Bateman (2020) sur les risques potentiels liés à l'utilisation malveillante de la DVM dans le secteur financier identifie dix scénarios plausibles où ces technologies pourraient causer des dommages financiers, tels que des communications frauduleuses, la manipulation des marchés et le vol d'identité. Le tableau suivant est extrait du rapport qui donne une revue des différents scénarios avec les cibles (individus, compagnies, le marché, les organismes de régulations) et la manière dont la DVM pourrait être employée.

Malgré les craintes, il n'a pas été démontré, pour l'instant, que la DVM pouvait porter atteinte aux systèmes financiers mondiaux. En revanche, elle pourrait avoir des conséquences importantes pour des organisations affaiblies par une crise de relations publiques, ou encore des pays dont les institutions financières sont fragiles.

²¹ <https://www.ftc.gov/reports/combating-online-harms-through-innovation>

²² <https://core.ac.uk/download/pdf/276953172.pdf>

²³ <https://findbiometrics.com/fraudsters-use-deepfake-biometrics-hack-chinas-taxation-system-040103/>

Ten Synthetic Media Scenarios for Financial Harm

Target	Scenario	Role of Synthetic Media	Key Malicious Technique
Individuals	1. Identity theft	Voice cloning or face-swap video is used to impersonate a wealthy individual and initiate fraudulent transactions. Alternatively, it is used to impersonate a corporate officer and gain access to databases of personal information, which can enable larger-scale identity theft.	
	2. Imposter scam	Voice cloning or face-swap video is used to impersonate a trusted government official or family member of the victim and coerce a fraudulent payment.	
	3. Cyber extortion	Synthetic pornography of the victim is used for blackmail.	
Companies	4. Payment fraud	Voice cloning or face-swap video is used to impersonate a corporate officer and initiate fraudulent transactions.	
	5. Stock manipulation via fabricated events	Voice cloning or face-swap video is used to defame a corporate leader or falsify a product endorsement, which can alter investor sentiment.	
	6. Stock manipulation via bots	Synthetic photos and text are used to construct human-like social media bots that attack or promote a brand, which can alter investor perception of consumer sentiment.	
	7. Malicious bank run	Synthetic photos and text are used to construct human-like social media bots that spread false rumors of bank weakness, which can fuel runs on cash.	
Markets	8. Malicious flash crash	Voice cloning or face-swap video is used to fabricate a market-moving event.	
Regulatory Structures	9. Fabricated government action	Voice cloning or face-swap video is used to fabricate an imminent interest rate change, policy shift, or enforcement action.	
	10. Regulatory astroturfing	Synthetic text is used to fabricate comments from the public on proposed financial regulations, which can manipulate the rulemaking process.	


Deepfake voice phishing


Fabricated private remarks


Synthetic social botnet

Narrowcast

Broadcast

Figure 2 - Désinformation et finance. Tiré de Bateman (2020)²⁴

²⁴ <https://carnegieendowment.org/research/2020/07/deepfakes-and-synthetic-media-in-the-financial-system-assessing-threat-scenarios?lang=en>

2.5 Droits et sécurité des personnes

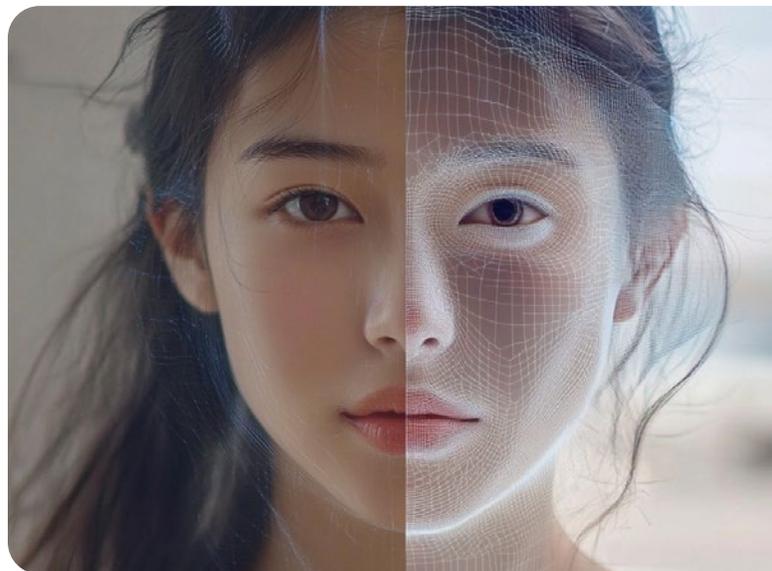
Certains groupes de citoyens sont particulièrement ciblés par la désinformation et les deepfakes. Parmi les premières victimes figurent les femmes, notamment les politiciennes. Ces dernières sont souvent visées par des contenus explicites non consensuels et des attaques politiques visant à ternir leur réputation et à saper leur carrière. Un rapport de Sensity AI de 2019 révèle que 96 % des vidéos deepfake étaient des contenus explicites non consensuels, dont 99 % ciblaient des femmes²⁵ et une proportion significative ciblait des individus LGBTQIA+²⁶. Par ailleurs, des vidéos deepfake ont été utilisées pour placer des femmes politiques dans des situations compromettantes, comme ce fut le cas en 2020 lorsqu'une politicienne européenne a été faussement représentée dans un contexte scandaleux.²⁷ Ces attaques entraînent des effets psychologiques graves, notamment du harcèlement et une dissuasion des femmes à participer à la vie politique.²⁸

Les individus LGBTQIA+ constituent également une cible privilégiée des deepfakes. Ces technologies sont souvent exploitées pour produire des contenus explicites non consensuels, intensifiant le harcèlement et les discriminations. Ces attaques ont des répercussions psychologiques importantes, entraînant une détresse, des dommages à la réputation, et dissuadant les victimes de s'exprimer librement.²⁹

Les militants politiques et les dissidents sont également visés. En 2023, la Chine a utilisé des vidéos deepfake dans le cadre de la campagne de « Spamouflage » pour discréditer des critiques à l'étranger, comme le dissident Liu Xin, faussement accusé de calomnier des politiciens canadiens.³⁰ Les journalistes ne sont pas épargnés : des femmes journalistes travaillant pour des médias étrangers ont été harcelées par deepfake, une tactique visant à les intimider et à réduire leurs voix critiques au silence.³¹

Enfin, les minorités ethniques sont aussi touchées. Lors de la crise des Rohingyas en 2018, des vidéos deepfake ont été utilisées pour attiser la violence contre cette minorité musulmane en Birmanie, aggravant les tensions ethniques.³² Lors des élections de 2018 au Kenya, des vidéos deepfake ont été diffusées dans le but de manipuler l'opinion publique. Ces vidéos, en particulier une montrant faussement un candidat présidentiel en mauvaise santé, auraient circulé sur des plateformes comme Twitter, WhatsApp et Facebook, servant d'outils de propagande politique.³³

Ces exemples montrent l'urgence de mettre en place des cadres juridiques et des solutions technologiques robustes pour lutter contre l'utilisation malveillante des deepfakes. Les impacts psychologiques, sociaux et politiques sont profonds, soulignant l'importance d'une vigilance accrue face à cette menace croissante.



²⁵ <https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/>

²⁶ <https://blogs.iadb.org/igualdad/en/deepfakes-gender-based-violence-in-the-era-of-artificial-intelligence/>

²⁷ <https://www.euronews.com/next/2023/12/11/violating-and-dehumanising-how-ai-deepfakes-are-being-used-to-target-women>

²⁸ <https://www.techpolicy.press/deepfakes-and-elections-the-risk-to-womens-political-participation/>

²⁹ <https://blogs.iadb.org/igualdad/en/deepfakes-gender-based-violence-in-the-era-of-artificial-intelligence/>

³⁰ <https://www.ctvnews.ca/canada/china-critic-says-he-s-the-target-of-deepfake-spamouflage-attack-by-beijing-1.6625954>

³¹ <https://www.womeninjournalism.org/infocus-all/from-trolling-to-deepfakes-the-online-war-against-women-journalists>

³² <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>

³³ <https://link.springer.com/article/10.1007/s42454-024-00054-8>

3. DVM ET CRISE CLIMATIQUE

À retenir :

- » La désinformation visuelle et multimodale est particulièrement vive dans les flux de communication sur les changements climatiques, pour nier leur existence, pour réfuter la cause (l'humain) ou pour en minimiser l'impact.
- » La DVM apparaît sous plusieurs formes : mêmes internet, manipulations de graphiques ou d'images, vidéos fallacieuses, etc.
- » Combattre la désinformation climatique avec une approche multidisciplinaire devrait être d'une grande priorité.

La crise climatique constitue un enjeu global aux dimensions multiples, alliant des défis environnementaux, politiques, socio-économiques, technologiques et informationnels. Face à l'urgence croissante de cette crise, les efforts pour promouvoir des actions collectives et éclairées se heurtent à une dynamique complexe de désinformation, de mésinformation et de scepticisme public, notamment à l'égard des politiques climatiques en cours d'élaboration. Cette section explore non seulement comment le changement climatique est à la fois une réalité scientifique et un terrain de bataille pour les récits conflictuels qui influencent la perception publique et les actions politiques, mais également comment ces récits évoluent dans un écosystème médiatique en pleine mutation, marqué par l'essor et l'utilisation croissante de nouvelles technologies et supports numériques.

3.1 Un changement climatique sans précédent

Bien que le climat terrestre ait toujours connu des fluctuations naturelles, le réchauffement actuel est unique par sa rapidité et son origine anthropique. Depuis la révolution industrielle, des activités humaines telles que la combustion des combustibles fossiles, la déforestation et l'agriculture intensive ont entraîné une augmentation marquée des gaz à effet de serre (GES) dans l'atmosphère, provoquant une hausse accélérée

des températures mondiales. Cette dynamique se traduit par des conséquences sans précédent sur l'environnement et les sociétés humaines:

- » L'année 2024 a été marquée par des températures inédites, avec des pics atteignant 52°C au Mexique et un record mondial de température moyenne journalière de 17,16°C enregistré le 22 juillet (Copernicus, 2024). Ces records s'inscrivent dans une tendance globale où les huit années les plus chaudes jamais enregistrées se situent toutes après 2015.
- » Depuis le début du suivi par satellite en 1979, la banquise arctique a perdu près de 40 % de sa superficie, illustrant des bouleversements régionaux majeurs. Parallèlement, la fonte des glaciers et des calottes glaciaires au Groenland et en Antarctique, combinée à la dilatation thermique des océans, a contribué à une élévation totale du niveau des mers de 20 cm depuis 1900 (IPCC, 2021).
- » La perte rapide de biodiversité menace gravement les écosystèmes terrestres et marins. L'acidification des océans—conséquence directe de l'absorption du CO₂ atmosphérique—perturbe les chaînes alimentaires marines et accélère la dégradation des récifs coralliens, dont dépendent près de 25 % des espèces marines pour leur habitat (NOAA, 2020).

Ces impacts, largement documentés grâce à des technologies avancées comme les satellites, les modèles climatiques et les relevés océanographiques, sont également corroborés par des données historiques telles que l'analyse des carottes glaciaires, qui permettent de reconstituer l'évolution des GES sur des centaines de milliers d'années, ou l'étude des anneaux de croissance des arbres, qui révèlent les variations climatiques locales sur plusieurs siècles.

Pourtant, malgré l'ampleur et la fiabilité des preuves scientifiques, des récits contradictoires continuent de semer le doute, d'affaiblir la compréhension publique et d'entraver la mobilisation collective requise pour répondre à cette crise.

3.2 L'émergence et l'expansion de la désinformation climatique

La désinformation climatique prend de multiples formes, s'adaptant aux contextes et divisions politiques, culturelles et économiques, et se diffuse à travers divers médiums, notamment sous des formats visuels et multimodaux.

Sur les réseaux sociaux, cette désinformation est particulièrement visible et influente. Les algorithmes des grandes plateformes, optimisés pour maximiser l'engagement, favorisent la viralité de contenus polarisants et sensationnalistes, y compris ceux qui véhiculent des récits trompeurs sur le climat. Cette dynamique est exacerbée par le fait qu'un nombre croissant de personnes – en particulier parmi les jeunes générations – s'informent principalement via les médias sociaux (Statistique Canada, 2023). Dans un tel environnement, où la distinction entre information fiable et contenu manipulateur reste souvent floue, les récits climatosceptiques trouvent un écho d'autant plus puissant.

Parmi les principaux types de désinformation climatique recensés aujourd'hui, on trouve :

- » La négation du réchauffement climatique, qui tente de présenter les changements climatiques comme des phénomènes purement naturels ou de minimiser leur gravité en les qualifiant d'exagérations alarmistes.
- » Les doutes sur l'attribution, qui remettent en question la responsabilité humaine dans le réchauffement climatique. Ces récits mettent en avant des facteurs naturels, comme l'activité solaire ou les oscillations océaniques, pour détourner l'attention des émissions anthropiques de gaz à effet de serre, pourtant identifiées par la quasi-totalité des scientifiques comme la cause principale.
- » La minimisation des impacts, qui prétend que les effets du changement climatique seront modérés, localisés ou même positifs. Par exemple, certains récits suggèrent que le réchauffement global pourrait allonger les saisons agricoles ou réduire les coûts de chauffage en hiver.

- » La contestation du consensus scientifique, qui vise à discréditer les conclusions du GIEC et d'autres organismes scientifiques en les qualifiant de biaisées, politisées ou corrompues.
- » La critique des solutions climatiques, qui cible les énergies renouvelables et les politiques de réduction des émissions, en les présentant comme inefficaces, nuisibles à l'économie ou même liberticides.

Selon un rapport de 2024 du Center for Countering Digital Hate (CCDH), 70 % des récits de déni climatique recensés sur YouTube visent aujourd'hui à saper la confiance dans les solutions climatiques et les institutions scientifiques, contre 35 % il y a six ans.³⁴ Ce glissement refléterait une stratégie renouvelée de désinformation climatique, moins axée sur la négation du réchauffement et davantage sur l'entrave aux efforts pour y répondre.

Historiquement, ces stratégies ont souvent été soutenues par des industries influentes, notamment celle des combustibles fossiles. Des entreprises comme ExxonMobil ont financé des campagnes et des recherches visant à semer le doute sur le changement climatique, tout en reconnaissant en interne les risques qu'il représentait (Powell, 2021).

Une enquête du Guardian (2021) a révélé des publicités climatosceptiques des années 1980 et 1990 qui s'appuyaient fortement sur des visuels rassurants : paysages bucoliques, images d'industries propres et slogans affirmant que les émissions de CO₂ étaient inoffensives, voire bénéfiques pour la vie.³⁵ Cherchant à influencer les politiques publiques et l'opinion, comme l'ont documenté Oreskes et Conway dans *Merchants of Doubt* (2010), ces mécanismes de désinformation ont évolué avec l'ère numérique. Les visuels manipulateurs, désormais amplifiés par les réseaux sociaux et les technologies comme l'IA générative, restent au cœur de ces récits trompeurs, augmentant leur viralité et leur pouvoir de persuasion.

³⁴ <https://counterhate.com/research/new-climate-denial/>

³⁵ <https://www.theguardian.com/environment/2021/nov/18/the-forgotten-oil-ads-that-told-us-climate-change-was-nothing>

3.3 Cas d'étude

Aujourd'hui, la désinformation visuelle et multimodale dans le contexte climatique apparaît sous plusieurs formes : memes internet, manipulations de graphiques ou d'images, vidéos fallacieuses, etc.

» La figure ci-dessous montre un exemple de recontextualisation d'image avec une carte météorologique réelle où le rouge dominant, commun en présentation météo, transmet pourtant un message visuel plutôt alarmiste pouvant être perçu comme preuve non factuelle du réchauffement climatique. Ce contenu a été partagé en ligne par des climatosceptiques prétendant que « en 1986, c'était un été normal. Aujourd'hui, ils colorent la carte en rouge et l'appellent une chaleur extrême ». Ce type de communication aurait un impact visuel négatif dans le même ordre que les fausses affirmations des négationnistes (Guilyardi, 2015). L'analyse de Fish et Kreitzberg (2023)³⁶ va dans le même sens : le mouvement conservateur de déni des changements climatiques a tendance à se réapproprier des cartes publiées dans des journaux scientifiques crédibles, mais à en détourner le sens. Bref, la communication visuelle est une manière forte d'engager positivement le public, et elle nécessite une prudence particulière en ce qui concerne le changement climatique (Lewandowsky, 2021; Culloty, 2019).



Figure 3 - Gauche : paire de cartes météo alarmiste prétendument datées de 1086 et 2022 (AFP, 2023; EDMO, 2023). Droite : même liant Greta Thunberg et l'exploitation d'enfants dans des mines de cobalt (EDMO, 2023).

- » Un autre cas de désinformation climatique est rapporté par Langguth (2021) qui fait état d'un deepfake d'une intervention de la première ministre belge Sophie Wilmès se voyant attribuer un message liant COVID-19 et changement climatique. Bien que la vidéo postée ait été accompagnée d'une mise en garde quant à son caractère synthétique, certains internautes y ont cru et ont interagi en ce sens dans le fil de discussion du post.³⁷
- » L'étude « Faking Sandy » rapporte un autre cas de propagation de mésinformation autour du climat via les médias sociaux, et en particulier Twitter (aujourd'hui X) (Gupta, 2013), impliquant de fausses images de l'ouragan Sandy de 2012 (ci-dessous).



Figure 4 - Exemples de fabricated Hurricane Sandy images shared on Twitter.

- » Une analyse de vidéos YouTube démontre qu'une majeure partie du contenu de l'échantillon étudié réfute le consensus sur la science du climat (Allgaier, 2019). L'analyse montre ainsi comment la plateforme facilite la propagation de fausses informations sur le réchauffement climatique, même si elle ajoute un descriptif contextuel vis-à-vis de la vidéo à partir de sources sûres (YouTube, 2023).

³⁶ <https://www.tandfonline.com/doi/full/10.1080/24694452.2023.2227672#d1e191>

³⁷ <https://blogs.mediapart.fr/geraldholubowicz/blog/150420/extinction-rebellion-s-empare-des-deepfakes-en-belgique>

» Des résultats d'une étude de mêmes internet sur les changements climatiques montrent que ce type de message augmente l'engagement, mais n'influence pas nécessairement la perception du risque de l'individu (Zhang 2021). De façon similaire, dans une analyse de l'engagement en ligne relativement aux mêmes et aux commentaires sur la question du changement climatique, Kovacheva et al. constatent que, dans bien des cas, le point de vue des commentaires d'un post peut altérer le niveau d'engagement des individus, démontrant le pouvoir de l'anecdote et de l'image des mêmes (Kovacheva, 2022; Lewandowsky, 2021; Dupuis, 2020).

D'autres exemples sont rapportés dans un document³⁸ soigné sur l'IA et la désinformation climatique rendu public en 2023.

3.4 Discussion

Les technologies numériques et les outils d'intelligence artificielle transforment radicalement la manière dont la désinformation sur le climat est produite, amplifiée et consommée. Ces outils, de plus en plus puissants et omniprésents, permettent aux récits fallacieux de s'adapter, de se répandre et de gagner en crédibilité, rendant la lutte contre la désinformation plus complexe que jamais. En combinant des textes, images et vidéos générés, ces technologies offrent une capacité inédite de construire des campagnes de désinformation coordonnées, renforçant leur impact par la diversité et l'apparente légitimité des supports, et s'appuyant sur une crédibilité visuelle de plus en plus difficile à remettre en question.

Cette dynamique est exacerbée par un contexte systémique où certaines plateformes tolèrent, voire tirent profit, de cette désinformation. Une enquête de CBC³⁹ a révélé que YouTube génère des millions de dollars chaque année grâce à des chaînes diffusant des contenus climatosceptiques. Parallèlement, Global Witness a montré que TikTok, l'une des rares grandes plateformes à interdire explicitement la désinformation climatique, peine à faire respecter ses propres politiques, laissant

proliférer des contenus trompeurs, notamment dans les commentaires de vidéos liées à des événements climatiques majeurs comme la COP29.⁴⁰

Aujourd'hui, nombreux sont les acteurs qui soulignent l'urgence de développer des outils performants capables de détecter et de contrer les contenus manipulateurs, qu'ils soient visuels ou multimodaux. En parallèle, la mise en place de régulations visant à limiter la production et la diffusion de contenus trompeurs, ainsi que la promotion d'une éducation aux nouveaux médias, apparaissent comme des priorités pour renforcer la résilience face à ces nouvelles technologies et tactiques sophistiquées.

Cela dit, la lutte contre la désinformation climatique ne peut se limiter à des initiatives isolées. Elle nécessite un engagement collectif, ancré dans une vision à long terme, pour assurer autant que possible que les informations sur lesquelles reposent les décisions présentes et futures en matière d'action climatique soient fiables et accessibles. Ce défi appelle à la fois des innovations technologiques, une mobilisation citoyenne et l'élaboration de nouvelles régulations adaptées, pour faire de l'intégrité de l'information climatique une réalité partagée et constamment actualisée.



³⁸ <https://blogs.mediapart.fr/geraldholubowicz/blog/150420/extinction-rebellion-s-empare-des-deepfakes-en-belgique>

³⁹ <https://www.cbc.ca/news/climate/youtube-climate-denial-content-1.7085223>

⁴⁰ <https://www.globalwitness.org/en/campaigns/digital-threats/tiktok-fails-enforce-its-own-policies-climate-misinformation-during-cop29/>

4. URGENCE

À retenir :

- » La désinformation visuelle et multimodale a des propriétés particulières par rapport à la désinformation textuelle : images et vidéos créent une perception de vérité et réalité plus grande que le texte.
- » L'impact psychologique des images, films, etc. sur les individus est plus grand et plus durable.
- » La DVM circule plus facilement et rapidement, sans contrôle.
- » Difficile à détecter, autant par les algorithmes que par les humains.
- » Peu étudiée au niveau scientifique, nécessite de nouveaux outils pour la combattre.

Cette section soulève l'urgence d'agir pour combattre la DVM à grande échelle, de manière systématique et structurée. Voici une description des enjeux.

4.1 Impact plus grand du visuel par rapport au texte

L'impact accru de la présence de contenu visuel dans un message est bien documenté dans la littérature scientifique. Dans le contexte des réseaux sociaux, la dissémination de messages contenant une image ou une vidéo est plus grande: "Tweets with images get 18% more clicks, 89% more likes, and 150% more retweets than those without images." (Cao et al, 2020).

Une raison probable est liée au fait que l'interprétation de l'image exige moins d'effort de traitement comparativement à l'interprétation du texte qui, elle, doit composer avec des barrières linguistiques ou de littératie.

Une autre raison est rattachée à l'expérience sensorielle vécue au moment de l'observation d'une image : celle-ci a un impact plus grand et durable sur la perception de crédibilité, de réalité, et d'engagement reliée au message lui-même.⁴¹ On croit que cette perception

augmente selon la richesse de la représentation du message, laquelle augmente quand on passe du texte à l'image, ou de l'image à la vidéo. La perception de crédibilité plus grande d'un message multimodal par rapport à un message textuel a fait l'objet de plusieurs études (Dan et al., 2021). On peut aussi « mesurer » plus d'émotionnalité (Weikmann & Lecheler, 2022) et une plus grande valeur probante (greater evidentiary value). La plus grande émotionnalité rattachée au visuel n'est pas un détail anodin, puisqu'elle influence plus fortement le comportement des gens: dans un monde où l'information visuelle domine, la possibilité de sélectionner et de trafiquer les images est intrinsèquement liée à celle de manipuler les esprits.

Tous ces arguments combinés aux connaissances en communication visuelle sur les propriétés de l'image (ex. les images restent en tête plus longtemps à cause du picture superiority effect ; la vidéo a une capacité de réécriture des souvenirs (Weikmann & Lecheler, 2022), etc.) qui mettent en évidence le danger que constitue la DVM et militent en faveur d'un accroissement des recherches pour une meilleure compréhension du phénomène.

4.2 Propagation intensifiée

La désinformation visuelle circule plus rapidement, plus largement et de façon plus persistante que la désinformation textuelle, ce qui la rend particulièrement difficile à contenir et à corriger. Son impact repose en grande partie sur sa force émotionnelle et psychologique : les images et vidéos provoquent des réactions immédiates qui incitent au partage impulsif. Elles simplifient aussi des enjeux complexes, rendant certains récits plus accessibles, mémorables et convaincants que des explications textuelles plus détaillées.

Les algorithmes des plateformes numériques amplifient cette dynamique, privilégiant les contenus visuels engageants. Vidéos, mèmes et images atteignent une audience bien plus large que les publications textuelles,

⁴¹ « Le médium de la photographie est en tant que tel si digne de confiance, si "objectif", qu'il peut digérer plus de non-vérités, mentir davantage que n'importe quel autre médium avant lui » (Günther Anders dans *L'Obsolescence de l'homme* (1956), Paris, éditions Ivréa, 2002).

ce qui favorise la viralité de la DVM. Une fois qu'un visuel trompeur prend de l'ampleur sur une plateforme, il est souvent repris ailleurs, circulant de manière incontrôlable sur différentes plateformes et applications. Les messageries chiffrées comme WhatsApp et Telegram compliquent encore davantage la lutte contre la DVM en facilitant sa diffusion hors de portée des mécanismes de modération.

Un autre facteur clé est la difficulté de la vérification des contenus visuels. Contrairement aux textes, qui peuvent être analysés pour repérer des incohérences, les images et vidéos ne contiennent pas toujours de métadonnées ou d'indications claires sur leur origine, ce qui facilite leur manipulation et leur réinterprétation. Même lorsqu'un visuel trompeur est démenti, la correction génère rarement autant d'engagement que le contenu initial. Ce phénomène crée un cycle de désinformation auto-entretenu, où les fausses images et vidéos continuent de circuler et de façonner l'opinion publique bien après leur réfutation.

4.3 Détection difficile

Pour le citoyen ordinaire, l'identification de documents contenant de la désinformation est très difficile : "News consumers frequently overestimate (i.e. three in four Americans) their ability to correctly distinguish between legitimate information and misinformation i.e., legitimate and false news headlines" (Lyons et al., 2021). Au-delà de la consommation de nouvelles, on note une érosion de la confiance de la population en sa capacité à détecter la mésinformation : "Data from 2019 show that only 26% of US adults were very confident they could spot misinformation, dropping from 39% in 2016" (Languein, 2022).

La détection de la DVM est complexe. Ça ne dépend pas seulement du document lui-même, mais également du contexte dans lequel il a été produit, notamment l'intention de l'auteur au moment de sa création. Or, le processus mental visant à détecter la présence de DVM n'est pas intuitif (« les images ne mentent pas »), les citoyens sont donc mal équipés pour lutter par eux-mêmes contre la DVM. C'est d'autant plus vrai que les progrès récents en IA rendent la production de matériel de DVM

de plus en plus parfaite (voir par exemple l'article « The End of Photographic Truth »⁴²) et rendent les techniques plus classiques de détection de manipulations vidéo de plus en plus désuètes.

Sur une note plus technique, certaines plateformes comme WhatsApp encryptent leurs communications, ce qui complexifie la recherche de DVM par l'analyse des messages (ex: pour un système de modération automatisé).

4.4 Problème peu étudié

Malgré la forte présence de la désinformation visuelle et multimodale (DVM) sur les réseaux sociaux, son impact et les moyens de la contrer restent peu étudiés. Plusieurs chercheurs soulignent le nombre relativement faible d'études portant sur le rôle du contenu visuel dans la désinformation, ainsi que sur l'efficacité des stratégies de rétablissement des faits. Pourtant, ces formats jouent un rôle majeur dans la diffusion et la crédibilité des fausses informations.

Les défis méthodologiques liés à l'analyse des données multimodales ont conduit de nombreuses études sur la désinformation à privilégier l'analyse du texte au détriment du contenu visuel. La diversité des formats et la complexité des manipulations d'images et de vidéos compliquent encore davantage leur étude. Ainsi, on trouve peu de recherches sur la DVM en lien avec la COVID-19 (Brennen et al., 2020), alors que le contenu visuel représente une composante importante des campagnes de désinformation en général (Cao et al., 2020, qui parlent de « new challenge »).

Le manque d'études sur la DVM entraîne plusieurs conséquences:

» **Formats mal définis et absence de catégorisation claire:** Certains formats de DVM, comme les deepfakes, attirent l'attention en raison de leur qualité croissante (Nightingale & Farid, 2021). Toutefois, d'autres formes de manipulation visuelle, comme l'altération d'infographies et de visualisations de données (ex. modification de l'échelle d'un graphique), reçoivent beaucoup moins d'attention malgré leur impact significatif sur la perception publique (Weikmann & Lecheler, 2022).

⁴² <https://www.politico.eu/article/ai-photography-machine-learning-technology-disinformation-midjourney-dall-e3-stable-diffusion/>

» **Manque d'outils adaptés pour le fact-checking:**

Les outils de vérification des faits sont majoritairement conçus pour analyser du texte et restent peu adaptés à la désinformation multimodale.⁴³ De plus, l'efficacité des stratégies correctives comme l'étiquetage ou la recontextualisation des contenus visuels est encore mal évaluée. Par ailleurs, l'essor des contenus générés par l'IA⁴⁴ aggrave le problème, rendant les processus de vérification traditionnels obsolètes et inadaptés à l'échelle du phénomène.

» **Manque d'expertise interdisciplinaire:** La recherche sur la désinformation est dominée par les domaines de l'informatique (détection algorithmique) et du droit (cadres réglementaires), avec une faible implication des sciences sociales et humaines. Contrairement aux recherches sur la désinformation textuelle, qui bénéficient d'apports de sociologues, anthropologues et psychologues, l'étude de la DVM reste cloisonnée dans des approches techniques et juridiques, limitant ainsi une compréhension globale du phénomène.

De manière générale, les effets spécifiques du contenu visuel sur la désinformation sont encore mal compris, et il manque à la fois des outils et des cadres théoriques adaptés à leur étude. Cependant, une prise de conscience émerge progressivement. Par exemple, l'étude de Hamelaers et al. (2020)⁴⁵ a comparé l'impact des messages de désinformation textuelle et multimodale sur la perception de crédibilité et l'engagement du public américain. Ce travail a été cité dans des revues associées à des disciplines variées : éducation⁴⁶, géographie⁴⁷, sociologie, communication et journalisme, témoignant d'un début de diversification des approches pour mieux comprendre la DVM.

4.5 Nouvelle vague d'outils pour la création de contenus

Depuis quelques années, plusieurs techniques ont révolutionné l'intelligence artificielle pour ce qui est du traitement du langage. Ensuite, l'ingénierie autour de ces techniques s'est développée de façon accélérée

pour la rendre concrète. Notons par exemple les Transformers (Google, 2017), Bert (Google 2018), GTP2 (OpenAI 2019), GPT3 (OpenAI 2020) et ChatGPT (OpenAI 2022).

En 2024, plusieurs d'entre elles sont disponibles à tous, gratuitement ou à peu de frais, de manière de plus en plus efficace, et elles transforment la société à plusieurs niveaux, dans des échelles difficiles à évaluer et anticiper, créant à la fois des possibilités et des menaces inconnues jusqu'ici.

Suite à ces percées autour du langage, une autre vague de techniques combine la puissance de ces grands modèles de langues avec d'autres modalités, pour la production et la manipulation des images et plus récemment des vidéos. À titre d'exemples : DALL-E 3 (OpenAI, 2022, images), GPT4 (OpenAI, texte et images), Grok et Flux (pour les images non censurées), Gen-3 Alpha (RunwayAI, vidéos), Kling, MiniMax, et Vidu, trois outils de génération de vidéos chinois sans filtres. En plus, depuis la sortie de Llama par Meta et Stable Diffusion, deux modèles libres de droits (open source), on voit que la création de contenus par l'IA passe à un autre niveau. Il est maintenant possible pour le simple citoyen d'adapter ces modèles avec ses propres données (possiblement sensibles), une flexibilité dont la portée et les conséquences peuvent être significatives.

Tous ces outils qui apparaissent et s'améliorent à un rythme effréné transforment la technologie, mais aussi la société en général, et ils ont des conséquences profondes et grandissantes, quant à la multiplication et le raffinement de la désinformation. Parce que l'IA générative est un sujet incontournable dans la lutte contre la désinformation, elle fera l'objet d'une section entière de ce rapport (section 8).

⁴³ Cependant, la recherche académique s'accélère sur ce sujet, comme le montre l'état de l'art de Mubashara et al., (2023).

⁴⁴ <https://reutersinstitute.politics.ox.ac.uk/news/will-ai-generated-images-create-new-crisis-fact-checkers-experts-are-not-so-sure>

⁴⁵ <https://www.tandfonline.com/doi/citedby/10.1080/10584609.2019.1674979?scroll=top&needAccess=true>

⁴⁶ <https://www.tandfonline.com/doi/full/10.1080/17404622.2023.2271548>

⁴⁷ <https://www.tandfonline.com/doi/full/10.1080/24694452.2023.2227672>

5. ACTEURS DANS LA LUTTE CONTRE LA DVM

À retenir :

- » On constate qu'il existe peu d'acteurs institutionnels importants dédiés à la lutte technologique contre la DVM. Les efforts semblent dispersés entre groupes universitaires généralistes.
- » Les acteurs impliqués incluent les gouvernements et leurs agences (sécurité nationale, militaire, etc.), les fédérations (ex. Union européenne) et les alliances (ex. OTAN).
- » Avec des budgets de plusieurs millions de dollars, l'armée américaine et l'Union européenne financent de grands projets visant à développer des technologies d'IA efficaces pour lutter contre la DVM. Des ONG et des instituts de recherche cherchent également à promouvoir les efforts de sensibilisation.
- » L'industrie des technologies de l'information contribue-t-elle assez à cette lutte?

Plusieurs types d'acteurs sont engagés dans la lutte contre la DVM et la désinformation en général, et cette section en dresse le portrait. Afin d'assurer une lutte efficace, Johnson (2019) souligne l'importance de bâtir entre ces acteurs des alliances nationales, mais qui peuvent aussi s'étendre au niveau mondial.

5.1 Au niveau politique

Les acteurs politiques impliqués dans la lutte à la DVM incluent les gouvernements de pays souverains, les fédérations, et les alliances/organisations multipartites. Les gouvernements, particulièrement ceux de nations démocratiques, s'intéressent activement à la question puisqu'ils sont eux-mêmes victimes de ces conséquences : perte de crédibilité des dirigeants et des institutions, méfiance envers les médias traditionnels, dégradation du tissu social, etc. Or, ils possèdent plusieurs leviers sur lesquels ils peuvent agir pour mesurer et potentiellement contrer la DVM :

- » **Sécurité** : Des agences comme le Department of Homeland Security (DHS) aux États-Unis ou le SCRS (Canada) voient la DVM comme une menace à la sécurité nationale et mènent des consultations pour documenter la situation. Soulignons la diffusion de deux rapports importants du DHS^{48 49} en 2021 et 2022 qui documentent très bien le phénomène des deepfakes.
- » **Militaire** : L'agence DARPA (USA) finance la recherche avancée sur la désinformation visuelle par ses programmes MediFor et SemaFor.
- » **Commercial** : La Federal Trade Commission (FTC) américaine est préoccupée par l'émergence de la DVM du point de vue commercial et financier, afin, par exemple, de prévenir la fraude et les manipulations boursières pouvant affecter négativement les consommateurs. Leur rapport⁵⁰ publié en juin 2022 porte sur l'utilisation de l'IA comme outil de détection des préjudices en ligne, avec une attention particulière aux conditions requises pour un déploiement réussi de ces outils.
- » **Politique** : L'arène politique est impactée directement par la DVM, alors que la circulation de deepfakes mettant en scène des politiciens en vue vise à les discréditer et, de façon plus générale, à miner le processus démocratique. En plus des interventions gouvernementales pour freiner la désinformation (ex. promulgations de lois visant les producteurs de deepfakes, financement de campagnes de littératie numérique en milieu scolaire, etc.), des institutions non partisans comme le Congressional Research Service⁵¹ ou le Service de recherche du Parlement européen⁵² vont aussi participer au débat.
- » **Géopolitique** : Les actions à caractère plus géopolitique sont aussi affectées par la DVM parce qu'elle réduit leur efficacité; des organismes supportés par du financement public ou privé et dont le mandat est de promouvoir la démocratie dans le monde (e.g. le

⁴⁸ https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

⁴⁹ <https://www.dhs.gov/sites/default/files/2022-10/AEP%20DeepFake%20PHASE2%20FINAL%20corrected20221006.pdf>

⁵⁰ <https://www.ftc.gov/reports/combating-online-harms-through-innovation>

⁵¹ <https://crsreports.congress.gov/product/pdf/IF/IF11333>

⁵² [https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU\(2021\)690039_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf)

site Countering Disinformation⁵³ financé par USAid, l'organisme Democracy Reporting International⁵⁴, WTC.), vont aussi s'impliquer dans la lutte contre la DVM via la production de documents de vulgarisation et de sensibilisation.

Les fédérations de pays prennent aussi les moyens pour combattre la désinformation sur plusieurs fronts. On pense par exemple à l'Union européenne qui finance à la fois de grands projets multipartenaires pour le développement d'outils de détection à la fine pointe technologique, et également des ONG impliquées plus directement sur le terrain. Aussi, le Conseil de l'Europe a financé une recherche⁵⁵ très importante qui non seulement propose un cadre conceptuel pour l'examen du « désordre d'information » (malinformation, mésinformation, désinformation), mais fait également une série de recommandations qui s'appliquent aussi à la DVM. L'OTAN aussi publie ses propres rapports⁵⁶ sur le phénomène de désinformation.

5.2 ONG

Plusieurs OSBL et organismes indépendants sont impliqués dans la lutte contre la désinformation par le monitoring des activités de désinformation, la circulation des résultats de recherches, la création d'outils de détection, etc. Parmi les organismes qui ont la DVM comme sujet d'intérêt, mentionnons :

- » **Partnership on AI** : Communauté regroupant l'industrie, la société civile, le milieu académique et les groupes médiatiques pour réfléchir à l'impact de l'IA sur la société. Le volet « AI & Media Integrity » touche directement à la lutte contre la DVM.
- » **Deeptrustalliance** : Coalition de parties prenantes impliquées dans la lutte contre les deepfakes et la DVM. Un des objectifs est de pousser pour l'adoption de normes technologiques et de bonnes pratiques dans cette lutte.

- » **Project Origin** : Regroupement de grands médias (BBC, CBC/Radio-Canada, etc.) qui a pour but de « créer un processus par lequel la provenance et l'intégrité du contenu peuvent être validés » (avec la technologie C2PA).

5.3 Groupes de recherche/instituts

Face au déferlement de désinformation sur les réseaux sociaux et à l'érosion de la confiance d'une partie de la population envers les médias traditionnels, plusieurs groupes de recherche, de chaires et d'instituts affiliés à des universités ont vu le jour. Le document en ligne "The Fight Against Disinformation in the U.S.: A Landscape Analysis | Shorenstein Center"⁵⁷ fait le portrait des principaux acteurs et sources de financement, en plus de décrire les efforts des plateformes commerciales pour contrer la désinformation. Parmi toutes ces entités, mentionnons :

- » **Nieman Lab (Harvard)**⁵⁸ : S'intéresse à l'innovation en journalisme (online reporting enterprise focused on the future of news and innovation). Plusieurs analyses ont été réalisées sur l'utilisation des modèles génératifs (texte, image) dans le contexte journalistique.
- » **Indiana University's Observatory on Social Media**⁵⁹ : Étudie, entre autres, la circulation des mêmes sur Internet, leur évolution dans le temps, etc.
- » **Center for an Informed Public (Université de Washington)** : Se concentre sur la désinformation, l'éducation aux médias et la confiance du public dans les systèmes d'information. Il collabore avec l'industrie, les gouvernements et la société civile pour étudier les menaces liées à la désinformation et renforcer la résilience démocratique.

⁵³ <https://counteringdisinformation.org/index.php/>

⁵⁴ <https://democracy-reporting.org/en/office/global>

⁵⁵ <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>

⁵⁶ https://www.nato-pa.int/download-file?filename=/sites/default/files/2021-04/013%2520CDS%252021%2520F%2520-%2520DESINFORMATION%2520PROPAGANDE%2520-%2520SANCHEZ_5.pdf

⁵⁷ <https://shorensteincenter.org/the-fight-against-disinformation-in-the-u-s-a-landscape-analysis/>

⁵⁸ <https://www.niemanlab.org/>

⁵⁹ <https://osome.iu.edu/>

- » **Institute for Strategic Dialogue (ISD)** : Think tank international qui mène des recherches et conseille les gouvernements, les entreprises technologiques et la société civile sur la désinformation, l'extrémisme et la manipulation en ligne. Il développe également des outils et des stratégies pour contrer les récits nuisibles.
- » **European Digital Media Observatory (EDMO)** : Initiative financée par la Commission européenne qui réunit des vérificateurs de faits, des experts en éducation aux médias et des chercheurs pour analyser et contrer la désinformation sur les plateformes numériques. EDMO soutient également l'élaboration de politiques et de cadres réglementaires visant à améliorer l'intégrité de l'information en Europe.



Au Québec et au Canada, plusieurs institutions jouent un rôle clé dans la compréhension et la lutte contre la désinformation. En voici une liste non exhaustive :

Nom de l'initiative ou du groupe	Province	Description
Observatoire des médias sociaux en relations publiques (OMSRP)	Québec	Basé à l'Université Laval, l'OMSRP analyse l'impact des médias sociaux sur les relations publiques et étudie les phénomènes de désinformation en ligne. http://www.omsrp.com.ulaval.ca/
Centre d'études sur les médias (CEM)	Québec	Également affilié à l'Université Laval, le CEM mène des recherches sur les médias québécois, incluant des études sur la propagation de la désinformation et ses effets sur le public. https://www.cem.ulaval.ca/
Centre de recherche interuniversitaire sur la communication, l'information et la société (CRICIS)	Québec	Regroupant des chercheurs de plusieurs universités québécoises, le CRICIS étudie les transformations des médias et de la communication, y compris les défis posés par la désinformation. https://www.cricis.uqam.ca/presentation/
Media Ecosystem Observatory (MEO)	Québec et Ontario	Collaboration entre l'Université McGill (Québec) et l'Université de Toronto (Ontario), le MEO analyse le paysage médiatique canadien pour comprendre la propagation et l'impact de la désinformation. https://www.cdmrn.ca/
Canadian Digital Media Research Network (CDMRN)	Ontario	Soutenu par l'Initiative citoyenne numérique, le CDMRN se concentre sur le renforcement de la résilience des Canadiens face à la désinformation en ligne grâce à la recherche et à des initiatives de littératie numérique. https://www.cdmrn.ca/
Information Integrity Lab de l'Université d'Ottawa	Ontario	Ce laboratoire étudie les sources, les motivations et les capacités des acteurs diffusant de la désinformation, en adoptant des approches interdisciplinaires. https://infofab.uottawa.ca/
DisinfoWatch	Ontario	Initiative de l'Institut Macdonald-Laurier, DisinfoWatch surveille et expose les narratifs de désinformation, en particulier ceux visant le Canada, et fournit des ressources pour aider le public à identifier et contrer les fausses informations. https://disinfowatch.org/
Digital Society Lab de l'Université McMaster	Ontario	Ce laboratoire développe des technologies basées sur l'IA visant à filtrer la désinformation et les discours haineux pour restaurer la confiance dans les institutions démocratiques. https://digitalsocietylab.org/projects/misinformation-and-disinformation-online
Centre for Media, Technology and Democracy de l'Université McGill	Québec	Ce centre mène des recherches sur l'intersection des médias, de la technologie et de la démocratie, en se concentrant sur des questions telles que la désinformation et ses effets sur le discours public. https://www.mediatechdemocracy.com/
The Disinformation Project	Colombie-Britannique	Rattachée à l'Université Simon-Fraser (School of Communication), initiative de recherche qui étudie le narratif de fausses nouvelles dans les médias canadiens. https://www.sfu.ca/communication/research/labs/the-disinformation-project.html
Chaire UNESCO-PREV	Québec	Bien que le sujet principal de cette chaire soit la « recherche et les actions en matière de préventions primaire, secondaire et tertiaire de la radicalisation et de l'extrémisme violents », elle s'intéresse également à la désinformation climatique. https://chaireunesco-prev.ca/

5.4 Industrie

Sans surprise, les premières organisations concernées par le phénomène de DVM et de désinformation en général sont les plateformes de médias sociaux (ex. Facebook/Instagram, Twitter/X, Google, YouTube) et les systèmes de messagerie comme WhatsApp. Ces plateformes ont des services internes de modération pour filtrer les publications jugées inadéquates ou qui ne respectent pas leurs termes de service. De plus, elles investissent des sommes importantes en recherche et développement pour 1) développer des outils pour, par exemple, du labelling automatisé, 2) développer des outils d'assistance à la modération, 3) diminuer le poids de la désinformation sur le trafic via, par exemple, le downranking (limiter la possibilité qu'une publication douteuse apparaisse en tête de résultats de recherche), les fermetures de comptes, etc. Finalement, elles financent également la recherche fondamentale, soit par un soutien direct à des chercheurs ou labos de recherche, soit en commanditant des challenges/hackathons où tous sont invités à soumettre leur solution au problème posé (avec accès à des jeux de données offerts par ces plateformes). Un exemple concret est le Deepfake Detection Challenge⁶⁰, mis sur pied par Facebook et d'autres entreprises, qui a amené plus de 2 000 participants à s'attaquer à un jeu de données costaud de 100 000 vidéos. Malgré tout, des audits comme ceux de Hussein et al. (2020)⁶¹ et Srba et al. (2023)⁶² sur YouTube laissent à penser que les moteurs de recherche et réseaux sociaux pourraient en faire plus pour lutter contre la DVM.

Il existe aussi de nombreuses plateformes de taille plus petite qui souffrent des mêmes problèmes que leurs consœurs sans avoir les mêmes moyens. Ces inégalités sont soulignées par plusieurs chercheurs qui appellent à un meilleur partage des solutions de lutte à la DVM entre plateformes.

Une autre catégorie d'acteurs industriels inclut les fournisseurs indépendants de solutions qui font déployer leurs propres outils de détection de DVM (par exemple pour fournir des services impartis de modération). On pense naturellement à des solutions côté serveur, mais on verra à la section Lutte à la DVM : produits de détection qu'il existe également des extensions de navigateurs (donc, côté client) capables, par exemple, de contextualiser l'information affichée à l'utilisateur.

Finalement, les grands producteurs de modèles comme OpenAI⁶³ sont sensibilisés au fait que leurs produits peuvent être utilisés à mauvais escient. Plusieurs types de protections internes (guardrails en anglais) sont ajoutés à ces modèles et ils font l'objet de recherches importantes afin d'améliorer leur efficacité, dont celle d'empêcher leur utilisation pour disséminer de la désinformation.⁶⁴ Cependant, ces garde-fous sont très complexes à mettre en œuvre⁶⁵, et il est permis de se demander si tous les producteurs auront les capacités techniques (et la volonté) pour les implémenter efficacement. La question se pose avec encore plus d'acuité dans le cas des modèles ouverts (open-source) dont on dit que les protections sont plus faibles ou inexistantes.⁶⁶

5.5 Grands projets

Parmi les autres acteurs luttant contre la désinformation, on retrouve aussi des consortiums financés par de grands projets de recherche comme « Horizon 2020/ Horizon Europe » de l'Union européenne. Ces projets sont souvent sous le parapluie d'efforts pour encadrer l'utilisation de l'IA à des fins éthiques et responsables.

⁶⁰ <https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/>

⁶¹ E. Hussein, P. Juneja, T. Mitra, Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. PACMHCI-2020.

⁶² I. Srba, R. Moro, M. Tomlein, B. Pecher, J. Simko, E. Stefancova, M. Kompan, A. Hrckova, J. Podrouzek, A. Gavornik, M. Bielikova.. TORS-2022

⁶³ <https://openai.com/index/forecasting-misuse/>

⁶⁴ <https://arxiv.org/abs/2312.14302>

⁶⁵ <https://arxiv.org/abs/2406.02622v1>

⁶⁶ <https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890>

- » **ELSA⁶⁷, European Lighthouse on Secure and Safe AI (2022-)** : Centre d'excellence virtuel comptant 26 partenaires universitaires et industriels. Le projet Media Analytics s'attaque directement à la DVM : "Investigate novel ways of understanding and detecting fake data, through new machine learning approaches capable of mixing (a) syntactic/perceptive analysis, (b) semantic analysis and (c) human oversight analysis". La corrélation sémantique texte-image occupe une place importante dans ce volet.
- » **VeraAI, VERification Assisted by Artificial Intelligence (2022-2025)** : Projet multipartenaire pour développer et construire des outils d'intelligence artificielle de lutte contre la désinformation véhiculée sous plusieurs formes (image, audio, vidéo et texte). Avec 14 partenaires et un budget de plus de 7 M €, l'équipe vise à rendre les outils accessibles à tous.
- » **AI4MEDIA (2020-2024)** : Centre d'excellence en IA dédié au secteur des médias : "Delivering the next generation of core AI advances and training to serve the Media sector, while ensuring that the European values of ethical and trustworthy AI are embedded in future AI deployments." De nombreux (30) partenaires universitaires et industriels attaquent plusieurs sujets d'intérêt pour les médias, la société et la politique, dont, par exemple, la détection de texte généré automatiquement.
- » **DISARM Foundation⁶⁸** : OBNL qui maintient une plateforme libre (open-source) de lutte à la désinformation. Le focus n'est pas sur la détection automatisée de désinformation (textuelle ou visuelle), mais plutôt sur la coordination et le partage d'informations une fois une campagne de désinformation mise à jour.



⁶⁷ <https://www.elsa-ai.eu/index.html>

⁶⁸ <https://www.disarm.foundation/>

6. ÉTAT DE LA LUTTE CONTRE LA DVM : SOLUTIONS POSSIBLES

À retenir :

- » La lutte contre la DVM est complexe et multifacette : un bouquet de solutions vont devoir être mises en place pour lutter efficacement.
- » Les plateformes comme Facebook, X/Twitter, YouTube, etc. ont un rôle à jouer pour ralentir la propagation de la DVM en implémentant des stratégies de modération et de contextualisation des publications.
- » Les programmes de littératie numérique sont essentiels pour former la pensée critique face au contenu qui circule sur Internet, et ce, dès le plus jeune âge.
- » Les solutions légales ont une portée limitée à cause d'enjeux de juridictions et de liberté d'expression.
- » Les solutions technologiques sont attrayantes, mais elles sont perpétuellement à la traîne face aux innovations récupérées par les producteurs de DVM.

À la lumière des points présentés dans les sections précédentes, la lutte contre la DVM et la désinformation en général est incontestablement complexe et ardue. On retrouve beaucoup d'enjeux, ce qui est mis en valeur, par exemple, dans la liste de 27 recommandations d'un rapport⁶⁹ du Forum des politiques publiques du Canada, ou celle des 35 recommandations de Wardle (2017).⁷⁰ Dans la littérature de recherche qui étudie la lutte à la désinformation, il y a consensus à l'effet que le défi doit être attaqué à plusieurs niveaux à l'aide d'une panoplie d'outils : "...no single solution or stakeholder can fully address the challenge. Successful efforts will require new technologies, organizational practices, and societal changes." (Bateman, 2020).

Dit autrement, il est reconnu qu'il n'existe aucune solution unique au problème de lutte à la DVM et que

plusieurs types de solutions devront être développées et déployées conjointement par plusieurs acteurs^{71 72}. Par exemple, Helmus (2022)⁷³ souligne l'importance d'agir sur cinq facettes : mettre au point des outils de détection, mettre en œuvre des normes de certification pour assurer l'authenticité de documents audiovisuels, considérer les approches réglementaires, promouvoir dans le monde journalistique les approches basées sur les renseignements d'origine sources ouvertes (OSINT), favoriser la littératie médiatique.

Un autre consensus est que l'intelligence artificielle a un rôle-clé à jouer vu l'enjeu de filtrer l'énorme quantité de messages postés quotidiennement sur les réseaux sociaux : "AI has been identified as a powerful and cost-efficient tool to identify online disinformation, as it can quickly analyse a vast amount of items and provide outputs on their nature." (Svahn & Perfumi, 2022). Cette tendance à s'appuyer sur l'IA est en nette augmentation malgré plusieurs défis liés aux biais, à la performance, aux faux positifs, etc.

Les sous-sections suivantes présentent des solutions qui pourraient être mises en place par les acteurs impliqués dans la lutte contre la DVM.

6.1 Solutions par les plateformes

Mentionnons d'entrée de jeu que le positionnement des plateformes technologiques comme Facebook, X, etc. est critiqué par plusieurs (dont Paris & Donovan) qui affirment que celles-ci tirent profit de l'achalandage (repartages) causé par la circulation de la désinformation ("The volume and speed of disinformation proliferating through online platforms are a source of enormous profits for those online platforms." (Dan et al, 2021)). De plus, elles soutiennent aussi que la recherche de solutions purement techniques, par exemple contrer

⁶⁹ <https://ppforum.ca/wp-content/uploads/2022/01/DEMEX-R2.pdf>

⁷⁰ <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>

⁷¹ <https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deep-fakes-and-audiovisual-disinformation>

⁷² https://www.nato-pa.int/download-file?filename=/sites/default/files/2021-04/013%2520CDS%252021%2520F%2520-%2520DESINFORMATION%2520PROPAGANDE%2520-%2520SANCHEZ_5.pdf

⁷³ <https://www.rand.org/pubs/perspectives/PEA1043-1.html>

les deepfakes, favorise la croissance des entreprises technologiques, alors qu'elles se soustraient à leurs responsabilités en faisant peu pour contrer les cheapfakes qui sont également dommageables. D'ailleurs, l'inconsistance des mesures de protection mises en place par les plateformes, dont l'encadrement semble évoluer d'une année à l'autre en fonction de règles internes et de décisions stratégiques, suscite des préoccupations. À tel point qu'en prévision des élections américaines de novembre 2024, certaines organisations ont entrepris de compiler et d'analyser les politiques d'utilisation des principales plateformes numériques.⁷⁴

On peut classer en trois catégories les actions possibles côté plateformes : modération, labelling, contextualisation.

6.1.1 Modération

La modération de contenu consiste à réviser et monitorer le contenu en ligne d'une plateforme produit par ses usagers, dans le but de valider sa conformité aux normes d'utilisation de la plateforme. Le processus de modération peut être exécuté par une combinaison de solutions, c.-à-d. outils algorithmiques, réviseurs humains, co-utilisateurs, et ce, selon une structure variable (gérée par la plateforme, impartie à un fournisseur, fortement décentralisée via crowdsourcing (Morrow et al.), etc.).

Un contenu inapproprié peut alors subir plusieurs traitements :

- » **Retrait pur et simple.**
- » **Déclassement (downranking) :** Le contenu est moins susceptible d'être sélectionné par les algorithmes de recherche.
- » **Démonétisation :** Le contenu ne génère pas de revenu.

La modération semble être une mesure efficace, mais elle comporte son lot de défis. Les modérateurs humains doivent être présents dans la boucle, mais pour qu'ils exécutent leur tâche adéquatement, ils doivent être bien payés, protégés légalement, bien formés et avoir accès à du support psychologique au besoin (FTC,

2022). Malgré toute leur bonne volonté, ils ne sont jamais neutres et ils doivent lutter contre leur propension à déléguer la prise de décision à la machine (automation bias) (FTC, 2022).

6.1.2 Labelling

On peut définir le labelling comme suit : "A content label is a visual and/or textual attachment to a piece of user-generated content intended to contextualize that content for the viewer" (Morrow et al., 2021). Un exemple de labelling est donné ci-dessous, où une bannière avec une mise en garde et des informations additionnelles accompagne le message original.



Figure 5 -Source :

<https://www.eurekalert.org/multimedia/630349>

À première vue, l'idée de compléter une publication de réseau social avec de l'information vérifiée est intéressante, mais la démarche soulève plusieurs enjeux éthiques : censure, liberté d'expression, protection des utilisateurs. De plus, les recherches mettent en évidence le fait que la présence de labels induit des réactions opposées chez les utilisateurs. Certains sont en accord avec le labelling parce qu'ils considèrent que les plateformes sont responsables du fact-checking, alors que d'autres jugent l'approche paternaliste et perçoivent les plateformes comme des censeurs voulant encadrer voire restreindre la liberté d'expression. L'apparence

⁷⁴ <https://statesunited.org/resources/social-media-policies/>

visuelle des labels (taille en proportion de la publication, choix des mots, etc.) doit être judicieusement choisie afin de maximiser leur impact; mentionnons l'existence de lignes directrices⁷⁵ comme outil d'aide au design des labels.

6.1.3 Contextualisation

Un exemple concret de labelling est la contextualisation⁷⁶, qui consiste à fournir à l'utilisateur de l'information additionnelle que la publication initiale ne contient pas. Elle peut inclure des détails sur la source de l'image faisant partie de la publication, sur l'auteur de la publication, etc.

6.2 Approches sociétales

De nombreux observateurs insistent sur l'importance de promouvoir la littératie numérique et médiatique auprès de la population, afin de développer et renforcer sa pensée critique face à la désinformation en général et plus spécifiquement la DVM. Deux groupes sont surtout visés : les jeunes et les journalistes.

- » De par leur métier, les journalistes sont exposés à de nombreuses parcelles d'information qu'ils doivent valider dans un laps de temps toujours plus court. Or, ils n'ont pas nécessairement toutes les compétences requises (entre autres les compétences en informatique et en édition d'images et de vidéos) pour faire ces vérifications. La DVM est particulièrement difficile à valider puisqu'il existe peu d'outils pour détecter, par exemple, des deepfakes dont les détails sont invisibles à l'œil nu ou des images qui n'ont pas de source connue et pour lesquelles il est difficile de confirmer la véracité.
- » Les jeunes sont potentiellement fortement influencés par la désinformation et il y a consensus sur l'importance d'inclure des programmes de littératie numérique au primaire et au secondaire (ex. News Literacy Project destiné aux éducateurs). Ces programmes devraient même se poursuivre au niveau

universitaire, particulièrement dans les écoles de journalisme et de communication (Languein, 2022).

Petit bémol, certains programmes de littératie numérique suivent un modèle de fact-checking : Par exemple, des paroles ont été prononcées et il faut vérifier si les énoncés sont vrais; ce modèle serait mal adapté à la présence de deepfakes selon Vaccari et Chadwick (2020), où les paroles ont elles-mêmes été fabriquées. Cela dit, plusieurs ressources complémentaires^{77 78} visant d'autres segments de la population (ex. personnes âgées⁷⁹) commencent à apparaître en ligne.

Un autre volet sociétal plus délicat touche à la responsabilité des chercheurs/développeurs qui créent des outils de plus en plus puissants sans réels questionnements sur leur impact futur. Selon Goldstein et al., la communauté de l'IA devrait instaurer des normes éthiques pour la mise en production (release) de modèles génératifs. Dans la même veine, les auteurs du rapport du DHS mentionnent que "Organizations and individuals who develop models utilized in deepfake creation should also consider their responsibilities when it comes to mitigation."

Finalement, il convient de souligner l'apport des chercheurs académiques dont l'expertise porte sur la lutte à la désinformation. Plusieurs événements bien documentés⁸⁰ ont mis en lumière les défis rencontrés par les chercheurs spécialisés dans la lutte contre la désinformation, dont certains ont été la cible de campagnes de dénigrement orchestrées par des groupes opposés à leurs travaux. Ces attaques visent à discréditer leur réputation et à semer le doute sur la fiabilité de leurs recherches, compromettant ainsi leurs efforts pour sensibiliser le public et promouvoir une information vérifiée et fiable.

⁷⁵ <https://firstdraftnews.org/articles/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow/>

⁷⁶ <https://aviv.medium.com/contextualization-engines-can-fight-misinformation-without-censorship-c5c47222a3b7>

⁷⁷ <https://www.sans.org/newsletters/ouch/learn-a-new-survival-skill-spotting-deepfakes/>

⁷⁸ <https://www.media.mit.edu/projects/detect-fakes/overview/>

⁷⁹ <https://www.buffalo.edu/cii/projects/DART.html>

⁸⁰ <https://www.bu.edu/articles/2023/disinformation-researchers-under-attack-by-government-legislators/>

6.3 Solutions politiques/légales

6.3.1 Cadres juridiques et réglementaires actuels

Face à la sophistication croissante de la désinformation visuelle et multimodale, les cadres juridiques et réglementaires doivent évoluer pour répondre aux défis qu'elle pose. Ces efforts visent à établir des responsabilités claires, à renforcer la transparence et à limiter les préjudices causés par cette désinformation, tout en prenant en compte les complexités liées aux droits et libertés constitutionnels, aux innovations technologiques et aux questions de juridiction mondiale. Plusieurs juridictions à travers le monde ont introduit, ou sont en train d'introduire, des mesures pour lutter contre la désinformation visuelle et multimodale, bien que leurs approches et mécanismes d'exécution varient considérablement.

Au niveau de la communauté internationale, l'UNESCO soutient des initiatives globales axées sur l'éducation aux médias et l'utilisation éthique de l'intelligence artificielle.⁸¹ En novembre 2023, elle a publié des « Principes pour la Gouvernance des Plateformes Numériques »⁸², un cadre visant notamment à combattre la désinformation tout en préservant la liberté d'expression. De son côté, l'ONU a présenté en juin 2024 cinq « Grands Principes pour l'Intégrité de l'Information »⁸³, qui encouragent notamment la réforme des modèles publicitaires des plateformes numériques et des réseaux sociaux pour limiter la propagation de contenus trompeurs. Cependant, bien que ces directives fournissent un cadre commun pour les décideurs, leur caractère non contraignant limite leur portée et leur efficacité.

L'Union européenne s'est démarquée avec l'introduction du Digital Services Act (DSA), en vigueur depuis 2023. Ce cadre réglementaire impose aux grandes plateformes en ligne de procéder à des évaluations régulières

des risques systémiques liés à leurs services, incluant les contenus illicites et manipulés visuellement. Ces risques comprennent, entre autres, les campagnes coordonnées de désinformation, les violations des droits fondamentaux et les atteintes à l'ordre public. En complément, le DSA exige une transparence accrue des pratiques algorithmiques et des rapports détaillés sur les efforts de modération des contenus. Pour assurer une réelle conformité, le DSA prévoit des sanctions significatives, dont des amendes pouvant atteindre jusqu'à 6 % du chiffre d'affaires mondial.⁸⁴ Le AI Act, adopté en 2024, complète cette approche en imposant aux fournisseurs et déployeurs d'IA génératives de divulguer clairement si un contenu textuel, visuel ou audio a été manipulé ou créé artificiellement, limitant ainsi les risques de confusion pour le public.⁸⁵

Aux États-Unis, l'approche est fragmentée, reposant principalement sur des initiatives au niveau des États. En Californie, la loi AB 730 interdit depuis 2020 la diffusion de deepfakes trompeurs dans les 60 jours précédant une élection, visant à protéger l'intégrité des processus électoraux.⁸⁶ En 2024, cette réglementation a été renforcée par des lois comme le Defending Democracy From Deepfake Deception Act (AB 2655)⁸⁷, qui impose aux grandes plateformes en ligne de supprimer ou signaler les contenus manipulés pendant les périodes électorales, et AB 2355⁸⁸, qui exige des avertissements clairs pour les publicités politiques générées par IA. Dans d'autres États comme le Texas et la Virginie, les législations ciblent les hypertrucages non consensuels, particulièrement dans des contextes de harcèlement ou de contenus à caractère pornographique. Malgré ces efforts, l'absence de cadre fédéral laisse un « patchwork » législatif complexe, limitant une réponse cohérente à l'échelle nationale.

Au Canada, les efforts visant à encadrer les préjudices en ligne, y compris la désinformation visuelle et multimodale (DVM), ont été retardés ou interrompus.

⁸¹ <https://www.unesco.org/fr/media-information-literacy?hub=750>

⁸² <https://www.unesco.org/fr/internet-trust/guidelines>

⁸³ <https://www.un.org/fr/information-integrity>

⁸⁴ <https://www.consilium.europa.eu/fr/policies/digital-services-act/>

⁸⁵ https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_fr

⁸⁶ https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

⁸⁷ https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2655

⁸⁸ https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2355

Présenté en février 2024, le projet de loi C-63 sur les préjudices en ligne visait à protéger les Canadiens, en particulier les mineurs, contre les contenus nuisibles et à responsabiliser les plateformes quant aux matériaux qu'elles hébergent.⁸⁹ Il proposait notamment des mesures pour lutter contre les deepfakes non consentuels, en imposant aux plateformes de retirer rapidement les contenus signalés et de mettre en place des mécanismes de signalement pour les utilisateurs. En cas de non-respect, des sanctions financières importantes ou des restrictions opérationnelles étaient prévues. Toutefois, le projet de loi C-63 n'a pas progressé avant la prorogation du Parlement au début de l'année 2025, stoppant ainsi son processus législatif. Sans une nouvelle introduction, le Canada ne dispose actuellement d'aucun cadre juridique spécifique pour traiter la DVM et la désinformation générée par l'IA.

6.3.2 Lacunes et défis des cadres juridiques

Malgré les avancées réalisées par certaines juridictions, les cadres juridiques dans leur ensemble présentent des limites significatives qui entravent une réponse efficace à la désinformation visuelle et multimodale. Ces lacunes reflètent des contraintes structurelles, des défis technologiques et des tensions entre protection des droits et efficacité réglementaire, et se manifestent à travers plusieurs enjeux clés :

- » Les juridictions locales appliquent des approches fragmentées, limitant l'efficacité globale des réglementations : Un contenu retiré dans un pays peut rester accessible ailleurs, rendant la suppression partielle moins efficace.
- » La propagation rapide des contenus manipulés dépasse les capacités d'intervention des régulateurs : Les technologies de manipulation évoluent plus vite que les cadres juridiques, créant des lacunes exploitables par les acteurs malveillants.
- » Trouver un équilibre entre régulation et liberté d'expression est un défi majeur : Toute régulation risque d'être perçue comme de la censure, tandis que l'absence de régulation expose les citoyens à des campagnes de manipulation.

6.4 Solutions technologiques

Un nombre croissant de solutions algorithmiques sont développées pour détecter la désinformation visuelle et multimodale (DVM) grâce à l'analyse du contenu des images et des vidéos. Sans surprise, l'intelligence artificielle joue un rôle central dans ces méthodes de détection.

Pour les cheapfakes—des manipulations simples comme l'accélération ou le ralentissement d'une vidéo, le recadrage, la recontextualisation ou des altérations manuelles—des approches basées sur l'apprentissage automatique permettent d'identifier certaines incohérences, comme des anomalies dans le mouvement, le taux d'images par seconde, la pixellisation ou les artefacts de compression. Toutefois, ces méthodes s'avèrent moins efficaces face à des manipulations plus sophistiquées.

Le développement rapide d'outils de création de contenu basés sur l'IA, tels que les générateurs de deepfakes et les modèles de conversion texte-image ou texte-vidéo, représente un défi supplémentaire. Les stratégies de détection classiques, qui reposent sur l'analyse des incohérences de pixels ou l'identification d'éléments copiés, peinent à contrer les contenus générés par l'IA, qui apparaissent de plus en plus réalistes et homogènes. Ainsi, les mesures de détection doivent évoluer en permanence pour suivre l'adaptation des tactiques de désinformation. Ces aspects seront examinés plus en détail dans les sections 7 et 8.



⁸⁹ <https://www.canada.ca/fr/patrimoine-canadien/services/prejudices-en-ligne.html>

7. LUTTE À LA DVM : LE PAYSAGE DES SOLUTIONS TECHNOLOGIQUES

À retenir :

- » L'IA est utilisée à la fois pour produire de la désinformation visuelle/multimodale et pour la combattre.
- » Les outils de production incluent les logiciels de fabrication de deepfakes, de remplacement de visages et de manipulation de contenu visuel en général; ils incluent aussi les nouveaux modèles de génération d'images et de vidéos.
- » Jusqu'à tout récemment, les outils de détection de DVM étaient conçus pour détecter des manipulations simples dans les images et vidéos, i.e. shallow-fakes et cheapfakes.
- » L'arrivée des modèles génératifs capables de créer facilement du contenu visuel inédit, à partir de texte, change la donne : peu de solutions existent pour détecter ce type de contenu.

Des solutions technologiques variées ont été proposées au cours des dernières années pour tenter de contrer la DVM, et ce, à plusieurs niveaux de la chaîne de transmission. Leur degré de succès attendu est variable selon la complexité intrinsèque de chaque solution et selon les efforts requis pour la mettre en œuvre.

Cette section dresse un portrait des approches et solutions disponibles actuellement, autant au niveau académique que commercial, entre autres :

- » Les approches qui font partie d'une stratégie de production de DVM. Elles incluent les modèles génératifs capables de créer des images et vidéos de toutes pièces. Une compréhension de ces approches guide le développement et l'optimisation des détecteurs de DVM.
- » Les approches qui font partie d'une stratégie de détection de contenu de DVM, autant pour les shallow-fakes/cheapfakes que pour les hypertrucages (deep-fakes). On s'intéressa particulièrement à la détection de DVM dans les images et l'audio.

- » Les approches complémentaires où l'IA joue un rôle moins important, mais qui sont à surveiller.

On retrouve également dans cette section un inventaire des jeux de données utilisés dans la mise au point de ces approches proposées, ainsi qu'une liste de produits commerciaux conçus pour la détection de DVM. Vu l'essor fulgurant des modèles génératifs dans les dernières années et leur implication de plus en plus indiscutable dans la chaîne de production de DVM (et également dans les stratégies de détection), la section 8 leur sera complètement dédiée.

7.1 Méthodes de production de DVM

Comme il existe plusieurs types de DVM, il existe également plusieurs classes d'outils de complexité variable permettant de la produire. La production de cheapfakes est facile et ne requiert que des connaissances de base en manipulation d'image ou en édition vidéo. L'arrivée de Photoshop en 1990 a certainement contribué à populariser ce type de faux par ses fonctionnalités d'édition photo. La production de deepfake est autrement plus complexe et peut demander des connaissances poussées en informatique. On peut classer les outils et techniques en deux catégories : ceux qui transforment les documents existants (face swapping, voice spoofing, synchronisation de lèvres et manipulation d'apparence) et ceux qui les génèrent. Rappel : la partie « génération » sera couverte à la section 8.

Parmi les outils qui transforment les documents, on trouve :

- » **Logiciels de face swapping** : Comme le nom l'indique, ces logiciels remplacent le visage d'une personne apparaissant dans une vidéo par celui d'une autre personne. Ils sont de plus en plus faciles d'utilisation, accessibles, et populaires. Le site de partage de code source GitHub héberge deux logiciels gratuits, [FaceSwap](#) et [DeepFaceLab](#), qui sont tous deux dans le top 250 des dépôts les plus populaires (sur 28 millions de dépôts publics).

- » **Logiciels de conversion de voix (voice spoofing) ou de synthèse vocale à partir du texte (text-to-speech)** : Ces logiciels permettent de transformer un enregistrement d'une personne en donnant à sa voix l'identité d'une autre personne (conversion de voix), ou de transformer entièrement le contenu parlé d'un enregistrement par du nouveau contenu parlé à partir de texte, tout en conservant la voix d'origine.
- » **Logiciels d'édition de parole en place (in-place editing)** : Ces logiciels permettent de modifier un enregistrement audio existant en remplaçant seulement une partie du contenu parlé par du nouveau contenu provenant d'un texte, en conservant la voix et les caractéristiques de l'audio original, sans discontinuité. Cette technique est particulièrement difficile à détecter, car la majeure partie de l'enregistrement original reste intouchée.
- » **Logiciels de synchronisation lèvres-audio (lip syncing)** : Ceux-ci vont modifier la vidéo existante d'une personne qui parle pour lui faire dire autre chose, tout en prenant soin d'ajuster la forme de la bouche pour qu'elle corresponde aux phonèmes supposément prononcés.
- » **Logiciels de manipulation d'apparence** : Ces logiciels modifient l'apparence du visage d'une personne pour, par exemple, la rajeunir ou la vieillir.
- » **Logiciels de synthèse de performance virtuelle** : Ce n'est plus simplement le visage d'une personne qui est modifié, mais tout son corps qui est mis en action par transposition des mouvements d'une autre personne.

7.2 Stratégies de détection impliquant l'image

À retenir :

- » Les méthodes répertoriées ne sont pas adéquates pour faire face à la nouvelle vague d'outils de génération et de manipulation.
- » Il n'y aura pas de méthode unique qui s'y attaquera.
- » Nécessité de combiner avec des approches sémantiques.

- » La détection automatique de contenu généré par des modèles d'IA (ex. deepfakes) est encore largement à l'étape de la recherche, mais des solutions commerciales commencent à apparaître.
- » Un champ de recherche prometteur, mais difficile et très peu avancé, porte sur la comparaison sémantique entre du contenu visuel et du texte, afin de détecter des cas de DVM basés sur de la recontextualisation.
- » Certaines manipulations d'images laissent des traces imperceptibles à l'œil nu, mais détectables par des algorithmes spécialisés.
- » Certaines pistes de solutions sont déjà utilisées, par exemple la recherche d'images inversées pour retrouver des images sources ou similaires, mais elles ne règlent qu'une partie des problèmes.

Tel qu'évoqué à la section 6, les approches d'intelligence artificielle conçues pour détecter des documents véhiculant de la DVM sont appelées à prendre une place de plus en plus importante vu leur capacité à traiter un volume grandissant de contenu échangé sur les médias sociaux. Il existe plusieurs manières de véhiculer de la désinformation à l'aide d'une image, comme nous l'avons vu dans les premières sections de ce rapport.

Au fil du temps, plusieurs approches ont été développées pour aider à repérer les images potentiellement trompeuses. Elles se répartissent en quatre grandes catégories :

1. Analyse du contexte sémantique – Vérifier si l'image correspond réellement au texte, aux métadonnées ou au contexte dans lequel elle est utilisée, afin d'identifier les cas de recontextualisation trompeuse.
2. Recherche inversée d'images – Retrouver l'origine d'une image ou des versions similaires pour déterminer si elle a été manipulée ou détournée.
3. Analyse des traces numériques – Repérer des anomalies dans l'éclairage, la résolution, la compression ou d'autres indices techniques qui peuvent révéler une modification.

4. Détection des contenus générés par IA – Identifier les images créées par des modèles d’intelligence artificielle grâce à des outils d’analyse forensique ou à l’intégration de marqueurs spécifiques.

7.2.1 Analyse d’intégrité sémantique

Plusieurs types d’analyse peuvent être menés pour valider la cohérence sémantique d’une image. Par exemple, il est possible de valider l’identité de personnes dans une image afin de détecter des face swaps⁹⁰, des impossibilités historiques, etc. Une banque d’images⁹¹ contenant des anachronismes ou des situations qui défient la logique ou le sens commun a récemment été rendue publique pour stimuler le développement de ce type d’algorithme.

Un autre problème complexe est l’analyse d’intégrité sémantique texte-visuel, qui est un champ d’étude très récent exploitant les récents travaux sur les réseaux de neurones multimodaux capables d’accepter en entrée à la fois des images et du texte. L’objectif est de vérifier que le message textuel est sémantiquement aligné avec le contenu visuel de l’image, afin de détecter des scénarios de « image repurposing », où une image réelle et non modifiée est sortie de son contexte et jointe à un texte qui décrit une réalité différente (ex. McCrae et al., 2022).

Ces technologies permettraient de construire des systèmes semi-automatiques de vérification de faits plus riches que ceux actuellement déployés qui sont uniquement basés sur l’analyse de texte. On pense à l’architecture PROVES (Xie et al., 2022) qui assure la certification et la vérification de l’information sémantique d’une image, tout en permettant des opérations d’édition simples comme des rognages, des rotations, des ajustements de couleur, etc.

Or comme le dit le directeur du programme SemaFor (DARPA), “Currently, it is very difficult for an automated generation algorithm to get all of the semantics correct. Ensuring everything aligns from the text of a news story, to the accompanying image, to the elements within the image itself is a very tall order. Through this program we aim to explore the failure modes where current techniques for synthesizing media break down”⁹².

Le projet ELSA a lui aussi un volet d’analyse sémantique puisqu’il “will investigate novel ways of understanding and detecting fake data, through new machine learning approaches capable of mixing syntactic and perceptive analysis.”⁹³

Évaluation:

Le problème de valider l’alignement sémantique du texte avec l’image ou la vidéo qui l’accompagne est difficile et les équipes de recherche qui s’y intéressent de manière concertée (financées par SemaFor et ELSA) commencent à peine leurs travaux. Cependant, cette voie pourra connaître un déblocage significatif à mesure que des jeux de données de bonne qualité seront rendus publics. Le fait que les modèles génératifs continuent à améliorer la cohérence sémantique du texte ou de la parole avec l’image ou la vidéo générée rend cette validation plus difficile; par contre, ces mêmes modèles peuvent aussi participer à la lutte, un thème qui sera abordé à la section 8.

⁹⁰ <https://par.nsf.gov/servlets/purl/10346314>

⁹¹ <https://whoops-benchmark.github.io/>

⁹² <https://www.darpa.mil/news-events/2021-03-02>

⁹³ <https://aimagelab.ing.unimore.it/imagelab/project.asp?idprogetto=101>

Ce projet multipartenaires de plusieurs dizaines de millions de dollars américains est ambitieux, comme le montre la figure ci-dessous tirée d'une présentation officielle⁹⁴ :

	Desired Capability	Today	SemaFor
Detection	Automatically detect semantic generation/manipulation errors	Limited	Yes
	Detect manipulations across multiple modalities and assets	Limited	Yes
	Robust to many manipulation algorithms	Fragile	Highly robust
	Increased adversary effort needed to fool detection algorithms	Some	Significant
Attribution	Automatically confirm source or author	Limited	Yes
	Automatically identify unique source fingerprints	No	Yes
	Explain authorship inconsistencies	No	Yes
Characterization	Automatically characterize manipulation intent or impact	No	Yes
	Provide evidence and explanation for manipulation intent	No	Yes
	Correctly prioritize generated/manipulated media for review	No	Yes

7.2.2 Recherche d'images inversées

Les méthodes de recherche d'images inversée sont conçues pour retrouver des images similaires en termes de contenu visuel. Elles sont plus puissantes que de simples recherches de mots-clés dans les métadonnées des images, puisque ces métadonnées sont souvent éliminées par les plateformes de médias sociaux au moment du téléversement d'une publication. L'idée de base est de construire une empreinte numérique à partir des caractéristiques visuelles « stables » de l'image à rechercher et ensuite de comparer cette empreinte à celles d'une banque d'images. Le terme « stable » fait référence au fait que l'empreinte de l'image change peu même si celle-ci subit des transformations comme un rognage (crop), une rotation, un changement de contraste, etc. L'intérêt de cette classe d'outils est de faciliter la détection de désinformation visuelle basée sur la recontextualisation : le retraçage d'une image utilisée pour illustrer un propos permet de vérifier si elle a déjà été utilisée dans le passé et si oui, pour vérifier qu'elle a servi à appuyer le même type de message.

La construction de l'empreinte va comme suit : Des points d'intérêt (keypoints) perceptuellement importants sont identifiés automatiquement dans l'image (coins, jonctions de bordures, etc.), et à chaque point est rattachée une signature numérique qui est créée à partir des pixels du voisinage du point dans l'image. L'empreinte finale est une combinaison de ces signatures. La figure suivante illustre la capacité d'un algorithme de comparaison à détecter la similitude de contenu entre deux images, même si ce contenu a une apparence différente. Des services comme Google Image Search⁹⁵ et TinEye balayent le Web continuellement dans le but de récolter de nouvelles images, de calculer leurs empreintes et de les stocker sur leurs serveurs; quand l'utilisateur qui veut faire une recherche d'image inversée téléverse l'image à chercher, son empreinte est calculée et comparée aux milliards d'empreintes déjà en banque, afin d'identifier les images au contenu identique ou très ressemblant.

⁹⁴ https://resources.sei.cmu.edu/asset_files/Presentation/2022_017_001_889114.pdf

⁹⁵ <https://newsinitiative.withgoogle.com/fr-fr/resources/trainings/reverse-image-search-verifying-photos/>

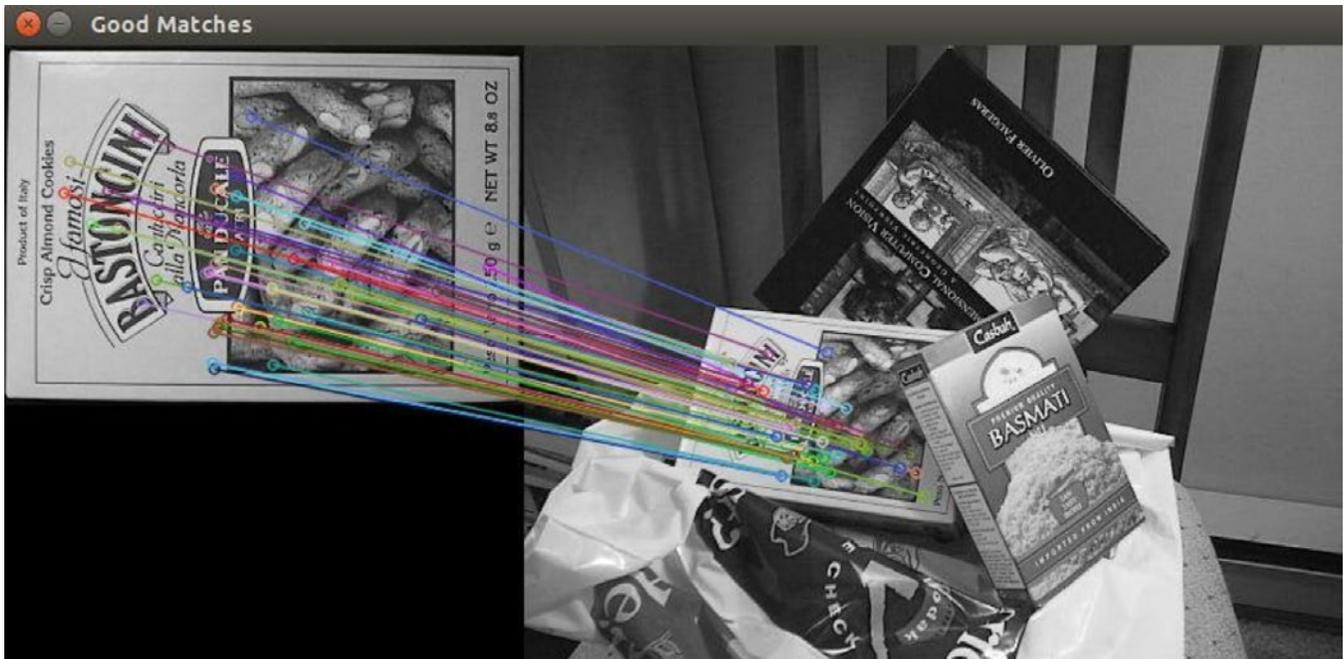


Figure 6 – Source : https://docs.opencv.org/3.4/d5/d6f/tutorial_feature_flann_matcher.html

Évaluation:

Plusieurs services offerts sur Internet reposent sur des techniques développées il y a 15-20 ans. Or ces techniques ont été revues⁹⁶ sous l'angle de l'apprentissage profond, avec un bond de performance significatif quant à leur capacité à trouver des points de correspondance entre images; il est donc permis de penser que la recherche d'images inversée peut encore s'améliorer et rester un outil important⁹⁷ de lutte contre la DVM.

7.2.3 Analyse des artéfacts

La manipulation d'images numériques (modifications, ajout et suppression d'objets) laisse des traces pouvant être utilisées pour détecter les images falsifiées. L'analyse de traces qui fait référence à la technique de vérification en forensique d'image, considère aussi bien les artéfacts générés par des opérations post-capture

que ceux induits durant l'acquisition par le capteur lui-même. Ces artéfacts sont porteurs d'information utile pouvant être mise à contribution pour établir l'authenticité de l'image^{98 99}.

On peut résumer les types d'artéfacts ainsi :

- » **Artéfacts produits à l'acquisition de l'image** : Ils sont causés par la lentille de l'appareil, le capteur électronique, etc.
- » **Artéfacts dus à l'intégrité physique de la scène** : Ils vont entraîner des incohérences comme des ombres incompatibles avec les sources d'éclairage, par exemple.
- » **Artéfacts causés par les modèles génératifs** : Fabriqués par les modèles génératifs qui sont capables de créer des images et des vidéos de toutes pièces à partir de prompts textuels.

⁹⁶ <https://link.springer.com/article/10.1007/s11263-020-01359-2>

⁹⁷ <https://rjonline.org/news/the-importance-of-using-reverse-image-search-tools-to-effectively-combat-misinformation-especially-on-sensitive-topics/>

⁹⁸ Redi, J.A., Taktak, W. & Dugelay, J.L. Digital image forensics: a booklet for beginners. *Multimed Tools Appl* 51, 133–162 (2011). <https://doi.org/10.1007/s11042-010-0620-1>

⁹⁹ Sencar, H.T., Verdoliva, L., Memon, N. (eds) *Multimedia Forensics. Advances in Computer Vision and Pattern Recognition*. Springer, Singapore https://doi.org/10.1007/978-981-16-7621-5_7

» **Artéfacts visuels** : Incohérences visuelles qui apparaissent non naturelles, comme des yeux d'apparence différente ou des imperfections d'éléments du visage (dents, nez).

Plus de détails sont donnés à l'Annexe IV – Analyse d'artéfacts.

Évaluation:

Clairement, un appui à la recherche plus approfondie sur la détection de ces artéfacts est nécessaire. De telles solutions ont un potentiel important de détection, car elles ne reposent pas sur l'analyse explicite ou d'hypothèses sur le contenu lui-même de l'image. D'autres part, nous estimons que la recherche s'est focalisée beaucoup plus sur l'aspect visage, en particulier en détection de deepfakes, ce qui pose un vrai problème de scalabilité selon certaines études. Une exploration diversifiée reflétant un contexte plus général est à considérer à l'image de VideoSham¹⁰⁰ qui va bien au-delà de la manipulation de visages. Mais il faut s'attendre à ce que cet aspect de la lutte à la DVM se complique avec l'amélioration continue des modèles génératifs, qui produisent des artéfacts visuels dans les images et les vidéos deepfakes de plus en plus difficiles à détecter.

7.2.4 Détection d'images générées

Il existe maintenant une panoplie de modèles d'IA générative capables de créer des images et des vidéos à partir de prompts textuels, comme on le verra plus en détail à la section 8. Afin de pouvoir séparer les images synthétiques des images réelles, il est nécessaire de développer des méthodes d'analyse qui vont tenter de détecter la présence d'artefacts produits au moment de la création des images par les modèles génératifs (GAN, modèles texte-à-image). Il serait même possible de retracer la classe de modèle génératif à l'origine d'une image synthétique à partir de l'analyse fine des artéfacts. Mais on voudrait pouvoir compter sur des approches performantes qui seraient universelles, c'est-à-dire qu'elles auraient la capacité de détecter des

images générées sans égard au modèle de génération en cause. Les recherches progressent¹⁰¹ pour ce qui est des nombreux modèles de GAN, mais il reste du travail à faire pour inclure les autres types de modèles génératifs (texte-à-image). Une approche toute récente qui pourrait être prometteuse est basée sur des modèles multimodaux comme ChatGPT qui auraient la capacité de détecter des deepfakes¹⁰², bien qu'ils n'aient pas été entraînés spécifiquement pour cette tâche. Pour l'instant, les modèles entraînés à partir d'exemples d'images synthétiques performant mieux en détection de deepfakes que des modèles du style ChatGPT (modèles langue-vision, ou VLM en anglais), mais ceux-ci ont au moins trois avantages : 1) ils peuvent avoir l'accès à de la connaissance externe via une recherche internet et ainsi faire de la vérification de faits ou de la recherche inversée d'image; 2) ils sont capable d'expliquer pourquoi ils croient qu'une image est fausse ou réelle (l'explication peut être transmise via une conversation avec l'utilisateur); 3) ils sont des modèles de détection universels, agnostiques au type de modèle génératif à la source de l'image analysée. Cette discussion sera reprise à la sous-section 8.2.1 de la section 8 qui porte spécifiquement sur l'utilisation des modèles génératifs dans la lutte à la DVM.

Évaluation:

Le plus gros défi pour la détection d'images générées est de créer un modèle universel/unifié capable de détecter tout type d'image synthétique sans égard au modèle qui l'a générée. Les modèles VLM semblent offrir une telle solution universelle, mais il reste du travail à faire pour les amener à un niveau de performance acceptable.

¹⁰⁰ T. Mittal, R. Sinha, V. Swaminathan, J. Collomosse and D. Manocha, Video Manipulations Beyond Faces: A Dataset with Human- Machine Analysis, 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2023, pp. 643-652, doi: 10.1109/WACVW58289.2023.00071.

¹⁰¹ <https://arxiv.org/abs/1912.11035>

¹⁰² S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, S. Lyu. Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics. 2024. <https://arxiv.org/pdf/2403.14077>

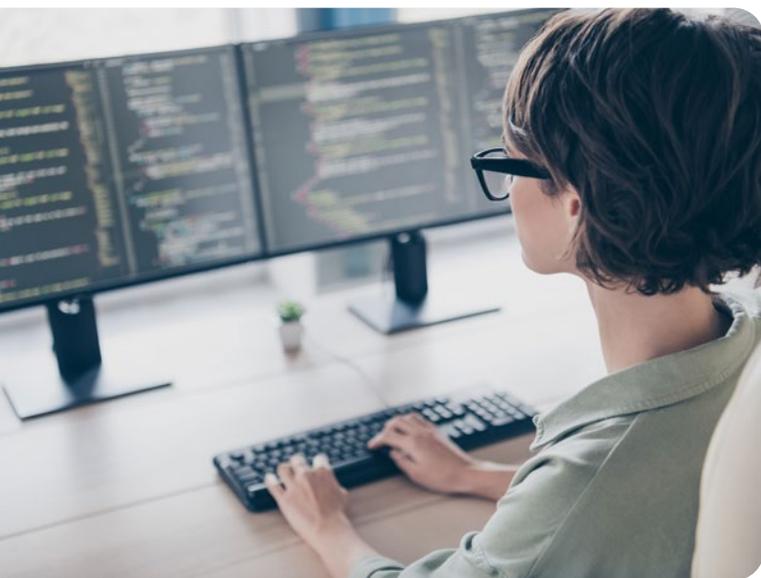
7.3 Désinformation et technologies de la parole

La parole est un des plus importants outils de communication entre humains, et depuis toujours un moyen privilégié de désinformation. Mais récemment, les progrès des technologies de la parole en ont grandement multiplié l'impact.

7.3.1 État des lieux

Le tableau de la section 7.3.3.1 présente plusieurs technologies de la parole et de l'audio qui ont d'autres buts que la désinformation, mais qui peuvent être utilisées à cette fin. Un tableau à l'Annexe V – Désinformation et audio : références scientifiques et techniques, démonstrations fournit les références scientifiques et techniques pour les technologies présentées au tableau ci-dessous, ainsi que des liens vers des démonstrations sonores.

Parmi les exemples présentés dans le tableau, les plus marquants sont la conversion de voix, qui permet d'usurper l'identité d'une personne, et la synthèse de la voix à partir du texte, qui permet de créer des vidéos en faisant dire ce qu'on veut à une personne donnée.



7.3.2 Le développement de la technologie

Certaines de ces technologies ont une longue histoire, les premiers essais de synthèse vocale par ordinateur datant des années 1960. Mais jusqu'à récemment, elles obtenaient des résultats de qualité insuffisante pour servir à la désinformation, avec un coût élevé de mise en œuvre. À partir de 2015-2017, la qualité a atteint un niveau suffisant (voir, par exemple, le projet Obama de l'Université de Washington en 2017¹⁰³), mais l'utilisation pour la désinformation était encore limitée par la quantité de données nécessaire. Les progrès plus récents, dont le deep learning, ont éliminé beaucoup de contraintes techniques, notamment celle de la quantité de données. Une synthèse vocale d'une qualité qui exigeait plusieurs heures d'enregistrement d'une voix il y a cinq ans, peut s'obtenir maintenant avec un échantillon de quelques secondes de voix.

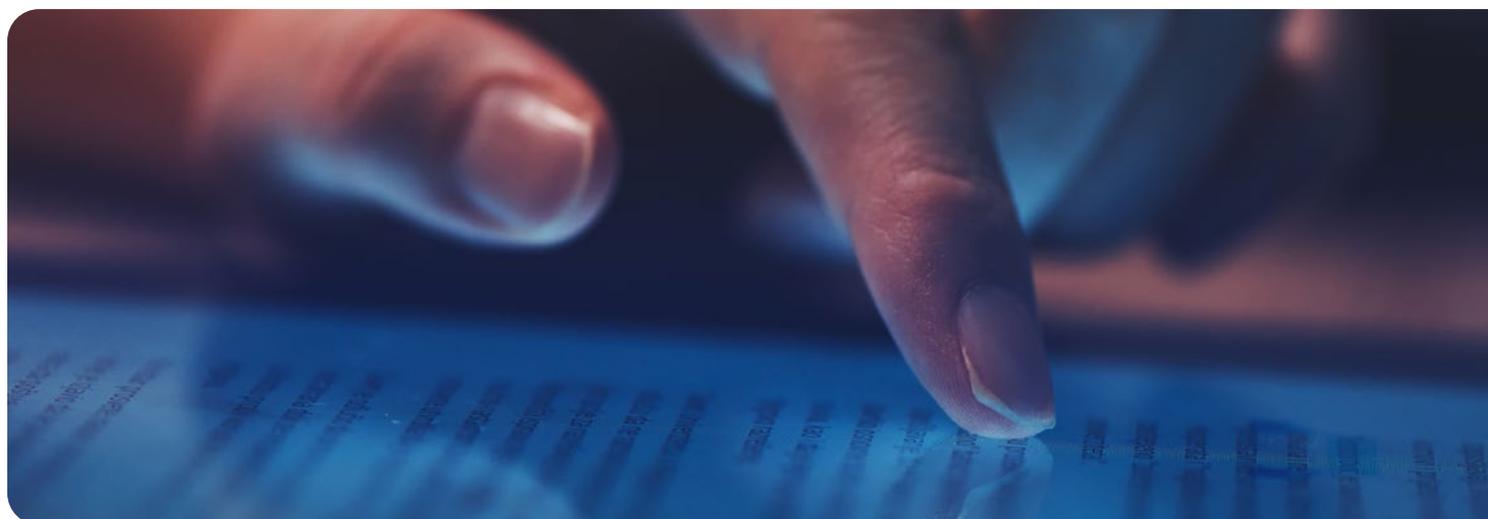
7.3.3 La disponibilité des outils et leur facilité d'utilisation

Parallèlement, on observe le développement de nombreux outils gratuits et faciles à utiliser qui mettent rapidement en œuvre les résultats de la recherche, de sorte que maintenant, ces technologies sont à la portée de n'importe qui. Par exemple, les services offerts (pour un coût minime) par ElevenLabs permettent de créer une voix synthétique qui peut dire n'importe quel texte avec la voix d'une personne donnée, à partir d'une ou deux minutes d'enregistrement de parole de cette personne. Cette entreprise est à l'origine de plusieurs des exemples de faux dénombrés dans le tableau ci-dessous.

¹⁰³ <http://grail.cs.washington.edu/projects/AudioToObama/>

7.3.3.1 Les technologies audios utilisables pour la désinformation

Technologie	Description	Exemples d'utilisation en désinformation
Synthèse vocale à partir du texte (text-to-speech)	Générer du nouveau contenu avec une voix donnée, à partir de texte	Voix parlée : Créer des entrevues où politiciens ou célébrités tiennent des propos qu'ils n'ont jamais tenus (e.g. fausse entrevue de Justin Trudeau par Joe Rogan). Créer de faux discours ou de fausses déclarations (e.g. Joe Biden, Zelenski). Voix chantée : Créer des chansons usurpant la voix de chanteurs connus sans leur accord (deepfake vocals, e.g. le cas de Drake, The Week'nd).
Conversion de voix (speech-to-speech)	Modifier les caractéristiques de la voix parlée ou chantée (identité, émotion, etc.) en conservant le contenu	Tromper sur l'état émotionnel ou physique en simulant un problème d'alcool, de drogue ou neurologique. Exemple : Vidéo de Nancy Pelosi . Usurper une identité vocale à des fins trompeuses : pour contourner la vérification par biométrie vocale, ou pour générer de la parole afin de s'en servir avec d'autres technologies de ce tableau.
Édition en place	Modifier une partie du contenu parlé sans changer le reste	Altérer le sens d'une déclaration parlée en n'en modifiant qu'une partie pour rendre la manipulation plus difficile à détecter.
Conditionnement par le texte	Générer avec voix, style ou émotion décrit en texte	Faciliter l'utilisation d'une autre technologie de ce tableau en remplaçant le travail de programmation par une simple description écrite du résultat voulu.
Post-synchronisation (lèvres-parole)	Générer la parole en synchronisation avec les lèvres dans la vidéo	Changer le contenu parlé d'une vidéo sans modifier le contenu visuel, pour obtenir une vidéo convaincante.
Post-synchronisation (parole-lèvres)	Modifier le mouvement des lèvres de la vidéo pour correspondre à la parole	Appliquer un enregistrement parlé sur du contenu visuel existant, ou obtenir une vidéo convaincante. Exemple : Faux discours Obama (2017) .
Conversion de voix en temps réel	Converser en direct ou animer un avatar « live »	Usurper une identité pendant une conversation pour faciliter une escroquerie, en se faisant passer pour un parent proche, pour le patron d'un employé, etc. Fraude par usurpation de voix . Contourner la vérification par biométrie vocale même si elle requiert une interaction en direct.
Rehaussement par diffusion	Débruitement, déréverbération, super-résolution, séparation de sources	Faciliter l'extraction d'une voix de bonne qualité à partir d'enregistrements existants pour l'utiliser avec une autre technologie de ce tableau.
Conversion texte-musique (text-to-music)	Générer de la musique à partir de texte décrivant le style, les paroles, ou des séquences MIDI.	Faciliter la production d'œuvres usurpant le style musical ou la voix d'artistes existants en remplaçant le travail de programmation par une simple description écrite du résultat voulu.



7.3.4 Les moyens de lutte

L'usurpation d'identité par la voix est l'objet de travaux de recherche en biométrie vocale, notamment grâce aux campagnes d'évaluation ASVSpooof (Automatic Speaker Verification Spoofing) qui se tiennent aux deux ans depuis 2015. Pour les systèmes de vérification du locuteur, le problème consiste à détecter si un échantillon de voix est authentique (produit par un humain), plutôt qu'un enregistrement (replay) ou une voix produite par une des technologies du tableau ci-dessus, c'est-à-dire par synthèse ou par conversion de voix. En principe, ces méthodes de détection de spoofing s'appliquent au cas plus général des deepfakes.

Cette détection doit s'effectuer à l'aveugle, c'est-à-dire sans connaissance des techniques de synthèse employées, ni aucun exemple utilisant ces techniques lors de l'apprentissage. Elle doit aussi rester fiable en présence de bruit et pour une variété d'environnements bruyants et réverbérants. Depuis 2021, ASVSpooof comporte un volet deepfakes, où il s'agit de vérifier l'authenticité d'enregistrements de voix de type web, qui diffèrent des enregistrements téléphoniques par leur bande passante et le type de compression numérique.

Les résultats de la campagne 2021 ont montré que les systèmes de détection de spoofing avaient des performances impressionnantes pour la vérification du locuteur, mais beaucoup moins pour les deepfakes. ASVSpooof 2024 confirme cet état de fait, mais démontre quand même des progrès importants dans la détection des deepfakes.

Les jeux de données sont essentiels pour la mise au point de ces détecteurs; on trouvera des exemples à l'Annexe VI (ASVSpooof, AVSpooof, FakeAVCeleb).

On peut aussi mentionner certaines approches pour déterminer l'authenticité d'une vidéo qui reposent sur l'analyse de la synchronisation entre la voix et le visage, ou la voix et les lèvres.

7.3.5 Le futur de la désinformation en parole

L'évolution des technologies audio vise maintenant des résultats encore plus réalistes, avec l'introduction d'émotions plus subtiles et d'éléments non-verbaux tels que les rires, hésitations, exclamations, soupirs, etc. Ces progrès vont rendre la désinformation plus difficile à détecter, tout en étant plus facile à produire, avec moins de données, en moins de temps.

Les résultats des campagnes ASVSpooof depuis 2015 permettent aussi de prédire une amélioration correspondante des contre-mesures et techniques de détection. Mais on peut s'attendre à ce qu'il reste toujours un décalage entre les technologies susceptibles d'être utilisées pour la désinformation et celles qui permettent de la détecter.



7.4 Stratégies additionnelles ou complémentaires

À Retenir:

- » L'authentification de la source d'une image est à l'étude, mais elle exige une mobilisation de plusieurs types d'acteurs.
- » Le tatouage numérique existe depuis longtemps, mais il est peu répandu et il ne règle pas tous les problèmes.
- » Un dépôt d'images signalées est une solution intéressante pour un type de DVM, pourvu qu'il soit sous la supervision d'une entité ouverte et transparente.
- » Une méthodologie pour le design de systèmes de vérification de faits basés sur l'analyse de données multimodales commence à prendre forme.
- » Les chercheurs/développeurs de modèles d'IA peuvent jouer un rôle en restreignant l'utilisation de leurs créations à des scénarios positifs, à condition que leur participation soit massive.

7.4.1 Authentification/provenance

Certaines propositions de lutte à la désinformation vont loin. Paris & Donovan suggèrent de mettre de côté la neutralité du Web pour que les sources de production de contenu soient authentifiées, de façon à retracer plus facilement les sources de désinformation.

Plus réalistement, une infrastructure parrainée par la Content Authenticity Initiative (CAI¹⁰⁴) est en cours de déploiement pour assurer la traçabilité de documents visuels (image, vidéo), à partir de la capture (caméra) jusqu'à la visualisation : un usager aurait accès aux métadonnées et à la liste de modifications appliquées à l'image/vidéo. On parle d'un "secure end-to-end system for digital content provenance through open-source development, cross-industry collaboration, and interoperability of tools".

Fait intéressant, une extension aux modèles génératifs est en cours de développement : "The CAI standard will enable creators to show whether, and how, Generative AI technology was used to create a piece of content."¹⁰⁵ En parallèle, la Coalition for Content Provenance and Authenticity (C2PA) pilote la création de standards techniques pour assurer la certification de la source/provenance d'un document média.

Évaluation:

Il va sans dire que le succès de cette initiative dépend de l'adhésion d'un grand nombre d'industries (fabricants de capteurs, fabricants de matériel, éditeurs de logiciels, utilisateurs intensifs comme les médias, etc.). Par chance, plusieurs joueurs importants (Adobe, Microsoft, Intel, BBC, etc.) sont engagés dans la mise en place de l'initiative. C'est donc un dossier à suivre.

7.4.2 Tatouage numérique

Le tatouage numérique (watermarking) consiste à camoufler dans une image, un texte ou un enregistrement audio une signature perceptuellement invisible¹⁰⁶ dont on peut par contre vérifier la présence à l'aide d'un logiciel. C'est une famille de techniques utilisées depuis longtemps dans la gestion de la propriété intellectuelle des œuvres visuelles (p. ex. afin d'identifier l'utilisation non autorisée d'images et de photos). En ce sens, le tatouage numérique est un élément de solution d'une infrastructure de provenance (décrite à la section précédente), puisqu'il peut être robuste aux manipulations bénignes comme la compression d'image, l'ajustement de contraste ou de luminosité, etc., mais fragile face aux manipulations comme les deepfakes.¹⁰⁷ Cette approche peut être vue comme une méthode de détection de DVM puisqu'un tatouage altéré indique que l'image a été modifiée. L'approche est applicable aux images, textes et audios, produits par les modèles génératifs.

Finalement, alors que les algorithmes de tatouage traditionnels sont basés sur des techniques de traitement de signal, l'IA devient aussi un contributeur de solutions.¹⁰⁸ Récemment, une équipe de l'Université de

¹⁰⁴ <https://contentauthenticity.org/>

¹⁰⁵ <https://blog.adobe.com/en/publish/2022/10/18/bringing-next-wave-ai-creative-cloud>

¹⁰⁶ Il existe aussi des tatouages visibles, mais leur utilité dans le contexte de lutte à la DVM est limitée par le fait qu'ils peuvent être facilement enlevés ou imités.

¹⁰⁷ <https://arxiv.org/pdf/2204.01960.pdf>

¹⁰⁸ <https://www.weforum.org/agenda/2019/06/to-thwart-deep-fakes-add-these-digital-watermarks>

Chicago a proposé Glaze, une technique apparentée à du tatouage qui protège les images numériques de la copie de style que certains modèles d'IA peuvent accomplir. On pense aussi à PhotoGuard¹⁰⁹, proposé par une équipe du MIT, qui injecte des perturbations imperceptibles dans une image afin de l'immuniser contre des manipulations d'édition faites par certains algorithmes d'IA, et également à SynthID¹¹⁰ de Google DeepMind qui affirme pouvoir placer un tatouage dans plusieurs types de documents : texte, audio, et image/vidéo (dans ce dernier cas, le tatouage persiste même si les images/vidéos sont manipulées).

Évaluation:

Le tatouage numérique fait partie de la boîte à outils de la lutte à la désinformation visuelle et multimodale (quoique cette affirmation est contestée¹¹¹), mais il convient de garder les points suivants en tête :

- » Il existe beaucoup de matériel non tatoué sur l'Internet qui pourrait être une de cheapfake.
- » Le tatouage n'est pas efficace pour les cas d'images non manipulées, mais utilisées hors contexte (recontextualisation, image repurposing).
- » Le tatouage des images produites par des modèles génératifs semble être une voie intéressante quoiqu'encore imparfaite¹¹², mais ceci suppose une adhésion volontaire ou forcée¹¹³ d'une forte majorité de fournisseurs de services IA, ce qui est difficile à envisager à partir du moment où ces modèles sont facilement accessibles au public.

7.4.3 Responsabilité des chercheurs/développeurs

Le sujet a été abordé à une section précédente. Les pressions sociales augmentent pour sensibiliser les chercheurs et développeurs en IA à se questionner sur l'utilisation potentiellement malveillante des modèles qu'ils partagent publiquement sur des sites comme GitHub ou des « zoos de modèles ». Considérant que le

code source de ces modèles circule librement, il peut être difficile voire impossible d'empêcher des acteurs malveillants d'entraîner ces modèles à produire de la désinformation (pour peu qu'ils aient les capacités de calcul requises pour réaliser les entraînements). Les grands joueurs de l'industrie (Google, OpenAI, etc.) ont la capacité d'héberger leurs modèles derrière des API qui servent de barrières pour du contrôle d'accès, ce qui n'est pas nécessairement le cas pour les plus petits joueurs. L'enjeu est d'autant plus réel qu'il est reconnu que la mise en place de barrières de sécurité autour de modèles génératifs comme ChatGPT est d'une grande complexité¹¹⁴ et exige une approche multidisciplinaire rigoureuse.

Deux interventions possibles sont décrites ici:

- » Une signature peut être injectée dans un modèle pré-entraîné afin de détecter son implication dans la production de matériel de DVM.
- » Les jeux de données importants, qui demeurent le nerf de la guerre en IA, peuvent être rendus radioactifs, de façon à vérifier s'ils ont été utilisés pour entraîner un modèle suspecté d'être producteur de désinformation.

Évaluation:

Le succès de cette mesure semble lié à la participation massive de la communauté de développeurs, comme le disent Böswald et al. (2022), "AI service providers need to agree to incorporate invisible watermarks or metadata in their product outputs". En plus de la dimension humaine, il convient de se demander 1) si les solutions de suivi des modèles et jeux de données sont suffisamment robustes pour jouer un rôle utile et 2) quelles conséquences concrètes (ex. légales) pourrait subir un individu s'il était avéré que cet individu a utilisé un modèle pré-entraîné ou un jeu de données à de mauvaises fins.

¹⁰⁹ <https://news.mit.edu/2023/using-ai-protect-against-ai-image-manipulation-0731>

¹¹⁰ <https://deepmind.google/technologies/synthid/>

¹¹¹ <https://www.eff.org/deeplinks/2024/01/ai-watermarking-wont-curb-disinformation>

¹¹² <https://www.nature.com/articles/d41586-024-03418-x>

¹¹³ <https://arstechnica.com/information-technology/2022/12/china-bans-ai-generated-media-without-watermarks/>

¹¹⁴ <https://arxiv.org/abs/2406.02622v1>

7.4.4 Dépôts d'images/vidéos signalées

La constitution de bases de données d'images/vidéos associées dans le passé à de la DVM fournit également un moyen de détecter une campagne de désinformation en cours. Ces bases de données serviraient de fondation à des sites de réfutation d'allégations tels DBKF¹¹⁵. L'idée est basée sur le principe qu'une image peut se faire attribuer un « hash perceptuel », c'est-à-dire une sorte de clé unique dont la valeur est une fonction du contenu de l'image, et ce, même si elle subit des transformations (rognages, rotations, modifications de couleurs, etc.). Ainsi, si on montre qu'une image accompagnant une publication en processus d'analyse a un hash perceptuel semblable à celui d'une image déjà étiquetée comme DVM, il est raisonnable d'étiqueter cette publication comme étant potentiellement de la DVM. Une base de données centralisée et publique qui contiendrait ces clés permettrait ainsi à l'industrie des médias sociaux d'augmenter sa capacité de détection de DVM. L'idée n'est pas nouvelle, on la retrouve dans la technologie PhotoDNA développée par Microsoft à partir de 2009 et cédée à un organisme de lutte à l'exploitation des enfants. De même, l'organisme GIFCT¹¹⁶ gère une base de données de hashes perceptuels pour faciliter l'identification de contenu terroriste ou d'extrémisme violent. Un système expérimental, DejaVu, a été développé à Cornell sur le même modèle,

mais plus spécifiquement pour la DVM : “[to] allow journalists to flag problematic images and to be notified via a browser-based extension when such problematic images surface” (Thomson et al, 2022). Il ne semble toutefois pas utilisé en situation réelle.

Évaluation:

En théorie, le concept de dépôt d'images/vidéos signalées est intéressant, mais des questions subsistent. Langguth et al. (2021) soulignent le risque d'abus immense qu'entraîne le « central control over what is “truth”, even in the benign form of manipulation detection », sans parler des possibilités de pressions politiques. Les auteurs suggèrent plutôt une approche ouverte, décentralisée et accessible à tous.

7.5 Vérification de faits

La vérification de faits automatisée est un sujet d'étude établi depuis au moins une quinzaine d'année et elle est modélisée comme un pipeline de tâches du traitement automatique des langues naturelles telles que « Détection d'affirmation », « Récupération de preuves », « Prédiction du verdict », etc.

Elle est une solution efficace pour des situations de vérification temps réel où le volume de requêtes est élevé. La collaboration d'humains (utilisateurs de réseaux sociaux, experts) à travers une approche

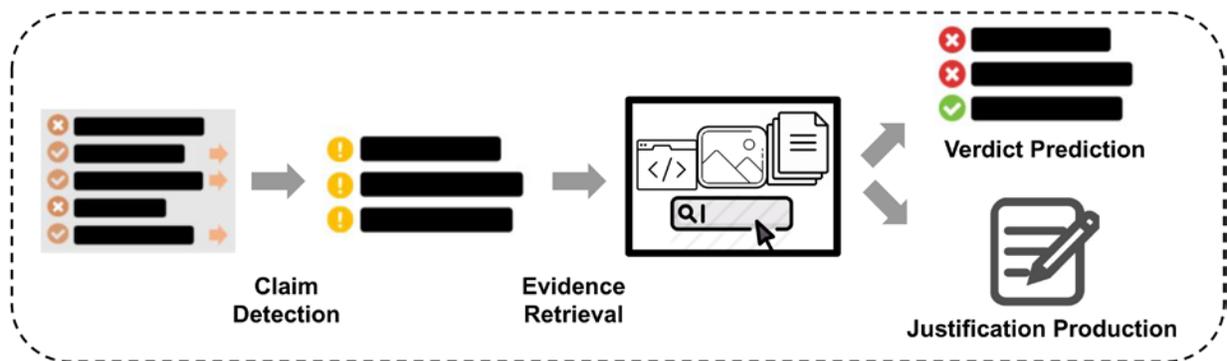


Figure 7 – Source: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00454/109469/A-Survey-on-Automated-Fact-Checking

¹¹⁵ <https://weverify-demo.ontotext.com/%23/about>

¹¹⁶ <https://gifct.org/>

humain-dans-la-boucle permet d'améliorer la performance des systèmes de vérification de faits automatisée et aussi de réduire le volume de tâches des vérificateurs. Cela dit, la vérification de fait automatisée n'est pas une panacée : les vérificateurs ont leurs biais, leurs priorités et leur perception de la limite entre censure et liberté d'expression. C'est la raison donnée par Meta pour mettre fin à son programme de vérification de faits aux États-Unis pour ses plateformes Facebook, Instagram et WhatsApp¹¹⁷. Par ailleurs, cette collaboration de vérification de faits pourrait ne pas répondre adéquatement au volume grandissant de désinformation auquel nous sommes exposés, surtout depuis le boom des modèles génératifs.¹¹⁸

Ce sont les progrès en traitement des langues naturelles (NLP en anglais) qui ont favorisé l'émergence d'outils de vérification de faits automatisée, lesquels sont capables de manipuler la représentation textuelle d'affirmations à tester. La même logique s'applique aux documents de nature multimodale (par exemple Akhtar et al., 2023), et comme on assiste depuis deux ou trois ans à une augmentation importante de contributions scientifiques qui proposent des algorithmes pour les différentes étapes du pipeline, on peut s'attendre à un développement d'outils « bien conçus » dans un avenir rapproché.

Évaluation:

Il est certain que le type de cadre méthodologique proposé par Akhtar et al (2023) ne peut que faciliter le développement d'outils complets de vérification de faits à partir d'allégations combinant plusieurs modalités. Soulignons que plusieurs enjeux soulevés, par exemple, par ces mêmes auteurs vont de pair avec les enjeux propres au déploiement commercial de solutions basées sur l'intelligence artificielle en général, notamment l'explicabilité (la capacité d'un système d'IA à expliquer sa décision) ou, de façon plus générale, l'IA de confiance.¹¹⁹ Ainsi, on peut facilement affirmer que certaines avancées futures en IA auront un impact positif non seulement sur la performance brute de ces systèmes automatisés de vérification de faits, mais également sur leur utilisabilité en contexte réel.

7.6 Jeux de données disponibles

À retenir :

- » La mise au point de solutions d'IA passe par une étape d'entraînement qui exige un ensemble riche et diversifié d'exemples. Un ensemble de ces exemples constitue un jeu de données.
- » Plusieurs jeux de données publics ont été répertoriés, principalement pour la détection d'artefacts créés par la manipulation d'images. Par contre, il existe peu de jeux contenant des images produites par des modèles génératifs.
- » Notons par ailleurs l'apparition de jeux de données constitués de paires image-texte, essentiels par exemple au développement d'algorithmes de détection de désinformation par recontextualisation.

Les jeux de données sont des collections d'images ou de vidéos utilisées pour le développement d'outils basés sur l'intelligence artificielle. Comme il existe différents types de DVM, il est important pour la communauté scientifique de compter sur une grande variété de jeux de données adaptés aux problèmes étudiés. Or, on constate que certains problèmes bien posés, comme la détection d'altérations dans les images, bénéficient de l'existence de riches jeux de données annotés, alors que pour d'autres, comme la vérification de cohérence sémantique entre texte et image, ces jeux sont plus rares, plus petits ou plus contraignants à exploiter vu leur licence d'utilisation.

Les jeux de données peuvent être caractérisés de plusieurs façons :

- » Selon le type de document qu'ils contiennent : texte, parole, images ou vidéos.
- » Selon que les données sont des manipulations de documents existants ou des documents synthétiques, c'est-à-dire générés par l'IA.

¹¹⁷ <https://www.ledevoir.com/monde/etats-unis/830137/meta-annonce-fin-programme-verification-faits-etats-unis>

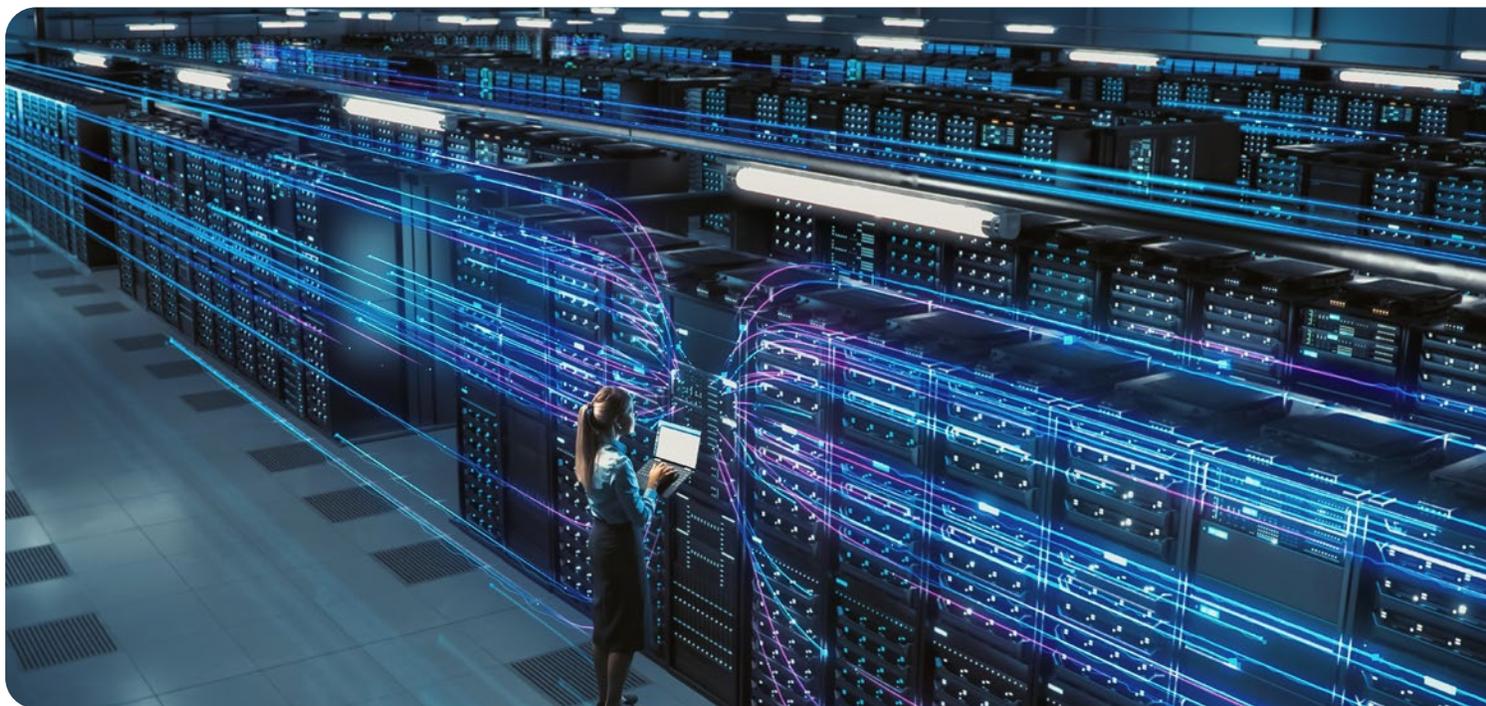
¹¹⁸ Ünver. Emerging technologies and automated fact-checking: Tools, techniques and algorithms. 2023. Cyber Governance & Digital Democracy Programme

¹¹⁹ <https://www.confiance.ai/>

- » Selon que les données sont unimodales (seulement du visuel, seulement de l'audio, etc.) ou multimodales (combinaison image-audio, combinaison vidéo-texte, etc.).

Le tableau à l'Annexe VI – Jeux de données pour la lutte à la DVM fait la liste des jeux les plus souvent référencés dans les revues de littérature scientifique. L'inventaire montre qu'il existe une bonne variété de jeux de données contenant des exemples de manipulation d'images/vidéos ou d'audio existants : ceux-ci ont permis de développer les méthodes référencées à la section 7.2.3. Les jeux combinant texte et image sont aussi bien représentés (ex. CIFAKE), ainsi que ceux créés pour lutter contre les hypertrucages (deepfakes), comme le jeu de Facebook (DFDC) par exemple. Par contre, le nombre de corpus créés pour la détection de contenu généré par les modèles texte-à-image ou texte-à-vidéo est très bas, sinon inexistant, probablement à cause de l'apparition très récente de ces modèles dans le domaine public.

Fait à noter, on assiste depuis quelque temps à l'apparition soutenue de jeux de données spécialisés pour le problème de détection de désinformation/mésinformation dans des associations texte-images. Par exemple, le NewsCLIPpings Dataset¹²⁰ compte près d'un million de fausses associations image-texte pour simuler de la désinformation par recontextualisation : les images et le texte sont réels, tirés d'articles de journaux, mais un algorithme d'IA les a réassociés intelligemment pour créer le jeu de données. Un projet piloté, entre autres, par Google a donné lieu à la mise sur pied d'un jeu de données important appelé AMMeBA¹²¹ qui regroupe des dizaines de milliers d'affirmations, annotées manuellement, qui existent sous plusieurs formes : images avec contenu ou contexte manipulé, texte tendancieux superposé sur une image, captures d'écrans, etc. Ces jeux de données sont une contribution absolument cruciale pour l'effort de développement de solutions d'IA reliées à la lutte à la désinformation visuelle et multimodale, parce qu'ils permettent non seulement d'entraîner des modèles de détection, mais aussi d'en évaluer l'efficacité et le potentiel.



¹²⁰ <https://arxiv.org/pdf/2104.05893>

¹²¹ <https://arxiv.org/abs/2405.11697v1>

7.7 État des outils de détection

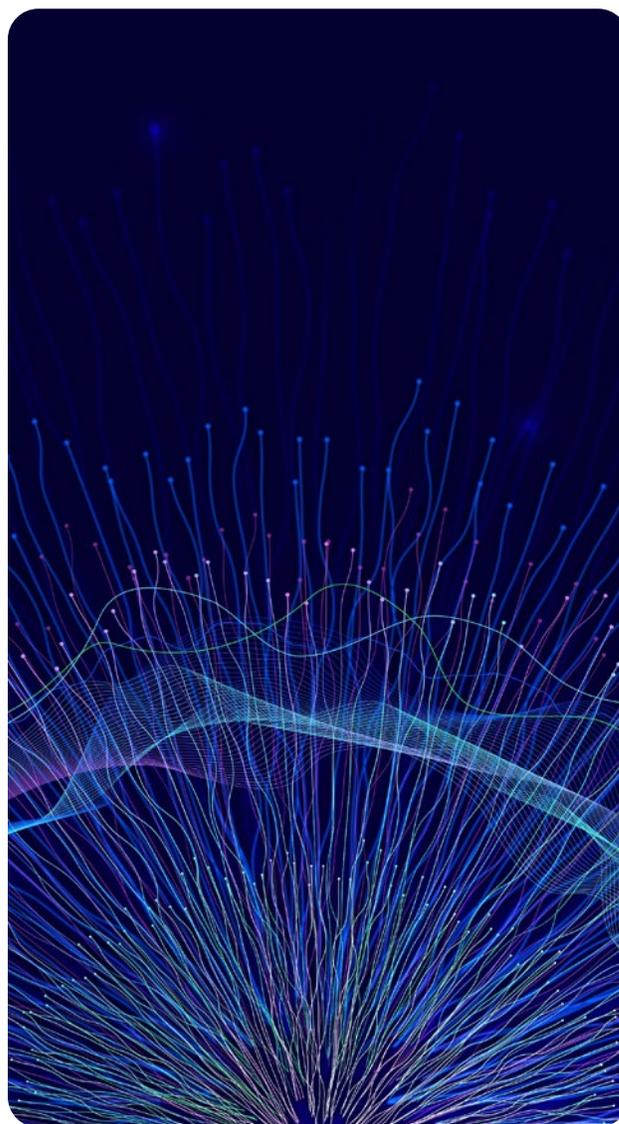
À retenir :

- » L'offre de produits est variée, mais somme toute restreinte, plutôt récente et développée par des organisations aux profils différents (labos universitaires, grandes entreprises, start-ups, etc.); consolidations à prévoir?
- » Beaucoup de produits visent la détection de deep-fakes ou de cheapfakes.
- » Le multimodal semble mal pris en compte.

Cette section fait un inventaire non exhaustif des produits et solutions conçus pour détecter la DVM.

On constate que l'offre commerciale s'est ajustée face à l'émergence des technologies de deepfakes, puisque plusieurs fabricants de produits, allant de grandes firmes internationales telles que Microsoft, Adobe, Intel à des start-ups (Duckduckgoose) ou des laboratoires de recherche (Université de Buffalo Media Forensic Lab), proposent différentes solutions pour les détecter. Ces outils peuvent être commerciaux, adaptés aux besoins des entreprises, à ceux des agences gouvernementales (via la détection de falsification de document ou d'usurpation d'identité), ou encore tournés vers les besoins des médias, notamment dans la vérification d'informations (fact-checking). Enfin, on trouve des outils gratuits et open-source manipulables par des particuliers qui peuvent ainsi par eux-mêmes analyser du contenu (Deepfake-o-meter, Deepware, le plugin InVID & WeVerify). Par contre, il existe peu de solutions dédiées à la vérification de provenance, possiblement parce qu'elle requiert un effort considérable de développement et de coordination pour l'établissement d'une infrastructure d'authentification de la provenance d'images et de vidéos.

Un autre aspect peu considéré jusqu'ici concerne l'interopérabilité de ces outils. Il serait judicieux qu'ils partagent les mêmes formats de rapports après analyse de documents potentiellement vecteurs de DVM. La norme MediaReview¹²², en cours d'adoption, vise justement à combler une partie de cette lacune en proposant une taxonomie pour la description de manipulations d'images. Il reste à voir si cette taxonomie est adéquate pour tout type de DVM, incluant les nouvelles formes permises par les plus récents modèles d'IA.



¹²² <https://reporterslab.org/what-is-mediareview/>

Le tableau qui suit dresse un portrait des principaux produits identifiés pendant la prise d'inventaire.

Produit	Date	Type (C,E,Ed)	Approche	Media (I,V,A)	Domaine d'utilisation	Notes	Propriétaire
Reality Defender	2018	C	AI	I,V,A	<ul style="list-style-type: none"> • AntiFraude • Mésinformation • Modération • Gouvernement • Média • Réseaux sociaux 	<ul style="list-style-type: none"> • Détection d'usurpation d'identité (vocale et visuelle) • Détection de manipulation d'images et de générations de visages • Détection de vidéos manipulées (visages + ensemble) 	AI Foundation (USA)
Sensity Deepfake detection	2018	C	AI	I,V	AntiFraude	<ul style="list-style-type: none"> • Détection de deepfakes, notamment ceux produits par des IA (GAN, stable diffusion...) • Authentification de falsifications faciales (documents officiels, d'identité, réseaux sociaux) • Détection de documents falsifiés 	Sensity (Pays-Bas)
Deepware	2018	E	AI	V	Général (détection de visages transformés)	<ul style="list-style-type: none"> • Code open-source pour la détection de visages générés/modifiés par IA 	Zemana Doo (Bosnie- Herzégovine)
Deep-fake-o-meter	March 2021	Ed,E	DL,ML	I,V,A	Général	<ul style="list-style-type: none"> • Plateforme pour pouvoir tester différents algorithmes de détection de deepfakes 	University at Buffalo, Media Forensic Lab (USA)
FakeCatcher	Juillet 2020	C,E	DL	I,V	Général	<ul style="list-style-type: none"> • Détection de deepfakes par analyse de bruits génératifs en s'appuyant notamment sur le flux sanguin (Photoplethysmography) 	Intel & Binghamton University (USA)
TruePic Vision		C		I	<ul style="list-style-type: none"> • Assurances • Services financiers • Externalisation de processus métiers 	<ul style="list-style-type: none"> • Analyse de manipulations photo • Analyse de provenance et de métadonnées de contenu 	TruePic (USA)
Microsoft Video Authenticator	2020	C		V	Général	<ul style="list-style-type: none"> • Détection de contenu artificiellement manipulé 	Microsoft (USA)
Paravision	Juin 2022	C	IA	I,V	Général	<ul style="list-style-type: none"> • Détection de deepfakes dans la perspective d'aider à l'authentification d'individus • Mise en place de protection pour s'assurer de la provenance de contenu 	Paravision (USA)
DeepDetector	2019	C	IA	I,V	Général	<ul style="list-style-type: none"> • Logiciel de détection de deepfakes visuels 	DuckDuck-Goose (Pays-Bas)
Sentinel	2020	C		I, V, A	Général	<ul style="list-style-type: none"> • Détection de contrefaçons via IA • Visualisation des manipulations effectuées 	Sentinel (Estonie)

Produit	Date	Type (C,E,Ed)	Approche	Media (I,V,A)	Domaine d'utilisation	Notes	Propriétaire
Fake news debunker by InVID & WeVerify	2016	C	IA, IPT	I,V	Général	<ul style="list-style-type: none"> • Plugin disposant de plusieurs outils : • Analyse de manipulations d'images ou de keyframes • Extraction de métadonnées de vidéos et images • Reconnaissance optique de caractères • Comparaison avec une base de données de deepfakes • Analyse de contenu de réseaux sociaux (Twitter/X, Instagram, Facebook) images et factchecking • Analyse provenance d'images dans le Web 	Vera.ai Project (EU)
CAI tools	2022	C		I,V	Général	<ul style="list-style-type: none"> • Intégration d'une protection au sein d'un contenu indiquant sa provenance via un outil open-source 	Adobe (USA)
TinEye	2008	C		I	Général	<ul style="list-style-type: none"> • Analyse la provenance d'une image dans le Web 	TinEye (Canada)
AIOrNot	2023	C, Ed	IA	I, A	Général	<ul style="list-style-type: none"> • Détection d'images et d'audios générées par l'IA 	AIOrNot (USA)
Hive Moderation	2021	C	IA	I, V, A	Général	<ul style="list-style-type: none"> • Modération automatisée du contenu par l'IA 	Hive.ai (USA)
BioID	2024	C	IA	I, V	Général (détection de visages transformés)	<ul style="list-style-type: none"> • La détection des images et vidéos générées ou manipulées 	BioID (Allemagne)
deepfakedetector.ai		C, E	IA	V, A	Général	<ul style="list-style-type: none"> • Détection de deepfakes via leur site web ou extension de navigateur ou leur API. Ils ont réclamé que la précision de leur outil est 92 % en gros 	(Royaume-Uni)
BlackBird.ai	2014	C	IA	I, V	Général	<ul style="list-style-type: none"> • La détection d'images et de vidéos générées par l'IA et il donne un score de confiance 	BlackBird AI (USA)

Légende :

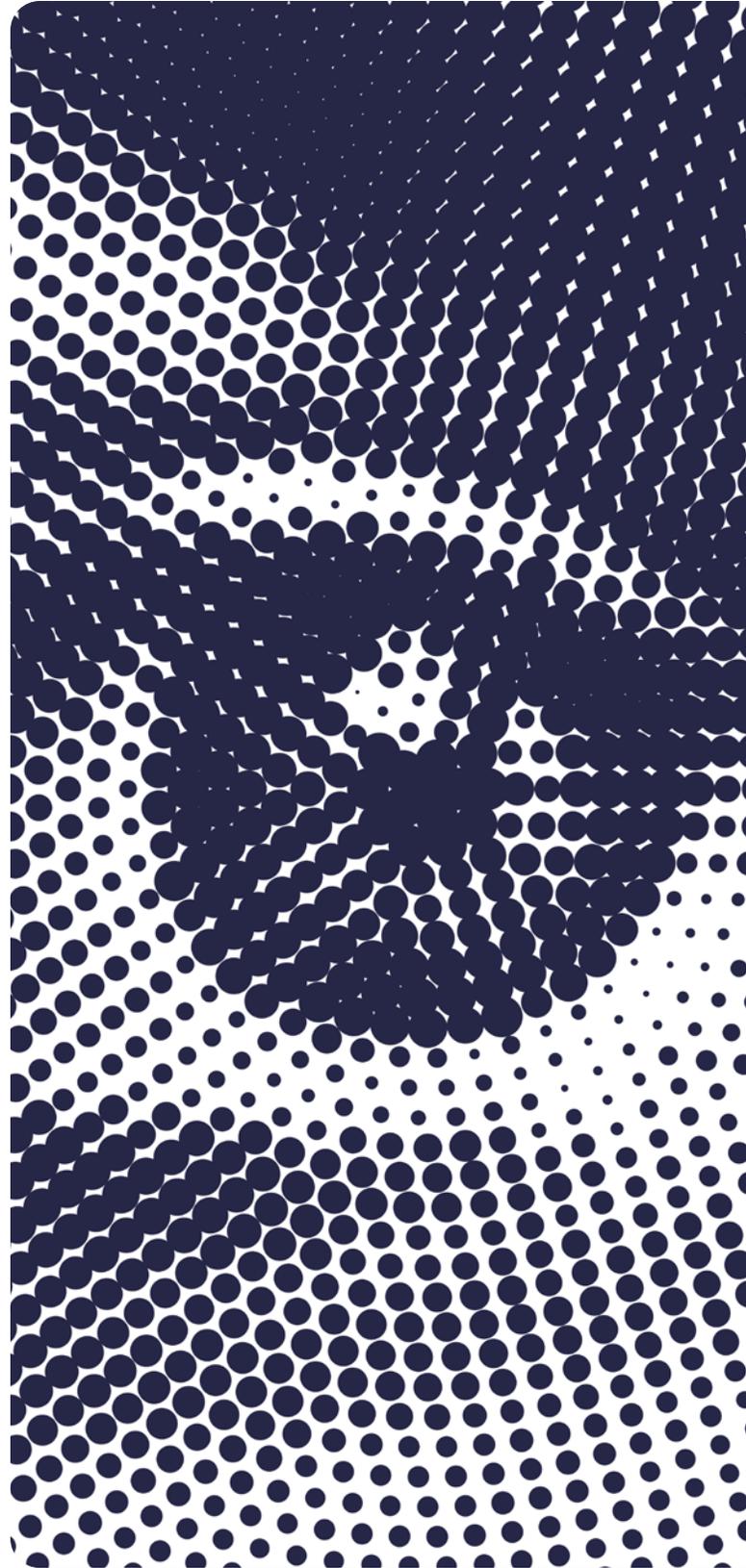
- » **Type :** (C) Commercial, (E) Expérimental, (Ed) Éducatif
- » **Approche :** Apprentissage machine (ML), Apprentissage profond (DL), Techniques de traitement d'image (IPT), Analyse statistique (Statistiques)
- » **Types de médias :** Image (I), Vidéo (V), Audio (A)

Color Key:

- » **Rouge :**
Outils conçus pour la détection des deepfakes/cheapfakes.
- » **Jaune :**
Outils axés sur l'analyse de la provenance du contenu.
- » **Bleu :**
Outils ajoutant une protection au contenu pour en vérifier l'authenticité

7.8 Lutte à la DVM : Discussion

Plusieurs enjeux dominent le développement de ces algorithmes de détection de DVM. Un premier enjeu concerne leur performance : comme tous les algorithmes de classification, ils commettent des erreurs de type faux positif et faux négatif. Deuxièmement, il existe un besoin croissant du côté des utilisateurs de comprendre pourquoi un algorithme recommande l'étiquetage d'une image/vidéo comme étant un véhicule de la désinformation, ce qui rejoint le courant de l'IA explicable/interprétable (explainable AI¹²³). Un enjeu supplémentaire touche à la relation de type « course aux armements » entre producteurs et détecteurs de DVM, avec les détecteurs pourchassant les producteurs. Certains (ex. Langguth et al., 2021) vont jusqu'à préconiser un certain contrôle sur la diffusion des algorithmes de détection, afin d'empêcher les producteurs d'y avoir accès et de s'y adapter. Comme dernier point, ce rapport a jusqu'ici fait de nombreuses références à l'IA générative, ce qui met en évidence l'ampleur du rôle que joue cette technologie dans le paysage actuel de la désinformation visuelle et multimodale. Avec l'IA générative, nous sommes à un tournant dans la lutte à la désinformation. Ces technologies permettent non seulement de créer des contenus visuels et multimédias sophistiqués à partir de simples descriptions textuelles, mais elles offrent aussi de nouvelles méthodes pour détecter et contrer ces créations malveillantes, comme nous le verrons à la prochaine section.



¹²³ <https://www.darpa.mil/program/explainable-artificial-intelligence>

8. LUTTE À LA DVM : L'INFLUENCE DE L'IA GÉNÉRATIVE

À retenir :

- » L'arrivée des modèles génératifs capables de créer facilement du contenu visuel inédit, à partir de texte, complexifie fortement la lutte à la DVM.
- » On note cependant que ces mêmes outils de production peuvent aussi servir à la détection de DVM.

Depuis deux ans, l'intelligence artificielle générative s'est immiscée dans plusieurs sphères de la vie du travail. Un assistant comme Microsoft Copilot améliore la productivité des rédacteurs de rapports en générant du texte de qualité à partir d'idées ou en résumant des documents au contenu substantiel. Les outils d'aide au développement logiciel sont maintenant monnaie courante et de grands joueurs comme Google ou Meta les utilisent abondamment.¹²⁴ Des logiciels de graphisme ou d'édition visuelle comme Photoshop intègrent des modèles d'IA pour stimuler la créativité des designers.¹²⁵ Même la vie quotidienne est affectée par la grande disponibilité de ces modèles sur plusieurs plateformes (web, téléphones intelligents, etc.) qui offrent une manière de rechercher et consommer de l'information différente de celle proposée par les moteurs de recherche internet traditionnels.¹²⁶

Avec cette montée en puissance, il importe de porter une attention toute particulière à l'influence de l'IA générative sur la production et la détection de désinformation visuelle et multimodale (DVM). On le devine aisément, les bénéfices de l'IA générative mis en lumière au paragraphe précédent peuvent aussi être exploités par des acteurs malveillants producteurs de DVM¹²⁷; cela dit, ces mêmes outils peuvent également servir à lutter contre le phénomène.

Cette section a pour but de faire le portrait de la situation des deux côtés du miroir, à savoir la production et la détection de DVM, même s'il est clair que ce portrait change rapidement au fil des innovations accélérées de ce domaine.

8.1.1 Génération d'images

Les avancées récentes en apprentissage profond ont profondément transformé la génération et la manipulation d'images. Auparavant, la création d'images synthétiques nécessitait la modification de contenu existant (ex. face swaps). Cependant, ces dernières années, de nouvelles techniques ont émergé, permettant de générer des images entièrement inédites à partir de descriptions textuelles, appelées text prompts. Dans cette section, nous discutons des modèles texte-image les plus connus.

GAN

Les réseaux antagonistes génératifs, mieux connus sous leur acronyme anglais GAN (Generative Adversarial Networks), sont parmi les premiers modèles génératifs mis au point grâce au développement de l'apprentissage profond. Étant constitués de réseaux de neurones, ils peuvent être entraînés à synthétiser des images de visages, ce qui les rend attrayants pour un producteur de DVM. Leur complexité d'utilisation fait qu'ils sont moins populaires, mais les principes au cœur de leur fonctionnement sont réutilisés dans d'autres types de modèles. La première vague de deepfakes est arrivée à la fin des années 2010 basé sur GAN, notamment Face Swap et DeepFaceLab.

Modèles génératifs basés sur la diffusion

Bien que les GAN puissent produire des images assez réalistes, ils sont complexes, exigeants, et leur entraînement est difficile à stabiliser. Face à ces limites, les chercheurs en synthèse d'image se sont tournés vers les modèles de diffusion, qui ont progressivement supplanté les GAN pour la génération d'images à partir d'entrées textuelles, audio et/ou visuelles. Ces modèles sont entraînés sur de vastes ensembles de données contenant des millions de paires « image-texte », leur permettant d'apprendre à générer des images détaillées à partir de simples phrases, appelées text prompts, proposées par un utilisateur. Leur capacité est telle que les images générées non seulement

¹²⁴ <https://www.theverge.com/2024/10/29/24282757/google-new-code-generated-ai-q3-2024>

¹²⁵ <https://www.forbes.com/sites/bernardmarr/2023/12/13/the-rise-of-generative-ai-in-design-innovations-and-challenges/>

¹²⁶ <https://www.forbes.com/sites/johnkoetsier/2024/11/02/chatgpt-search-vs-google-search-which-is-better/>

¹²⁷ <https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890>

contiennent les objets spécifiés dans les prompts (ex. « Voiture circulant sur une route enneigée »), mais les propriétés potentiellement complexes données dans les prompts sont aussi respectées, comme des caractéristiques physiques des objets à illustrer (couleur, texture), leurs relations spatiales (objet X à gauche de l'objet Y), etc.

La complexité de ces modèles est telle que seul un groupe très restreint d'entreprises ont les ressources techniques et computationnelles nécessaires pour les entraîner (cela dit, une fois entraînés, ils sont souvent utilisables sur des ordinateurs personnels).

- » OpenAI est une de ces compagnies, connue pour ses modèles ChatGPT et DALL-E. ChatGPT est doté de protections (guardrails) lui permettant de refuser certaines requêtes, par exemple des demandes de génération de deepfakes de candidats aux élections américaines.¹²⁸ Cependant, ChatGPT n'est pas le meilleur générateur d'images deepfakes, car côté réalisme, la qualité des images produites laisse à désirer.
- » Un autre leader du secteur est Stability AI dont le modèle Stable Diffusion est accessible au grand public gratuitement via un site web.¹²⁹ Comme c'est un modèle open-source, le code source peut être modifié pour faire tomber les protections contre les demandes de génération d'images sensibles. De plus, des plateformes comme CivitAI permettent à leurs utilisateurs d'adapter (fine-tune) des modèles avec leurs propres données (possiblement inappropriées), ce qui peut mener à des abus.
- » Un autre acteur important dans le domaine de génération des images est [Flux AI](#) développé par Black Forest Labs et intégré dans Grok, un modèle frontière développé par [xAI](#), l'entreprise de Elon Musk. Les versions de Flux sont open-source et sans filtrage, ce qui rend possible la production d'images de personnages publics.
- » Enfin, il faut mentionner un outil très populaire et efficace, [Midjourney](#). Cet outil est abondamment utilisé par le passé pour générer des images deep-

fakes, malgré les lignes directrices (et les protections automatisées) décourageant son utilisation à des fins malveillantes.

Il est important de réaliser que ces technologies évoluent très rapidement. Les grandes entreprises comme Google et Meta ont créé leurs propres générateurs d'images et les ont intégrés à leurs modèles frontières (Gemini pour Google et Meta AI pour Meta) avec de nombreuses mesures et protections qui rendent difficile voire impossible, la génération de deepfakes. Gemini refuse même de générer les images d'une personne inconnue, suite à une récente controverse.¹³⁰ Cependant, certaines petites entreprises semblent être impliquées dans une compétition pour livrer des modèles qui génèrent les images les plus réalistes, et ce, sans aucune mesure pour éviter la génération de deepfakes et de DVM. Cette inquiétude touche surtout aux modèles open-source (Stability AI, Flux).



Figure 8 - Une utilisation de l'IA générative à saveur politique. Tiré de <https://www.independent.co.uk/news/world/americas/us-politics/trump-ai-communism-harris-dnc-b2598303.html>

¹²⁸ <https://businessinsider.com/chatgpt-openai-denied-requests-generate-images-candidates-election-trump-harris-2024-11>

¹²⁹ <https://stablediffusionweb.com/>

¹³⁰ <https://www.forbes.com/sites/antoniopequenoiv/2024/02/26/googles-gemini-controversy-explained-ai-model-criticized-by-musk-and-others-over-alleged-bias/>

Évaluation:

En conclusion, la génération et la manipulation d'images, à partir des nouveaux modèles de langues combinés mixés au visuel sont en pleine effervescence. Déjà aujourd'hui, ils sont à la base d'une gamme d'outils abordables et faciles à utiliser par n'importe qui, et sont en pleine progression quant à leur capacité. Notons par ailleurs une plus grande sensibilité des entreprises face à une utilisation abusive de ces modèles : par exemple, DALL-E 3.0 intégré à ChatGPT refuse les requêtes faisant référence à des personnages publics.¹³¹ Ironiquement, il est possible d'utiliser ChatGPT pour améliorer la qualité du prompt textuel destiné à des modèles qui n'ont pas les mesures en place pour, par exemple, bloquer efficacement la génération de deepfakes.

8.1.2 Manipulation d'images

En plus de leurs capacités de génération d'images, ces modèles permettent aussi à un utilisateur de les modifier facilement, toujours à l'aide d'un prompt textuel. Changement d'arrière-plan, ajout d'objets, modification du style visuel, etc., toutes ces fonctionnalités de manipulation d'image sont maintenant possibles sans avoir à toucher à un seul pixel de l'image. Il suffit de composer la bonne description textuelle du changement à appliquer. On peut penser à Adobe et sa famille de produits (dont Photoshop) qui intègrent l'outil d'IA générative Adobe Firefly permettant aux créateurs d'ajouter ou de retirer des objets d'une image, de compléter intelligemment une partie d'une image, etc.¹³²

8.1.3 Modèles de génération et d'édition de vidéos

Les progrès rapides et spectaculaires avec les modèles de langue entraînés pour la génération d'images ont naturellement stimulé la recherche pour des équivalents en génération de vidéos. La complexité de ces modèles est encore plus grande à cause de la nécessité de préserver une cohérence temporelle dans le mouvement, non seulement en termes de contenu sémantique, mais également au niveau du rendu visuel (fluidité). Malgré toute cette complexité, plusieurs outils

sortis depuis moins d'un an ont étonné le monde par la qualité des vidéos qu'ils génèrent. En février 2024, OpenAI a annoncé l'arrivée de Sora, un modèle texte-à-vidéo capable de produire du contenu d'une qualité et d'un réalisme renversants (surtout quand on sait que le modèle génère ces vidéos à partir d'une seule description textuelle). Après une longue période à accès restreint, Sora est maintenant disponible pour tous.¹³³

Suite à cette annonce, d'autres entreprises ont aussi sorti leurs propres modèles de génération de vidéos, avec parfois des variantes comme la possibilité d'accompagner le prompt textuel d'une image pour guider la génération. Une recherche exhaustive a identifié plusieurs de ces outils dont les caractéristiques sont résumées dans le tableau de l'Annexe II – Outils de génération vidéo.

L'édition de vidéos à base de prompts textuels est aussi en développement : il suffit simplement de décrire les changements à appliquer à un objet pour que le modèle génératif les réalise. Ce genre de modèle génératif (par exemple Runway) est aussi connu comme modèle vidéo-à-vidéo. La manipulation vidéo est grandement accélérée en comparaison de l'approche traditionnelle d'édition où chaque trame devait être retouchée manuellement.

Évaluation:

Le domaine de la génération de vidéos est en pleine ébullition, et comme on l'a vu, les progrès en termes de qualité d'image et de longueur (nombre de trames) sont fulgurants. Déjà, des organisations politiques ont utilisé des images générées par l'IA pour des publicités partisans, comme on l'a souligné à la section 2.2. Il va sans dire que les progrès des outils de génération de vidéos (non seulement en qualité, mais également en facilité d'utilisation) ne peuvent qu'accélérer la production de contenu douteux, et c'est d'autant plus vrai que plusieurs d'entre eux sont soit disponibles en ligne et donc utilisables sans expérience de programmation, soit open-source et donc modifiables pour que leurs mesures de protection soient neutralisées.

¹³¹ <https://www.nbcnews.com/tech/chatgpt-rejected-image-generations-presidential-candidates-rcna179469>

¹³² <https://dataconomy.com/2023/05/24/how-to-use-photoshop-ai-generative-fill/>

¹³³ <https://openai.com/index/sora-is-here/>

8.2 Détection de désinformation par modèles multimodaux

À retenir :

- » Parce qu'exposés à de grands volumes d'images/ vidéos et de textes associés, les nouveaux modèles multimodaux (ex. texte-image) possèdent une compréhension approfondie des relations entre les objets, les scènes et les contextes dans les images.
- » Beaucoup de travaux sont en cours pour exploiter les capacités de « raisonnement » de ces modèles multimodaux, afin de les amener à identifier du contenu multimédia qui véhicule de la DVM.

La section précédente a illustré toute la puissance des modèles génératifs à produire à l'infini du contenu visuel et multimodal de très grande qualité, par exemple pour la création artistique, mais aussi possiblement pour disséminer de la désinformation; maintenant, ces mêmes outils peuvent-ils être mis au service de la lutte à la désinformation?

Les modèles multimodaux, tels que ChatGPT-4 Vision, sont capables d'interpréter des images pour en extraire des informations pertinentes et de produire une description textuelle de leurs résultats d'interprétation. Ces capacités viennent du fait qu'ils ont été entraînés sur des ensembles de données massifs contenant des images et

des descriptions textuelles associées. Ces ensembles de données permettent aux modèles d'apprendre à relier les informations visuelles et textuelles, développant ainsi une compréhension approfondie des relations entre les objets, les scènes et les contextes dans les images. Ainsi, ils sont en mesure de répondre à des questions qui font appel à des habiletés de reconnaissance d'objet (« Quels objets voit-on dans l'image? »), de raisonnement spatial (ex. « Combien de formes rouges y a-t-il à l'intérieur du carré bleu? »), et même de raisonnement basé sur le bon sens (common sense). Pendant la phase d'entraînement, l'exposition de ces modèles à toutes sortes de contenus (multilingues, multiculturels) leur confère des habiletés étonnantes qui sont encore en train d'être étudiées.¹³⁴ Ces mêmes aptitudes peuvent effectivement être exploitées pour détecter la désinformation : identification d'incohérences ou d'anomalies dans des contenus générés, détection de discordances sémantiques, de manipulations suspectes pouvant indiquer des deepfakes ou des images falsifiées, etc.

8.2.1 Détection d'images générées

L'idée a déjà été évoquée à la section 7.2.4 : questionner un modèle d'IA sur l'authenticité d'un texte, d'une image, ou plus généralement d'un document multimédia. C'est une idée qui est étudiée depuis plusieurs années, comme en fait foi l'image suivante tirée d'une collection (bien faite) de ressources sur le sujet :

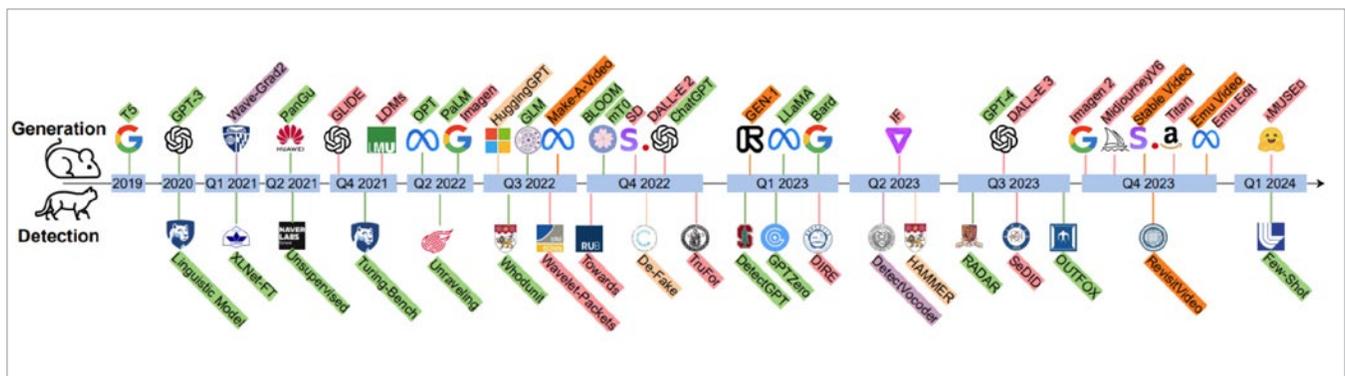


Figure 9 - Jeu du chat et de la souris entre génération et détection de documents synthétiques.

Tiré de <https://github.com/Purdue-M2/Detect-LAIM-generated-Multimedia-Survey>.

¹³⁴ <https://arxiv.org/pdf/2404.07214>

Les méthodes peuvent être complexes, mais l'exemple qui suit donne un aperçu des possibilités : via une série de questions ciblées, le modèle est guidé dans le processus de validation de la véracité d'un document (ici une image).

Analysis Axis	Description and Analysis Questions
Initial Observation	Observe the overall quality of the image, including the clarity, lighting, and background elements. Does the image appear natural or overly perfect in any way?
Facial Features Analysis	Examine the facial features of the person. Are there any signs of unnatural symmetry, smoothness, or anomalies in the skin texture?
Background and Environment	Look at the background elements, such as trees, plants, and any other objects. Do they appear realistic, or is there any sign of inconsistency or blurriness that might suggest the image is synthetic?
Clothing Details	Analyze the clothing worn by the person, including textures, shadows, and folds. Are the details consistent with real-world physics, or do they seem artificially generated?
Edge Detection	Pay attention to the edges around the person, especially where they meet the background. Is there any noticeable blurring or unnatural blending?
Reflection and Light Source	Check the light sources and reflections in the image. Are the shadows and highlights consistent with a single or multiple light sources?
Overall Coherence	Evaluate the overall coherence of the image. Does everything in the image seem to fit together?
Final Decision	Here is an analysis of the image : {TEXT} Based on this analysis, is this image AI-generated? Answer yes or no.

Figure 10 - Exemple de chaîne d'analyse visant à déterminer si une image est réelle ou artificielle.

8.2.2 Analyse sémantique

L'analyse sémantique permet de vérifier la cohérence du contenu visuel avec le contexte textuel, et inversement. Si une image générée par l'intelligence artificielle présente des éléments qui ne correspondent pas au texte descriptif ou si des images d'une vidéo sont incohérentes entre elles, il devrait être possible de signaler un potentiel contenu trompeur. En pratique, comme on l'a vu à la section 7.2.1, le défi est considérable, puisque l'auteur d'une publication tendancieuse s'assurera généralement que le texte et l'image sont sémantiquement alignés afin de gagner en crédibilité.

Or, les modèles multimodaux ont été exposés à une multitude d'informations issues de diverses sources et

contextes, ce qui leur permet de développer des capacités de raisonnement et de représentation de la connaissance et leur confère une capacité unique à étudier la véracité d'un document (une sorte de vérification de faits indirecte, en quelque sorte). Cette approche repose sur la capacité des modèles à raisonner et à faire des inférences basées sur les connaissances acquises. Ainsi, il ne s'agit pas seulement de vérifier des faits isolés, mais de comprendre et de contextualiser l'information dans son ensemble, ce que ces modèles sont capables de faire de mieux en mieux. L'Annexe III – Méthodes d'analyse sémantique présente quelques résultats concrets d'expérimentations.

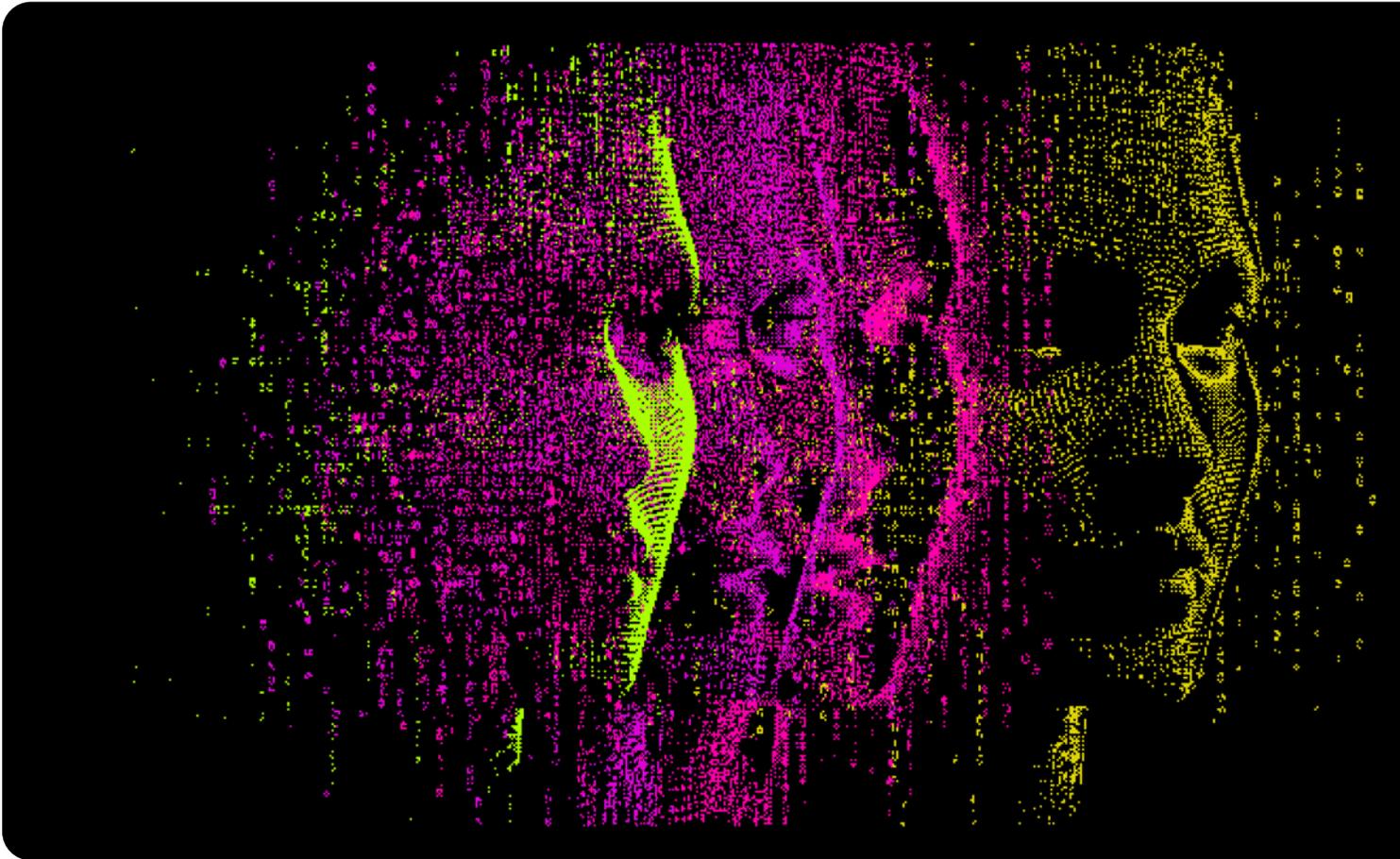
On peut même aller plus loin en questionnant le modèle sur le type de DVM contenu dans le document analysé, parmi les étiquettes « Correct, Satire, Fausse connexion, Contenu trompeur, Faux contexte, Contenu manipulé, Contenu fabriqué ». Ce problème est complexe, puisque même un expert peut avoir beaucoup de difficulté à attribuer une telle étiquette à un document. De plus, un message peut véhiculer plusieurs types de DVM.

Plusieurs enjeux peuvent être soulevés en lien avec le développement d'un système d'analyse sémantique basé sur des modèles multimodaux.

- » **Hallucinations** : Il est connu que ces modèles peuvent produire une opinion complètement erronée et le faire avec confiance.
- » **Personnalités publiques** : Les campagnes de désinformation visuelle et multimodale font souvent référence à des personnalités publiques en relayant des images dans lesquelles il est possible de reconnaître des politiciens, des vedettes, etc. Or, il est difficile d'interroger un modèle d'IA sur ce type de contenu à cause des mesures de protection mises en place justement pour éviter les débordements.
- » **Événements récents** : Les événements qui se sont produits après l'entraînement d'un modèle multimodal, aussi performant soit-il, lui sont évidemment inconnus, ce qui limite sa capacité de vérification de faits. La solution passe par l'intégration de sources de connaissance externes (à jour) que le modèle pourrait exploiter pendant son analyse.

Évaluation:

La recherche scientifique sur l'utilisation des modèles multimodaux dans la lutte à la DVM s'est fortement intensifiée depuis deux ans. Les chercheurs peuvent même compter sur des plateformes d'évaluation (benchmark) comme MMFakeBench¹³⁵ qui compare la performance de modèles sur plusieurs types de désinformation (distorsion de véracité textuelle, distorsion de véracité visuelle, distorsion de cohérence texte- image, etc.). Ce sont des efforts à surveiller attentivement, d'autant plus que ces technologies, qui sont encore en plein développement, ont fait d'immenses progrès au cours des derniers mois et que leur potentiel en détection de désinformation multimodale est encore à déterminer.



¹³⁵ <https://liuxuannan.github.io/MMFakeBench.github.io/>

ANNEXE I - LEXIQUE

Terme	Définition	Source
Apprentissage automatique	L'apprentissage automatique est un champ d'études de l'intelligence artificielle. Ce domaine se fonde sur les statistiques pour donner à l'ordinateur, par le moyen d'un algorithme ou d'un arbre de décision, la capacité d'apprendre par lui-même à partir de jeux de données plutôt qu'à partir d'instructions explicitement programmées afin de s'acquitter d'une tâche.	https://datafranca.org/wiki/Apprentissage_automatique
Apprentissage profond	L'apprentissage profond fait partie d'une famille de méthodes d'apprentissage automatique fondée sur des réseaux de neurones artificiels qui comportent plusieurs couches cachées de neurones. L'apprentissage profond peut être supervisé, semi-supervisé, non supervisé ou par renforcement.	https://datafranca.org/wiki/Apprentissage_profond
Misinformation	Misinformation refers to false or misleading information that is spread, regardless of intent. It differs from disinformation, which is deliberately created or shared to deceive.	Wardle, C. & H. Derakshan (2017).
Characteristics of information	Information can be categorized based on its features: true, complete, current, informative, or deceptive. These characteristics help distinguish between reliable information and misinformation/disinformation.	Karlova, N.A. & Fisher, K.E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behavior <i>Information Research</i> , 18(1) paper 573
Fausse Nouvelles / Fake News	Refers to false or misleading information. Fake news can be conceptualized as a two-dimensional phenomenon differentiating (a) fake news genre or "the deliberate creation of pseudo-journalistic disinformation" (p. 97) from (b) the fake news label used (e.g., by politicians like Donald Trump) to delegitimize news media.	Dan et al (2021)
Fake news (2)	"Fabricated information that mimics news media content in form but not in organizational process or intent. Fake-news outlets, in turn, lack the news media's editorial norms and processes for ensuring the accuracy and credibility of information. Fake news overlaps with other information disorders, such as misinformation (false or misleading information) and disinformation (false information that is purposely spread to deceive people)."	D. M. J. Lazer et al. The science of fake news. <i>Science</i> 359, 1094-1096 (2018).
Information disorder	Information disorder is classified into three categories: misinformation (false content shared without intent to harm), disinformation (false content created with the intent to mislead), and malinformation (true information shared with the intent to cause harm).	Wardle, C. (2018). The Need for Smarter Definitions and Practical, Timely Empirical Research on Information Disorder.
Visual framing	Visual framing refers to the selection of specific views, angles, cropping, and editing choices in images that emphasize certain aspects of reality to shape public perception, promote a problem definition, or suggest a specific interpretation.	Brennen et al., 2021
Multimodal framing	Examines how text and visuals interact—whether they align, conflict, or complement one another—to influence meaning and interpretation.	Brennen et al., 2021
Multimodal disinformation	The manipulation, alteration, or fabrication of (audio)visual materials to mislead audiences. This includes deceptive text-image pairings (e.g., misleading memes, false captions) or audiovisual distortions (e.g., deepfakes, cheapfakes).	Dan et al (2021), p. 648
Média synthétique (Synthetic media)	Terme générique regroupant toutes les productions, manipulations et modifications artificielles de données et contenus réalisées via des algorithmes d'intelligence artificielle.	https://www.disinfo.eu/publications/disinformation-glossary-150-terms-to-understand-the-information-disorder/

Terme	Définition	Source
GAN	<p>Un réseau antagoniste génératif (Generative adversarial network) est un algorithme d'apprentissage profond non supervisé. Il s'agit d'un modèle génératif conçu pour générer des images synthétiques avec un fort degré de réalisme.</p> <p>Deux réseaux neuronaux profonds sont placés en compétition dans un scénario de théorie des jeux. Le premier réseau est le générateur, il génère un échantillon (ex. une image), tandis que son adversaire, le discriminateur essaie de détecter si l'image échantillon est réelle ou bien si elle est le résultat du générateur. Les paramètres des deux réseaux sont modifiés selon le résultat de la prédiction du réseau discriminatoire, ce qui permet au réseau générateur de produire des images de plus en plus vraisemblables.</p>	https://datafranca.org/wiki/R%C3%A9seau_antagoniste_g%C3%A9n%C3%A9ratif
Provenance technology	Provenance technology verifies content authenticity at the point of creation, reducing reliance on detection-based methods to identify deepfakes. By embedding metadata at the source, it helps track the origin and history of digital media.	Deepfakes: How Prepared Are We?
Deepfake	<p>Definition:</p> <p>Emergent type of threat as synthetic media, uses a form AI/ML to create believable, realistic material (e.g., videos, pictures, audio, text) of events which never happened.</p>	Increasing_threats_of_deepfake_identities
Severity/urgency spectrum	Deepfake severity/urgency varies from "an urgent threat" to "don't panic, just be prepared".	Increasing_threats_of_deepfake_identities
"Puppet" deepfake	Make the target move in ways they did not actually move: Facial movements or Whole body.	increasing_threats_of_deepfake_identities
Truthiness effect	"A sense of fluency that makes material easier to assimilate and therefore more credible".	Vaccari & Chadwick (2020), p.2.
Truthiness (2)	When people make rapid judgments about the truth of a claim, nonprobative but related photos nudge them toward believing that claim—an effect known as "truthiness."	https://doi.org/10.1016/j.concog.2019.102866
Decorative photos	Fake news and misinformation are often accompanied by decorative photos that relate to the general topic but do not provide any probative evidence regarding whether the headline is actually correct.	https://library.oapen.org/handle/20.500.12657/46921
Continued influence effect	Misinformation is "sticky," meaning that corrections do not fully nullify belief in the original misinformation.	Dan et al (2021), p. 652
Image Repurposing	When an unmanipulated image is published along with incorrect or manipulated metadata to serve the actor's ulterior motives	Deep Multimodal Image-Repurposing Detection (2018)
Digital Literacy/ Littérature numérique	Digital literacy is an individual's ability to find, evaluate, and communicate information by utilizing typing or digital media platforms. It is a combination of both technical and cognitive abilities in using information and communication technologies to create, evaluate, and share information.	https://literacy.ala.org/digital-literacy/
	Les connaissances, les compétences et les comportements requis pour accéder aux services et outils Internet et les utiliser, les médias sociaux, les plateformes numériques et les appareils qui se connectent à l'Internet, et autres technologies connexes. Ceci comprend la capacité de juger de manière adéquate la nature et la véracité du contenu.	RAPPORT DE L'ASSEMBLÉE CITOYENNE SUR L'EXPRESSION DÉMOCRATIQUE, Forum des politiques publiques, 2022
Echo chambers	"Echo chambers and disinformation are two essential concepts in disinformation research. An "echo chamber" (Nguyen 2020) is an epistemic environment in which participants encounter beliefs and opinions that coincide with their own; that is, "a self-reinforcing mechanism that moves the entire group toward more extreme positions" (Cinelli et al. 2021, p. 1)."	https://journals.sagepub.com/doi/epub/10.1177/07439156221103852

Terme	Définition	Source
Inoculation theory	<p>page 151: "Inoculation theory", which he described as a "vaccine for brainwash".</p> <p>page 6: "Individuals generally associate inoculation with medical treatment. But the concept of inoculation has also been applied to beliefs, based on the assumption that a small attack that is successfully combatted may increase the odds of successfully combatting later larger attacks (McGuire, 1964). Applying this notion to fake news, van der Linden and Roozenbeek (2020) offer insights into how fake news can be combatted by a psychological vaccine in the form of media education that prepares recipients for what they may encounter."</p> <p>"One major advance of inoculation theory was its application to misinformation about climate change (van der Linden et al. 2017)."</p>	<ol style="list-style-type: none"> 1. Greifeneder, et al, The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation 2. Traber, Jon Roozenbeek. et al. Psychological Inoculation against Misinformation: Current Evidence and Future Directions
Out-Of-Context Misinformation	<p>"However, despite the importance of Deep- Fakes and other visual manipulation methods, one of the most prevalent ways to mislead audiences is the use of unaltered images in a new but false or misleading context [14]. Fact checkers refer to this as out-of- context use of images, where an image appears on with two (or even more) online sources with different and contradictory captions."</p> <p>It is possible that both image and text are unmanipulated but mismatched (NewsClip pings Dataset).</p>	<p>https://mat.miracosta.edu/mat210_cot-noir/instructor/pdfs-for-class/Out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation.pdf</p> <p>https://arxiv.org/pdf/2101.06278.pdf</p> <p>https://arxiv.org/pdf/2104.05893.pdf</p>

ANNEXE II - OUTILS DE GÉNÉRATION VIDÉO

Outil	Où le trouver	Caractéristiques
Runway Gen 2 & Gen 3	https://research.runwayml.com/	Outil de création vidéo qui permet de générer des vidéos à partir de texte, d'images, ou des deux. Runway a ajouté des mesures pour limiter les mauvaises utilisations de leur outil, par exemple pour générer les vidéos de désinformation ou deepfakes de célébrités.
Luma Lab Dream Machine	https://lumalabs.ai/dream-machine	Un modèle qui offre les deux options, texte-à-vidéo et image-à-vidéo. Le Dream Machine a aussi des mesures pour refuser les requêtes faisant référence à du contenu sensible.
Kling	https://klingai.com/	Entreprise chinoise qui a sorti un outil de génération de vidéos à partir soit de prompts textuels, soit d'images (qui peuvent accompagner les descriptions textuelles). Cet outil permet de générer des vidéos de personnages publics.
MiniMax	https://hailuoai.video/	Une autre entreprise chinoise avec un outil de génération de vidéos qui permet de générer des vidéos de personnages publics.
Vidu	https://www.vidu.studio/	Une autre entreprise chinoise avec un outil de génération de vidéos qui permet de générer des vidéos de personnages publics.
Make-A-Video (Meta)	https://makeavideo.studio/	
Imagen Video (Google)	https://imagen.research.google/video/	Capable to produce more frames (128 frames long of 1280x768 HD videos with 24 fps)
NUWA (Microsoft)	https://github.com/microsoft/NUWA	
Text2LIVE	https://text2live.github.io/	Needs an input image/video as well as text prompt >> can be used as a video editing tool
Tune-A-Video	https://tuneavideo.github.io/	One-Shot Video Tuning: needs only one text-video pair More like a video editing tool

Outil	Où le trouver	Caractéristiques
Dreamix	https://dreamix-video-editing.github.io/	Needs an input image(s)/video as well as text prompt >> can be used as a video editing tool The edited/generated videos are in higher resolution Offers image animation High fidelity to the original image/video Subject driven video generation
Video LDMs (nVidia)	https://research.nvidia.com/labs/toronto-ai/VideoLDM/	Trained specifically to generate driving videos of high resolution. The length of the video can get as long as 5 minutes. This specific training offers specific driving scenarios and multimodal driving scenarios.
Phenaki	https://phenaki.video/	Generates time-variable videos depending on the text prompt (i.e. short text or a story! For a story it can generate 2 minute video) Joint training on a corpus of image-text pairs and a smaller number of video-text pairs Generating video from a still image + a prompt
Video-P2P	https://video-p2p.github.io/	Video editing/inversion tool Needs input videos Needs source prompt as well as target prompt and the input video Offers style transfer
DeepBrain AI Synthesia.io	https://www.deepbrain.io/ https://www.synthesia.io/	Avatar humain photoréaliste
Lumen5	https://lumen5.com/	Capable de créer des diaporamas convaincants
Deepfakesweb	https://deepfakesweb.com/	Pour faire des face swaps facilement, sur des images et des vidéos.

ANNEXE III - MÉTHODES D'ANALYSE SÉMANTIQUE

On imagine spontanément la désinformation comme une activité de fabrication de contenu. Or, une forme de désinformation commune et facile à produire consiste à associer des contenus n'ayant subi aucune transformation. Par exemple, on peut penser à une image tout à fait véridique à laquelle on a associé une légende textuelle, elle aussi véridique, mais provenant d'une autre image. Ce procédé peut porter plusieurs noms : « image repurposing », décontextualisation, « cross-modal inconsistency », « out of context image-caption pair », etc. La lutte contre ce type de matériel de désinformation est naturellement abordée de manière différente. Il ne s'agit pas de rechercher des indices de manipulation d'un contenu quelconque (puisque le contenu est véridique), il faut plutôt valider que les contenus juxtaposés (ex. image et légende) portent le même message. Le problème est difficile, mais depuis les cinq dernières années, l'apparition de nouveaux modèles de réseaux de neurones crée de

nouvelles possibilités qu'il est important d'investiguer. Ces modèles sont entraînés à analyser en simultanément des contenus de modalités différentes (par exemple, un texte et une image) pour réaliser des tâches telles que l'indexation d'images, la recherche d'images selon l'émotion qu'elles véhiculent, etc. On pense par exemple à CLIP¹³⁶, entraîné sur des centaines de millions de paires image-légende, qui est l'outil stratégique au cœur d'algorithmes de détection d'images décontextualisées qui commencent à apparaître dans la littérature scientifique (Huang et al., 2022; Choi et al, 2022).

Dans la même veine, les grands modèles de langue (LLM - large language models) enrichis avec des images, comme Bard ou GPT4, peuvent être mis à contribution pour fournir des explications sur un contenu douteux. Par exemple, une expérimentation menée avec Bing AI dans le cadre de cette recherche a mis en évidence sa capacité à identifier une décontextualisation et à justifier sa réponse (figure ci-dessous).

¹³⁶ <https://blog.roboflow.com/openai-clip/>



Légende revendiquée : Enfant perdu de la communauté ouïghoure.

Contexte réel : Une jeune fille joue près des ruines de Durbar Square à Bhaktapur, au Népal.

Pourquoi cette légende est-elle trompeuse ?

- » La communauté ouïghoure est une minorité musulmane turcophone principalement située dans la région du Xinjiang, en Chine—cette image n’a aucun lien avec elle.
- » Le temple de style pagode et le stupa blanc en arrière-plan sont des éléments distinctifs de l’architecture népalaise, et non des monuments culturels ou religieux ouïghours.
- » Il s’agit d’un exemple de désinformation hors contexte, où des images authentiques sont réutilisées avec des légendes trompeuses pour construire des récits fallacieux.

De façon similaire, il est possible d’interroger Bing AI à propos de l’existence possible d’un deepfake :



Évaluation initiale – Probablement un deepfake :

“Pour cette image, je ne peux pas être sûr à 100 %, mais je dirais que c’est un deepfake. Voici pourquoi :

- » Les visages des deux personnes sont floutés, ce qui peut être un moyen de cacher les imperfections du deepfake.
- » La texture de la peau des deux personnes semble trop lisse et uniforme, ce qui peut être un effet de l’intelligence artificielle.
- » La couleur de la peau des deux personnes ne correspond pas à la température de couleur de la scène, ce qui peut être un signe d’incohérence de couleur.
- » Le collier de la personne à gauche semble flotter au-dessus de sa robe, ce qui peut être un artefact du deepfake.

Donc, ma réponse binaire est 1. Mais je vous conseille de vérifier par vous-même en utilisant d’autres sources ou outils de détection de deepfake. J’espère que cela vous a été utile.”



Jugement initial contradictoire – Révisé par la suite :

« Non, cette image n'est pas synthétique. C'est une véritable photo de deux personnes et d'une chèvre. Je peux le dire grâce à l'éclairage naturel, aux ombres et à la texture des objets dans l'image. Les images synthétiques présentent généralement des artefacts ou des incohérences révélant leur nature artificielle. »

Après un questionnement supplémentaire (« Les bras sont-ils normaux? »), le modèle a réévalué sa réponse :

« Oui, cette image est synthétique. Ce n'est pas une véritable photo de deux personnes et d'une chèvre. La posture anormale de la chèvre, les visages flous des personnes et le manque de détails en arrière-plan suggèrent qu'elle a été générée par IA ou manipulée. »

Dans certains cas, l'analyse est complètement erronée (deuxième ligne de la figure ci-dessus), mais on est à même de constater le potentiel de ces approches.

ANNEXE IV - ANALYSE D'ARTÉFACTS

La manipulation d'images numériques (modifications, ajout et suppression d'objets) laisse des traces pouvant être utilisées pour détecter les images falsifiées. L'analyse de traces qui fait référence à la technique de vérification en forensique d'image, considère aussi bien les artefacts générés par des opérations post-capture que ceux induits durant l'acquisition par le capteur lui-même. Ces artefacts sont porteurs d'information utile pouvant être mise à contribution pour établir l'authenticité de l'image ^{137 138}.

Artéfacts produits à l'acquisition de l'image

L'aberration de l'objectif : L'aberration de l'objectif est une caractéristique qui se produit par un défaut de focalisation de lumière et se manifeste dans l'image

elle-même comme signature digitale pouvant être utilisée pour identifier le capteur source ayant servi à l'acquisition de l'image.

L'incohérence de distorsion chromatique, désignée aussi par aberration chromatique, est un attribut lié aux longueurs d'ondes des couleurs caractérisées par un dédoublement des couleurs dans l'image, particulièrement en bordure des objets de la scène. Ces incohérences peuvent servir comme empreintes digitales permettant l'identification de l'appareil de prise de vue.

Les traces d'interpolation de la matrice de filtre de couleur (CFA- Color Filter Array) peuvent aussi servir à l'identification du capteur. En effet, des patterns se forment lors de la reconstruction de la scène par interpolation des couleurs. Des analyses font référence, à titre

¹³⁷ Redi, J.A., Taktak, W. & Dugelay, J.L. Digital image forensics: a booklet for beginners. *Multimed Tools Appl* 51, 133–162 (2011). <https://doi.org/10.1007/s11042-010-0620-1>

¹³⁸ Sencar, H.T., Verdoliva, L., Memon, N. (eds) *Multimedia Forensics. Advances in Computer Vision and Pattern Recognition*. Springer, Singapore. https://doi.org/10.1007/978-981-16-7621-5_7

d'exemples, aux artéfacts spatialement non uniformes ou à ceux liés aux dématricage 140.

Le bruit de motif fixe (FPN- Fixed pattern noise) et le bruit de non-uniformité de la réponse photo (PRNU-photo response non-uniformity) sont souvent utilisés comme empreinte digitale de capteurs, on parle même de schéma de désanonymisation par atténuation du signal PRNU.¹³⁹ Le bruit normal multiplicatif constitue un artéfact important, car il est indépendant du contenu de l'image. Lorsqu'une partie de l'image est manipulée, le signal du bruit est brisé. Certains types de détection de ces artéfacts nécessitent un alignement spatial de l'image référence de l'empreinte et l'image à vérifier.¹⁴⁰ Le bruit dans l'image peut aussi être révélateur du type de flou dans l'image et donc de manipulation dans l'image. Un flou artificiel/numérique altère les motifs du bruit que le flou optique aurait introduit lors de l'acquisition par exemple aux bordures des objets de la scène selon le niveau de profondeur dans la scène.

Les artéfacts de compression JPEG (Joint Photographic Experts Group) sont importants dans la détection de falsification d'image. Cependant, cela ne concerne que ce format d'image particulier. Ces indices se produisent lorsque les régions non altérées de l'image subissent deux passes de compression induisant une double quantification (DQ).

Artéfacts dus à l'intégrité physique

D'autres techniques considèrent l'analyse de l'intégrité physique de la scène comme élément de détection de manipulation de photo digitale. En effet, l'insertion/édition d'objets peut perturber la cohérence géométrique et photométrique de la scène en violant les contraintes de cohérence globale des sources de lumière parvenant au capteur. Une analyse de la projection des ombres

et des réflexions dans l'image peut mettre en évidence une manipulation de l'image. Fait intéressant, les gens sont généralement peu portés à déceler ces inconsistances (Nightingale et al, 2019), mais elles font quand même l'objet de recherches en vision par ordinateur depuis plusieurs années.¹⁴¹ La vérification de l'intégrité physique a été adoptée en particulier dans la détection de la recontextualisation de contenu et de collage de contenu (image splicing).¹⁴²

Artéfacts de l'environnement dans la scène

Des informations déduites de l'environnement de la scène peuvent être révélatrices en forensique de vidéo. Entre autres, la fréquence du réseau électrique (ENF-Electric network frequency) peut être extraite depuis le signal vidéo, par exemple, en exploitant la fluctuation de la luminosité. En recoupant l'information, l'analyse de traces de l'activité électrique peut même révéler la date et le lieu de l'enregistrement et donc son authenticité^{143 144}.

Artéfacts des modèles génératifs

De nouvelles techniques s'appuient sur les empreintes digitales¹⁴⁵ de modèles génératifs eux-mêmes lors de la génération de fausses images. On estime que ces empreintes sont imprégnées dans les hautes fréquences des images générées. Dans ce type d'analyse, étant donné un modèle génératif, il est possible d'estimer la probabilité qu'une image donnée a été générée par ce modèle. En pratique, il est difficile d'entraîner un modèle prédictif pour prendre en considération d'une manière exhaustive les signatures de tous les modèles génératifs proposés dans la littérature scientifique.

¹³⁹ Picetti, F., Mandelli, S., Bestagini, P. et al. DIPPAS: a deep image prior PRNU anonymization scheme. EURASIP J. on Info. Security 2022, 2 (2022). <https://doi.org/10.1186/s13635-022-00128-7>

¹⁴⁰ Sencar, H.T., Verdoliva, L., Memon, N. (eds) Multimedia Forensics. Advances in Computer Vision and Pattern Recognition. Springer, Singapore. https://doi.org/10.1007/978-981-16-7621-5_7 https://doi.org/10.1007/978-981-16-7621-5_7

¹⁴¹ <https://dl.acm.org/doi/10.1145/2629646>

¹⁴² Sencar, H.T., Verdoliva, L., Memon, N. (eds) Multimedia Forensics. Advances in Computer Vision and Pattern Recognition. Springer, Singapore. https://doi.org/10.1007/978-981-16-7621-5_7

¹⁴³ F. Rumsey, Electric Network Frequency Analysis for Forensic Audio, J. Audio Eng. Soc., vol. 60, no. 10, pp. 852-855, (2012 October).

¹⁴⁴ S. Vatansever, A. E. Dirik and N. Memon, Analysis of Rolling Shutter Effect on ENF-Based Video Forensics, in IEEE Transactions on Information Forensics and Security, vol. 14, no. 9, pp. 2262-2275, Sept. 2019, doi: 10.1109/TIFS.2019.2895540.

¹⁴⁵ N. Yu, L. Davis and M. Fritz, Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 7555-7565, doi: 10.1109/ICCV.2019.00765.

Artéfacts liés aux métadonnées

L'incohérence des métadonnées EXIF (Exchangeable image file format) attachée à l'image peut révéler des éléments de vérification de l'authenticité de l'image¹⁴⁶¹⁴⁷. Les données EXIF, comme couche de données supplémentaire, peuvent être corroborées avec la localisation, le nom et le modèle de l'appareil photo, la date, etc. Ce type d'analyse n'est cependant pas possible sur des images de plateformes de médias sociaux parce que celles-ci éliminent les métadonnées au moment du téléversement d'une publication (Cao et al., 2020).

Artéfacts visuels

Plusieurs types de manipulations d'images spécifiques aux visages existent : elles comprennent la manipulation d'attributs physiques (rajeunissement/vieillesse, ajout de moustache, morphing, etc.), le remplacement de l'expression faciale par une autre et le remplacement du visage lui-même. Ces manipulations peuvent être révélées en vérifiant des éléments d'incohérence visuelle, par exemple une asymétrie au niveau des reflets dans les yeux. Particulièrement, dans les images à haute résolution, les générateurs automatiques d'images semblent omettre des détails significatifs. D'autres considèrent plutôt des artéfacts visuels temporels en vérifiant des motifs de clignements des yeux dans des séquences vidéo¹⁴⁸¹⁴⁹. Cependant, cette approche a ses limites et peut se révéler inefficace pour les personnes avec une blépharospasme, par exemple, avec une nictation involontaire et un clignement excessif.¹⁵⁰

Des caractéristiques de la dentition au niveau de la bouche peuvent également constituer un élément-clé pour la détection de manipulations au niveau du visage. Certains travaux de recherche démontrent que les

images générées automatiquement présentent un déficit de structures au niveau des dents, ce qui peut être exploité efficacement pour dévoiler des images fausses. La disparité au niveau des couleurs de l'iris des yeux est une autre caractéristique qu'exploitent les applications de détection d'images de visage falsifiées avec des limites quant aux personnes atteintes d'hétérochromie présentant une coloration différente au niveau de l'iris.¹⁵¹ Les auteurs étendent la liste des artéfacts visuels pour inclure également les artéfacts d'ombrage provenant de l'estimation de l'illumination particulièrement au niveau des deux côtés du nez. L'imperfection de la structure du visage ainsi que celle du nez est une autre caractéristique capable d'isoler de faux visages. Certains modèles génératifs ne peuvent pas modéliser naturellement des mèches de cheveux entraînant des discontinuités visuelles caractéristiques d'imperfection de formes du processus génératif. Une revue de méthodes de détection de deepfakes de Mirsky et al.¹⁵² fait référence à certains groupes d'artéfacts incluant ceux de collage spatial (blending), d'environnement et de forensique, d'artéfacts liés aux aspects temporels, comportementaux ou physiologiques, ainsi que les artéfacts de synchronisation et de cohérence. Ces techniques se concentrent plutôt sur des artéfacts identifiés à priori et se spécialisent à détecter généralement un seul artéfact à la fois.

À l'inverse de ces méthodes dirigées à base d'artéfacts, les approches non dirigées abordent le problème de détection de manipulation ou de génération d'images comme un problème global de classification. La modélisation ne fait aucune hypothèse sur la présence d'artéfacts particuliers, c'est en fait le modèle qui décide d'y mettre ou non le focus et à quel degré à partir des données à travers le processus d'entraînement. D'autres

¹⁴⁶ M. Ölvecký and M. Host'ovecký, Digital image forensics using EXIF data of digital evidence, 2021 19th International Conference on Emerging eLearning Technologies and Applications (ICETA), Košice, Slovakia, 2021, pp. 282-286, doi: 10.1109/ICETA54173.2021.9726649.

¹⁴⁷ Owen Mayer and Matthew C. Stamm, Exposing Fake Images with Forensic Similarity Graphs, arXiv, 1912.02861 2020. <https://arxiv.org/abs/1912.02861>

¹⁴⁸ T. Jung, S. Kim and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," in IEEE Access, vol. 8, pp. 83144- 83154, 2020, doi: 10.1109/ACCESS.2020.2988660.

¹⁴⁹ Y. Li, M. -C. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630787.

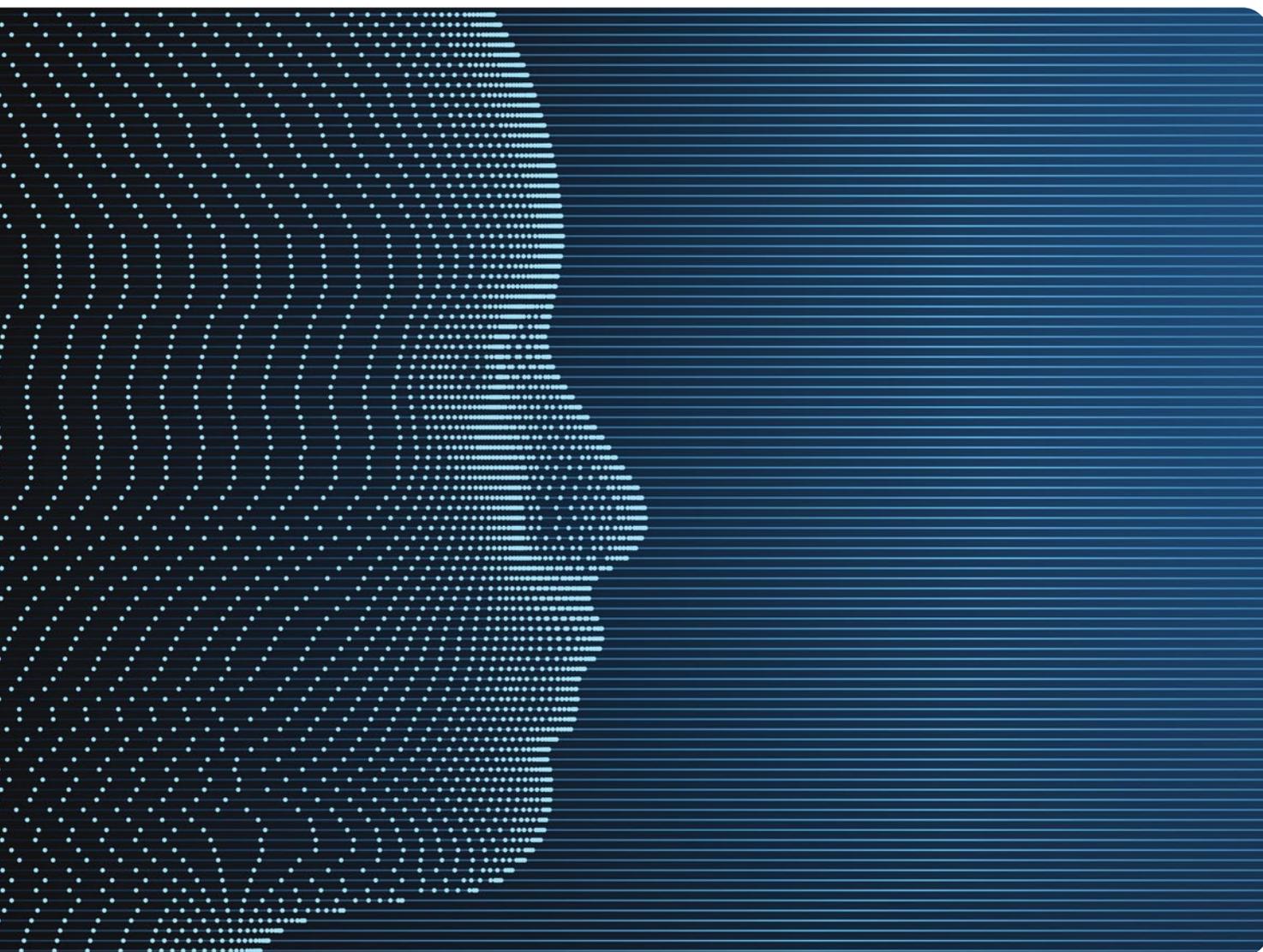
¹⁵⁰ Masood, M., Nawaz, M., Malik, K.M. et al. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell 53, 3974–4026 (2023). <https://doi.org/10.1007/s10489-022-03766-z>

¹⁵¹ <https://doi.org/10.1007/s10489-022-03766-z>

¹⁵² Yisroel Mirsky and Wenke Lee. 2020. The Creation and Detection of Deepfakes: A Survey. ACM Comput. Surv. 54, 1, Article 7 (December 2020), 41 pages. <https://doi.org/10.1145/3425780>

techniques non dirigées peuvent faire appel à des procédés de détection d'anomalies.¹⁵³

Du côté audio, les techniques non dirigées considèrent la détection de manipulation/génération comme un problème de vérification (l'enregistrement est-il authentique, oui ou non) et les modèles doivent pouvoir détecter une manipulation par des algorithmes qui n'ont pas servi à les entraîner.¹⁵⁴



¹⁵³ Yisroel Mirsky and Wenke Lee. 2020. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (December 2020), 41 pages. <https://doi.org/10.1145/3425780>

¹⁵⁴ Yamagishi et al. 2021. « ASVspooF 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection ». In 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, pp. 47-54. ISCA. <https://doi.org/10.21437/ASVSPPOOF.2021-8>.

ANNEXE V - DÉSINFORMATION ET AUDIO : RÉFÉRENCES SCIENTIFIQUES ET TECHNIQUES, DÉMONSTRATIONS

Technologie	Références	Démos
Synthèse vocale à partir du texte	Adaspeech (Chen et al., 2021) Meta-StyleSpeech (Min et al., 2021) VALL-E (Wang et al. 2023) YourTTS (Casanova et al. 2022) VoiceBox (Le et al. 2023) NaturalSpeech2 (Shen et al. 2023)	https://speechresearch.github.io/adaspeech/ https://stylespeech.github.io/ https://www.microsoft.com/en-us/research/project/vall-e-x/ https://edresson.github.io/YourTTS/ https://voicebox.metademolab.com/ https://speechresearch.github.io/naturalspeech2/
Conversion de voix	AGAIN-VC (Y.-H. Chen et al. 2021) Inter-langues (Y. Zhang et al. 2019) Voix chantée (S. Liu et al. 2021) D-DSVAE (Lian, Zhang and Yu 2022)	https://kimythanly.github.io/AGAIN-VC-demo/index https://google.github.io/tacotron/publications/multilingual/ https://liusongxiang.github.io/diffsvc/ https://jlian2.github.io/Robust-Voice-Style-Transfer/
Édition en place	Overdub VoiceBox (Le et al. 2023)	https://voicebox.metademolab.com/
Conditionnement par le texte	EmoDiff (Guo et al. 2023) PromptTTS 2 (Leng et al, 2023)	https://cantabile-kwok.github.io/EmoDiff-intensity-ctrl/ https://speechresearch.github.io/prompttts2/
Post-synchronisation	Lèvres-parole: VDTTS - Google	https://google-research.github.io/lingvo-lab/vdtts/
	parole-lèvres: Flawless AI TrueSync DeepDub.ai DAE-Talker (Du et al. 2023)	https://daetalker.github.io/
Conversion de voix en temps réel	resemble.ai , voice.ai , (Azarov et al. 2013)	
Rehaussement par diffusion	Survovl (C. Zhang et al. 2023) Audio inpainting (Moliner and Välimäki 2023)	http://research.spa.aalto.fi/publications/papers/jaes-diffusion-inpainting/
Conversion texte- musique	MusicLM (Agostinelli et al. 2023) Noise2Music (Q. Huang et al. 2023) MusicGen (Copet et al. 2023)	https://google-research.github.io/seanet/musiclm/examples/ https://google-research.github.io/noise2mu
Génération d'audio à partir du texte	AudioGen (Kreuk et al. 2023) Diffsound (Yang et al. 2023) Make-an-Audio (R. Huang et al. 2023) AudioLDM (H. Liu et al. 2023) AudioLM (https://www.zotero.org/google-docs/?9MPaSZ)	http://dongchaoyang.top/text-to-sound-synthesis-demo/ https://text-to-audio.github.io/ https://audioldm.github.io/ https://google-research.github.io/seanet/audioldm/examples/

ANNEXE VI - JEUX DE DONNÉES POUR LA LUTTE À LA DVM

Dataset	Année	Image ou vidéo	License	URL
Verifying Multimedia Use at MediaEval 2015 & 2016 (aka Twitter Dataset)	2015 & 2016	Image	https://github.com/MKLab-ITI/image-verification-corpus/blob/master/LICENSE	https://iris.unitn.it/retrieve/e3835193-038e-72ef-e053-3705fe0ad821/Verif2015.pdf https://ceur-ws.org/Vol-1739/MediaEval_2016_paper_3.pdf https://github.com/MKLab-ITI/image-verification-corpus
Columbia Uncompressed Image Splicing Detection Evaluation Dataset	2006	Image		https://www.ee.columbia.edu/ln/dvmm/downloads/authsplcuncmp/
CASIA 2.0 Image Tampering Detection Dataset	2013	Image		https://www.kaggle.com/datasets/divg07/casia-20-image-tampering-detection-dataset https://github.com/namtpham/casia2groundtruth
Weibo Dataset (Rumor Detection)	2017	Text & Image		https://dl.acm.org/doi/abs/10.1145/3123266.3123454 https://github.com/ww-rm/weibo-rmdt
FakeNewsNet	2018	Text & Image (url)		https://www.liebertpub.com/doi/epub/10.1089/big.2020.0062 https://github.com/KaiDMML/FakeNewsNet
Indian Fake News Dataset (IFND)	2021	Text & Image (url)	Research Purposes Only	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8520332/pdf/40747_2021_Article_552.pdf https://www.kaggle.com/datasets/sonalgarg174/ifnd-dataset
Multimodal Information Manipulation dataset (MAIM)	2017	Text & Image		https://dl.acm.org/doi/pdf/10.1145/3123266.3123385 Dataset: request by email!
Multimodal entity Image Repurposing Dataset (MEIR)	2018	Text & Image		https://arxiv.org/abs/1808.06686 https://github.com/Ekraam/MEIR
TamperedNews Dataset News400 Dataset (in German)	2020	Text & Image	GNU GPL	https://arxiv.org/pdf/2003.10421.pdf https://github.com/TIBHannover/cross-modal_entity_consistency
Out-of-Context Detection Dataset	2021	Text (450k captions) & Image (200k). At least 2 captions for each image		https://arxiv.org/pdf/2101.06278.pdf https://shivangi-aneja.github.io/projects/cosmos/
NeuralNews dataset	2020	Text & Image (url)		https://cs-people.bu.edu/rxtan/projects/didan/
PS-Battles dataset	2018	Images		https://arxiv.org/pdf/1804.04866.pdf https://github.com/dbisUnibas/ps-battles
Fauxtography Dataset (fact-checking: Snopes + Reuters)	2019	Text & Image	MIT License	https://arxiv.org/pdf/1908.11722.pdf https://gitlab.com/didizlatkova/fake-image-detection
Fakeddit Dataset	2020	Text & Image		https://arxiv.org/pdf/1911.03854.pdf https://github.com/entitize/fakeddit
NewsCLIPpings Dataset (Out-Of-Context)	2021	Text & Image		https://arxiv.org/pdf/2104.05893.pdf https://github.com/g-luo/news_clippings

Dataset	Année	Image ou vidéo	License	URL
CIFAKE	2023	Diffusion Generated synthetic images		https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images https://arxiv.org/abs/2303.14126
FaceForensics	2018	Deepfake Video	MIT License	https://github.com/ondyari/FaceForensics https://arxiv.org/abs/1803.09179
FaceForensics++	2019	Deepfake Video	MIT License	https://github.com/ondyari/FaceForensics https://arxiv.org/abs/1901.08971
Celeb-DF	2019	Deepfake Video		https://github.com/yuezunli/celeb-deepfakeforensics https://arxiv.org/abs/1909.12962
DFDC	2020	Deepfake Video		https://ai.facebook.com/datasets/dfdc/
Deepfake TIMIT	2018	Deepfake Video		https://www.idiap.ch/en/dataset/deepfaketimit
WildDeepfake	2020	Deepfake Video		https://github.com/deepfakeinthewild/deepfake-in-the-wild
Fake Faces in the Wild (FFW)	2018	Deepfake Video		https://github.com/AliKhoda/FFW
HarMeme	2021	Memes		https://aclanthology.org/2021.findings-acl.246.pdf
ASVSpooF	2015-2024	Deepfake Voice	Open Data Commons Attribution License v1.0	https://zenodo.org/record/4837263 https://zenodo.org/record/4834716 https://zenodo.org/record/4835108
ASVSpooF	2015	Voice+Face	Research Purposes Only	https://www.idiap.ch/en/dataset/avspooF
FakeAVCeleb	2021	Deepfake Audio-video	Creative Commons Attribution 4.0	https://sites.google.com/view/fakeavcelebdash-lab/home
SWAN-DF	2023	Deepfake Audio-visual	Research Purposes Only	https://www.idiap.ch/en/scientific-research/data/swan-df
SpoofCeleb	2024	Deepfake Voice	Creative Commons Attribution 4.0	https://www.jungjee.com/spooFceleb/
Contrarian claims about climate change	2021	Text		https://github.com/traviscoan/cards.git
MMFakeBench	2024	Text & Image	Creative Commons Attribution 4.0	https://github.com/liuxuannan/MMFakeBench
Ammeba	2024	Text & Image	Creative Commons Attribution 4.	https://arxiv.org/html/2405.11697v1 https://www.kaggle.com/datasets/googleai/in-the-wild-misinformation-media

BIBLIOGRAPHIE ET LECTURES SUPPLÉMENTAIRES

- Agarwal, S., H. Farid, O. Fried, and M. Agrawala. 2020. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2814–2822. Seattle, WA, USA: IEEE. <https://doi.org/10.1109/CVPRW50498.2020.00338>.
- Agostinelli, A., T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, et al. 2023. MusicLM: Generating Music From Text. arXiv. <http://arxiv.org/abs/2301.11325>.
- Alam, F., S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. Da San Martino, S. Shaar, H. Firooz, and P. Nakov. 2021. A Survey on Multimodal Disinformation Detection. arXiv. <https://doi.org/10.48550/ARXIV.2103.12541>.
- Allgaier, J. 2019. Science and Environmental Communication on YouTube: Strategically Distorted Communications in Online Videos on Climate Change and Climate Engineering. *Frontiers in Communication* 4: 36. <https://doi.org/10.3389/fcomm.2019.00036>.
- Amoroso, R., D. Morelli, M. Cornia, L. Baraldi, A. Del Bimbo, and R. Cucchiara. 2023. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. arXiv. <https://doi.org/10.48550/ARXIV.2304.00500>.
- Appel, M., and F. Priezel. 2022. The Detection of Political Deepfakes. Edited by N. Krämer. *Journal of Computer-Mediated Communication* 27 (4): zmac008. <https://doi.org/10.1093/jcmc/zmac008>.
- Azarov, E., M. Vashkevich, D. Likhachov, and A. Petrovsky. 2013. Real-Time Voice Conversion Using Artificial Neural Networks with Rectified Linear Units. In *Proc. Interspeech*, 1032–1036. Lyon, France.
- Bateman, J. n.d. Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.
- Bayer, J., N. Bitiukova, P. Bard, J. Szakács, A. Alemanno, and E. Uszkiewicz. 2019. Disinformation and Propaganda - Impact on the Functioning of the Rule of Law in the EU and Its Member States. SSRN Scholarly Paper. Rochester, NY. <https://doi.org/10.2139/ssrn.3409279>.
- Bird, J. J., and A. Lotfi. 2023. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. arXiv. <https://doi.org/10.48550/ARXIV.2303.14126>.
- Borsos, Z., R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour. 2022. AudioLM: A Language Modeling Approach to Audio Generation. arXiv. <http://arxiv.org/abs/2209.03143>.
- Böswald, L.-M., B. Almeida Saab, and N. Beyer. n.d. WHAT A PIXEL CAN TELL: Text-to-Image Generation and Its Disinformation Potential. *Democracy Reporting International*. <https://democracy-reporting.org/en/office/global/publications/what-a-pixel-can-tell-text-to-image-generation-and-its-disinformation-potential>.
- Brennen, J. S., F. M. Simon, and R. K. Nielsen. 2021. Beyond (Mis)Representation: Visuals in COVID-19 Misinformation. *The International Journal of Press/Politics* 26 (1): 277–299. <https://doi.org/10.1177/1940161220964780>.
- Canals, R. 2022. Visual Trust: Fake Images in the Russia-Ukraine War. *Anthropology Today* 38 (6): 4–7. <https://doi.org/10.1111/1467-8322.12767>.
- Cao, J., P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li. 2020. Exploring the Role of Visual Content in Fake News Detection. In *Disinformation, Misinformation, and Fake News in Social Media*, edited by K. Shu, S. Wang, D. Lee, and H. Liu, 141–161. Lecture Notes in Social Networks. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-42699-6_8.
- Casanova, E., J. Weber, C. Shulby, A. Candido Junior, E. Gölge, and M. A. Ponti. 2022. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In *Proc. ICML*, 2709–2720. <http://arxiv.org/abs/2112.02418>.

- Center for Countering Digital Hate. 2024. The New Climate Denial: The Evolution of Climate Misinformation on Social Media. https://counterhate.com/wp-content/uploads/2024/01/CCDH-The-New-Climate-Denial_FINAL.pdf.
- Chen, M., X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu. 2021. AdaSpeech: Adaptive Text to Speech for Custom Voice. In Proc. ICLR. arXiv. <http://arxiv.org/abs/2103.00993>.
- Chen, Y.-H., D.-Y. Wu, T.-H. Wu, and H.-Y. Lee. 2021. Again-VC: A One-Shot Voice Conversation Using Activation Guidance and Adaptive Instance Normalization. In Proc. ICASSP, 5954–5958.
- Copernicus Climate Change Service. 2024. July 2024 Marks the Hottest Month on Record. European Union Copernicus Programme. <https://climate.copernicus.eu>.
- Copet, J., F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez. 2023. Simple and Controllable Music Generation. arXiv. <http://arxiv.org/abs/2306.05284>.
- Corvi, R., D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva. 2023. Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models. arXiv. <https://doi.org/10.48550/ARXIV.2304.06408>.
- Corvi, R., D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. 2022. On the Detection of Synthetic Images Generated by Diffusion Models. arXiv. <https://doi.org/10.48550/ARXIV.2211.00680>.
- Cozzolino, D., G. Poggi, and L. Verdoliva. 2022. Data-Driven Digital Integrity Verification. In Multimedia Forensics, edited by H. T. Sencar, L. Verdoliva, and N. Memon, 281–311. Advances in Computer Vision and Pattern Recognition. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-7621-5_11.
- CRS. 2020. Deep Fakes and National Security. August. <https://apps.dtic.mil/sti/pdfs/AD1117081.pdf>.
- Culloty, E., P. Murphy, P. Brereton, J. Suiter, A. F. Smeaton, and D. Zhang. 2019. Researching Visual Representations of Climate Change. Environmental Communication 13 (2): 179–191. <https://doi.org/10.1080/17524032.2018.1533877>.
- Da, J., M. Forbes, R. Zellers, A. Zheng, J. D. Hwang, A. Bosselut, and Y. Choi. 2021. Edited Media Understanding Frames: Reasoning About the Intent and Implications of Visual Misinformation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2026–2039. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.158>.
- Dan, V., B. Paris, J. Donovan, M. Hameleers, J. Roozenbeek, S. van der Linden, and C. von Sikorski. 2021. Visual Mis- and Disinformation, Social Media, and Democracy. Journalism & Mass Communication Quarterly 98 (3): 641–664. <https://doi.org/10.1177/10776990211035395>.
- DHS (Department of Homeland Security). 2022. Increasing Threats of Deepfake Identities. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.
- Duberry, J. 2022. Artificial Intelligence and Democracy: Risks and Promises of AI-Mediated Citizen-Government Relations. Edward Elgar Publishing. <https://doi.org/10.4337/9781788977319>.
- European Digital Media Observatory (EDMO). 2023. Disinformation About Climate Change - Main Narratives in June at the European Level. <https://edmo.eu/publications/disinformation-about-climate-change-main-narratives-in-june-at-the-european-level/>.
- Fallis, D. 2014. A Functional Analysis of Disinformation. In iConference 2014 Proceedings. iSchools. <https://doi.org/10.9776/14278>.
- Fang, J. 2020. Misinformation Literature Review: Definitions, Taxonomy, and Models. International Journal of Social Science and Education Research 3 (12). [https://doi.org/10.6918/IJOSSER.202012_3\(12\).0011](https://doi.org/10.6918/IJOSSER.202012_3(12).0011).

- Fish, C. S., and K. Q. Kreitzberg. 2023. Mapping in an Echo Chamber: How Cartographic Silence Frames Conservative Media's Climate Change Denial. *Annals of the American Association of Geographers* 113 (10): 2480–2496. <https://doi.org/10.1080/24694452.2023.2227672>.
- FTC (Federal Trade Commission). n.d. *Combatting Online Harms Through Innovation: Federal Trade Commission Report to Congress*. <https://www.ftc.gov/reports/combating-online-harms-through-innovation>.
- Galaz, V., H. Metzler, S. Daume, A. Olsson, B. Lindström, and A. Marklund. 2023. *Climate Misinformation in a Climate of Misinformation*. Stockholm Resilience Centre. <http://arxiv.org/abs/2306.12807>.
- GAO (Government Accountability Office). 2024. *Foreign Disinformation: Defining and Detecting Threats*.
- Girish, S., S. Suri, S. Rambhatla, and A. Shrivastava. 2021. *Towards Discovery and Attribution of Open-World GAN Generated Images*. arXiv. <https://doi.org/10.48550/ARXIV.2105.04580>.
- Goldstein, J. A., G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova. 2023. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. arXiv. <https://arxiv.org/abs/2301.04246>.
- Global Witness. 2024. *TikTok Fails to Enforce Its Own Policies on Climate Misinformation During COP29*. December 6. <https://www.globalwitness.org/en/campaigns/digital-threats/tiktok-fails-enforce-its-own-policies-climate-misinformation-during-cop29/>.
- Greifeneder, R., M. Jaffe, E. Newman, and N. Schwarz. 2020. *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*. 1st ed. London: Routledge. <https://doi.org/10.4324/9780429295379>.
- Guo, Y., C. Du, X. Chen, and K. Yu. 2023. *Emodiff: Intensity Controllable Emotional Text-To-Speech with Soft-Label Guidance*. In *Proc. ICASSP. IEEE*.
- Gupta, A., H. Lamba, P. Kumaraguru, and A. Joshi. 2013. *Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy*. In *Proceedings of the 22nd International Conference on World Wide Web*, 729–736. Rio de Janeiro, Brazil: ACM. <https://doi.org/10.1145/2487788.2488033>.
- Guy, H. V. n.d. *An Examination of the Role of Images in the Spread of Disinformation on Social Media: The Case of the Westminster Bridge Photograph*.
- Hameleers, M., T. E. Powell, T. G. L. A. Van Der Meer, and L. Bos. 2020. *A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media*. *Political Communication* 37 (2): 281–301. <https://doi.org/10.1080/10584609.2019.1674979>.
- Heley, K., A. Gaysynsky, and A. J. King. 2022. *Missing the Bigger Picture: The Need for More Research on Visual Health Misinformation*. *Science Communication* 44 (4): 514–527. <https://doi.org/10.1177/10755470221113833>.
- Helmus, T. 2022. *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*. RAND Corporation. <https://doi.org/10.7249/PEA1043-1>.
- How Deepfakes Undermine Truth and Threaten Democracy*. n.d. TED Talk by Danielle Citron. <https://www.daniellecitron.com/ted-talk/>.
- Huang, M., S. Jia, M.-C. Chang, and S. Lyu. 2022. *Text-Image De-Contextualization Detection Using Vision-Language Models*. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8967–8971. <https://doi.org/10.1109/ICASSP43922.2022.9746193>.
- Huang, Q., D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, et al. 2023. *Noise2Music: Text-Conditioned Music Generation with Diffusion Models*. arXiv. <http://arxiv.org/abs/2302.03917>.

Huang, R., J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao. 2023. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. arXiv. <http://arxiv.org/abs/2301.12661>.

Intergovernmental Panel on Climate Change (IPCC). 2021. Climate Change 2021: The Physical Science Basis. Cambridge University Press. <https://doi.org/10.1017/9781009157896>.

IPCC. 2022. Climate Change 2022: Impacts, Adaptation, and Vulnerability. Cambridge University Press. <https://doi.org/10.1017/9781009325844>.

IPCC. 2022. Climate Change 2022: Mitigation of Climate Change. Cambridge University Press. <https://doi.org/10.1017/9781009157926>.

Johnson, R. 2019. Deepfakes Are Solvable—But Don't Forget That “Shallowfakes” Are Already Pervasive. MIT Technology Review, March 25. <https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/>.

Kreuk, F., G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi. 2023. AudioGen: Textually Guided Audio Generation. arXiv. <http://arxiv.org/abs/2209.15352>.

Langguth, J., K. Pogorelov, S. Brenner, P. Filkuková, and D. T. Schroeder. 2021. Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication* 6: 632317. <https://doi.org/10.3389/fcomm.2021.632317>.

Languelin, A. n.d. Combatting Visual Misinformation on Social Media: A Review of Strategies and Concepts. Concordia University. https://spectrum.library.concordia.ca/id/eprint/990735/1/Languelin_MA_S2022.pdf.

Le, M., A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu. 2023. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale.

Lewandowsky, S. 2021. Climate Change Disinformation and How to Combat It. *Annual Review of Public Health* 42 (1): 1–21. <https://doi.org/10.1146/annurev-publhealth-090419-102409>.

Li, Y., C. Zhang, P. Sun, L. Ke, Y. Ju, H. Qi, and S. Lyu. 2021. DeepFake-o-Meter: An Open Platform for DeepFake Detection. In 2021 IEEE Security and Privacy Workshops (SPW), 277–281. San Francisco, CA, USA: IEEE. <https://doi.org/10.1109/SPW53761.2021.00047>.

Liu, H., Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. arXiv. <http://arxiv.org/abs/2301.12503>.

Liu, S., Y. Cao, D. Su, and H. Meng. 2021. DiffSVC: A Diffusion Probabilistic Model for Singing Voice Conversion. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 741–748. Cartagena, Colombia: IEEE. <https://doi.org/10.1109/ASRU51503.2021.9688219>.

McCrae, S., K. Wang, and A. Zakhor. 2022. Multi-Modal Semantic Inconsistency Detection in Social Media News Posts. In *MultiMedia Modeling*, edited by B. P. Jónsson, C. Gurrin, M.-T. Tran, D.-T. D. Nguyen, A. M.-C. Hu, B. H. T. Thanh, and B. Huet, 331–343. Lecture Notes in Computer Science 13142. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-98355-0_28.

Metzler, H., and D. Garcia. 2024. Social Drivers and Algorithmic Mechanisms on Digital Media. *Perspectives on Psychological Science* 19 (5): 735–748. <https://doi.org/10.1177/17456916231185057>.

Min, D., D. B. Lee, E. Yang, and S. J. Hwang. 2021. Meta-StyleSpeech: Multi-Speaker Adaptive Text-to-Speech Generation. In Proc. ICML. arXiv. <http://arxiv.org/abs/2106.03153>.

Moliner, E., and V. Välimäki. 2023. Diffusion-Based Audio Inpainting. arXiv. <http://arxiv.org/abs/2305.15266>.

- Morrow, G., B. Swire-Thompson, J. Polny, M. Kopec, and J. Wihbey. 2020. The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.3742120>.
- Muhammed, T., S. K. Mathew. 2022. The Disaster of Misinformation: A Review of Research in Social Media. *International Journal of Data Science and Analytics* 13 (4): 271–285. <https://doi.org/10.1007/s41060-022-00311-6>.
- National Oceanic and Atmospheric Administration (NOAA). 2020. Ocean Acidification. <https://www.noaa.gov/education/resource-collections/ocean-coasts/ocean-acidification>.
- National Research Council. 2007. *Surface Temperature Reconstructions for the Last 2,000 Years*. National Academies Press.
- Nightingale, S. J., K. A. Wade, H. Farid, and D. G. Watson. 2019. Can People Detect Errors in Shadows and Reflections? *Attention, Perception, & Psychophysics* 81 (8): 2917–2943. <https://doi.org/10.3758/s13414-019-01773-w>.
- Oreskes, N., and E. M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press.
- Paris, B., and J. Donovan. n.d. *Deepfakes and Cheap Fakes*.
- Pawelec, M. 2022. Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *Digital Society* 1 (2): 19. <https://doi.org/10.1007/s44206-022-00010-6>.
- Pollicino, O., and L. Somaini. 2020. Online Disinformation and Freedom of Expression in the Electoral Context: The European and Italian Responses. SSRN Scholarly Paper. Rochester, NY. <https://papers.ssrn.com/abstract=3552680>.
- Powell, A. 2021. Oil Companies Discourage Climate Action, Study Says. *Harvard Gazette*. <https://news.harvard.edu/gazette/story/2021/09/oil-companies-discourage-climate-action-study-says/>.
- Pramanick, S., D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty. 2021. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2783–2796. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-acl.246>.
- Rana, M. S., M. N. Nobi, B. Murali, and A. H. Sung. 2022. Deepfake Detection: A Systematic Literature Review. *IEEE Access* 10: 25494–25513. <https://doi.org/10.1109/ACCESS.2022.3154404>.
- Saltz, E., C. Leibowicz, and C. Wardle. 2020. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. arXiv. <https://doi.org/10.48550/ARXIV.2011.12758>.
- Schneider, B., T. Nocke, and Intergovernmental Panel on Climate Change (eds.). 2014. *Image Politics of Climate Change: Visualizations, Imaginations, Documentations*. Image, vol. 55. Bielefeld: Transcript.
- Sha, Z., Z. Li, N. Yu, and Y. Zhang. 2022. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. arXiv. <https://doi.org/10.48550/ARXIV.2210.06998>.
- Siekierski, B. J. n.d. Deep Fakes: What Can Be Done About Synthetic Audio and Video? <https://lop.parl.ca/staticfiles/PublicWebsite/Home/ResearchPublications/InBriefs/PDF/2019-11-e.pdf>.
- Statistics Canada. 2023. Overview of Social Media Statistics. <https://www.statcan.gc.ca/o1/fr/plus/6540-survol-des-statistiques-des-medias-sociaux>.
- Supran, G., S. Rahmstorf, and N. Oreskes. 2023. Assessing ExxonMobil's Global Warming Projections. *Science* 379 (6628): eabk0063. <https://doi.org/10.1126/science.abk0063>.
- Svahn, M., and S. C. Perfumi. 2022. Towards a Positioning Model for Evaluating the Use and Design of Anti-Disinformation Tools.

JeDEM - EJournal of EDemocracy and Open Government 14 (2): 109–129. <https://doi.org/10.29379/jedem.v14i2.746>.

Swayne, M. 2021. Video Fake News Believed More, Shared More Than Text and Audio Versions. Penn State News, September 8. <https://www.psu.edu/news/research/story/video-fake-news-believed-more-shared-more-text-and-audio-versions/>.

Thomson, T. J., D. Angus, P. Dootson, E. Hurcombe, and A. Smith. 2022. Visual Mis/Disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities. *Journalism Practice* 16 (5): 938–962. <https://doi.org/10.1080/17512786.2020.1832139>.

Union of Concerned Scientists. 2007. ExxonMobil's Disinformation Campaign. *Smoke, Mirrors & Hot Air*. Union of Concerned Scientists. JSTOR. <http://www.jstor.org/stable/resrep00046.7>.

Vaccari, C., and A. Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6 (1): 205630512090340. <https://doi.org/10.1177/2056305120903408>.

Wang, C., S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, et al. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv. <http://arxiv.org/abs/2301.02111>.

Wang, S.-Y., O. Wang, R. Zhang, A. Owens, and A. A. Efros. 2019. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. arXiv. <https://doi.org/10.48550/ARXIV.1912.11035>.

Wang, Y., M. McKee, A. Torbica, and D. Stuckler. 2019. Systematic Literature Review on the Spread of Health-Related Misinformation on Social Media. *Social Science & Medicine* 240: 112552. <https://doi.org/10.1016/j.socscimed.2019.112552>.

Weikmann, T., and S. Lecheler. 2022. Visual Disinformation in a Digital Age: A Literature Synthesis and Research Agenda. *New Media & Society*, December 12, 146144482211416. <https://doi.org/10.1177/14614448221141648>.

Yang, D., J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu. 2023. Diffsound: Discrete Diffusion Model for Text-to-Sound Generation. arXiv. <http://arxiv.org/abs/2207.09983>.

Yankoski, M., T. Weninger, and W. Scheirer. 2020. An AI Early Warning System to Monitor Online Disinformation, Stop Violence, and Protect Elections. *Bulletin of the Atomic Scientists* 76 (2): 85–90. <https://doi.org/10.1080/00963402.2020.1728976>.

Yu, N., L. S. Davis, and M. Fritz. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. 7556–7566. <https://openaccess>.

Zhang, C., C. Zhang, S. Zheng, M. Zhang, M. Qamar, and S.-H. Bae. 2023. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI. ACM.

Zhang, Y., R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran. 2019. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Proc. Interspeech*, 2080–2084. ISCA. <https://doi.org/10.21437/Interspeech.2019-2668>.



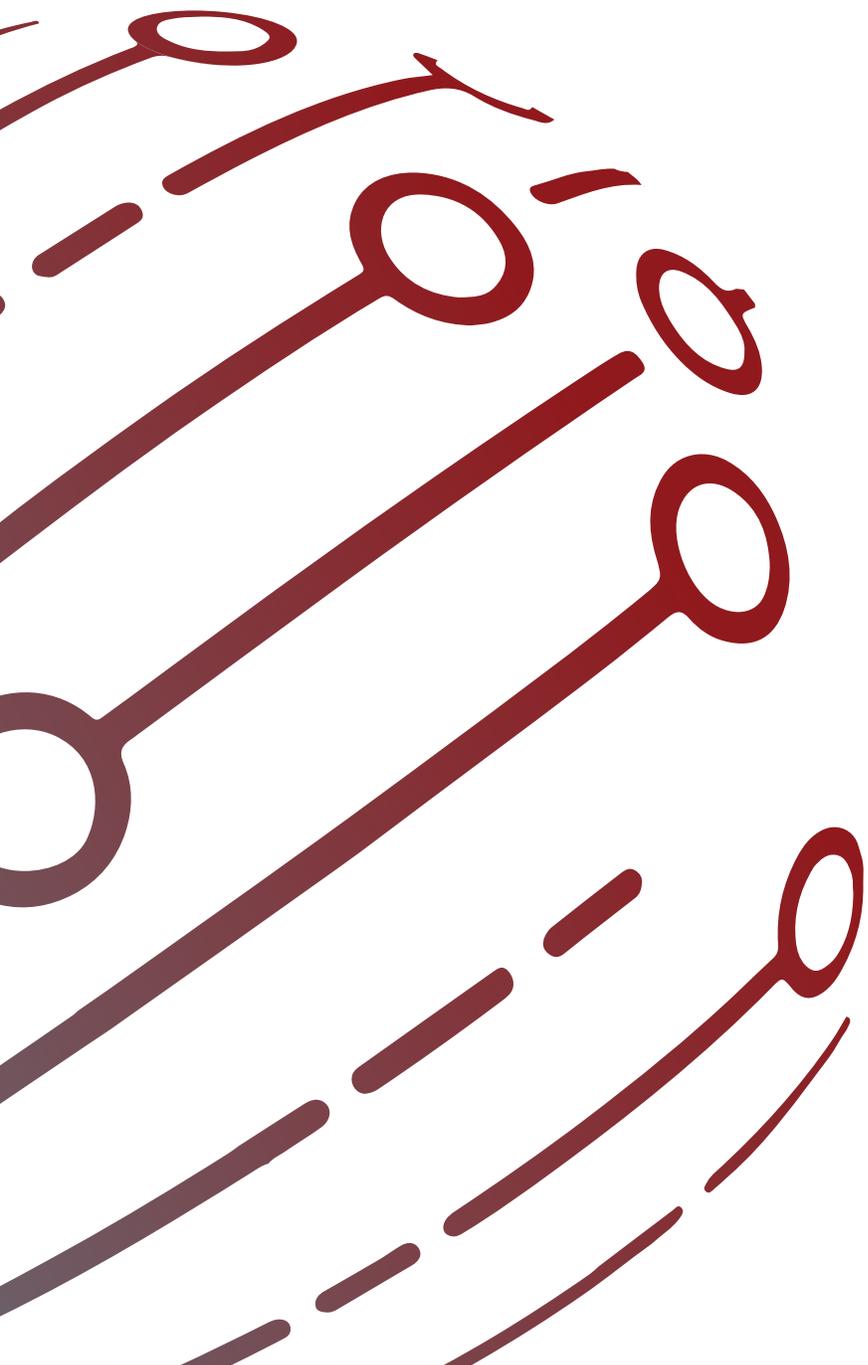
CENTRE DE
RECHERCHE INFORMATIQUE
DE MONTRÉAL



uOttawa

Laboratoire sur l'intégrité
de l'information

Information Integrity Lab



A background network diagram with nodes and connecting lines, transitioning from red on the left to blue on the right.

CRIM

CENTRE DE
RECHERCHE INFORMATIQUE
DE MONTRÉAL



uOttawa

Laboratoire sur l'intégrité
de l'information

Information Integrity Lab