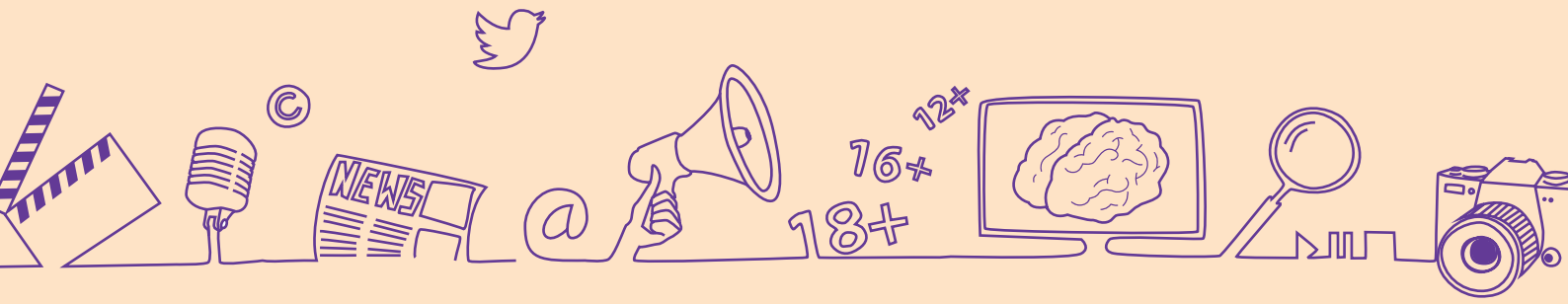
2024 | Étude | # 50

Médias & Actions citoyennes | Philippe Courteille

Deepfakes

Le mensonge à l'ère de l'intelligence artificielle



Citoyenneté & Participation

: lien consultable ou téléchargeable

TABLE DES MATIÈRES

INTRODUCTION : DEEPFAKES, OU HYPERTRUCAGES, APPRENDRE À NE PLUS CROIRE CE QU'ON VOIT.

« Mentez, mentez, il en restera toujours quelque chose ». Cette célèbre, et peut-être apocryphe, citation de Voltaire n'est malheureusement pas dénuée de vérité. Le souci est que les quelques rumeurs que partageaient nos aïeux en petit comité se sont transformées, depuis internet et les réseaux sociaux, en flots continus et mondialement diffusés des fameuses fake news, ou désinformations intentionnelles, qui furent l'objet de notre dernière étude[1]. Celles-ci sont particulièrement rentables car plus partagées sur les réseaux que les informations avérées[2]. Elles peuvent aller des plus légères aux plus déstabilisantes, des plus basiques aux plus élaborées comme désormais les deepfakes ou hypertrucages. Ringards les textes pernicieux bien ficelés ou les trucages photos staliniens, désormais, les logiciels offrent la possibilité d'imiter l'écriture[3] et la voix de n'importe qui, tout comme de changer le mouvement des lèvres, le visage ou le corps d'une personnalité sur une vidéo pour lui faire dire et/ou faire ce qu'elle n'a jamais dit et/ou fait. Et l'intelligence artificielle accélère leur perfectionnement. Déjà des sites proposent d'en réaliser en quelques minutes. Ces technologies, qui évoluent à une vitesse vertigineuse, sont désormais à la portée de tout individu, mais aussi de toute équipe de communication ou de dirigeants malintentionnés.

Glenn Kessler, le rédacteur en chef de la chronique de vérification des faits du Washington Post, soulignait déjà en 2019 : « *Nous avons vu une explosion de vidéos qui sont délibérément déformées, ou qui sont en train d'être montées d'une manière ou d'une autre pour changer la façon dont les gens voient ce qui s'est passé, cela va jusqu'aux deepfakes (…) Au cours des deux dernières années, nous avons étendu le factchecking*[4] *aux vérifications de faits vidéo. Ils obtiennent cinq fois plus de vues que nos vérifications des faits de texte. C'est une indication du nombre de personnes supplémentaires qui obtiennent leurs informations par vidéo plutôt que par écrit* »[5].

Depuis, les spéculations vont bon train quant aux dérives hypothétiques d'un tel outil mis entre les mains du premier venu. Cela flirte parfois avec la science-fiction mais nous allons tenter d'énumérer les plus vraisemblables et d'en analyser les risques potentiels.

L'un des premiers deepfakes européens a eu lieu en Belgique dès 2019 et même la Première ministre Sophie Wilmes en a été victime l'année suivante, sans conséquence sur la crédibilité de celle-ci. Mais en 2023 un rapport de la société Sumsub[6], entreprise anglaise de sécurité en ligne, annonce que la Bel-

1 Courteille P., « Fakeland, un nouvel et obscur continent », *Citoyenneté & Participation*, Étude n°31, Avril 2020, [en ligne :] http://www.cpcp.be/publications/fakeland.

2 Selon une étude de la revue Science parue en 2018. Démontrant que ces chiffres sont davantage dus aux humains qu'aux bots, ces logiciels réalisant seuls des opérations sur internet. Temming M., « On Twitter, the lure of fake news is stronger than the truth », *Science News*, 8 mars 2018, [en ligne :] (https://www.sciencenews.org/article/twitter-fake-news-truth, consulté le 6 mars 2020.

3 Cellan-Jones R., « Rory Cellan-Jones », *BBC News Technology*, 12 août 2016, [en ligne :] https://www.bbc.com/news/technology-37046477, consulté le 2 avril 2020.

4 Vérification des faits, le plus souvent faite par des professionnels de l'info, comme des journalistes.

5 Kessler G., « introducing fact checkers guide manipulated video », *Washington Post*, le 25 juin 2019, [en ligne :] https://www.washingtonpost.com/politics/2019/06/25/introducing-fact-checkers-guide-manipulated-video/, consulté le 2 avril 2020.

6 La rédaction de Sumsub, « Sumsub Research : Le nombre d'incidents liés au Deepfake a décuplé entre 2022 et 2023 », *Sumsub*, le 28 novembre 2023, [en ligne :] https://sumsub.com/newsroom/sumsub-launches-advanced-deepfakes-detector, consulté le 2 avril 2024.

gique fait partie des pays les plus touchés par l'explosion des deepfakes voyant le nombre de fraudes par deepfake exploser de 2 950 % cette année-là, presque autant que les Américains, avec +3 000 % de cas.[7]

Le sujet était relativement peu abordé en Europe avant 2022-23, en comparaison avec les craintes qu'il suscite aux États-Unis depuis cinq, six ans. Il faut dire que ce pays est encore sous le coup des élections de 2016 et du scandale Cambridge Analytica ou encore des immixtions russes dans ce processus électoral. Depuis l'élection de Donald Trump et le Brexit, les fake news cumulées aux données personnelles des citoyens sont vues comme une arme de propagande très tentante pour des équipes politiques à travers le monde. Beaucoup d'élites américaines se demandent quelles seraient dès lors les conséquences de vidéos truquées dans les campagnes à venir, d'autant que les dernières étaient de plus en plus nauséabondes. Rappelons que Donald Trump n'hésitait pas à déclarer qu'Obama n'était pas américain, à qualifier Hillary Clinton de « crooked »[8], « crapule » en français, Joe Biden de « sleeping Joe », Joe l'endormi, et Kamala Harris de « folle », de « stupide comme un roc » et de « clocharde ». Dès 2019, Trump était capable de partager à chaud sur Twitter des photos et des vidéos de désinformations, notamment à l'encontre de Nancy Pelosi, présidente démocrate de la Chambre des représentants. Mais l'usage des hypertrucages était resté anecdotique en politique jusqu'en 2024, année qui a vu leur multiplication et des élections pour la moitié de l'Humanité. D'autant que Trump est désormais soutenu par Elon Musk, un homme influent et peu regardant sur la véracité des informations qui circulent sur son réseau X (ex Twitter), tout comme sur l'utilisation de son IA Grok pour propager de la désinformation et des deepfakes. Le camp démocrate ne manque pas non plus de mordant et est capable, dans sa communication, de flirter avec les limites de la bienséance voire de la légalité.

Désormais ces trucages deviennent bluffant et se multiplient dans nombres de propagandes, d'arnaques ou d'intimidations. Et nous le verrons, les femmes en subissent de particulièrement perverses et violentes.

Internet est devenu l'empire de la désinformation et beaucoup de responsables et de spécialistes s'en inquiètent, allant jusqu'à parler « l'infocalypse »[9] . Mais est-il vraiment raisonnable d'imaginer qu'un jour une majorité de citoyens, excités par la lumière bleue de leurs écrans, espérant y trouver la lune tels des papillons de nuit devant une ampoule incandescente, prendraient le risque d'y brûler les ailes de leur liberté démocratique ? Rien n'est moins sûr, même si pour beaucoup mieux vaut prévenir que guérir.

En revanche, ce qui est clair, c'est que le terrain numérique, source de débats polarisés et de croissance des partis extrémistes, est de plus en plus propice à une désinformation par l'image et/ou le son, avec un réalisme déconcertant. Le tout propagé à grande vitesse par des bots, des robots, chargés de les disséminer par millions. « *Un mensonge répété dix fois reste un mensonge, répété mille fois, il*

7 Gaudiaut T., « Intelligence artificielle : les deepfakes explosent », *Statista*, le 18 mars 2024, [en ligne :] https://fr.statista.com/infographie/31929/pays-ayant-connu-les-plus-fortes-hausses-de-cas-de-deepfakes, consulté le 23 juillet 2024.

8 AFP, « Face au tweet de Trump, Hillary Clinton répond par une réplique culte du film "Lolita malgré moi" », *Le Soir*, 6 mars 2019, [en ligne :] https://www.lesoir.be/210738/article/2019-03-06/face-au-tweet-de-trump-hillary-clinton-repond-par-une-replique-culte-du-film, consulté le 2 avril 2020.

9 Ovadya A., « Quoi de pire que de fausses nouvelles ? La distorsion de la réalité elle-même. », *The Washington Post*, 22 février 2018, [en ligne :] https://www.washingtonpost.com/news/theworldpost/wp/2018/02/22/digital-reality/?noredirect=on, consulté le 18 septembre 2020.

devient alors une vérité », cette phrase attribuée à Joseph Goebbels, l'un des plus implacables propagandistes de l'Histoire, résume assez bien une méthode qui a fait ses preuves.

Nous tentons dans cette publication d'évaluer les risques qui peuvent en découler et les facteurs pouvant favoriser leur croissance, que ce soit au niveau sociétal, économique ou politique. Car si cet outil peut offrir une part d'amusement ou de sensibilisation, on constate d'ores et déjà qu'il perfectionne principalement la tromperie, la fraude, la vilenie et le chantage.

À l'aune des inquiétudes d'experts de la question, nous analysons ensuite les pistes de solutions.

Mais commençons par comprendre ce que sont les deepfakes et quelle est leur origine.

I. DEEPFAKE, LE PROFONDÉMENT TROMPEUR

Comme Saint-Thomas, beaucoup ne croient que ce que qu'ils voient. Et dans notre monde de l'image toute puissante, les deepfakes risquent d'en déstabiliser plus d'un. Le terme est un mélange de fake news, soit une désinformation intentionnelle, et de *deep learning*, qui désigne un type d'intelligence artificielle où la machine « apprend » par elle-même, à partir de sa propre observation de divers phénomènes. On appelle ces derniers des algorithmes d'apprentissage, par opposition aux algorithmes de programmation qui se contentent d'exécuter des ordres donnés. « *Ce sont des faux, quelle que soit la nature du contenu – vidéo, photo, audio ou texte – conçus grâce à l'intelligence artificielle (…) Pour l'heure, les deepfakes les plus couramment diffusés sur internet sont des vidéos truquées dans lesquelles le visage et la voix d'une personne connue sont falsifiés, lui faisant dire ou faire ce qu'elle n'a jamais dit ou jamais fait* »[10]. Attention il s'agit bien d'un *hypertrucage* et non d'une astuce de montage, de type ralenti ou coupure d'une partie du discours, que beaucoup ont tendance à englober dans le terme deepfake et que d'autres nomment *cheapfake* (littéralement « le faux bon marché »). Par exemple si on coupe une partie du discours d'un homme politique pour lui faire dire autre chose.

On savait que les hypertrucages vidéos étaient possibles après avoir vu au cinéma Forrest Gump serrer la main de JFK ou lorsque, dans le film *Rogue One, une histoire de Star Wars*, sorti en 2016, avec le personnage de Grand Moff, réapparu sous les traits de Peter Cushing, l'acteur qui l'avait incarné dans un épisode précédent de la saga et mort… en 1994, soit vingt-deux ans auparavant. Tout cela était fait par des studios professionnels avec d'énormes puissances de calcul pour modifier chaque image d'une vidéo. Ça demandait aussi de gros investissements ce qui limitait leur nombre et leur impact. Mais ça, c'était avant.

10 Laugée F., « Deepfake », *La revue européenne des médias et du numérique*, automne 2019, [en ligne :] https://la-rem.eu/2019/11/deepfake, consulté le 23 avril 2020.

A. GAN, un ping-pong cognitif

Les évolutions technologiques, susceptibles de faire évoluer les deep-fakes, explosent depuis quelques années et ne cessent de surprendre par leur réalisme grandissant. Ces progrès ont pu être réalisés, à la base, grâce à une technique appelée GAN (Generative Adversarial Networks) soit des Réseaux Antagonistes Génératifs en français. Il s'agit en fait d'une classe d'algorithmes d'apprentissages non-supervisés par l'homme. En clair, deux réseaux sont placés en compétition. Le premier réseau est le « générateur », il génère par exemple une image, tandis que son adversaire, le « discriminateur » essaie de détecter si l'image est réelle, à partir de sa base de données d'images, ou bien si elle est le résultat du générateur. Ces deux réseaux s'entraînent l'un l'autre dans le cadre d'une relation contradictoire, s'échangeant les données et les résultats de leurs analyses. Les deux algorithmes entretiennent donc une relation gagnant-gagnant d'amélioration continue.[11]

Wintics, start-up parisienne qui travaille sur l'intelligence artificielle et le deep learning au service notamment de la mobilité urbaine explique assez bien cette technologie : « *Prenons l'exemple des faussaires de billets de banque traqués par les policiers. Le Générateur joue le rôle d'un faussaire qui produit une liasse de 100 faux billets de banque (dont les designs sont tous différents). Il la présente à un policier (le Discriminateur) qui, grâce à l'observation d'une base de données de billets authentiques qui lui a été transmise, a des connaissances basiques en identification de billets contrefaits. Le policier va donc analyser les billets du faussaire et les classer en deux catégories : ceux qu'il pense être vrais et ceux qu'il pense être faux. À chaque fois qu'un faux billet est identifié par le policier, celui-ci est renvoyé au faussaire. Cela va permettre à ce-dernier de connaitre les designs qui n'ont pas été capables de tromper la police et par symétrie, ceux qui ont été assez réalistes pour passer à travers les contrôles. Par cette logique d'apprentissage, le faussaire va pouvoir créer de nouveaux billets plus réalistes et les représenter au policier. Celui-ci donnera une nouvelle fois son verdict et ainsi de suite. Le processus s'arrête lorsque le faussaire (le Générateur) est capable de créer des billets qui trompent le policier (le Discriminateur) à tous les coups* »[12].

Les GAN peuvent ainsi par exemple faire évoluer des designs en fonction de contraintes physiques ou augmenter la résolution d'une image.

Et Wintics de conclure : « *Avec l'apparition des GAN, la Data Science s'est dotée d'un formidable outil de création et s'attaque ainsi à ce qui semblait être un des derniers prés carrés de l'intelligence humaine* ».

11 Wintics, « quand la data science devient creative avec les gan », *Wintics*, juin 2020, [en ligne :] http://wintics.com/fr/quand-la-data-science-devient-creative-avec-les-gan, consulté le 17 juin 2020.

12 *Ibid.*

B. Holly GAN et jeu de dupes

Comme l'expliquait l'un des spécialistes belges de la question, Charles Cuvelliez, professeur à l'École polytechnique de Bruxelles (ULB), sur les ondes de la RTBF, les deepfakes peuvent être déclinés en trois catégories [13] :

1. Face Swapping : Qu'est-ce que mon visage fait sur ce corps ?

Fin 2017, un développeur, se faisant appeler Deepfakes sur le forum Reddit [14], avait réussi à insérer des visages de célébrités dans des films pornographiques. Il a ainsi conçu « un programme capable d'automatiser ce processus, en se basant notamment sur une technologie d'IA [15] mise à disposition gratuitement par Google, nommée Tensorflow. Son système, « nourri » de centaines de photos et de vidéos de la star choisie glanées sur le Web, est ensuite capable de déformer suffisamment le visage d'une actrice de film pornographique pour qu'elle ressemble au modèle qu'a « appris » le programme [16]. Puis ce fut un autre internaute qui mit en ligne « *un programme similaire, ne nécessitant pas de compétences pointues. C'est alors l'emballement : les internautes s'emparent de ce logiciel nommé FakeApp, gratuit, et se mettent à publier en masse leurs créations, aidés par des modes d'emploi détaillés* » [17]. Au fil des expériences de chacun, le programme se perfectionne et des bases de données d'images de stars, indispensables à la création de ces vidéos, sont partagées. Car pour arriver à un résultat satisfaisant, il fallait disposer d'un grand nombre d'images, les personnalités publiques étaient donc particulièrement visées. Déjà « *en 2018 une entreprise du secteur avait même annoncé pouvoir insérer ses clients dans leurs vidéos favorites accompagnés des actrices "de leur choix"* » [18].

Le principe, appelé « face swapping » (échange de visages) ne se limitera bien sûr pas aux films érotiques ou pornographiques. Des visages, dans des scènes de films cultes, seront par exemple remplacés par d'autres et les déclinaisons vont se multiplier.

Et il ne faut pas aller jusqu'aux États-Unis pour trouver des amateurs de ce type de trucage vidéo. L'un des premiers deepfake de l'histoire aurait été fait en Belgique. Dans un reportage de M6 de mars 2019, l'Anversois Sven Charleer présente ainsi ses vidéos dans lesquelles il s'amuse à placer le visage de sa femme

13 Cuvelliez C., Emission radio « Les deepfakes : 5 questions pour comprendre comment faire dire n'importe quoi à n'importe qui », La Première – Culture (RTBF), lundi 17 juin 2019, [en ligne :] https://www.rtbf.be/lapremiere/article/detail_les-deepfakes-5-questions-pour-comprendre-comment-faire-dire-n-importe-quoi-a-n-importe-qui?id=10248164, consulté le 10 avril 2020.

14 Reddit est un site web communautaire américain d'actualités sociales fonctionnant via le partage de signets permettant aux utilisateurs de soumettre leurs liens et de voter pour les liens proposés par les autres utilisateurs. Ainsi, les liens les plus appréciés du moment se trouvent affichés en page d'accueil. Fondé en 2005, Reddit contenait alors surtout du contenu sur la programmation et la science. Il ne cesse depuis de se diversifier et de s'ouvrir à du contenu plus grand public. Ayant connu une explosion de croissance en 2010, en 2020, il se place comme le vingtième site web le plus populaire au monde et le sixième aux États-Unis selon Alexa Internet.

15 Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine.

16 Tual M., « Du porno aux fausses informations, l'intelligence artificielle manipule désormais la vidéo », *Le Monde* (Pixels), 04 février 2018, [en ligne :] https://www.lemonde.fr/pixels/article/2018/02/04/du-porno-aux-fausses-informations-l-intelligence-artificielle-manipule-desormais-la-video_5251535_4408996.html, consulté le 30 mars 2020.

17 Ibid

18 Maillé P., « "Deepfake", quand le faux gagne la vidéo », *Libération*, 31 août 2018, [en ligne :] https://www.liberation.fr/futurs/2018/08/31/deepfake-quand-le-faux-gagne-la-video_1675699, consulté le 18 juillet 2019.

sur le corps de l'actrice américaine Anne Hathaway. Il déclare lui-même : « *C'est un outil très puissant, entre de mauvaises mains il peut être destructeur. Je pense qu'on ne réalise pas son pouvoir* » [19]. Heureusement les deepfakes qu'il réalise sont encore décelables pour qui s'y attarde un peu.

Puis la technique n'a cessé d'évoluer. L'application Zao a fait parler d'elle à l'été 2019. C'est une application mobile qui permettait de réaliser très facilement du *face swapping*. Il suffisait à l'application d'une seule photo de notre visage, ou d'une petite vidéo pour un meilleur effet, pour que notre tête soit transposée sur celle d'un acteur, dans une séquence vidéo déterminée, et ce en quelques secondes. Des millions de gens se sont ensuite amusés à placer des visages sur des clips, des films, des vidéos grâce à FaceApp, Zao, Morphin ou Reface app. Cette dernière application, lancée par l'entreprise ukrainienne Reface AI, a fait le buzz à l'été 2020. En quelques jours, elle s'est hissée à la première place du classement des meilleurs téléchargements sur le PlayStore. Il suffisait de faire un selfie et de sélectionner le clip dans lequel vous vouliez intégrer votre visage. Bien sûr ces quelques minutes d'amusement s'échangent contre une utilisation par ces sociétés, qu'elles soient chinoises ou ukrainiennes, des informations et photos fournies. De quoi fournir une belle base de données de reconnaissance faciale, mais c'est un autre débat.

Désormais, des sites comme Undress AI ou PTool proposent de placer des visages sur des photos de femmes nues. Des sites qui ventent leur facilité d'utilisation. Un jeu d'enfants. Nous verrons que cela peut représenter de sérieux dangers démocratiques et humains.

2. Ne me faites pas dire ce que je n'ai pas dit !

Ici, on ne se contente pas de mettre un visage sur le corps d'un autre, c'est le discours qui est modifié. L'une des premières techniques était celle du Lip sync, pour synchronisation labiale, qui consistait à ne modifier que les lèvres, et leurs contours, d'une personne pour les adapter à un autre discours. L'une des plus connues est celle où on voit Barak Obama insulter Donald Trump. Le trucage, révélé au public en avril 2018, a été fait par le réalisateur et comédien Jordan Peele, avec Adobe After Effects, un logiciel vidéo facilement disponible, et FakeApp. L'objectif poursuivi était d'éveiller les consciences aux problèmes des deepfakes. [20]

Les hypertrucages ne sont bien sûr pas l'apanage de pirates. De nombreux débouchés leur sont trouvés comme corriger le bafouillement d'un acteur dans une prise de tournage ciné. Dès 2019, une équipe de chercheurs de l'Université de Stanford, de l'Institut Max Planck, de l'Université de Princeton et d'Adobe Research avait déjà mis au point « *une version simplifiée de ces trucages, avec l'aide des développeurs des logiciels Adobe* » [21]. A l'aide d'une vidéo existante et suffisamment longue de quelqu'un en train de parler, on peut donc lui faire dire d'autres choses de façon assez naturelle. Même des transitions de mouvements de mains ou de corps sont bluffantes. Pour les besoins de leur démonstration,

19 Lemega C. et Ben Sassi P., « Deepfake, les visages du danger », *M6 - Le Mag*, 12h45, 12 mars 2019, [en ligne :] https://www.rtl.fr/actu/futur/video-les-deepfakes-ces-fake-news-video-ultra-realistes-7797188493, consulté le 1 avril 2020.

20 Buzzfeedvideo, « You Won't Believe What Obama Says In This Video! », *Youtube*, 17 avril 2018, https://www.youtube.com/watch?v=cQ54GDm1eL0&feature=emb_logo, consulté le 20 mai 2020.

21 Lesaffre C., « Attention, les « deepfake », des vidéos truquées, se multiplient sur le Net », *Europe 1*, 17 juin 2019, [en ligne :] https://www.europe1.fr/technologies/attention-les-deepfake-des-videos-truquees-se-multiplient-sur-le-net-3904942, consulté le 1 avril 2020.

les scientifiques ont transformé la célèbre phrase tirée du film *Apocalypse Now* « *j'aime l'odeur du napalm au petit matin* » en « *j'aime l'odeur du pain grillé au petit matin* ». *Impossible de distinguer l'original de la supercherie* » [22].

Plus fort encore, le « face2face » a permis, dès 2016 déjà, de falsifier une vidéo quasiment en direct. Grâce à l'approche conçue par des étudiants allemands et américains [23], on utilise un acteur source comme vous et moi pour faire faire les mêmes mouvements et expressions à une vidéo d'un acteur cible. N'importe qui peut ainsi servir d'acteur source pour faire faire des grimaces à une vidéo de Vladimir Poutine, par exemple. « *Il ne s'agit plus de coller son visage sur celui d'une star dans un blockbuster hollywoodien, mais d'animer le visage d'une personnalité avec des mimiques et des paroles inventées, ce qui pourrait par exemple permettre de produire une fausse conférence de presse d'un chef d'État, le tout en direct* », s'inquiète Camille Toussaint, journaliste à la RTBF [24].

Soulignons aussi l'arrivée d'Avatarify, un programme qui superpose le visage de quelqu'un d'autre au vôtre en temps réel, lors de visioconférences. Tout le monde peut ainsi avoir par exemple le visage d'Elon Musk pendant une conférence sur Zoom ou Skype. [25]

Diverses technologies, susceptibles de faire gagner en qualité les deepfakes, évoluent ainsi en parallèle comme l'application Lyrebird qui développe son propre outil de clonage des voix. [26] En avril 2018, Lyrebird annonçait avoir développé une technologie d'intelligence artificielle capable d'imiter n'importe quelle voix en se basant sur un enregistrement « d'une minute seulement ».

Depuis, l'IA a révolutionné l'accessibilité aux deepfakes. Le célèbre ChatGPT a ouvert le bal et des applications comme Midjourney ou Dall-E ont connu un succès colossal. Nous n'avons pas tardé à voir de fausses images d'Emmanuel Macron en train de ramasser des poubelles dans Paris, du pape François en doudoune blanche ou encore de Donald Trump en tenue orange de prisonnier US. Des photos satiriques qui ont permis à beaucoup de monde de prendre connaissance du phénomène deepfake en Belgique, et ça n'est pas plus mal. On a pu voir également quantité de photos de quidams, sur lesquelles ils s'étaient transformés en rockeur, en viking, en personnage de série culte ou en train de faire un selfie avec Albert Einstein. Des applications plutôt ludiques de manière générale et essentiellement de photos. Car des sites nous proposent de créer nos deepfakes quasi en direct, comme deep-fake.ai qui nous annonce « *Deep-Fake.ai propose deux fonctionnalités intéressantes : l'image deepfake et la vidéo deepfake. Alors que la fonctionnalité d'image deepfake est déjà disponible, la fonctionnalité*

22 Desaunay D., « La menace du "deepfake" se précise », *RFI*, 14 juin 2019, [en ligne :] http://www.rfi.fr/fr/emission/20190615-menace-deepfake-precise, consulté le 31 mars 2020.

23 Dont le professeur Michael Zollhöfer, professeur assistant invité à l'Université de Stanford, et ses collègues de l'Université technique de Munich, de l'Université de Bath, de Technicolor et d'autres institutions. Thies J., Zollhöfer M., Stamminger M., Theobalt C., Niebner M., « Face2face: Real-time Face Capture and reenactment of RGB Videos », *Youtube*, 2016, [en ligne :] https://www.youtube.com/watch?v=ohmajJTcpNk, consulté le 9 avril 2020.

24 Toussaint C. avec AFP, « Les vidéos "deepfakes" se perfectionnent et inquiètent les chercheurs », *RTBF*, 11 septembre 2019, [en ligne :] https://www.rtbf.be/info/societe/detail_les-videos-deepfakes-se-perfectionnent-et-inquietent-les-chercheurs?id=10312839, consulté le 9 avril 2020.

25 Cole S., « Ce programme open-source vous permet de profiter des réunions Zoom, en temps réel », *Vice*, 16 avril 2020, [en ligne :] https://www.vice.com/en_us/article/g5xagy/this-open-source-program-deepfakes-you-during-zoom-meetings-in-real-time, consulté le 20 avril 2020.

26 Grand H., « Après les fake news, la menace du "deep fake" prend de l'ampleur sur le web », *Le Figaro Tech&Web*, 2 janvier 2019, [en ligne :] https://www.lefigaro.fr/secteur/high-tech/2019/01/02/32001-20190102ARTFIG00162-apres-les-fake-news-la-menace-du-deep-fake-prend-de-l-ampleur-sur-le-web.php, consulté 10 avril 2020.

vidéo deepfake est actuellement en préparation et sera bientôt publiée. Restez à l'écoute pour cet ajout passionnant qui permettra aux utilisateurs de créer des vidéos deepfake réalistes et captivantes en un rien de temps ».

Nous verrons par la suite les éventuels risques que cela peut représenter, notamment une fausse vidéo jointe à une fausse voix qui peuvent tromper bien des gens inattentifs.

3. Qui est cette personne ?

Certains se sont peut-être déjà amusés à chercher lequel des deux visages proposés en photo sur www.whichfaceisreal.com était réel. À chaque fois un des deux individus est créé de toute pièce. « *Les résultats du jeu Which Face Is Real ?, mis en ligne par deux professeurs de l'université de Washington, Jevin West et Carl Bergstrom, afin de tester la technologie de Nvidia, ne sont pas rassurants. Sur 6 millions de parties jouées par 500 000 personnes, le taux de réussite est de 60 % dès le premier essai mais ne dépasse pas 75 % avec de l'entraînement* ».[27]

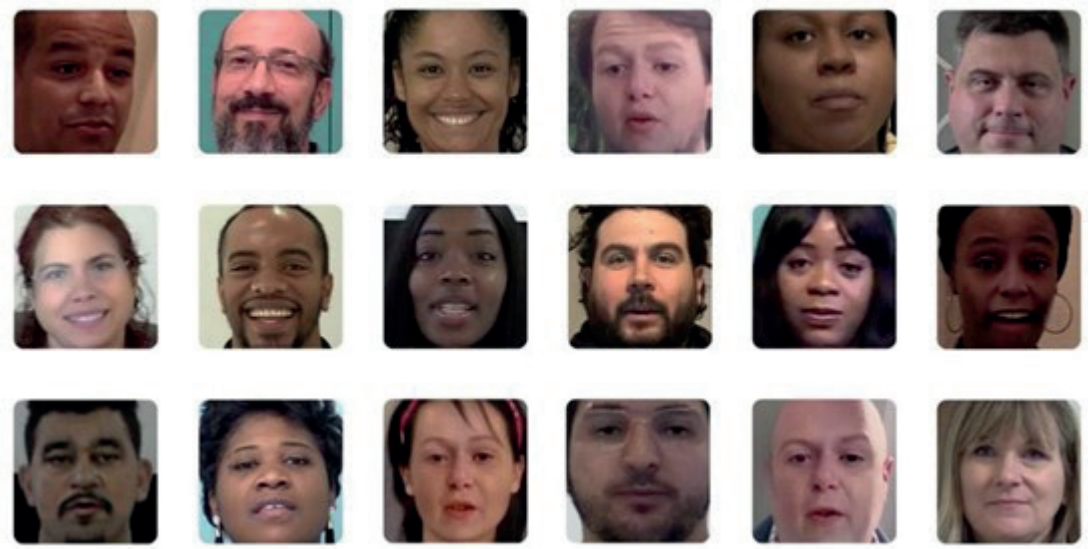


Image : Facebook (voir si whichfaceisreal.com)

Et ces personnes ultra réalistes fabriquées de toutes pièces grâce au GAN, peuvent désormais être animées en vidéo. Et d'après Charles Cuvelliez, ce sont les deepfakes les plus difficiles à déceler.

27 Laugée F., « Deepfake », *La revue européenne des médias et du numérique*, automne 2019, *op. cit.*

II. LES DÉBOUCHÉS DU « FAUX PROFOND »

La technologie de pointe derrière ces vidéos, a une gamme d'applications commerciales, en particulier dans les industries créatives telles que la publicité, le champ artistique ou la réalisation de films. Les publicités et les campagnes peuvent être plus facilement doublées, les contraintes de temps des acteurs peuvent être réduites et les effets spéciaux peuvent être créés plus rapidement et en toute sécurité. Des campagnes de sensibilisation peuvent également mettre en image une réalité future ou une situation hypothétique.

Citons cette jolie initiative du musée Salvador Dali de Floride qui a redonné vie à l'artiste espagnol sur écran pour accueillir les visiteurs. Le résultat est bluffant. On peut même faire un selfie avec lui.[28] Côté belge, on n'est pas en reste. Christopher Ume, et sa société Metaphysic, qui s'est fait connaître avec des deepfakes de Tom Cruise devenues virales sur Tik Tok[29], a créé l'événement en 2022 dans la célèbre émission de télé crochet américaine America's got Talent.[30] Il y a notamment fait chanter les membres du jury sur grand écran aux côtés d'Elvis Presley, le tout via la technique deepfake.[31] Le public et le jury ont adoré. Et il y a fort à parier qu'on reverra ce style de show mêlant des personnalités du passé à d'autres actuelles, voire même à des personnages tout à fait imaginaires. L'IA n'a pas fini de nous surprendre.

Mais d'étranges déclinaisons voient également le jour. Exemple récent, en février 2024, un article du journal Le Monde titrait *« TikTok et le business des récits de faits divers dopés au deepfake … Un visage, appartenant à une personne réelle ou fictive et animé par l'IA, vous fait de la voyance, raconte un fait divers glauque ou un récit incroyable. Le Monde a ainsi recensé une centaine de comptes TikTok spécialisés dans ce format, dont certains affichent plusieurs centaines de milliers d'abonnés, et des vidéos en français, arabe, anglais, espagnol, allemand ou italien qui cumulent parfois des millions de vues »*[32]. Le deepfake est ainsi utilisé pour ajouter de la crédibilité au récit. Voir un criminel raconter ses horreurs paraît ainsi intéresser les amateurs du genre, même si cela reste une initiative quelque peu glauque.

« TikTok travaille également sur une nouvelle option qui permettrait aux marques de déployer des influenceurs virtuels pour promouvoir leurs articles. Ces derniers seraient en mesure de vendre les produits via des vidéos et des diffusions en direct »[33]. Mais cette tendance semble mieux acceptée par les consommateurs asiatiques qu'européens, pour le moment.

28 The Dali Museum, « Behind the Scenes: Dalí Lives », *Youtube*, [en ligne :] https://www.youtube.com/watch?v=-BIDaxl4xqJ4&t=178s, consulté le 20 août 2024.

29 Tondeur C., « Un Belge, derrière les deepfake de Tom Cruise », *RTBF*, le 8 mars 2021, [en ligne :] https://www.rtbf.be/article/un-belge-derriere-les-deepfake-de-tom-cruise-10714393, consulté le 20 août 2024.

30 Dekock C., « Chris Umé, ce Belge qui a bluffé l'Amérique en faisant chanter un air d'opéra aux jurés d'America's got Talent », *RTBF*, le 2 septembre 2022, [en ligne :] https://www.rtbf.be/article/chris-ume-ce-belge-qui-a-bluffe-l-amerique-en-faisant-chanter-un-air-d-opera-aux-jures-d-america-s-got-talent-11058573, consulté le 20 août 2024.

31 Voir extrait sur Youtube, [en ligne :] https://www.youtube.com/watch?v=rJfwx8iZel0, consulté le 20 août 2024.

32 Reynaud F. et Croquet P., « TikTok et le business des récits de faits divers dopés au deepfake », *Le Monde*, le 27 février 2024, [en ligne :] https://www.lemonde.fr/pixels/article/2024/02/27/tiktok-et-le-business-des-recits-de-faits-divers-dopes-au-deepfake_6218894_4408996.html, consulté le 20 aout 2024.

33 Rochefort M., « Des influenceurs générés par l'IA pourraient bientôt envahir TikTok », *Siècle Digital*, le 12 avril 2024, [en ligne :] https://siecledigital.fr/2024/04/12/des-influenceurs-generes-par-lia-pourraient-bientot-envahir-tiktok, consulté le 21 août 2024.

Il faut dire que des deepfakes, présentateurs de journaux télévisés ou influenceurs créés de toute pièce peuvent travailler vingt-quatre heures sur vingt-quatre sans se fatiguer. La technique peut également rendre les jeux vidéo hyperréalistes et plus immersifs. Avoir son visage sur un héros de jeu pourfendant le mal sera certainement prisé par les amateurs. Ces personnages pourraient même être transposés à nos profils sur les réseaux sociaux, donnant une image idéalisée de ce que nous voulons montrer de nous.

Du côté de la mode et des achats en ligne, les débouchés sont nombreux. Vous pouvez essayer virtuellement les articles qui vous plaisent. « *Afflelou, Optic 2000 ou encore Atol offrent à leurs clients la possibilité d'essayer leurs lunettes depuis chez eux. De son côté, l'entreprise SuitUs propose une cabine d'essayage en ligne pour les marques de vêtements. Cette technologie crée un double corporel pour diminuer de 50% le nombre de retours lié au e-commerce* »[34].

Si les possibilités utiles sont nombreuses, nous allons nous intéresser à présent aux divers dangers que peuvent représenter les hypertrucages.

III. QUELQUES DÉRIVES POUR LA NAVIGATION

Le succès des fake news sur le net a surpris par son ampleur et par l'irrationalité de nombreuses d'entre-elles. Des réactions émotionnelles ont permis à certaines de faire un buzz à plusieurs millions de vues. Avec la possibilité de créer de la vidéo et du son, beaucoup craignent encore plus les dérives à l'échelle de la planète, qu'elles soient sociétales (incitations au désordre social, influence sur les pratiques et les pensées des citoyens), criminelles (la falsification de preuves, l'extorsion, la fraude ou encore les problèmes de droits d'auteur), sociétales (harcèlement, l'intimidation). Voici les scénarios de risques possibles, certains plus réalistes que d'autres. À chacun d'en juger.

A. Nouveaux marchés et appâts du GAN

Qui dit nouvel outil, dit nouvelles possibilités. Parmi celles-ci, se trouvent souvent des moyens de se faire de l'argent illégalement ou d'être tenté de flirter avec des lois inadaptées et/ou archaïques.

1. Faire du clic, faire du fric :

Pour l'instant, il s'agit surtout d'amusement, en mettant par exemple le visage de Silvester Stallone sur le corps d'Arnold Schwarzenegger ou celui de Di Caprio sur notre propre corps.

[34] IA School, « Quels sont les dangers du phénomène deepfake ? », *IA School*, [en ligne :] https://www.intelligence-artificielle-school.com/ecole/technologies/quels-sont-les-dangers-deepfake, consulté le 17 août 2024.

Mais, à l'instar des fake news, on peut aisément imaginer des deepfakes créés pour faire du clic [35] sur le dos de personnalités, et donc ramener des revenus publicitaires intéressants tout en profitant d'un bon référencement sur Google. On a ainsi vu nombre de fake news être bien plus rentables que des vraies dans le domaine politique. Des vidéos, sorties de leur contexte, ont ainsi surfé sur la polarisation des débats sur les réseaux sociaux et connu un gros succès. Pendant la campagne présidentielle américaine de 2016, n'a-t-on pas vu Paul Horner, appelé le roi des fake news [36], avouer détester Trump, tout en propageant nombre d'absurdités à propos d'Obama ou d'Hillary Clinton. Son explication était simple, pour lui les anti-Démocrates étaient ceux qui relayaient le plus d'infos sans les vérifier. [37] Il produisait donc les infos qui rapportaient et tant pis si des gens étaient assez idiots pour les croire. Pour lui, il n'y avait pas d'intention de nuire dans ses actes. On peut imaginer une transposition du phénomène, faux audios et/ou vidéos à l'appui.

D'ailleurs on a aussi pu voir, début 2023, dans un hypertrucage devenu viral (10,5 millions de vues), Elon Musk avouer, face caméra, s'être drogué et expliquer être prêt à imaginer de « *nouvelles voitures spatiales* » et à conquérir Mars [38]. C'était en fait un canular d'un habitué du genre, lancé sur Twitter, qui aura quand même fait 7,5 millions de vues. Mais combien de personnes y auront cru, impossible de le savoir. Le plaisantin aura en tout cas fait le buzz et un peu d'argent.

Par ailleurs, les influenceur·se·s deepfakes permettent d'aller encore plus loin pour vendre et attirer des followers. Début 2024, on a beaucoup parlé d'Adrianna Avellino, influenceuse générée par une intelligence artificielle (IA). Elle cumulait plus de 94 000 abonnés sur son compte. Sur son profil, un lien vers une page Fanvue, concurrent d'Onlyfans, avec des photos d'elle dénudée, moyennant un abonnement de cinq dollars par mois. Pour ce faire, « *selon le média américain 404media, des dizaines de vidéos d'utilisatrices Instagram ont été volées sans leur consentement et détournées pour remplacer leur visage, par celui d'Adriana. Avec, à la clé, des millions de likes et d'abonnés* » [39]. Des cas qui se multiplient avec des comptes totalisant des centaines de milliers de followers et des millions de vues, « *en utilisant "presque exclusivement" du contenu volé. Selon le média américain Manofmany, les influenceurs IA gagnent en moyenne 3.200 à 11.000 dollars sur Onlyfans* ». Des deepfakes tellement bien réalisés qu'ils sont devenus difficiles à détecter pour les followers. Et quand ces comptes sont signalés, puis supprimés par les plateformes, d'autres apparaissent rapidement.

Il est à parier que, pour faire de l'argent, nombre d'idées vont encore voir le jour, deepfakes à l'appui.

35 Comme nous le soulignions dans notre précédente étude de Philippe Courteille, « Fakeland, un nouvel et obscur continent », *Citoyenneté & participation*, Étude n°31, Avril 2020, [en ligne :] http://www.cpcp.be/publications/fakeland.

36 De Buissiere Z., Thiebaut S. et Collot G., « Fake news : fausses infos et vrais bénéfices », *France 2 - Complément d'enquête*, le 23 mars 2017, [en ligne :] https://www.francetvinfo.fr/monde/usa/video-fake-news-fausses-infos-et-vrais-benefices_2107586.html?fbclid=IwAR3JEmkRmNeqH_oI8pxrzmxOj1KaYLmZAfc92y-NL7hl0k-gbgpnn1qufUH4, consulté le 16 juillet 2019.

37 À sa mort en 2017, beaucoup ont cru à un fake. Il serait mort dans son lit d'une overdose de médicament, de quoi alimenter encore une quelconque théorie du complot. Kulwin N., « Un cocktail de fentanyl a tué le faux journaliste qui prétendait faussement avoir élu Trump », *Vice*, le 6 déc.2017, [en ligne :] https://news.vice.com/en_ca/article/d3xx7v/a-fentanyl-cocktail-killed-fake-news-writer-paul-horner, consulté le 16 juillet 2019.

38 Saint-Léger A., « Elon Musk drogué ? Un deepfake donne de la voix sur les réseaux sociaux », *France 24*, le 12 janvier 2024, [en ligne :] https://www.france24.com/fr/%C3%A9missions/info-ou-intox/20230112-un-deepfake-d-elon-musk-drogu%C3%A9-devient-viral, consulté le 17 août 2024.

39 Ferraris S., « Sur Instagram, des influenceurs IA utilisent des deepfakes pour gagner des abonnés », *BFM Tech&Co*, le 10 avril 2024, [en ligne :] https://www.bfmtv.com/tech/instagram/sur-instagram-des-influenceurs-ia-utilisent-des-deepfakes-pour-gagner-des-abonnes_AV-202404100592.html, consulté le 12 juillet 2024.

2. Manipulations, abus de confiance, vols d'identité et autres arnaques

Durant la pandémie de Covid-19, les solutions numériques aux problèmes ont explosé et, par la force des choses, la fraude en ligne aussi. La démocratisation des deepfakes a provoqué un boum des manipulations et des arnaques. Imaginez, par exemple, un parent recevoir un coup de fil de son enfant en difficulté et qui a besoin qu'on lui envoie de l'argent immédiatement. Par ailleurs, un militaire ou un employé pourrait-il refuser d'obéir à l'ordre d'un faux supérieur hiérarchique en vidéo conférence ? L'exemple, bien que déjà ancien, de Gilbert Chikli est significatif dans ce cas de figure. En 2005-2006 l'homme s'est fait passer pour le PDG de grandes entreprises auprès de cadres et leur demandait, par téléphone, de lui transmettre des centaines de milliers d'euros. Il a ainsi réussi à dérober plusieurs millions d'euros à des dizaines de grands groupes bancaires et industriels. Appelé « arnaque au président », la technique a fait de nombreux émules depuis, révolution de l'IA et imitation de voix deepfake à l'appui. Et en mars 2019, le Wall Street Journal annonçait que des criminels avaient utilisé pour la première fois un logiciel basé sur l'intelligence artificielle pour usurper l'identité d'un chef de direction et exiger un transfert frauduleux de 220 000 euros [40]. Début 2020, ce sont trente-cinq millions de dollars qui semblent avoir été dérobés à une banque émiratie de Hong-Kong [41]. Faux mails, faux papiers, et surtout… fausse voix de directeur d'entreprise simulée par ordinateur. Plus fort encore, et toujours à Hong-Kong, début 2024 : le salarié d'une grande multinationale reçoit un mail de son directeur financier, basé à Londres, lui demandant d'effectuer de gros transferts. L'employé est méfiant mais sera vite rassuré par une visioconférence dans laquelle il reconnaît ses collègues et effectuera pour vingt-six millions de dollars de transferts. On s'apercevra qu'il avait parlé à … des avatars. Les fraudeurs avaient trouvé des vidéos et des audios accessibles au public via YouTube, puis utilisé les technologies deepfake et IA pour imiter leurs voix.

On a vu pendant la pandémie le nombre d'arnaques, d'hameçonnages et autres usurpation d'identité se multiplier sur le net. Les hypertrucages et l'IA ont considérablement perfectionné leur crédibilité apparente.

Le phénomène des *brouteurs* [42], qui font de l'arnaque aux sentiments, souvent depuis l'Afrique, pourraient avoir de beaux jours devant lui avec les deepfakes. Se faire passer pour quelqu'un d'autre devient de plus en plus facile. Un homme inscrit dans un CPAS hennuyer, nous expliquait récemment avoir perdu près de mille euros en croyant parler à une ravissante jeune femme, alors qu'il était dans une détresse affective, suite à un divorce.

Désormais les usurpations d'identité sont légions. Des personnes se font soutirer de l'argent en pensant avoir affaire à Florent Pagny ou à Conner Rousseau. Deux espagnoles ont ainsi perdu 325 000 euros, pensant entretenir une relation privilégiée avec Brad Pitt. Les malfrats avaient profilé et ciblé les deux

40 Stupp C., « Des fraudeurs ont utilisé l'IA pour imiter la voix du PDG dans une affaire de cybercriminalité inhabituelle », *The Wall Street Journal*, 30 août 2019, [en ligne :] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, consulté le 14 mai 2020.

41 Brasseur T., « Des fraudeurs ont cloné la voix du directeur d'entreprise dans un braquage de banque de 35 millions de dollars, selon la police », *Foirbes*, 14 octobre 2021, [en ligne :] https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=4b47e9097559, consulté le 26 juillet 2022.

42 Un brouteur est un escroc opérant sur Internet, notamment sur les réseaux sociaux. L'idée du « broutage », consiste à se faire envoyer de l'argent en faisant croire à une relation amoureuse ou amicale, totalement feinte en réalité.

femmes. Ils ont heureusement été arrêté mais ce qui interpelle, c'est que « *Pour trouver leurs victimes, les cybercriminels ont étudié les réseaux sociaux des femmes. Ils ont même dressé un profil psychologique d'elles … Ils ont découvert que les deux femmes étaient des personnes vulnérables, en état de dépression et en manque d'affection* »[43]. Deepfakes et profilages rendent désormais l'arsenal des escrocs particulièrement efficace et déstabilisant.

3. Droits d'auteur, comme de faux airs de faussaires

C'est l'un des grands débats du moment. Utiliser le deepfake pour imiter la voix, parfois avec l'image en plus, d'une chanteuse ou d'un chanteur, en a fait bondir quelques-uns dans le monde musical. Le 14 avril 2023, on découvrait sur YouTube, TikTok et d'autres plateformes, la chanson *Heart on my sleeve*, chantée par les Canadiens Drake et The Weeknd. Elle a rapidement fait le buzz le temps d'un week-end, « *avant que les artistes concernés et leur maison de disques, Universal Music Group, dénoncent un "fake" et obligent les plates-formes à le retirer dare-dare. Une contrefaçon confirmée au même moment par un faussaire anonyme, du nom de Ghostwriter977, écrivant dans un commentaire vidéo avoir utilisé l'intelligence artificielle (IA) pour générer les voix des deux vedettes. "J'ai été pendant des années auteur anonyme payé des clopinettes pour le plus grand profit des majors. Voici l'avenir", narguait-il* »[44]. Depuis, les plagiats n'ont plus cessé, d'Angèle à Jay-Z, jusqu'à faire revivre des chanteurs décédés comme Frank Sinatra ou Michael Jackson. Comment gérer les droits d'auteur d'un·e chanteur·se si sa voix et son style sont plagiés par quiconque ? On peut éventuellement entendre le côté amusant de la parodie, et imaginer faire danser et chanter *La danse des canards* à Taylor Swift, mais faire un gros succès sur les réseaux avec une fausse chanson de Billie Eilish. Le principe flirte avec l'usurpation d'identité.

Au-delà du côté divertissant, The Guardian[45] s'interroge sur les conséquences de ces deepfakes pour l'industrie musicale. Car OpenAI est loin d'être le seul acteur à s'intéresser aux algorithmes générateurs de musique. Google a créé en 2016 le projet Magenta, dont le but est de mettre au point des intelligences artificielles créatives. Spotify s'est doté d'un Creator Technology Research lab, à l'origine de « Hello World », premier album composé avec une IA. En 2023, des fans d'Oasis, fatigués d'attendre une réconciliation des frères Gallagher et une reformation du groupe, ont tout simplement sorti eux-mêmes un album intitulé *AIsis* (Contraction d'Oasis et de AI) : *The lost tapes*, et produit à partir de mélodies emblématiques de Liam Gallagher et de voix générées par ordinateur[46]. Les fans sont ravis mais les Gallagher peuvent-ils réclamer des droits d'auteur sur cette imitation, respectueuse du travail des deux frères ennemis ?

43 La redaction de LLB, « Des faux Brad Pitt extorquent près de 325 000 euros à des femmes en situation de vulnérabilité », *La Libre Belgique*, le 25 septembre 2024, [en ligne :] https://www.lalibre.be/international/europe/2024/09/25/des-faux-brad-pitt-extorquent-pres-de-325-000-euros-a-des-femmes-en-situation-de-vulnerabilite-MWX7HLZPWVFG7B64RBQ6IFSG4M, consulté le 24 juillet 2024.

44 Davet S., « Quand l'intelligence artificielle crée des "deepfakes" musicaux », *Le Monde*, le 11 octobre 2023, [en ligne :] https://www.lemonde.fr/culture/article/2023/10/11/quand-l-intelligence-artificielle-cree-des-deepfakes-musicaux_6193851_3246.html, consulté le 2 août 2024.

45 Robertson D., « It's the screams of the damned!' The eerie AI world of deepfake music », *The Guardian*, le 9 novembre 2020, [en ligne :] https://www.theguardian.com/music/2020/nov/09/deepfake-pop-music-artificial-intelligence-ai-frank-sinatra, consulté le 2 août 2024.

46 La rédaction du Figaro, « Des fans créent un album d'Oasis grâce à l'intelligence artificielle », *Le Figaro*, avril 2023, [en ligne :] https://www.lefigaro.fr/musique/des-fans-creent-un-album-d-oasis-grace-a-l-intelligence-artificielle-20230420, consulté le 2 août 2024.

Et puis, si les stars ont les moyens de lancer des poursuites judiciaires, les artistes moins connus ne risquent-ils pas de se faire plagier facilement ? Sera-t-il encore judicieux de mettre sa maquette ou son morceau en ligne pour se faire connaître sans risquer de se la faire voler ?

Aujourd'hui, avec l'IA Uberduck, vous choisissez la voix d'une star de la musique puis vous saisissez le texte qu'elle doit prononcer et le tour est joué. Il faut juste promettre que ça ne sera pas utilisé à des fins commerciales.

À terme, chacun pourrait se créer sa, voire ses, propre·s chanson·s. Ou l'IA pourrait carrément vous créer une chanson personnalisée en fonction de votre humeur, ce qui est désormais réaliste au vu des masses de données de plus en plus disponibles pour alimenter les algorithmes d'apprentissages ? Les possibilités sont tellement vertigineuses, qu'il est difficile d'anticiper les réactions et acceptations du public, auxquelles s'adaptera inévitablement l'industrie et la technologie IA. Il faudra légiférer en conséquence mais ça ne sera pas simple au vu des évolutions permanentes.

4. Le Métavers, allégorie de la grotte platonicienne 3.0. ?

En 2021 était lancé le Métavers de Mark Zuckerberg, qui nous avait imaginé un monde virtuel avec des avatars et une économie parallèle où on pouvait s'acheter une maison ou des NFT (œuvres virtuelles), se faire de nouveaux amis et épater la galerie à grand renfort de bitcoins. Mais ce concept élaboré en 1992 par Neal Stephenson, dans le roman de science-fiction *Le Samouraï virtuel*, un livre culte pour les entrepreneurs de la Silicon Valley, fut un flop retentissant pour Mark Zuckerberg. Cela dit, à ce jour, plusieurs centaines de métavers peuvent déjà être recensés et les plus grands (Roblox, Second Life, Zepeto, Minecraft, Fortnite) regroupent des millions d'utilisateurs.

Le concept est aussi vague que fourre-tout. Le rapport interministériel français de la mission sur le développement des métavers, publié en octobre 2022[47], définit ce dernier comme « *un service en ligne donnant accès à des simulations d'espaces 3D en temps réel, partagées et persistantes, dans lesquelles on peut vivre ensemble des expériences immersives* ». Malgré l'échec de Meta, l'idée persiste et de nombreuses entreprises comme Microsoft, Amazon ou Google investissent dans le concept. Un rapport récent du cabinet McKinsey évalue à 5 000 milliards de dollars le marché du métavers à l'horizon 2030, soit l'équivalent de la troisième économie mondiale derrière les États-Unis et la Chine. Les investissements sont évalués à plus de 120 milliards de dollars. Le projet métavers dépasse les ambitions d'une seule entreprise, aussi grande soit-elle … de grandes marques, comme *« Nike, Balenciaga ou Louis Vuitton se sont positionnées dans ces espaces virtuels »*[48]. De son côté Mark Zuckerberg n'envisage pas le Métavers rentable avant 2030.

Ces projets permettraient de créer une économie parallèle, en bitcoins, difficile à surveiller pour les États, et de se créer une vie parallèle « plus fun » dans laquelle les deepfakes et l'IA auraient une grande place à jouer. Une belle opportunité pour se faire connaître, que l'on soit une grande marque ou un artiste inconnu.

47 Basdevant A., Camille François C., Ronfard R., « Mission exploratoire sur les métavers », *Mission exploratoire pour le Gouvernement Français*, octobre 2022, [en ligne :] https://www.economie.gouv.fr/files/files/2022/Rapport-interministeriel-metavers.pdf, consulté le 2 août 2024.

48 Perez C. et Derhy A., « Métavers : l'heure du premier bilan », *The Conversation*, janvier 2023, [en ligne :] https://theconversation.com/metavers-lheure-du-premier-bilan-197149, consulté le 3 août 2024.

D'autres y voient carrément l'avenir du télétravail[49]. En tout cas, le concept ne séduit pas encore beaucoup de monde, il est d'ailleurs encore assez méconnu voire flou.

Mais n'est-il pas plus simple de jouer à un jeu vidéo immersif avec un casque de réalité virtuelle ? D'autant que ces mondes parallèles du Métavers risquent d'augmenter un peu plus la dépendance à ces vies virtuelles, au détriment d'une vie sociale et d'une économie locale. Sans compter que la plongée dans ces univers trop parfaits peut entraîner déception, voire dégoût de son physique et de sa propre personne, des dérives déjà observées sur Instagram et sur TikTok[50]. Sans parler de la perte de perceptions des expressions faciales des autres êtres humains, par exemple et de l'empathie. Dans son fameux livre, *Alone together* (*Seuls ensemble*, L'Echappée, 2015), la chercheuse Sherry Turkle, psychologue et professeur au Massachusetts Institute of Technology, constate que la connexion incessante avec les robots et les ordinateurs dévore les relations humaines en face-à-face et affaiblit l'empathie[51].

Si le métavers devenait une réalité, il pourrait accélérer la déliquescence du tissu social, affaiblir l'empathie humaine et nous conduire à rêver notre vie plutôt que de la vivre vraiment et rompre les liens sociaux qui sont une des bases fondamentales de notre humanité.

5. Quand l'IA nous relia, jusqu'à ce que sa mort nous sépare

L'IA permet désormais, grâce aux milliers de données laissées sur le net par un proche, de faire « revivre » celui-ci après son décès. Le thème de l'immortalité est un vieux fantasme et le deuil d'un être cher est une des choses les plus difficiles à accepter, il suffit de voir le succès de certains voyants communiquant avec les esprits. Le sujet est pernicieux car, converser avec un aïeul pour comprendre ses racines peut être vraiment intéressant. C'est ce que propose HereAfter AI, une société basée aux États-Unis qui permet aux gens de *« télécharger leurs souvenirs, qui sont ensuite transformés en un "avatar d'histoire de vie" avec lequel les amis et la famille peuvent communiquer »*[52]. Cette méthode permet aux gens de laisser leurs souvenirs pour les générations futures. Mais que les datas laissées sur le net soient représentatives de qui était la personne semble excessif. N'oublions pas que, comme le précise Amélie Cordier, ingénieure et maîtresse de conférences en intelligence artificielle à l'Université Lyon I : « *On prête à l'intelligence artificielle des idées, des pouvoirs, des applications qui ne sont en fait ni plus ni moins que des désirs, des vœux d'êtres humains. Le mot important dans tout ça, c'est "simulation". Avec ces intelligences artificielles génératives, qu'on n'avait pas jusqu'à maintenant, effectivement, il est possible de simuler le fait qu'une personne*

49 Garnier A., « L'avenir du télétravail est dans le métavers », *Le Point*, le 30 septembre 2024, [en ligne :] https://www.lepoint.fr/debats/l-avenir-du-teletravail-est-dans-le-metavers-30-09-2022-2491914_2.php, consulté le 4 août, 2024.

50 Blackburn M.R. et Hogg R.C., « Pour vous, l'impact du contenu pro-ana TikTok sur l'insatisfaction de l'image corporelle et l'internalisation des normes de beauté sociétales », *Plos One*, le 7 août 2024, [en ligne :] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0307597, consulté le 12 août 2024.

51 Wiart L., « L'hyperconnexion au détriment de la conversation », *La revue des médias*, le 4 mars 2019, [en ligne :] https://larevuedesmedias.ina.fr/lhyperconnexion-au-detriment-de-la-conversation, consulté le 13 août 2024.

52 Lamort E., « "Parler" à un proche disparu, cette nouvelle illusion inquiétante de l'intelligence artificielle », *Ouest France*, le 2 mai 2023, [en ligne :] https://www.ouest-france.fr/leditiondusoir/2023-05-02/parler-a-un-proche-disparu-cette-nouvelle-illusion-inquietante-de-l-intelligence-artificielle-6abba44f-1fdb-4a22-a605-df54a9a712f5, consulté le 12 août 2024.

que l'on connaisse puisse parler. Et l'IA permet de faire ça avec une qualité de simulation qui est vraiment très intéressante. Mais ce n'est ni plus ni moins la même chose que de se dire : "Je vais parler avec un personnage de jeu vidéo" »[53].

Pour le « *Dr Kirsten Smith, chargée de recherche clinique à l'Université d'Oxford, le risque est que ces avatars de défunts nous maintiennent accrochés au passé, nous rendent incapables d'aller de l'avant et de mener normalement notre vie* »[54]. Le danger de conséquences psychologiques est bien réel.

Autre manière de prendre ses simulations pour des réalités, en 2022, une femme de quatre-vingt-sept ans a carrément assisté à ses propres funérailles au Royaume-Uni, grâce à une startup appelée StoryFile, qui enregistre des séquences vidéo et audio avant le décès d'une personne. Elle les rend ensuite interactives grâce à la puissance de l'intelligence artificielle et d'un avatar holographique[55].

Au royaume de l'IA et du deepfakes, la fantaisie humaine semble être un bel exemple de l'infini.

B. Enjeux de société

1. Le faux pour secouer le vrai

Diverses associations ont aussi compris tout l'intérêt des hypertrucages pour des campagnes de sensibilisation dites positives.

Exemple chez nous, en mars 2020, avec un discours vidéo fictif de Sophie Wilmès publié sur les réseaux sociaux pour relancer les actions du groupement d'activistes écologistes Extinction Rebellion. La Première ministre y déclarait : « *La pandémie actuelle de Covid-19 plonge ses racines dans la destruction écologique mondiale et ce sont les plus vulnérables dans nos sociétés qui sont le plus durement frappé·e·s* ». Une vidéo qui a été largement diffusée sur les médias sociaux et dans les médias main stream, sans doute en raison de sa nouveauté et de son originalité. En effet, jamais on n'avait détourné la vidéo d'un ou d'une Premier·e ministre belge via un deepfake, d'autant que le message allait plutôt à l'encontre du discours plutôt libéral de celle-ci. L'objectif a donc été atteint pour l'association écologiste.

Ces campagnes de sensibilisation dites positives sont aussi accusées de jouer avec le feu. En octobre 2019, l'association Solidarité Sida a mis en ligne une vidéo du président des États-Unis Donald Trump. « *Le sida, c'est fini* », semble déclarer ce dernier. Mais, il s'agissait en fait d'un acteur dont le visage avait été numériquement modifié pour laisser entrevoir la possibilité d'un « monde sans sida », comme l'explique Antoine de Caunes, le président d'honneur de l'association, à la fin de la vidéo. Beaucoup d'internautes s'y sont cependant laissé prendre et ont critiqué une campagne qui joue avec les limites du réel à des fins de communication. Le clip a été vu 3,5 millions de fois en quelques heures et a sans doute

53 *Ibid.*
54 *Ibid.*
55 Bryce A.L., « Communiquer avec les morts grâce à l'Intelligence artificielle », *EuroNews*, 16 mars 2023, [en ligne :] https://fr.euronews.com/next/2023/03/16/communiquer-avec-les-morts-grace-a-lintelligence-artificielle, consulté le 12 août 2024.

touché son public cible : les jeunes. Le directeur fondateur de l'association Solidarité Sida, Luc Barruet, prête à cette vidéo truquée « *des vertus pédagogiques, pour que les gens mesurent ce que l'on peut faire avec les deepfakes* »[56]. À ses yeux, les critiques émises sur le Net, quant au caractère « *douteux* » ou « *dangereux* » de cette fausse vidéo dont il faut attendre la fin pour en comprendre le sens, sont peu nombreuses en comparaison de son succès viral, précisant que « *ceux qui réagissent sur les réseaux sociaux sont toujours ceux qui ne sont pas contents, pas ceux qui trouvent ça formidable* »[57].

Le deepfake peut ainsi être très utile pour éveiller les consciences sur différents sujets mais ces vidéos fabriquées peuvent se retourner contre l'intention de départ ou être sorties de leur contexte initial, en tout ou en partie. Ainsi en 2018, en Inde, une vidéo a longtemps circulé dans tout le pays et effrayé les Indiens. « *Elle montre deux hommes à moto qui s'approchent d'un groupe d'enfants qui jouent au cricket. La moto ralentit, le passager attrape l'un des enfants et ils prennent tous deux la fuite. Un procédé rapide et terrifiant. Sauf que cette vidéo ne décrit en aucun cas un fait réel. C'est même l'opposé : ce clip fait partie d'une campagne de prévention contre les enlèvements d'enfants, et pas en Inde, mais au Pakistan. (…) La vidéo a circulé sur WhatsApp et dans l'État du Goujarat, à l'ouest du pays, accompagné du message d'alerte suivant, prétendument issu par la police : "un gang de 300 enleveurs d'enfants est arrivé dans la région, protégez vos enfants !"*[58] ». Des dizaines de personnes suspectées d'être kidnappeurs ont été tabassées par la population et de janvier à juillet 2018, une trentaine de personnes sont mortes, lynchées par des foules. La police a été dépassée par le phénomène, d'autant que l'info a été diffusée sur WhatsApp, qui est une messagerie cryptée et donc quasiment impossible à intercepter. Un simple détournement d'une vidéo de sensibilisation sur fond de tension sociale aura ainsi provoqué plusieurs dizaines de morts.

Une multiplication excessive de vidéos fictives pourrait également desservir la cause s'il n'est pas précisé systématiquement à l'image que la vidéo est un faux. Mais cela peut être camouflé pour être détourné, à l'instar de cet exemple indien.

2. Satires à tout va

Il est certain que le deepfake est une magnifique opportunité pour tous de faire de la satire. Elle sera de bon ou de mauvais goût et elle sera différemment acceptée selon son point de vue. À l'instar du deepfake qui a fait danser la princesse Leonor, héritière d'Espagne, sur TikTok. Beaucoup y ont cru et ont ensuite trouvé la blague de mauvais goût car l'utilisation de l'image de la princesse avait été réalisée de manière trompeuse. Il aurait mieux valu souligner qu'il s'agissait d'un faux.

56 Kahn S., « "Le sida, c'est fini": Solidarité Sida créé un deepfake de Donald Trump pour une campagne », *Le Figaro Tech&Web*, le 7 octobre 2019, [en ligne :] https://www.lefigaro.fr/secteur/high-tech/le-sida-c-est-fini-solidarite-sida-cree-un-deepfake-de-donald-trump-pour-une-campagne-20191007, consulté le 26 juillet 2024.

57 Tual M., « Fausse vidéo de Trump : pourquoi Solidarité sida a sorti un "deepfake" pour sa campagne », *Le Monde*, le 9 octobre 2019, https://www.lemonde.fr/pixels/article/2019/10/09/fausse-video-de-trump-pourquoi-solidarite-sida-a-sorti-un-deepfake-pour-sa-campagne_6014864_4408996.html, consulté le 26 juillet 2024.

58 Farcis S., « Les Indiens, terrorisés par les enlèvements d'enfants, alimentent les rumeurs », 16 juin 2018, [en ligne :] https://www.rfi.fr/fr/asie-pacifique/20180615-inde-reseaux-sociaux-rapt-enfants, consulté le 26 juillet 2024.

Et encore, lorsque la chaîne de télévision britannique Channel 4 suscite la controverse à Noël avec une vidéo « deepfake », diffusée vendredi 25 décembre, qui tourne en dérision le traditionnel discours de la reine Elizabeth II et le font terminer par un jerk endiablé de la Reine[59], les sujets de Sa Majesté ont crié à l'outrage, malgré les bonnes intentions clamées par la chaîne.

À l'heure où n'importe qui peut se prendre pour un caricaturiste de Charlie Hebdo, nul doute que la justice de nombreux pays va avoir fort à faire dans les prochaines années pour trancher et fixer les limites de l'amusement et de l'humiliation. Enfin, si elle en reçoit les moyens.

3. Je jure de dire la vérité

Au niveau judiciaire, la notion de preuve risque d'être quelque peu ébranlée. Ainsi, le premier cas de deepfake invoqué dans un procès en France, l'a été par le célèbre humoriste polémique Dieudonné. Dans une vidéo, publiée le 8 avril 2020 sur YouTube puis retirée, on le voyait critiquer vigoureusement les réquisitions d'une magistrate de Nanterre, la comparant notamment aux femmes ayant collaboré avec le régime nazi. Pour sa défense, il avait déclaré que cette vidéo était un deepfake. « *Aujourd'hui, avec les deepfake, l'image n'est plus une preuve* », avait plaidé son avocat David de Stefano, demandant sa relaxe. La cour a finalement quand même condamné Dieudonné à 30 000 euros d'amende ne jugeant pas cet argument crédible.

À l'inverse, ce faits-divers, aux États-Unis, qui souligne combien le deepfake s'immisce dans toutes les sphères de nos sociétés. Raffaela Spone est accusée d'avoir harcelé trois adolescentes, à l'aide d'images manipulées, pour leur faire abandonner leur rôle dans l'équipe locale de *cheerleading* (l'équipe de supportrice) du comté de Bucks en Pennsylvanie. La mère de famille encourait jusqu'à un an de prison et les médias ont largement contribué à populariser l'affaire. Après presqu'un an de procédure, Matthew Weintraub, le procureur du district du comté de Bucks a annoncé abandonner les poursuites contre Raffaela Spone, la police ayant été incapable de fournir les preuves d'une quelconque manipulation de la vidéo.

Faudra-t-il bientôt prouver, de manière irréfutable, qu'un faux est faux et qu'un vrai est un vrai ? La justice n'a pas fini de devoir adapter les lois à ces nouvelles technologies, en mutation perpétuelles.

4. À chacun son Histoire

Un outil deepfake, permettant aux utilisateurs d'animer d'anciennes photos de proches, a été largement utilisé sur le site MyHeritage, site de généalogie permettant de retrouver ses aïeux. Au vu du succès, le site a désormais ajouté LiveStory, qui permet d'y insérer des voix.

En télévision, Thierry Ardison propose des interviews d'artistes décédés comme Dalida ou Jean Gabin grâce aux hypertrucages et au talent de comédiens.

59 Honnet T., « La danse d'(une fausse) Elizabeth II, ou la vidéo de lèse-majesté qui scandalise les Britanniques », *Madame Figaro*, décembre 2020, [en ligne :] https://madame.lefigaro.fr/celebrites/la-danse-tiktok-elizabeth-ii-qui-indigne-le-royaume-uni-video-deepfake-291220-194274, consulté le 27 juillet 2024.

Ces exemples soulignent qu'on pourrait tenter de faire croire qu'une personnalité n'est pas morte à l'aide d'une fausse vidéo, du moins pendant quelques temps. Pourrait-on aussi faire dire n'importe quoi au passé ? Sans aucun doute. La preuve ? En juillet 1969, en cas d'échec tragique de la mission lunaire Apollo 11, le président américain Richard Nixon avait prévu un discours. Des chercheurs du Massachusetts Institute of Technology (MIT)[60] en ont fait une vidéo[61]. Objectif : démontrer les dangers des hypertrucages, qui permettent de faire dire à une personne des mots qu'elle n'a jamais prononcés. Cette expérience montre ainsi qu'on peut imaginer un faux discours de Nixon prétendant qu'Armstrong et Aldrin n'ont jamais mis le pied sur la lune et il est fort à parier que la vidéo serait largement partagée au regard du nombre de sceptiques face à ce moment de l'histoire. Une enquête IFOP[62] a ainsi démontré qu'en 2019, un Français sur dix croyait que les Américains n'avaient jamais marché sur la lune.

Nos sociétés ne se construisent-elles pas sur des mythes fondateurs comme le fait que Charlemagne aurait inventé l'école ou que son ami Rolland aurait été tué par des Sarrasins, alors que c'était par des Basques. Et que dire de « Nos ancêtres les Gaulois », idée répandue aux XIXe et XXe siècles par les nationalistes français, à une époque où les peuples d'Europe cherchaient à se donner une ascendance antique, pour justifier l'existence de leur État-Nation. Mais ces mythes étaient fondateurs d'un pays, d'une population acceptant de vivre ensemble et de croire en un récit commun, et non d'une vague croyance partagée par des personnes éparpillées aux quatre coins du monde et sans aucune cohérence existentielle.

En fait, les deepfakes risquent bien d'alimenter les prétendues preuves des complotistes et des négationnistes de tout bord, qu'elles soient sonores, scripturales ou visuelles.

5. Pour le journalisme, un risque d'y perdre des plumes

Au-delà de la difficulté pour les journalistes de dénouer le vrai du faux avec l'arrivée des deepfakes et de l'IA, comme nous le verrons plus tard, les hypertrucages peuvent devenir une tentation professionnelle pour ceux-ci. L'IA est effectivement de plus en plus utilisée par les journalistes, notamment pour la retranscription ou la traduction d'interviews ou encore pour une aide à la rédaction, à la correction des articles, au résumé ou à l'analyse de dossiers complexes. Mais à vouloir gagner du temps, ils doivent faire attention à la limite, parfois fragile, entre «*les super assistants*» et la manipulation.

En Belgique des journalistes débattent sur le sujet. Selon Yves Thiran, journaliste et membre d'un groupe de réflexion sur ces nouvelles technologies au sein de la RTBF, interrogé par Martin Bilterijs : « *C'est tout ce qui est la production de contenus. Que ce soit du contenu sous forme de texte, d'image ou de vidéo. A la RTBF, on s'abstient pour l'instant. Mais on voit bien que d'autres médias ont déjà commencé, notamment quand il s'agit d'illustrer un article sur un site web et qu'il n'y a pas l'image de l'événement. Traditionnellement, on utilise des images pré-*

60 MIT Open Learning, « Tackling the misinformation epidemic with "In Event of Moon Disaster" », *MIT News*, juillet 2020, [en ligne :] https://news.mit.edu/2020/mit-tackles-misinformation-in-event-of-moon-disaster-0720 , consulté le 16 août 2024.

61 MIT Open Learning, « In Event Of Moon Disaster Movie Trailer », *Youtube*, juillet 2020, [en ligne :] https://www.youtube.com/watch?v=kc_ufCSQLwI&t=56s, consulté le 16 août 2024.

62 Sondage IFOP, « Enquête sur le complotisme », *IFOP*, le 8 janvier 2018, [en ligne :] https://www.ifop.com/publication/enquete-sur-le-complotisme/, consulté le 12 août 2024.

textes, mais là on demande à l'ordinateur de fabriquer l'image. C'est une des questions qu'on se pose : doit-on ou pas s'autoriser ce genre de productions ? … Un site américain a d'ailleurs dénombré "49 médias désormais entièrement automatisés". Des sites qui génèrent des articles entièrement créés par l'intelligence artificielle. Ce sont "des médias de niche pas encore de grands médias", précise Yves Thiran [63].

Par ailleurs, le choix d'une image réelle, pour illustrer un article, reste un choix rédactionnel, en fonction d'une volonté de sensationnalisme par exemple. Mais créer une image en fonction de ce qu'on connaît d'un événement peut devenir beaucoup trop subjectif.

Il est également difficile de reconnaître une photo hyper truquée, même pour des professionnels. On se rappelle de cette photo qui avait gagné le prix du Sony World Photography Awards en 2023, avant que l'artiste ne révèle qu'il s'agissait d'un deepfake [64]. Les journalistes doivent garder la capacité de ne pas être trompés par des images qui leurs seraient envoyées par de faux témoins d'un événement.

Mais n'oublions pas ce que nous entendons en atelier d'éducation permanente ou en formation. Les participants nous révèlent préférer les articles courts et faciles à comprendre, voire même ne sélectionner que les bonnes nouvelles. L'info tend à devenir irréversiblement un produit de consommation comme un autre pour de plus en plus de citoyens. Pourtant l'actualité est souvent complexe et nuancée et ne peut être une simple distraction. Les journalistes ont sans doute une carte à jouer en restant gages de qualité et de vérification des faits. Gageons qu'on leur en laisse les moyens car il est difficile d'imaginer que l'IA puisse complètement remplacer un journaliste, avec son talent de plume, sa créativité, son recul, son expérience, son humour, sa perception de la vérité, sa probité, son éthique, et tout ce qui fait l'âme d'un être pensant neutre et indépendant.

6. Humiliations des femmes, quand la honte changera-t-elle donc de camp ?

En matière de harcèlement, les deepfakes sont malheureusement plus efficaces qu'ailleurs, surtout envers les femmes. « *D'après un rapport de l'entreprise hollandaise de cybersécurité Deeptrace Lab, sur les 14.000 vidéos hypertruquées mises en ligne en 2019, 96% d'entre elles étaient à caractère pornographique… De nombreuses célébrités comme Billie Eilish, Emma Watson ou encore* [65] *d'autres streameuses de la plateforme Twitch en ont été les victimes. Car l'autre constat alarmant de Deeptrace Lab est que le deepfake pornographique cible exclusivement des femmes. Les deepfakes non pornographiques analysés sur YouTube contenaient, quant à eux, une majorité de sujets masculins* » [66]. Et pour les femmes, cela est souvent dévastateur.

63 Biourge C., « L'intelligence artificielle : une menace de plus en plus importante pour l'information et un défi pour les journalistes », *RTBF Info*, le 3 mai 2023, [en ligne :] https://www.rtbf.be/article/l-intelligence-artificielle-une-menace-de-plus-en-plus-importante-pour-l-information-et-un-defi-pour-les-journalistes-11192380, consulté le 21 août 2024.

64 Wagon L., « Une œuvre générée par IA lauréate d'un prix de photographie », *Le Journal des Arts*, le 20 avril 2023, [en ligne :] https://www.lejournaldesarts.fr/creation/une-oeuvre-generee-par-ia-laureate-dun-prix-de-photographie-166032, consulté le 21 août 2024.

65 Une image deep porn de Taylor Swift a fait … 47 millions de vues. La chanteuse est connue pour ses positions anti Donald Trump et il y a fort à parier qu'il s'agit là d'une attaque des partisans de l'ancien président.

66 Jacquet T., « Deepfakes pornographiques, politiques, économiques : quelles sont les sanctions prévues par le droit belge contre ces pratiques ? », *RTBF*, le 14 février 2023, [en ligne :] https://www.rtbf.be/article/deepfakes-pornographiques-politiques-economiques-quelles-sont-les-sanctions-prevues-par-le-droit-belge-contre-ces-pratiques-11151913, consulté le 19 juillet 2024.

En Belgique, nous sommes loin d'être à l'abri car les exemples se multiplient. Fin 2023, Julia, vingt-et-un ans, étudiante belge en marketing et mannequin semi-professionnelle, reçoit des photos d'elle nue, pour lesquelles elle n'a jamais posé. Les photos sont tellement bien faites qu'elle est abasourdie : « *J'avais déjà entendu parler des deepfakes et des deepnudes (...) Mais je n'en avais pas conscience plus que ça avant que ça ne m'arrive à moi. C'était un événement un peu anecdotique qui se passait dans la vie des autres, mais qui ne se passerait pas dans la mienne* »[67]. La jeune femme dépose plainte au commissariat. Il est vrai que des lois peuvent être invoquées, comme celles sur le voyeurisme et la diffusion non consentie, ou comme la directive européenne sur les violences faites aux femmes[68]. Mais on prévient Julia « *que le "parquet est débordé" et qu'il y a "très peu de chances" que sa plainte aboutisse* »[69]. Autre exemple, à la même époque, c'est Céline Van Ouytsel, miss Belgique 2020 qui en était victime : « *Je savais que cela se faisait parfois avec des stars de Hollywood très célèbres, mais je n'aurais jamais pensé que cela puisse m'arriver un jour. Comme quoi, cela peut arriver à n'importe qui...* »[70].

Aucune femme publique n'est à l'abri. C'est par exemple arrivé à la célèbre influenceuse française Léna Situation. Elle explique : « *Ils ont mis un screen du vlog*[71] *et le reste du corps ne m'appartient pas. Et il y a tellement de meufs sur internet qui vivent ça. C'est vraiment dégueulasse. Ça fait longtemps que je n'ai pas eu un moment que j'ai kiffé en tapant mon nom sur les réseaux sociaux. De bon matin, voir ma tête sur un corps nu ?!* »[72]. Février 2023, dans un direct diffusé le 2 février dernier sur la plateforme Twitch, QTCinderella, une streameuse (une joueuse qui transmet et commente ses parties de jeux vidéo en direct) américaine de vingt-huit ans, revient en larmes sur le deepfake à caractère pornographique dont elle a été victime : « *C'est à ça que ressemble la douleur. Voilà ce que cela fait de se sentir violée, abusée et de se voir nue contre sa volonté partout sur internet* »[73]. Une victime australienne, va également déclarer à Euronews « *C'est une condamnation à vie ... Elle peut détruire la vie des gens, leurs moyens de subsistance, leur employabilité, leurs relations interpersonnelles, leurs relations amoureuses. Et il y a très, très peu de choses que l'on puisse faire une fois que quelqu'un est pris pour cible* ».

Il y a, d'ailleurs, fort à parier que des cas de chantages risquent de se multiplier pour un certain nombre de quidams. Combien préféreront payer cent, deux cents voire trois cents euros plutôt que de voir circuler sur le net des *deep nudes* ou *deep porns* ?

67 Hess A., « "La honte doit changer de camp" : une mannequin belge alerte sur les deepnudes », *Euronews*, 14 mars 2024, https://fr.euronews.com/next/2024/03/14/la-honte-doit-changer-de-camp-une-mannequin-belge-alerte-sur-les-deepnudes, consulté le 10 octobre 2024.

68 Journal officiel de l'Union européenne, « DIRECTIVE (UE) 2024/1385 DU PARLEMENT EUROPÉEN ET DU CONSEIL, sur la lutte contre la violence à l'égard des femmes et la violence domestique », *le Journal officiel de l'Union Européenne*, le 14 mai 2024, [en ligne :] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=OJ:L_202401385, consulté le 20 juillet 2024.

69 Hess A., *op. cit.*

70 La rédaction de RTL Info, « Une ex-Miss Belgique victime de "deepnude" témoigne : "*J'ai été profondément choquée*" », *RTL Info*, le 6 octobre 2023, [en ligne :] https://www.rtl.be/people/news/une-ex-miss-belgique-victime-de-deepnude-temoigne-jai-ete-profondement-choquee/2023-10-06/article/595323, consulté le 13 août 2024.

71 Blog Vidéo

72 La rédaction de Midi Libre, « "C'est vraiment dégueulasse" : Léna Situations victime de deepfake porno, l'influenceuse témoigne de son écœurement », *Midi Libre*, le 11 août 2023, [en ligne :] https://www.midilibre.fr/2023/08/11/cest-vraiment-degueulasse-lena-situations-victime-de-deepfake-porno-linfluence-temoigne-de-son-ecoeurement-11391227.php, consulté le 20 juillet 2024.

73 Bordecq C., « Victime de deepfake porno, la streameuse QTCinderella remet en lumière le phénomène des cyberviolences », *RTBF*, le 15 février 2023, [en ligne :] https://www.rtbf.be/article/victime-de-deepfake-porno-la-streameuse-qtcinderella-remet-en-lumiere-le-phenomene-des-cyberviolences-11153166, consulté le 20 juillet 2024.

Ironie supplémentaire, les deep nudes d'hommes sont non seulement rares, mais ils sont également de moindre qualité. « *Les algorithmes sont entraînés spécifiquement sur le corps des femmes … Cela révèle une inégalité plus profonde entre les sexes, un manque de respect pour les femmes et la violence sexuelle dans la société. La technologie ne fait qu'encourager l'objectivation des femmes. De plus, il ne s'agit pas toujours de satisfaire un fantasme. Une fausse vidéo porno peut aussi être utilisée pour réduire les femmes au silence* »[74]. Déjà, en 2018, l'histoire vécue en Inde par Rana Ayyub, a fait froid dans le dos des femmes engagées en politique indienne. Cette journaliste d'investigation a été la victime d'une vidéo « deepfake » à caractère pornographique qui a circulé sur les réseaux sociaux, notamment grâce à l'aide d'hommes politiques d'après une de ses sources. Très vite, elle reçoit des menaces de viol. Un tweet est diffusé sur les réseaux sociaux avec une capture d'écran de la vidéo et son numéro à côté, disant « *Salut, c'est mon numéro et je suis disponible ici* ». « *Les gens ont commencé à m'envoyer des messages WhatsApp me demandant mes tarifs pour le sexe* ». « *J'ai été envoyée à l'hôpital avec des palpitations cardiaques et de l'anxiété, le médecin m'a donné des médicaments. Mais je vomissais, ma tension artérielle a monté en flèche, mon corps avait réagi si violemment au stress* », a-t-elle déclaré[75]. « *Le pays tout entier regardait en boucle une vidéo porno que l'on m'attribuait et j'étais tétanisée* »[76].

« *"Si vous étiez le pire misogyne du monde, cette technologie vous permettrait de réaliser tout ce que vous voulez", a déclaré Mary Anne Franks, professeur de droit à l'Université de Miami et présidente de l'association Cyber Civil Rights Initiative dont la mission est de venir en aide aux personnes victimes d'abus sur Internet partout dans le monde* »[77].

L'histoire de Rana Ayyub fait ainsi craindre le pire pour les femmes qui mènent des combats politiques et/ou idéologiques à travers le monde, notamment face à une misogynie persistante. Cela dit, les hommes risquent de ne pas être épargnés non plus, quand on sait qu'une vie ou qu'une carrière politique peut dépendre d'une vidéo, à l'instar de celle, à caractère sexuel et partagée sur les réseaux sociaux, qui a conduit le candidat français Benjamin Griveaux à renoncer à la mairie de Paris en février 2019 avant même que l'authenticité des images n'ait pu être prouvée. Un deepfake pourrait avoir cet effet.

Heureusement, des femmes sont prêtes à se battre pour faire évoluer les lois. C'est le cas de la présentatrice tv hollandaise, Welmoed Sijtsma, qui s'est retrouvée dans un deep porn. En novembre 2023, le tribunal d'Amsterdam a condamné son auteur, un homme de trente-neuf ans, à cent quatre-vingts heures de travaux d'intérêt général avec sursis[78]. Son père avait refusé de regarder la vidéo, de peur de garder ces images en tête. Car n'oublions pas les dégâts collaté-

74 Hulstaert E., « Fake porno: comment les "deepnudes" détruisent la vie de Belges », *Le Vif*, le 13 août 2024, [en ligne :] https://www.levif.be/societe/fake-porno-comment-les-deepnudes-detruisent-la-vie-de-belges/, consulté le 16 août 2024.

75 Bureau Web d'INDIA TODAY, « I was vomiting: Journalist Rana Ayyub reveals horrifying account of deepfake porn plot », *India Today*, 21 novembre 2018, [en ligne :] https://www.indiatoday.in/trending-news/story/journalist-rana-ayyub-deepfake-porn-1393423-2018-11-21, consulté le 1 septembre 2020.

76 Ayyub R., « J'ai été victime d'une vidéo porno truquée et d'un complot destiné à me faire taire », *huffingtonpost*, 2 décembre 2018, [en ligne :] https://www.huffingtonpost.fr/rana-ayyub/j-ai-ete-victime-d-une-video-porno-truquee-et-dun-complot-destine-a-me-faire-taire_a_23603520, consulté le 2 avril 2019.

77 Bonte A., « "Deepfake" : Scarlett Johansson met en garde contre cet inquiétant phénomène », *RTL France*, 04 janvier 2019, [en ligne :] https://www.rtl.fr/girls/societe/deepfake-scarlett-johansson-met-en-garde-contre-cet-inquietant-phenomene-7796128802, consulté le 3 avril 2020.

78 Hulstaert E., « Fake porno: comment les "deepnudes" détruisent la vie de Belges », *op. cit.*

raux de ces images sur l'entourage des victimes, lui aussi souvent désemparé et frustré par un sentiment d'impuissance face à un phénomène nouveau, violent et excessivement intrusif.

7. « Déshabillez n'importe qui, déshabillez les filles gratuitement »

Un tiers des élèves de la FWB serait concerné par le harcèlement, et un « programme-cadre » de prévention du harcèlement et d'amélioration du climat scolaire a été lancé [79]. Il est effectivement difficile de protéger les jeunes des messages de haine, des harcèlement ou encore des images pornographiques, voire pédopornographiques, tout comme des arnaques et des réseaux de propagandes mensongères. La mode depuis peu : les deep nudes dans les écoles. En Belgique, « *une dizaine de jeunes filles de 12 à 16 ans, élèves du collège Saint-Remacle de Stavelot, ont été victimes de deepfakes. Des garçons de leur école et d'autres établissements scolaires ont récupéré les photos qu'elles postaient sur les réseaux sociaux. En utilisant l'intelligence artificielle, ils ont modifié ces photos pour faire apparaître ces adolescentes entièrement nues. Les images ont ensuite été partagées sur Snapchat* » [80]. Une « nouvelle mode » qui tourne parfois au racket, au chantage et/ou au rançonnage. D'autant que ces images sont d'une facilité désormais déconcertante à réaliser. « Déshabillez n'importe qui, déshabillez les filles gratuitement », c'est le slogan de l'application ClothOff, qui a fait parler d'elle en Espagne récemment. Elle permet aux utilisateurs « *de "déshabiller" n'importe quelle personne apparaissant dans la galerie de photos de leur téléphone. Il en coûte 10 euros pour créer 25 images de nus* » [81]. Une application qui a donné des idées de chantages à quelques ados d'une école d'Almendralejo, une ville du sud de l'Espagne, où une vingtaine de jeunes filles ont vu passer de fausses photos d'elles nues. En clair, si une adolescente ne payait pas une petite rançon, un faux nu d'elle était publié sur les réseaux. Un faux qui a des chances de devenir viral dans leur école, à minima. Ce cas est une *sextorsion* mais il peut tout à fait s'agir de revanche ou de volonté d'humiliation d'une jeune personne. Et quand il s'agit de jeunes, voire d'enfants, les lois y sont encore moins préparées. Dans une interview accordée à Euronews, relatant ces faits , Manuel Cancio, professeur de droit pénal à l'Université autonome de Madrid, « *souligne qu'il existe un vide juridique car l'utilisation du visage de mineurs sur des photographies porte atteinte à leur vie privée, mais lorsqu'il s'agit de crimes dans lesquels des images intimes sont diffusées, c'est l'image dans son ensemble qui porte atteinte à la vie privée* » [82] mais « *Comme elle est générée par deepfake, la vie privée de la personne en question n'est pas affectée. L'effet qu'elle produit (sur la victime) peut être très similaire à celui d'une vraie photo de nu, mais la loi est en retard* », ajoute-t-il. La loi, espagnole en l'occurrence, est en retard, une phrase qu'on a pas fini d'entendre. Ce cas de chantage entre ados, nous apprend également que la plus jeune des victimes

79 Huttin C., « Harcèlement scolaire : 280 écoles sont invitées à lutter contre ce "fléau" », *Le Soir*, le 10 février 2024, [en ligne :] https://www.lesoir.be/567489/article/2024-02-10/harcelement-scolaire-280-ecoles-sont-invitees-lutter-contre-ce-fleau, consulté le 28 août 2024.

80 Hildesheim M., « Une dizaine de jeunes filles élèves du collège Saint-Remacle de Stavelot victimes de deepfakes », *RTBF Info*, le 21 juin 2024, [en ligne :] https://www.rtbf.be/article/une-dizaine-de-jeunes-filles-eleves-du-college-saint-remacle-de-stavelot-victimes-de-deepfakes-11392755, consulté le 3 septembre 2024.

81 Llach L., « Des adolescents espagnols ont reçu des nus d'eux-mêmes générés par l'IA: s'agit-il d'un crime ? », *Euronews*, le 20 septembre 2023, [en ligne :] https://fr.euronews.com/next/2023/09/20/des-adolescents-espagnols-ont-recu-des-nus-deux-memes-generes-par-lia-sagit-il-dun-crime, consulté le 20 juillet 2024.

82 *Ibid.*

n'avait que onze ans et combien la manière de qualifier les faits ne fait pas l'unanimité parmi les avocats : « *il peut s'agir de pédopornographie, de crimes contre l'intégrité morale ou de distribution d'images à contenu sexuel non consensuel* » [83].

L'année dernière, Child Focus a ouvert pour la première fois un certain nombre de dossiers sur les deep nudes. « *L'augmentation du phénomène nous inquiète, déclare Niels Van Paemel, conseiller politique. Grâce à notre service d'assistance téléphonique, nous recevons de plus en plus de questions de la part de jeunes sur des images truquées, qui se sont effectivement retrouvées dans les écoles. C'est le point de basculement que nous redoutions. Les deep nudes sont au carrefour du genre, de la cyberintimidation, de l'exposition (NDLR: diffusion numérique d'images privées sans autorisation dans le but de nuire ou d'humilier la personne représentée), de l'exploitation sexuelle et de la maltraitance des enfants. Tout y est réuni* » [84].

Nous le verrons plus loin des solutions existent mais sont complexes à mettre en place.

C. Nouvelles armes de persuasion massives

1. Crédibilisation de la caricature et discréditation du vrai

Les attaques et discréditations politiques se multiplient déjà dans le monde. On a vu passer une vidéo de Mauricio Macri, président argentin de 2015 à 2019, avec le visage d'Adolf Hitler [85] ou un faux d'Inés Arrimadas, membre du Parlement de la Catalogne, superposé sur une vidéo pornographique. Est-ce une déclinaison moderne de la caricature ? Non, car on joue ici sur un hyper réalisme avec lequel un dessin n'essaie pas de rivaliser. En revanche pourrait-on présenter la vidéo de M. Macri comme de la satire ? En mai 1968 ne voyait-on pas une affiche présenter A. Hitler derrière le masque de Charles De Gaulle ? Il s'agirait ici d'une opinion politique tandis que la vidéo d'Inés Arrimadas ne peut être considérée comme telle.

Où mettre la limite ?

En mai 2019 un « deepfake » peu sophistiqué, plutôt appelé cheapfake (un faux « bon marché »), avait fait parler de lui aux États-Unis. On y voyait la présidente de la Chambre des représentants américaine, et farouche opposante à Donald Trump, Nancy Pelosi parler avec un phrasé hésitant qui donnait l'impression qu'elle était passablement éméchée, droguée, malade, voire hébétée. Une vidéo qui a eu le temps d'être partagée trois millions de fois avant que le *Washington Post* n'ait le temps de comparer la vidéo avec l'original et de constater que les images avaient simplement été ralenties de 25 %. Même Rudy Giuliani, le propre avocat du Président Trump, avait eu le temps de la partager [86]. Cela permet de souligner l'importance du « biais de confirmation » dans le phénomène des partages de vidéos truquées, même de façon grossière : lorsqu'une vidéo

83 Llach L., *op. cit.*
84 Hulstaert E., « Fake porno: comment les "deepnudes" détruisent la vie de Belges », *op. cit.*
85 En fait avec la tête de Bruno Ganz jouant Adolf Hitler dans le film de 2004 « Downfall ».
86 Mathiot C., Moullot P., Leboucq F., Pezet J., Andraca R., Coquaz V., « L'ivresse de la deepfake – Désintox », *ARTE*, Juin 2019, [en ligne :] https://www.youtube.com/watch?v=nrNzG2yQqOA, consulté le 19 juillet 2019.

semble prouver quelque chose à laquelle ils croient déjà, les internautes penseront plus volontiers que la vidéo est réelle, ou n'en auront cure, et la partageront. Quand bien même une vidéo aurait été mise en ligne comme satire, au départ.

Mais si les deepfakes étaient plutôt rares dans les campagnes électorales jusqu'ici, l'année 2024 a vu la moitié de la planète aller voter et le nombre de deepfakes politiques se multiplier.

Aux États-Unis, ce sont des hypertrucages audios qui ont fait parler d'eux. Dès janvier, un étonnant appel automatique a été reçu par 5000 électeurs du New Hampshire. « *Au bout du fil, la voix de Joe Biden qui, dans un message préenregistré, leur déconseillait d'aller voter aux primaires démocrates américaines du mardi 23 janvier. "Votre vote fera la différence en novembre, pas ce mardi (…) Voter ce mardi ne fera qu'aider les républicains à faire réélire Donald Trump". C'était un faux : la voix de Joe Biden a été imitée par intelligence artificielle. Un deepfake politique qui a généré beaucoup d'émoi aux États-Unis* ». L'auteur a été découvert un mois plus tard. Steve Kramer, consultant pour la campagne d'un rival de Joe Biden, s'est défendu en argumentant que son initiative avait justement pour objectif de mettre en lumière les dangers de l'intelligence artificielle en politique. « *C'était une façon pour moi de faire la différence, et ça a marché* », a-t-il déclaré à NBC. « *Pour 500 dollars, j'ai obtenu un effet équivalent à 5 millions de dollars* »[87].

Quelques mois plus tard c'est Elon Musk, ardent défenseur de Donald Trump qui partageait un deepfake de la candidate démocrate en train de dire des choses telles que « *Moi, Kamala Harris, je suis votre candidate démocrate à la présidence parce que Joe Biden a finalement révélé sa sénilité lors du débat* » contre Donald Trump, explique la voix générée par IA dans la vidéo. Dans cette campagne artificielle, l'on peut entendre la fausse Kamala Harris affirmer qu'elle est une « *recrue de la diversité* », qu'émettre la moindre critique contre elle est « *sexiste* » ou « *raciste* » et qu'elle ne connaissait « *rien à la gestion du pays* »[88]. La vidéo, qui arbore le logo « *Harris for President* », n'a jamais été signalée par Elon Musk comme étant générée par l'IA. L'affaire a remis au premier plan la responsabilité des réseaux sociaux dans la diffusion de ces contenus. Surtout que X déclare interdire « *le partage de médias synthétiques, manipulés ou hors contexte qui peuvent tromper ou dérouter les gens et entraîner des préjudices* », à l'exception des mèmes et de la satire « *à condition qu'ils ne provoquent pas de confusion significative quant à l›authenticité des médias* ». Même pour de la satire, il semble tout à fait déplacé d'utiliser une vidéo d'une personne sans son consentement, surtout avec une vidéo hyperréaliste.

À la veille des élections présidentielle, les autorités américaines ont pris très au sérieux ce genre d'affaires. Tout comme en Europe, comme nous le verrons plus loin.

Il faut dire qu'aux États-Unis, des clans complotistes s'affrontent. Les Q-Anon, sympathisants de Trump, qui croient que des personnalités influentes sont impliquées dans des réseaux pédophiles internationaux, qu'elles veulent créer un nouvel ordre mondial dans lequel les États auraient abandonné leur sou-

87 La rédaction du Monde, « Deepfake de Joe Biden : l'identité du commanditaire dévoilée », *Le Monde*, le 26 février 2024, [en ligne :] https://www.lemonde.fr/pixels/article/2024/02/26/deepfake-de-joe-biden-l-identite-du-commanditaire-devoilee_6218633_4408996.html, consulté le 17 août 2024.

88 Bensinger K., « Elon Musk Shares Manipulated Harris Video, in Seeming Violation of X's Policies », *The NY-Times*, le 27 juillet 2024, [en ligne :] https://www.nytimes.com/2024/07/27/us/politics/elon-musk-kamala-harris-deepfake.html, consulté le 18 août 2024

verainеté au profit de cette élite, et que seul Donald Trump pourrait les contrer, s'il est réélu. Comme réponse à l'absurdité de cette théorie, des sympathisants démocrates ont aussi créé une mouvance complotiste appelée BlueAnon, qui s'approprie l'univers et les codes de QAnon tout en inversant la cible. Ils entendent ainsi Lutter contre « *l'État profond* »[89] qui veut « *détruire la candidature du président Biden* » (de Kamala Harris désormais) et « *ramener Trump au pouvoir le 5 novembre* ». Affirmant que les médias refusent de relayer des révélations contenues dans de nouveaux documents rendus publics, impliquant Donald Trump dans le réseau de trafic sexuel de mineures du financier et prédateur sexuel Jeffrey Epstein. Le tout étayé par … une photo deepfake. Quelques incohérences au niveau des doigts de certaines mains et de la longueur du bras de la jeune fille de gauche ont confirmé qu'il s'agissait bien d'un deepfake.

Image deepfake, produite par les BlueAnon afin de discréditer Donald Trump et le camp des QAnon. La guéguerre entre complotistes risque d'alimenter les réseaux sociaux en deepfakes.

Côté européen, à l'approche des élections de 2024, on a pu découvrir Amandine Le Pen et Léna Maréchal sur TikTok. Deux sympathiques avatars sexy avec les visages rajeunis de Marine Le Pen et celui de sa nièce de Reconquête, Marion Maréchal. Deux profils deepfakes qui faisaient la promotion de l'extrême droite en présentant une image glamour, jeune et moderne du parti grâce à de faux contenus générés par l'IA. Très bien faites, ces vidéos reprenaient les codes TikTok qui plaisent à la jeune génération[90] : des musiques tendances, accompagnées de chorégraphies réalisées par des jeunes filles, bien souvent sur fond de paysages clinquants. Le tout en mettant en avant leur plastique avantageuse et accompagné de commentaire de type : « *Quand le RN sera élu* », « *Moi quand je vais voter Reconquête en juin* », « *La robe que je vais porter pour la victoire de Jordan (Bardella)* » … Si l'on peut croire à de simples comptes parodiques créés pour faire sourire les internautes, ils semblent bien avoir été créés pour les tromper et faire la publicité du Rassemblement national auprès des jeunes utilisateurs.

89 L'État profond (de l'anglais deep state) fait référence à l'idée qu'il existerait au sein d'un État une entité informelle détenant secrètement le pouvoir décisionnel sur la société.

90 L'équipe de « C quoi l'info ? », « Amandine Le Pen, Lena Maréchal Le Pen… Mais qui sont ces fakes ? », *Youtube*, le 12 avril 2024, [en ligne :] https://www.youtube.com/watch?v=nNyyPYYut1I&t=31s, consulté le 20 août 2024.

On pouvait même leur verser de l'argent. Les deux femmes politiques parodiées ont réfuté tout lien avec ces comptes. Car il est difficile de faire des liens entre ce genre de comptes et les campagnes de partis. Voilà une utilisation des deepfakes dans une campagne qui est simple, bon marché mais à l'efficacité difficilement quantifiable, mis à part les trente mille abonnés chacune [91].

Plus troublant, « *En Slovaquie, en mars 2024, une fausse conversation générée par IA mettait en scène la journaliste Monika Tódová et le dirigeant du parti progressiste slovaque Michal Semecka fomentant une fraude électorale. Les enregistrements diffusés sur les réseaux sociaux pourraient avoir influencé le résultat de l'élection. Le même mois, en Angleterre, une soi-disant fuite sur X fait entendre Keir Starmer, le leader de l'opposition travailliste, insultant des membres de son équipe. Et ce, le jour même de l'ouverture de la conférence de son parti. Un hypertrucage vu plus d'un million de fois en ligne en quelques jours* » [92].

Dans nos démocraties, le sujet des deepfakes est relativement bien pris au sérieux, que ce soit par nos politiques ou nos journalistes. Les alertes sont régulièrement lancées quand des photos ou des vidéos truquées sont mises en ligne. En revanche, certains gouvernements sont moins regardants.

Exemple avec cette utilisation beaucoup plus dérangeante, celle du pouvoir en place au Vénézuela. L'équipe du président Maduro en use et abuse. Les Vénézuéliens ont ainsi pu profiter de deux avatars, pseudo présentateurs de journal télévisé, vantant le renouveau de l'économie du pays. Une vision idéalisée du programme présidentiel pour contrer les déclarations des autres médias. « *Les présentateurs de House of News, Noah et Daren, ont été sélectionné parmi plus d'une centaine de visages multiethniques disponibles sur le software Synthesia … Mais ce vaste catalogue humain ne propose pas uniquement des journalistes comme Noah et Daren. On peut choisir un Dave avec un look de médecin ou de cadre supérieur, un Carlo avec des vêtements de travail et un casque et même le père Noël* » [93]. La propagande chez un dictateur n'est pas neuve, dans les années 1970 et 1980 les JT sur l'OZRT, l'Office Zaïrois de Radio-Télévision, étaient tout à la gloire du Maréchal Mobutu. Mais leur modernisation est bluffante, elles peuvent être adaptées et améliorées, et on peut les multiplier à l'envi pour discréditer le travail des journalistes, cibles très à la mode. Cela a permis également de montrer aux Vénézuéliens de faux JT étrangers, notamment américains et donc largement hostiles à Maduro, parlant positivement de leur pays, deepfakes à l'appui [94]. De quoi ajouter de la confusion à la confusion.

En Turquie, les élections présidentielles de mai 2023 ont été entachées par des deepfakes. Un candidat de l'opposition, Muharrem Ince, s'est ainsi retiré après avoir été la cible d'une campagne de dénigrement en ligne, qui comprenait notamment des images truquées de lui avec des femmes ou au volant de voitures

91 La rédaction de France Info, « Amandine Le Pen, Lena Maréchal… Quels sont ces deepfakes qui poussent à voter pour l'extrême droite ? », *France Info*, le 12 avril 2024, [en ligne :] https://www.francetvinfo.fr/l-actu-pour-les-jeunes/marine-le-pen-marion-marechal-mais-qui-sont-ces-fakes-d-extreme-droite-regardez-le-nouveau-numero-de-c-quoi-l-info_6482963.html, consulté le 20 août 2024.

92 Frau-Meigs D., « Deepfakes, vidéos truquées, n'en croyez ni vos yeux ni vos oreilles ! », *The Concersation et RFI*, le 10 juillet 2024, [en ligne :] https://www.rfi.fr/fr/connaissances/20240710-deepfakes-vid%C3%A9os-truqu%C3%A9es-n-en-croyez-ni-vos-yeux-ni-vos-oreilles, consulté le

93 Herrera Y., « Venezuela : de faux JT présentés par des avatars prêchent la parole du gouvernement chaviste », *Libération*, le 21 février 2023, [en ligne :] https://www.liberation.fr/international/amerique/venezuela-de-faux-jt-presentes-par-des-avatars-prechent-la-propagande-de-nicolas-maduro-20230221_EEFBBHGKT5ACTEBVKDOQAPRCZM/, consulté le 21 août 2024.

94 Kansara R. et Krygier R., « Venezuela: ‚I'm paid to tweet state propaganda' », *BBC*, le 26 mai 2023, [en ligne :] https://www.bbc.com/news/blogs-trending-65622685, consulté le 22 août 2024.

de luxe. Plus étonnant encore, en plein meeting, Recep Tayyip Erdogan a diffusé un clip de quatorze secondes, présenté comme la preuve que son principal rival, Kemal Kiliçdaroglu, « *avance main dans la main avec le groupe armé PKK* »[95]. Kemal Kilicdaroglu, principal rival du président, a ensuite affirmé que des pirates étrangers, recrutés par le camp Erdogan, préparaient des deepfakes, vidéos ou sons manipulés grâce à l'intelligence artificielle, afin de le discréditer, ciblant principalement la Russie. Autre exemple, en 2024, le président turc Recep Tayyip Erdogan menait « *une bataille pour reprendre le contrôle d'Istanbul lors d'élections locales très disputées … Mais alors que le parti AK d'Erdogan intensifie ses efforts pour reprendre le contrôle de Istanbul, une vidéo générée par l'intelligence artificielle du maire sortant Ekrem Imamoglu félicitant Erdogan pour ses réalisations à Istanbul, a circulé sur les médias sociaux … Les médias indépendants ont alors mis en garde contre la menace de fausses nouvelles, car les médias traditionnels, qui sont principalement sous le contrôle du gouvernement*[96]*, ne vérifiaient pas l'authenticité des vidéos* »[97]. Le secteur des médias indépendants ferait face à une pression des autorités turques et une grande partie de leurs nouvelles seraient bloquées sur les médias sociaux. Une tendance confirmée par Emma Sinclair-Webb, chercheuse principale en Turquie de Human Rights Watch : « *Ce que nous avons vu, c'est que très, très souvent du matériel, principalement des nouvelles sur les médias sociaux, est supprimé et bloqué en ligne … Il est très inquiétant de voir que les autorités sont prêtes à réprimer la liberté d'expression, mais les entreprises de médias sociaux elles-mêmes ne sont pas assez robustes pour résister à cette pression* »[98]. La Turquie est ainsi un bon exemple de guerre de récits, en période électorale, appuyé par des deepfakes. Mais ces derniers ne montrent une efficacité que grâce aux pressions gouvernementales sur les médias de presse et les médias sociaux. La pression sur ces derniers est bien réelle. Exemple avec le blocage en 2023 par Elon Musk de quatre comptes sur X, car il était effrayé par la menace du régime de bloquer son réseau social dans tout le pays. Car Erdogan ne plaisante pas et l'a montré en 2024, quand Instagram, pourtant utilisé par 57 millions de Turcs, a été complètement bloqué par les autorités qui n'ont pas donné de raison claire mais le président avait accusé le réseau social de censurer les critiques contre Israël et certains messages de soutien aux Palestiniens. Il a été jusqu'à parler de «fascisme numérique». Voilà donc un exemple de pays qui montre combien la censure de deepfakes par les plateformes n'est pas la panacée, ces dernières n'étant pas en mesure de tenir tête à un régime autoritaire, à la tête d'un juteux marché de quatre-vingts millions d'âmes.

De plus, l'efficacité de deepfakes dépend beaucoup des contextes politiques locaux. En 2023, au Bangladesh, pays en proie à de vastes émeutes contre le régime, des dizaines d'experts, de spécialistes et autres écrivains ont été créés par l'IA pour défendre le bilan du gouvernement ou accuser des pays étrangers d'être responsables des problèmes économiques du pays. Faux articles, fausses photos. Le tout propagé par des médias nationaux. Peine perdue, la première mi-

95 La rédaction numérique de France Inter et AFP, « Présidentielle en Turquie : la campagne polluée par les "deepfakes", les infox et les montages », *France Inter*, le 12 mai 2023, [en ligne :] https://www.radiofrance.fr/franceinter/presidentielle-en-turquie-la-campagne-polluee-par-les-deepfakes-les-infox-et-les-montages-5450979, consulté le 23 août 2024.

96 D'après Reporter Sans Frontière, le pouvoir turc contrôlerait 90 % des médias nationaux.

97 Jones D., « Deepfake videos used in local elections in Turkey as Erdogan battles for Istanbul », *RFI*, le 16 mars 2023, [en ligne :] https://www.rfi.fr/en/podcasts/international-report/20240316-deepfake-videos-used-in-local-elections-in-turkey-as-erdogan-battles-for-istanbul, consulté le 26 août 2024.

98 *Ibid.*

nistre Sheikh Hasina devra démissionner [99]. Il faut dire que cette autocrate était au pouvoir depuis quinze ans et que, quel qu'ait pu être sa communication, il était trop tard pour convaincre un peuple écœuré et révolté.

Comme le précise Danielle Citron, spécialiste des deepfakes et professeure à la Faculté de droit de l'Université de Virginie : « *La possibilité d'influencer le résultat d'une élection est réelle, en particulier si l'auteur est capable de programmer la diffusion de manière à ce qu'il y ait suffisamment de temps pour que le contenu trafiqué circule, mais pas assez pour que la victime puisse le démentir efficacement – à supposer qu'il puisse être démenti* » [100]. L'efficacité politique d'un deepfake dépend donc non seulement du contexte politique mais aussi du timing de sa publication et sans doute de la manière dont il est diffusé. L'Inde est ainsi un laboratoire pour les deepfakes publiés en période électorale. Dans cet immense pays, où la désinformation est largement utilisée, et où une grande partie de la population est peu éduquée aux subtilités des médias numériques, on a pu voir « *Des hommes politiques défunts ressuscités pour soutenir un candidat dans le Tamil Nadu ; un dirigeant d'un parti musulman entonnant des chants de dévotion hindous ; des stars de Bollywood d'habitude très discrètes, Ranveer Singh et Aamir Khan, critiquant ouvertement le premier ministre indien et apportant leur soutien au Congrès, le principal parti d'opposition… Les deepfakes (ou hypertrucages), ces vidéos reproduisant à s'y méprendre visages et voix et pouvant servir à propager de la désinformation, ont envahi les réseaux sociaux, comme les fake news avaient déjà marqué la campagne de 2019* » [101]. Plus efficace encore, un message vocal personnalisé du président nationaliste Narendra Modi adressé à chaque électeur en l'appelant par son nom, via Whatsapp. L'avatar du président leur parle des avantages gouvernementaux qu'ils ont reçus et demande leur vote. Le tout réalisé par l'équipe de communication du président, sans que ce dernier n'ait à y travailler une seconde. C'est une nouvelle méthode pour parler directement aux électeurs via des *chatbots* élaborés, qui leurs fait croire qu'il les connaît et connaît leurs problèmes. Pour Suhasini Raj, qui a écrit un article sur ces méthodes pour The New-York Times, « *Pour avoir un aperçu de l'avenir de l'intelligence artificielle dans les campagnes électorales, regardez ce qu'il se passe en Inde* » [102]. Car M. Modi n'est pas le seul à utiliser ce système qui pourrait faire des émules dans d'autres pays. La méthode, qui rappelle celle de Cambridge Analytica aux USA et en Grande-Bretagne en 2016, pourrait amener à la propagation de mensonges et de désinformations personnalisés juste pour gagner une élection. Pour Nicolas Obin, Maître de conférences à Sorbonne Université et chercheur à l'Ircam (Institut de recherche et coordination acoustique/musique), il existe des manipulations pernicieuses, « *comme la manipulation des émotions, qui s'adresse à nos affects. Par exemple, un assistant vocal pourrait avoir des interactions émotionnelles ou expressives et influencer nos comportements en infléchissant nos émotions, ou en nous incitant à*

99 AFP, « Fake experts drive disinformation before Bangladesh polls », *France 24*, le 7 septembre 2023, [en ligne :] https://www.france24.com/en/live-news/20230907-fake-experts-drive-disinformation-before-bangladesh-polls, consulté le 27 août 2024.

100 La redaction de l'Union Interpalementaire, « Dangers des "deepfakes" pour les parlementaires », *UIP*, le 21 février 2024, [en ligne :] https://www.ipu.org/fr/actualites/actualites-en-bref/2024-02/dangers-des-deepfakes-pour-les-parlementaires, consulté le 25 août 2024.

101 Landrin S., « En Inde, des élections dopées aux deepfakes », *Le Monde*, le 20 mai 2024, [en ligne :] https://www.lemonde.fr/pixels/article/2024/05/20/en-inde-des-elections-dopees-aux-deepfakes_6234465_4408996.html, consulté le 25 août 2024.

102 Raj S., « How A.I. Tools Could Change India's Elections », *New-York Times*, le 18 avril 2024, [en ligne :] https://www.nytimes.com/2024/04/18/world/asia/india-election-ai.html, consulté le 26 août 2024.

acheter quelque chose, etc. En politique, un même discours pourrait être adressé à chaque citoyen avec des variations de ton adapté pour obtenir un effet optimal de persuasion »[103]. Là on entrerait dans la manipulation caractérisée.

En février 2022, le candidat conservateur sud-coréen, Yoon Suk-Yeol, présentait son avatar, histoire d'attirer les jeunes, lassés par les politiciens qu'ils estimaient trop éloignés de leurs problèmes. Cet avatar, un deepfake, répondait aux questions des citoyens sur le net avec un langage humoristique et satirique, utilisé dans les jeux en ligne, avec des phrases calibrées[104] pour devenir virales. Exemple : « *Le président Moon Jae-in et Lee Jae-myung (le candidat du parti au pouvoir, ndlr) se noient. Lequel sauvez-vous?* », demande un internaute, « *Je leur souhaite bonne chance à tous les deux* » rétorque l'avatar. Des répliques faites pour devenir virales, en réalité données par une équipe de campagne et non par le candidat lui-même. Les adversaires ont eu beau reprocher au candidat de rabaisser le niveau de la campagne, des millions de questions ont été posées dans ce pays qui compte l'internet le plus rapide du monde. Dans une interview, le directeur de la campagne de l'avatar, Baik Kyeong-hoon a déclaré : « *Les mots prononcés le plus souvent par Yoon sont mieux reproduits dans l'avatar* » et « *Il est inévitable que cette technologie soit utilisée dans de futures élections* »[105].

Et encore une fois, internet aura été au centre d'une élection présidentielle, puisque le 10 mars dernier, Yoon Suk-yeol s'impose dans le scrutin le plus serré de l'histoire de la Corée du Sud. Avec une carrière politique éclair, un programme conservateur et une personnalité controversée, il est parfois surnommé le « Donald Trump coréen » ou K-Trump[106]. Même si le gendarme électoral de Corée du Sud autorise les avatars de candidats à condition qu›ils soient identifiés comme technologie deepfake et ne diffusent pas de la désinformation, on peut encore une fois se demander, comme pour les candidats en Inde, si on ne donne pas l'illusion aux citoyens que c'est le candidat qui répond directement à leurs questions avec un langage qui n'est pas du tout le sien ? N'est-ce pas une fausse impression de proximité avec le candidat ? D'autant qu'il est qualifié de piètre orateur et surnommé « Monsieur une gaffe par jour ». Cela ouvre un autre débat : une personnalité politique doit-elle avoir du charisme et être bonne oratrice pour diriger un pays, et les deepfakes ne sont-ils pas une solution à ce problème ? Nul doute que ces questions risquent de revenir à l'avant plan dans divers pays du monde.

Il est clair que la guerre de l'information est l'une des composantes de la guerre et, dans un contexte géopolitique international en pleine restructuration où les grandes nations comme la Chine, les États-Unis, l'Iran ou encore la Russie peuvent ne pas se faire de cadeaux, à quels types de vidéos doit-on s'attendre

---

103 Entretien avec Obin N., « Deepfake : menace ou opportunité ? », *Sorbonne Université*, le 29 février 2024, [en ligne :] https://www.sorbonne-universite.fr/actualites/deepfake-menace-ou-opportunite, consulté le 26 août 2024.

104 Le candidat Yoon de chair et d'os a enregistré plus de 3 000 phrases, soit vingt heures d'audio et de vidéo, pour fournir suffisamment de données à une entreprise sud-coréenne de technologie deepfake chargée de créer l'avatar.

105 La rédaction d'Euronews avec AFP, « En Corée du Sud, un avatar "deepfake" pour booster la campagne d'un candidat », *Euronews*, le 14 février 2022, [en ligne :] https://fr.euronews.com/2022/02/14/en-coree-du-sud-un-avatar-deepfake-pour-booster-la-campagne-d-un-candidat, consulté le 3 août 2022.

106 Rocca N., « Corée du Sud: Yoon Suk-yeol, prochain président, un K-Trump ? », *RFI*, 10 mars 2022, [en ligne :] https://www.rfi.fr/fr/asie-pacifique/20220310-cor%C3%A9e-du-sud-yoon-suk-yeol-prochain-pr%C3%A9sident-un-k-trump, consulté le 3 août 2022.

pour discréditer l'adversaire ? Rappelons-nous comme Chine, Russie et États-Unis se sont livrés à une bataille de propagandes acharnées quant à la responsabilité de l'un ou de l'autre face à la pandémie de Covid-19.

En 2022, les Ukrainiens, en pleine guerre, ont pu entendre et voir un deepfake de leur président annonçant *« Je vous recommande de déposer les armes et de retourner auprès de vos familles. Vous ne devriez pas mourir dans cette guerre. Je vous demande de vivre, et je compte faire de même ! »* [107]. Le site de la chaîne nationale Ukraine 24 avait été piraté et relayait la vidéo, comme pour lui donner plus de crédit. Mais le gouvernement avait préalablement et largement prévenu son armée de ce genre de risque et des alertes au fake ont rapidement été lancées par la chaîne nationale et par Meta. Il faut également ajouter que le deepfake n'était, étonnement, pas d'une grande qualité.

Ajoutons qu'un deepfake n'est pas souvent nécessaire pour convaincre. Souvenons-nous qu'avant l'invention de l'hypertrucage, des photos satellites de prétendues armes de destruction massive, avaient été présentées par les États-Unis devant le Conseil de Sécurité de l'ONU le 5 février 2003, afin de discréditer au maximum le régime de Saddam Hussein et de légitimer une attaque de l'Irak. On sait que tout cela était faux.

Politiquement, les deepfakes n'ont pas encore prouvé qu'ils étaient d'une efficacité considérable, comme beaucoup le craignent depuis plusieurs années. Ils sont un outil supplémentaire de propagande et de désinformation parmi des milliers d'autres, mais ils sont aussi très visuels et cela ajoute à la confusion. Antonin Descampe, professeur en journalisme et en innovation média à l'UCLouvain, parle carrément de «désordre informationnel massif» plutôt que de désinformation. *« On se retrouve aujourd'hui dans une situation dans laquelle effectivement tout contenu, qu'il s'agisse de texte mais également d'audio, de vidéo, d'image, peut potentiellement avoir été généré, éventuellement partiellement, par une intelligence artificielle. Et donc on se retrouve dans une situation où le citoyen est dans un état de confusion. Il ne sait pas si la photo a été prise véritablement ou fabriquée par une intelligence artificielle, et cela ne lui permet pas de faire des choix éclairés dans son environnement »* [108]. Il est actuellement impossible de connaître l'impact de toutes ces fausses informations mais on sait qu'elles peuvent influencer un jugement, comme démontré par Olivier Klein, professeur de psychologie sociale à l'ULB. Selon son étude, *« Même une information que l'on sait fausse nous influencera »* [109].

En revanche, ce qui nous semble être un danger provoqué par des deepfakes, c'est l'avenir de la preuve vidéo, audio ou autre, qui peut être présentée comme fausse. Ce fut déjà le cas en Afrique centrale fin 2018. Le président gabonais, Ali Bongo Ondimba, se faisait soigner à l'étranger depuis plusieurs semaines suite à un AVC. Son discours à la nation, prononcé le 31 décembre 2018, devait rassurer la population inquiète et faire taire toute rumeur de décès. Mais celle-ci

107 Beurnez V., « Piratée, une chaîne d'information ukrainienne diffuse un "deepfake" de Volodymyr Zelensky », *BFM Tech&Co*, le 17 mars 2022, [en ligne :] https://www.bfmtv.com/tech/piratee-une-chaine-d-information-ukrainienne-diffuse-un-deepfake-de-volodymyr-zelensky_AN-202203170296.html, consulté le 17 septembre 2024.

108 Biourge C., « L'intelligence artificielle : une menace de plus en plus importante pour l'information et un défi pour les journalistes », *op. cit.*

109 d'Otreppe B., « Les risques des "fake news": "Même une information que l'on sait fausse nous influencera" », *La Libre Belgique*, 29 octobre 2018, [en ligne :] https://www.lalibre.be/belgique/2018/10/29/les-risques-des-fake-news-meme-une-information-que-lon-sait-fausse-nous-influencera-H2M3NPDKENB4RL5CEXSNYTGYZM, consulté le 31 mars 2020.

fut accusée d'être un deepfake, par ses opposants politiques, et a servi de déclencheur, une semaine plus tard, à un coup d'État infructueux de l'armée gabonaise, à Libreville, le 7 janvier 2019. La première tentative de coup d'État au Gabon depuis 1964 [110].

En 2018, Joao Doria, le gouverneur de Sao Paulo au Brésil, marié, a affirmé qu'une vidéo qui le montrait lors d'une orgie sexuelle était un faux et personne n'a pu prouver de manière concluante que ce n'était pas le cas.

Les deepfakes posent ainsi une question très importante : pourrait-on désormais discréditer toute preuve de malversation politique, qu'elles soient vidéos, audios, scripturales ou autre ? Un commentaire déplacé d'un chef d'état comme Trump lui suffirait pour invoquer un deepfake, comme il a qualifié de fake news tout ce qui desservait sa cause [111]. Autre cas de figure, en mai 2019 le chancelier conservateur autrichien Sebastian Kurz avait annoncé la démission de l'ensemble des ministres d'extrême droite, à la suite de la diffusion d'une vidéo compromettante pour le FPÖ. On y voyait, en caméra cachée, le chef du FPÖ, Heinz-Christian Strache, se dire prêt à accepter des financements russes occultes [112]. Cet Ibizagate, comme on l'a surnommé, confirme, d'une part, combien les mandats ministériels restent fragiles face à une vidéo compromettante, mais on peut désormais aussi se demander s'il sera encore possible à l'avenir d'utiliser comme preuve une telle vidéo vu l'accessibilité croissante aux technologies du deepfake.

Pourra-t-on désormais croire une image de caméra de surveillance souvent de moindre qualité ? Ou celle d'un témoin de violences impliquant une personnalité ou un agent de l'état ?

2. Propagandes 4.0.

Nous l'avons vu, des manipulations de l'opinion par les gouvernements sont également envisageables, notamment depuis l'étranger. Si la Russie a clairement tenté d'influencer les élections américaines en 2016, d'autres de ses méthodes sont moins connues. Ainsi, un rapport [113] d'experts français souligne combien des continents comme l'Afrique ou l'Amérique latine, avec des langues communes à plusieurs pays et des populations moins averties – et néanmoins très connectées grâce à la démocratisation des technologies de l'information et de la communication – étaient susceptibles d'être traversées de passions faciles à instrumentaliser, dont des tensions ethniques et religieuses, et un ressentiment à l'égard des anciennes puissances coloniales. Le rapport cite le cas du Maghreb,

110 Owono J., « Les "deepfakes", arme de désinformation massive », *Jeune Afrique*, 11 avril 2019, [en ligne :] https://www.jeuneafrique.com/mag/760688/societe/tribune-les-deepfakes-arme-de-desinformation-massive/, consulté le 31 mars 2020.

111 À la veille des élections américaines de 2016, Donald Trump avait défrayé la chronique avec une vidéo de 2005 dans laquelle il tenait des propos obscènes. Juste avant une interview pour NBC, oubliant qu'il a un micro-cravate, il parle dans un bus de son rapport avec les femmes. Extraits : « "J'ai essayé de me la faire, elle était mariée", "Je suis automatiquement attiré par les belles, je les embrasse tout de suite" ... il explique notamment qu'être une star lui permet de faire "tout ce qu' [il] veut" avec les femmes, notamment les "attraper" par la "chatte" ». Cet enregistrement avait fait scandale. Aujourd'hui, les pro-Trump auraient désormais beau jeu de crier au deepfake pour discréditer l'info.

112 AFP, « Autriche: démission de tous les ministres d'extrême droite », Le Soir, 20 mai 2019, [en ligne :] https://www.lesoir.be/225482/article/2019-05-20/autriche-demission-de-tous-les-ministres-dextreme-droite, consulté le 18 juillet 2020.

113 Jeangène Vilmer J-B., Escorcia a., Guillaume M., Herrera J., « LES MANIPULATIONS DE L'INFORMATION - Un défi pour nos démocraties », *France Diplomatie*, 4 septembre 2018, [en ligne :] https://www.diplomatie.gouv.fr/fr/politique-etrangere-de-la-france/manipulations-de-l-information/rapport-conjoint-caps-irsem-les-manipulations-de-l-information-un-defi-pour-nos, consulté le 12 mai 2020.

où les populations sont largement exposées à la propagande des médias russes en arabe, qui véhicule des messages anti-européens, dont elles ne sont que la cible indirecte, le vecteur. L'objectif est que ces populations, qui sont en lien quotidien avec leurs familles et leurs proches vivant en Europe, leur transmettent ces messages et les convainquent que les médias européens leur mentent et que les Européens leur sont hostiles. La propagande anti-immigration que l'on voit en Europe visant à exciter les communautés nationalistes n'est donc qu'une face de l'opération. Pour diviser, monter les communautés les unes contre les autres, il faut aussi convaincre les populations issues de l'immigration qu'elles sont maltraitées et, de ce point de vue, le fait de passer par des relais en Afrique du Nord est particulièrement habile [114]. Et pour sa propagande, la Russie n'hésite pas à utiliser des deepfakes. On peut citer l'exemple d'une vidéo de JT de France 24, annonçant qu'Emmanuel Macron aurait annulé sa visite en Ukraine en raison d'un projet d'assassinat contre lui en Ukraine [115]. Il est à noter que les journalistes souvent traités de menteurs sont régulièrement utilisés comme sources fiables par les complotistes pour créditer leurs théories.

Beaucoup craignent ainsi que les deepfakes ne s'ajoutent à des manipulations d'opinion, à l'attisement de réactions émotionnelles, à la création de scandales politiques pour affaiblir un gouvernement, voire même à des dissensions entre deux États en les cumulant à des cyberattaques.

3. La tentation des « campagnes positives »:

Un article du *Vice Magazine* [116] soulève un aspect inquiétant d'une campagne électorale en Inde. Le 7 février 2020, un jour avant les élections à l'Assemblée législative à Delhi, le président du Parti Bharatiya Janata (BJP), Manoj Tiwari, critique face caméra le bilan de son adversaire politique dans trois vidéos identiques mais chacune dans une langue différente : hindi, anglais et haryanvi, un dialecte du nord de l'Inde. Mais seule la première a été réellement prononcée par le candidat, les autres sont des deepfakes. L'un des objectifs de l'homme politique était bien sûr de cibler particulièrement les électeurs migrants venant d'Haryana et qui travaillent à Dehli pour les convaincre de ne pas voter pour le ministre en chef sortant [117].

« *La technologie Deepfake nous a aidé à intensifier les efforts de campagne comme jamais auparavant* », a déclaré le co-responsable des médias sociaux et de l'informatique pour BJP Delhi. Selon lui, quinze millions de personnes ont vu ces vidéos diffusées via des milliers de groupes WhatsApp, application ultra populaire en Inde. Bien sûr l'intention est de se faire comprendre d'un maximum de gens, en évitant les sous-titres qui n'auraient pu être lus par une partie analphabète de la population, mais cela reste faux et laisse à penser que ce candidat, parlant l'haryana, est des leurs et les comprend dès lors mieux qu'un autre qui ne parle pas cette langue. Un deepfake, même avec le consentement de la personne et de bonnes intentions, reste un faux. Il doit donc être assumé, ou à tout le moins, signalé.

114 Jeangène Vilmer J-B., Escorcia a., Guillaume M., Herrera J., *op. cit.*

115 Saint-Léger A., « France 24 victime d'un "deepfake" : l'intox continue à circuler sur le web ! », *France 24*, le 16 février 2024, [en ligne :] https://www.france24.com/fr/%C3%A9missions/info-ou-intox/20240216-france-24-victime-d-un-deepfake-l-intox-continue-%C3%A0-circuler-sur-le-web, consulté le 24 juillet 2024.

116 Nilesh C., « Nous venons de voir la première utilisation de Deepfakes dans une campagne électorale indienne », *Vice Magazine*, 18 février 2020, [en ligne :] https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp, consulté le 30 avril 2020.

117 Arvind Kejriwal, qui a finalement remporté l'élection régionale de la capitale indienne.

D'un côté, normaliser ces « deepfakes positifs » pourrait ouvrir une boîte de pandore car l'IA permet des traductions désormais quasi simultanées, et promet de banaliser le phénomène, mais peut aussi faire parler un avatar aux électeurs. Jusqu'où peut-on utiliser cette technologie pour montrer une belle image d'un candidat ou conforter une idéologie ? Désormais, on voit par exemple, des fans de Trump créer leur propre réalité, notamment avec des photos « hypertruquées » de leur candidat préféré aux côté de nombreux Noirs-Américains, des avatars créés par l'IA, comme pour convaincre qu'il n'est pas si suprématiste qu'on le prétend. Alors que devoir créer une telle photo, n'est-ce pas démontrer qu'il n'en existe pas de réelles ?

[118]

Gageons que les citoyen·ne·s du monde entier aient le temps d'être mis au courant de l'existence de ces nouvelles possibilités technologiques qui jouent sur notre perception du monde.

4. Faux témoignages idéologiques

On pourrait imaginer des personnages, créés de toute pièce, faire de faux témoignages. Quels seraient les réactions à chaud du public dans une situation de crise par exemple ?

Citons l'exemple de Jenna Adams qui, de 2014 à 2017, « était une militante pro-Trump connue », icône de l'alt-right américaine [119], citée par les grands médias (dont *The Washington Post*, *The New York Times*, *The Independent* et France 24) et suivie par septante mille comptes sur Twitter. Mais Jenna Abrams n'existait pas :

118 La rédaction du Courrier International, « États-Unis. L'IA utilisée pour créer de fausses images de Trump entouré de Noirs », *Courrier International*, le 11 mars 2024, [en ligne :] https://www.courrierinternational.com/article/etats-unis-de-fausses-images-de-trump-entoure-de-noirs-pour-seduire-cet-electorat, consulté le 19 juillet 2024.

119 L'extrême-droite américaine.

son compte était une création de l'IRA, usine à trolls basée à Saint-Pétersbourg. L'intelligence artificielle rendra ces personnalités fictives plus sophistiquées, moins détectables. Elles pourront donner des interviews en visio-conférence, écrire des tribunes dans la presse, avant d'être découvertes.

Autre cas incroyable, celui d'au moins dix-neuf faux journalistes ou faux experts en géopolitiques qui ont été publiés dans de nombreux journaux anglophones conservateurs. Leurs articles et éditos faisaient l'éloge des Émirats arabes unis et dénonçaient la politique du Qatar, se montrant sceptiques envers la politique de Facebook. Et tous ces pseudo-spécialistes… n'existaient tout simplement pas. Les usurpateurs récupéraient par exemple la photo d'un véritable humain et la modifiaient de manière à ne pas pouvoir être retrouvés avec une recherche inversée. Dans certains cas, les photos de profils montraient des personnes créées par intelligence artificielle. En plus de cette fausse identité ils s'étaient fait un faux CV sur LinkedIn et ils étaient, pour la majorité, des contributeurs sur deux sites construits de toutes pièces : The Arab Eye et Persia Now. Certains articles ont même été relayés par des personnalités comme Ryan Fournier, le cofondateur de l'association Students for Trump, et suivi par près d'un million de personnes sur Twitter, ou encore la sénatrice française de l'Orne, Nathalie Goulet. L'IA va énormément faciliter et diversifier la fabrication de ces pseudo preuves d'existence d'experts, de journalistes, de témoins…

Les hypertrucages risquent ainsi de donner de nouvelles idées de faux témoignages aux *trollers* et aux partisans et idéologues de tout poil.

5. Huile sur le feu

En avril 2015, la police de Baltimore arrête et embarque violemment le jeune afro-américain Freddie Gray qui mourra de ses blessures une semaine plus tard. Sa mort conduira aux émeutes de Baltimore de 2015. Sans parler de Georges Floyd et de Jacob Blake en 2020 et des autres cas de violences policières abusives, voire mortelles, à caractère raciste.

Dans ce genre de climat de haute tension entre communautés, que se passerait-il si une vidéo trafiquée d'un policier tenant des propos racistes par exemple était lancée sur les réseaux sociaux ?

Les manifestations des gilets jaunes ont été un vivier de fake news. De tout temps les fausses informations ont attisé des émeutes voire des révolutions à l'instar de la fameuse phrase apocryphe de la reine de France Marie-Antoinette parlant du peuple : « *Ils n'ont pas de pain ? Eh bien qu'ils achètent de la brioche* » juste avant la Révolution française. Un deepfake risque ainsi d'être un catalyseur de violences en pleine tension sociale ? Et si les pouvoirs en place venaient à souligner qu'une vidéo est fausse, seraient-ils seulement cru par les manifestants ?

D'un point de vue politique, les partis extrémistes ont montré qu'ils étaient capables de mettre de l'huile sur le feu pour appuyer leurs thèses. Il suffit qu'un migrant commette un crime pour que toute la communauté, particulièrement africaine et/ou musulmane, soit mise dans le même sac. Exagération, manipulation, décontextualisation ont déjà été largement utilisés par ceux-ci. La correspondante pour France 24 en Italie écrivait quelques semaines avant les élections européennes de 2019 : « *En Italie, les "fake news" ont eu davantage de visibilité que*

les "vraies" informations selon l'autorité de régulation de l'information (AGCOM) »[120]. Soulignant l'exemple de cette vidéo montrant soi-disant des migrants en train de vandaliser une voiture de carabiniers qui a été vue près de dix millions de fois sur une page de soutien au leader de la Ligue Matteo Salvini. Il s'agissait en fait d'un extrait de film de fiction. Les deepfakes risquent ainsi d'amener encore plus de confusion à une confusion politique, voire envenimer une tension sociale.

6. Galéjades en cascade pour noyer le poisson

Autre danger : « *l'altération discrète d'une partie seulement d'un contenu audio ou vidéo, un discours par exemple. Ou encore la possibilité d'en faire un grand nombre de variations – diffuser une vingtaine de variantes du même discours, par exemple, pour diluer l'authentique dans la confusion* »[121]. Multiplier les faux pour noyer le vrai. Des deepfakes peuvent en effet être créés à l'envi. A la diffusion d'une vidéo compromettante pour une personnalité X, il serait facile d'en créer une autre en prétendant que celle-là est la vraie. Les adeptes de M. ou Mme X seraient tentés d'en produire même plusieurs pour transformer l'originale en faux parmi tant d'autres, accentué par l'effet polarisant des réseaux sociaux qui voit souvent s'affronter les pour et les contre. De quoi en perdre son décryptage.

En 2004, en pleine campagne présidentielle entre le démocrate John Kerry et Georges W. Bush, une « *photographie, faussement attribuée à Associated Press, combinait deux images distinctes pour donner l'impression que M. Kerry partageait une scène lors d'un rassemblement anti-guerre au début des années 1970 avec l'actrice, Jane Fonda* »[122]. Le but était de discréditer Kerry, ancien vétéran du Vietnam, en le mettant en présence d'une comédienne considérée comme traîtresse à la patrie après son voyage de 1972 à Hanoï, la capitale de la République démocratique du Viêt Nam d'Hô Chi Minh. Il est intéressant de constater ici que d'une part les trucages font partie des campagnes depuis longtemps, même dans les pays démocratiques, et que d'autre part, différents opposants à Kerry, ont affirmé que l'image originale avait été fabriquée et que c'était l'image combinée qui était l'image réelle avant que ne soit officiellement démontré le trucage.

Un ou plusieurs deepfakes pourraient ainsi jeter le discrédit sur une vraie info et une vraie information.

7. Du flou pour les géants du Net

Beaucoup de gouvernement incitent les GAFAM à réguler les deepfakes, les audios et vidéos mensongers. Mais comment concilier les points de vue ? Car certaines décisions sont inévitablement politiques. Ce qui est acceptable aux États-Unis, ne l'est pas forcément en UE ou aux Philippines. D'autant que s'il y a un paquet d'argent à la clé, on trouvera toujours un endroit pour produire des deepfakes. Citons quelques exemples pour percevoir la complexité des choix à effectuer par les plateformes.

120 Mendoza N., « Italie : ces "fake news" qui parasitent la campagne des Européennes », *France 24*, le 22 mai 2019, [en ligne :] https://www.france24.com/fr/20190522-italie-europeennes-fake-news-infox-ligue-salvini-migrants-facebook, consulté le 27 juillet 2024.

121 Jeangène Vilmer J-B., Escorcia A., Guillaume M., Herrera J., « LES MANIPULATIONS DE L'INFORMATION - Un défi pour nos démocraties », *op. cit.*

122 Goldenberg S., « Prenez une partie Kerry, une partie Fonda ... et essayez de susciter une polémique politique », *The Guardian*, 18 février 2004, [en ligne :] https://www.theguardian.com/media/2004/feb/18/newmedia.uselections2004, consulté le 20 mai 2020.

Rien qu'aux États-Unis, les GAFAM ont régulièrement des choix « *politiques* » à faire. Traditionnellement depuis le début de notre siècle, les Républicains reprochent aux patrons du numérique d'être plutôt démocrates. Juillet 2022, le service de vidéos de Google, Youtube, a « *ajouté les contenus sur l'avortement à ses règlements sur la désinformation médicale, qui interdisent déjà les contenus faux ou trompeurs sur la COVID-19 ou les vaccins. Par exemple, "les affirmations selon lesquelles les avortements sont très risqués ou causent souvent des infertilités ou des cancers", précise le groupe californien … Les plateformes craignent en effet que les informations personnelles de femmes qui ont avorté ou d'individus qui les auraient aidées (recherches en ligne, déplacements en Uber, etc.) ne soient retenues contre eux par les procureurs d'États conservateurs ayant interdit l'avortement* » [123]. Des informations de santé publique qui sont apparues comme ouvertement démocrates aux yeux de nombreux conservateurs.

À l'été 2022, Facebook a fait l'objet de vives critiques, aux États-Unis, pour avoir « *communiqué à la justice le contenu de conversations entre une mère et sa fille de 17 ans dans un dossier d'avortement illégal* » [124]. Face à la justice, les GAFAM ne peuvent garantir l'anonymat de leurs abonnés, qui peuvent devenir hors la loi suite à une décision de la Cour suprême.

Comment réguler universellement les différentes thématiques ? Difficile pour les amateurs d'art d'accepter la censure d'œuvres d'art, comme la fameuse « Origine du monde » de Courbet, sous des prétextes de pudeur voire de pudibonderie [125].

Ou que faire quand un Donald Trump, alors président des États-Unis, publiait jusqu'à 22 mensonges par jour, à une période de son mandat, selon le Washington Post, journal plutôt critique à l'égard de Trump en général ? Le président a été censuré puis exclu du réseau. Non seulement il a créé son propre réseau social de microblogage, du Trump Media & Technology Group (TMTG), mais cet acte a suscité nombre de critiques au pays de la liberté d'expression. Le résultat a été le rachat de twitter par Elon Musk et la réhabilitation de Trump sur le réseau, qui est également devenu entre-temps, comme souligné dans la publication *Numérique et démocratie*, le lieu d'expression de l'extrême-droite. 2024 semble être l'année de la question : « comment mettre au pas des réseaux sociaux à la puissance démesurée et utilisés par la majorité des citoyens ? ».

Un deepfake est un faux me direz-vous, il suffit de l'indiquer. Mais cette notion est toute relative. Penser qu'il ne peut exister qu'un point de vue à l'échelle de la planète, et sur tous les sujets, afin de légiférer, semble presque illusoire. Et puis nombre de deepfakes, risquent d'être présentés comme de la satire, ce qu'un juge devra apprécier. Oui mais un juge de quel pays ?

Nous le verrons, la plupart des réseaux sociaux y travaillent, même si d'autres beaucoup moins, comme X ou Telegram.

123 AFP, « YouTube instaure des mesures pour mettre fin à la désinformation sur les avortements », *Lapresse.ca*, 21 juillet 2022, [en ligne :] https://www.lapresse.ca/affaires/techno/2022-07-21/youtube-instaure-des-mesures-pour-mettre-fin-a-la-desinformation-sur-les-avortements.php, consulté le 8 août 2022.

124 Leloup D., « Avortement illégal aux Etats-Unis : Facebook critiqué pour avoir fourni à la justice des messages privés », *Le Monde*, 11 août 2022, [en ligne :] https://www.lemonde.fr/pixels/article/2022/08/11/avortement-illegal-aux-etats-unis-facebook-critique-pour-avoir-fourni-a-la-justice-des-messages-prives_6137767_4408996.html, consulté le 22 août 2022.

125 Signoret P., « Censure de "L'Origine du monde" : une faute de Facebook reconnue, mais pas sur le fond », *Le Monde*, le 15 mars 2018, [en ligne :] https://www.lemonde.fr/pixels/article/2018/03/15/censure-de-l-origine-du-monde-une-faute-de-facebook-reconnue-mais-pas-sur-le-fond_5271666_4408996.html, consulté le 8 août 2024.

Le débat sur la censure du fake est loin d'être résolu, d'autant qu'affiner les décisions des algorithmes prend du temps et que ceux-ci sont, en grande majorité, programmés aux États-Unis et en Chine, deux pays qui n'ont pas exactement les mêmes valeurs démocratiques que la Belgique, notamment sur le respect de certaines minorités ou des personnes précarisées. La mathématicienne américaine Cathy O'Neil met en garde contre « *les dangers de certains algorithmes, aux impacts destructeurs dans la justice, l'éducation, l'accès à l'emploi ou au crédit* » [126]. Les algorithmes reproduisent les inégalités sociales et leur programmation reste trop opaque pour le monde entier [127]. Un algorithme n'a aucune conscience des inégalités qu'il porte en lui, alors comment va-t-il estimer un message à tendance raciste, sexiste ou en défaveur des précarisés ?

Discussions et bras de fer font évoluer les débats, entre acteurs et décideurs, mais la technologie évolue bien trop et, désormais, on assiste à une course à l'IA planétaire et phénoménale, qui rend une régulation encore plus urgente mais encore plus complexe.

IV. JE NE CROIS QUE CE QUE JE VOIS, ENFIN JE CROIS

Selon Nicolas Obin : « *Par leur ultra-réalisme, il devient de plus en plus difficile voire impossible de distinguer un vrai d'un faux. Il peut cependant subsister des indices, comme par exemple des déformations ou des incohérences de synchronisation labiale ou entre les expressions du visage. Mais elles sont de plus en plus subtiles. Néanmoins, toute manipulation laisse une trace caractéristique, même imperceptible par un être humain. La détection de ces traces par des IA nécessite de les retrouver et de les identifier. Le problème est qu'il existe une grande variété d'algorithmes de génération, ce qui augmente considérablement la complexité pour les identifier. Et comme l'algorithme utilisé pour la génération est inconnu lorsque nous devons essayer d'identifier un deepfake, il devient extrêmement difficile de proposer une solution universelle de détection robuste à toutes les formes d'attaques* » [128].

Diverses personnalités politiques misent encore sur les initiatives visant à renforcer l'éducation aux médias pour cultiver un public averti. Mais il nous semble quelque peu illusoire d'espérer que l'éducation aux médias arrive à suivre, par exemple, les évolutions de l'IA qui s'affinent chaque jour et des possibilités multiples qu'elle offrira bientôt. Sans parler des inconnues sur son fonctionnement, qui reste flou même pour les experts. Autant demander aux mêmes politiques d'adapter leurs lois à cette vitesse.

En Éducation permanente, nous tentons d'informer, d'éveiller aux multiples dangers d'internet, mais les différents publics sont loin de se passionner pour le fonctionnement du Net, pour la logique des datas ou tout simplement pour un cookie. Rien que la technique de l'image inversée, qui permet de recontextualiser l'origine d'une photo, est excessivement peu utilisée. Dans un *skrolling* où

126 Cuny D. et O'Neil C., « Les algorithmes peuvent creuser les inégalités et saper la démocratie », *Nouvel Obs*, 21 novembre 2016, [en ligne :] https://www.nouvelobs.com/rue89/rue89-le-grand-entretien/20160826.RUE3093/les-algorithmes-peuvent-creuser-les-inegalites-et-saper-la-democratie.html, consulté le 8 août 2022.

127 Gillet E., « Numérisation du recrutement et de l'orientation. Promesses et conséquences des algorithmes », *Citoyenneté & Participation*, Analyse n°472, juin 2023, [en ligne :] https://www.cpcp.be/publications/numerisation-recrutement/, consulté le 10 juillet 2023.

128 Entretien avec Obin N., « Deepfake : menace ou opportunité ? », *op. cit.*

chacun fait défiler les images à des rythmes effrénés, demander de vérifier tout ce qu'on voit est absurde. Les personnes voient une image, l'apprécient ou non, la partagent ou non, le tout prend une demi-seconde. La plupart du temps ils n'ont cure de leur véracité. Ce qu'ils cherchent c'est de l'émotion, comme le joueur devant un bandit-manchot attend de voir des pièces tomber pour vibrer quelques instants.

De plus en plus de spécialistes nous disent que la sophistication des avancées technologiques est telle, qu'imaginer que tout citoyen puisse distinguer les vraies images des fausses est illusoire. Sam Gregory, Directeur exécutif de WITNESS, une organisation internationale à but non lucratif qui aide le public à utiliser la vidéo et les technologies pour protéger et défendre les droits de l'homme, explique ainsi que « *Les conseils qui consistent à "repérer les mains à six doigts", à inspecter les erreurs visuelles sur l'image du Pape en doudoune, à vérifier si une image suspecte cligne des yeux ou à écouter très attentivement le son dans l'espoir d'entendre une erreur ne sont pas suffisants et ne sont d'aucune utilité à long terme, ni même à moyen terme … C'est pourquoi il est urgent de mettre en place une législation ciblée, des techniques dynamiques relatives à la provenance et à la divulgation, ainsi que des outils de détection des dommages potentiels, afin de faire face aux menaces croissantes. Ces mesures doivent s'accompagner d'efforts de collaboration de la part de l'industrie des technologies, de la société civile et des décideurs politiques. Comme toujours, les parlements joueront un rôle essentiel pour mettre tous les acteurs sur la même longueur d'onde* »[129]. En 2023, la RTBF a proposé un jeu où il fallait dire, parmi 20 photos, si elles avaient été générées par l'IA ou non. Sur les 5-6 personnes qui l'ont testé au bureau, aucun n'a fait plus que 16/20, malgré une concentration bien plus appuyée que lors d'un scrolling sur Instagram[130].

V. DES CHIFFRES ALARMANTS ET DES RESPONSABLES ALARMÉS

Depuis les ingérences russes dans les élections américaines de 2016, le monde plonge inexorablement dans l'ère de la cyberguerre[131]. Et le Pentagone, comme l'UE, se montrent très inquiets face aux deepfakes alors même que les opérations militaires se digitalisent aussi[132]. Le Colonel américain Liam Collins citait ainsi diverses techniques modernes de guerre de l'information opérées par les Russes. Par exemple des attaques personnalisées envers des militaires en opération, comme ce fut le cas en Ukraine, déjà en 2015, où le Kremlin envoyait des SMS aux soldats ukrainiens « *visant à altérer leur moral ou leur cohésion, leur*

129 La redaction de l'Union Interpalementaire, « Dangers des "deepfakes" pour les parlementaires », *op. cit.*

130 N'hésitez pas à essayer : [en ligne :] https://www.rtbf.be/article/ia-or-not-ia-arriverez-vous-a-reperer-les-images-generees-par-l-intelligence-artificielle-dans-ce-quiz-interactif-11227306.

131 Sur le site internet de l'OTAN, on peut d'ailleurs lire : « Les cybermenaces et les cyberattaques deviennent de plus en plus fréquentes, sophistiquées et dommageables. L'Alliance est confrontée à un environnement de menaces complexes en pleine évolution. Lors de récents événements, des cyberattaques ont été utilisées dans le cadre d'actions de guerre hybride. L'OTAN et ses Alliés s'appuient sur des moyens de cyberdéfense forts et résilients pour remplir les tâches fondamentales de l'Alliance que sont la défense collective, la gestion de crise et la sécurité coopérative. L'Alliance doit être préparée à défendre ses réseaux et opérations contre les cybermenaces et les cyberattaques toujours plus complexes auxquelles elle est confrontée. ». OTAN, « Cyberdéfense », *OTAN*, 31 May. 2018, [en ligne :] https://www.nato.int/cps/fr/natohq/topics_78170.htm, consulté le 16 septembre 2019, consulté le 21 avril 2020.

132 Redaction De L'armee De Terre Française, « Entrez dans la nouvelle ère numérique de l'armée de Terre », *Défense française*, 2018, [en ligne :] https://www.defense.gouv.fr/web-documentaire/ere_numerique/index.html, consulté le 21 avril 2020.

signifiant par exemple qu'ils étaient « encerclés et abandonnés ». Puis, quelques minutes plus tard, leurs familles recevaient un message leur annonçant la mort de leur fils, leur frère ou leur père, tué par l'ennemi – ce qui suscitait généralement des appels des familles vers les soldats, et permettait, par la concentration de signaux, de détecter leur localisation pour ensuite les bombarder»[133]. Au vu de ce type de technique, on pourrait se demander comment un soldat réagirait sur le terrain à une vidéo truquée de sa famille ou répondrait à un faux ordre donné par un supérieur via un deepfake audio ou vidéo ? D'autant que l'armée se digitalise de plus en plus. Début 2019 déjà, *« Le directeur du Renseignement américain, Dan Coats, a déclaré devant le Congrès, s'attendre à ce que des puissances étrangères hostiles « militarisent des deepfakes » contre les États-Unis et leurs alliés, afin de semer chez eux le doute et la discorde »*[134]. L'agence de recherche du Pentagone, la Darpa, (Defense Advanced Research Projects Agency)[135], collabore avec plusieurs des plus grandes institutions de recherche des États-Unis afin de prendre de l'avance sur les dérives graves possibles. Exemple à l'Université du Colorado de Denver, où des chercheurs travaillent sur le programme de la DARPA pour créer des deepfakes convaincants. Ceux-ci seront ensuite utilisés par d'autres chercheurs qui développent des technologies pour détecter ce qui est réel et ce qui est faux[136]. Ce ping-pong entre chercheurs permet de suivre les évolutions technologiques et de tenter de devancer toute personne mal intentionnée.

Politiciens, journalistes et experts ne cessent de s'inquiéter du phénomène. En février 2024, des experts en intelligence artificielle, dont des sommités belges[137], et des patrons d'industrie ont signé une lettre ouverte[138] appelant à plus de réglementation autour de la création de deepfakes, citant les risques potentiels pour la société. La lettre a été rédigée par Andrew Critch, chercheur en IA à UC Berkeley. Nous allons en grande partie l'évoquer dans ce chapitre car elle illustre parfaitement les craintes et les propositions des experts de la question.

Selon eux, « *Les nouvelles lois devraient :*

1. Complètement criminaliser la pornographie enfantine deepfake, même lorsque seuls des enfants fictifs sont représentés.

133 Jeangène Vilmer J-B., Escorcia A., Guillaume M., Herrera J., « LES MANIPULATIONS DE L'INFORMATION - Un défi pour nos démocraties », Rapport du Centre d'analyse, de prévision et de stratégie (CAPS, ministère de l'Europe et des Affaires étrangères) et de l'Institut de recherche stratégique de l'École militaire (IRSEM, ministère des Armées), août 2018, [en ligne :] https://www.defense.gouv.fr/actualites/economie-et-technologie/lancement-du-rapport-conjoint-caps-irsem.-les-manipulations-de-l-information, consulté le 21 avril 2020.

134 Ripoche B., « États-Unis. Les deepfakes font encore plus peur que les fake news », *Ouest France*, 12/06/2019, [en ligne :] https://www.ouest-france.fr/monde/etats-unis/etats-unis-les-deepfakes-font-encore-plus-peur-que-les-fake-news-6394487, consulté le 9 septembre 2020.

135 La Defense Advanced Research Projects Agency (DARPA), (« Agence pour les projets de recherche avancée de défense ») est une agence du département de la Défense des États-Unis chargée de la recherche et développement des nouvelles technologies destinées à un usage militaire.

136 O'Sullivan D., « Quand voir n'est plus croire », *CNN Business*, 2016, [en ligne :] https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes, consulté le 21 avril 2020.

137 Comme Tony Belpaeme : Professeur en IA et Robotique à l'Université de Gent ; Tom Lenaerts : spécialiste de l'IA, vice-président du département d'informatique de l'ULB, Président de l'Association Benelux pour l'IA et expert au sein du Global Partnership on AI, Georg Riekeles Directeur associé du European Policy Centre, Risto Uuk : Responsable de recherche pour l'UE au Future of Life Institute ; Cristina Vanberghen : Professeur Dr., EUI et Commission européenne, Reconnue par le Service européen pour l'action extérieure (SEAE) pour son rôle dans l'avancement de la diplomatie numérique dans le cadre du G20 et du G7, en particulier pour ses efforts dans la promotion de l'agenda numérique et cybernétique de l'UE en collaboration avec les pays partenaires ; Lode Lauwaert : Professeur de Philosophie de la technologie, KU Leuven ; Frederic Heymans : Chercheur au Knowledge Centre Data & Society ; Gianluca Bontempi : Professeur à l'ULB et fondateur du ULB Machine Learning Group …

138 Critch A., « Disrupting the Deepfake Supply Chain », *FAccT Conférence*, le 21 février 2024, [en ligne :] https://openletter.net/l/disrupting-deepfakes, consulté le 5 août 2024.

2. Établir des sanctions pénales pour quiconque crée ou facilite sciemment la propagation de deepfakes nuisibles.
3. Exiger des développeurs de logiciels et des distributeurs que leurs produits audio et visuels interdisent de créer des deepfakes nuisibles, et d'être tenus responsables si leurs mesures préventives sont trop facilement contournées.

Si elles sont conçues à bon escient, ces lois pourraient nourrir des entreprises socialement responsables et n'auraient pas besoin d'être excessivement lourdes ».

En effet, tout va beaucoup trop vite. Demander à nos sociétés de fonctionner aussi rapidement que l'IA est une illusion et est peine perdue. Il faut absolument instaurer des garde-fous. Nous l'avons vu, aujourd'hui, les « deepfakes » concernent souvent l'imagerie sexuelle, la fraude ou la désinformation politique. Il y a urgence en ce qui concerne les deep-porn, notamment la pédopornographie.

La lettre des experts fait référence à différentes études aux chiffres édifiants.

1. La pornographie, reine du deepfake

L'étude d'une association américaine d'experts en sécurité en ligne, SecurityHero [139], obtient des résultats interpellant. Morceaux choisis :

1. Le nombre total de vidéos en ligne en 2023 est 95.820, représentant une augmentation de 550 % par rapport à 2019. Et rien qu'entre 2022 et 2023, la quantité de pornographie deepfake créée a augmenté de 464 %.
2. La pornographie représente 98 % de toutes les vidéos deepfake en ligne.
3. 99 % des personnes ciblées dans la pornographie deepfake sont des femmes. 58 % sont des chanteuses, 33% sont des actrices.
4. Il faut désormais moins de vingt-cinq minutes et zéro euros pour créer une vidéo pornographique de soixante secondes de quiconque, en utilisant une seule image nette de son visage.
5. 48 % des hommes américains interrogés ont vu de la pornographie deepfake au moins une fois. Et 74 % des utilisateurs de pornographie deepfake ne se sentent pas coupables à ce sujet. 20 % des participants à l'enquête ont envisagé d'apprendre à créer une pornographie deepfake.
6. 7 des 10 plus grands sites pornographiques hébergent des deepfakes.
7. Parmi les dix plus gros sites Web dédiés à la pornographie deepfake, les vues vidéo cumulatives totalisent un nombre de 303 640 207.

Enfin, l'étude souligne deux facteurs importants qui ont joué un rôle central dans l'explosion des deepfakes : la montée en puissance des Réseaux Génératifs Adversaires (GAN) et la disponibilité croissante d'outils, de logiciels et de communautés pour leurs créations. Les outils et logiciels sont désormais accompagnés d'interfaces graphiques intuitives et plus accessibles. Le nombre d'outils conviviaux rendant la génération de contenu deepfake plus accessible est en augmentation, l'étude en a repéré pas moins de quarante-deux, destinés à un large éventail d'utilisateurs.

139 La rédaction de Security Hero, « 2023 state of deepfakes », *Security Hero*, 2023, [en ligne :] https://www.securityhero.io/state-of-deepfakes/, consulté le 10 août 2024.

Autres chiffres instructifs : Les chanteurs et actrices sud-coréens constituent 53% des personnes représentées dans la pornographie deepfake au monde. C'est le groupe le plus souvent ciblé. Suivent les États-Unis avec 20%, le Japon : 10%, l'Angleterre : 6%, la Chine : 3%, l'Inde : 2%, Taïwan : 2%, Israël : 1%, Autres : 4%. Cette diversité géographique souligne combien la prise de décision doit être globale et concertée.

Chez nous, le site deepfuck.com présente un choix de vedettes, au visage incrusté dans des scènes pornographiques. Il est spécialement dédié à ce genre de vidéos manipulées.

2. La cybercriminalité, un argent tellement facile

Autre étude intéressante, celle d'Onfido, une société de lutte contre la fraude qui remet un rapport sur la fraude en matière d'identité chaque année depuis 2019. Quelques éléments de celui de 2024 [140] :

1. 71 Millions de personnes sont victimes de cybercrimes globalement chaque année.
2. 20% des Européens ont été victimes de vols d'identité de 2022 à 2024, sur la base des recherches de la Commission européenne.
3. Un Américain sur trois a été victime de fraude d›identité.
4. La fraude coûte à peu près six milliards de dollars de dommages-intérêts à l›économie mondiale, chaque année.

Cette étude montre l'explosion de l'utilisation des deepfakes dans la cybercriminalité. Et elle ne risque pas de s'arrêter là au vu des ce qu'elle rapporte. D'autant que jusqu'ici, on pouvait encore se rabattre sur la biométrie pour lutter contre la fraude, « *mais la facilité d'accès à l'IA générative et aux applications d'échange de visages a créé une nouvelle voie pour les fraudeurs* ».

3. Les élections

Nous avons déjà évoqué ce chapitre, notamment avec les faux coups de fil de Joe Biden, les deepfake sur Kamala Harris partagé par Elon Musk ou encore les exemples indiens, turcs ou bangladais. C'est également un aspect qui inquiète fortement les experts mais, à ce jour, ils n'ont pas encore démontré qu'ils pouvaient faire basculer une élection. Difficile pour nous-même de savoir parfois ce qui a influencé notre vote. Cela peut être la phrase d'un voisin, une info de presse, une image à la télé ou un deepfake. D'autant que les personnalités politiques sont très réactives lorsqu'il s'agit de dénoncer la diffusion d'un deepfake politique. C'est sans doute le type de fake news le plus surveillé au monde par ces derniers.

En revanche, ces outils utilisés par des dictateurs face à une population peu éduquée aux médias peut faire bien plus de dégâts. Idem dans des pays fracturés, comme le sont les États-Unis actuellement, où les deepfakes peuvent conforter un camp ou l'autre et renforcer les extrémismes. En revanche leur impact sur les électeurs indécis n'a jamais été mesuré clairement.

[140] La rédaction d'Onfido, « Identity Fraud Report 2024 », *Onfido*, 2024, [en ligne :] https://onfido.com/landing/identity-fraud-report, consulté le 17 août 2024.

4. La désinformation

Dans leur lettre ouverte, ces experts déclarent à raison que « *Pour qu'une société moderne fonctionne, les gens doivent avoir accès à des informations crédibles et authentiques. Induire le public en erreur en utilisant l'IA devrait être réglementé et appliqué par des lois spécifiques et formalisées. Il devient de plus en plus difficile de savoir ce qui est réel sur Internet, et des lignes doivent être tracées pour protéger notre capacité à reconnaître de vrais êtres humains* »[141]. Mais les chercheurs Simon, Altay et Mercier, relativisent en soulignant que les sources d'information traditionnelles continuent d'occuper le haut du pavé et que les médias traditionnels restent la référence pour s'informer. Selon eux, le « *public qui s'informe à partir de médias alternatifs et qui "consomme" des fausses informations est déjà abreuvé de telles sources et ne recherche pas tant une information précise que des informations qui confirment leurs idées, fondées sur une méfiance généralisée vis-à-vis des politiques et des médias* »[142]. Il est vrai que quand Trump dit, en plein débat télévisé, contre Kamala Harris, que les immigrés envahissent nos villes, encouragés par les Démocrates qui les font venir pour les faire voter pour eux, ajoutant qu'ils mangent nos chiens et nos chats domestiques, il cherche essentiellement à conforter les complotistes et les anti système.

Une photo truquée d'un meeting imaginaire de Kamala Harris, partagé par Donald Trump sur le réseau X.

Mais, au vu de l'augmentation exponentielle des deepfakes et de leur facilité de manipulation, on peut imaginer une explosion d'hypertrucages plus crédibles avec, par exemple, trois quarts de vrai et un quart de petits mensonges. Les fakes news les plus efficaces ont souvent utilisé une part de vrai. Dans ce cas, le mélange des genres risque de créer de la confusion malgré tout. En effet, en *scrollant* sur son réseau social, une personne peut voir défiler du vrai ou du faux ou les deux mélangés, sans prendre le temps de faire la part des choses. Démultiplier et décliner à l'infini le fake, à notre sens, c'est donner plus d'armes à la perte de repères informatifs.

141 Lettre ouverte, « Disrupting the Deepfake Supply Chain », *FAccT Conférence*, le 21 février 2024, *op. cit.*
142 Poibeau T., « L'IA générative, un acteur majeur dans une société de la désinformation ? », *The Conversation*, le 7 mars 2024, [en ligne :] https://theconversation.com/lia-generative-un-acteur-majeur-dans-une-societe-de-la-desinformation-225051, consulté le 20 août 2024.

En revanche, ce qui semble faire consensus, c'est la multiplication de deep propagande depuis l'étranger pour créer la confusion et entretenir la méfiance des citoyens d'un pays envers leurs pouvoirs politiques, juridiques ou médiatiques. Chine, Iran et Russie sont particulièrement actifs dans ce domaine. Sans parler des Évangélistes américains (91 millions de pratiquants aux USA[143]), grands défenseurs du créationnisme, d'un Donald Trump messianique, voire de l'Armageddon, que plusieurs millions[144] d'entre eux voient comme le lieu du combat final entre le Bien et le Mal à la fin du monde, lors du retour sur terre de Jésus-Christ en Israël.

5. Un flou artistique

« *En tant que membres du public, nous nous réjouissons des exploits de vrais artistes dans la danse, le cinéma, la magie, la musique, les sports et le théâtre. Si le divertissement diffusé devient saturé de deepfakes, le lien entre le public et les artistes s'érodera, et les deepfakes remplaceront injustement les artistes dont les œuvres ont été utilisées, à l'origine, pour "entraîner" l'IA* »[145]. Le risque est pris très au sérieux par les industries du disque et de la vidéo, qui vont certainement mettre des moyens importants pour défendre leurs droits d'auteurs sur des voix et des styles musicaux.

Et que vont devenir les musiciens quand il suffit de demander à l'IA de créer un morceau ou d'accompagner un air. Suffisamment alimenté en prompts, l'IA pourrait largement les remplacer, sans pour autant se substituer à la créativité ressentie d'un artiste. Aux États-Unis déjà, on tente de légiférer comme avec le « No Fakes Act », visant à créer une loi fédérale pour protéger les acteurs, les musiciens et les autres artistes contre des répliques numériques non autorisées de leur visage ou de leur voix. « *Le projet de loi prévoit une exception pour l'utilisation de copies numériques à des fins de parodie, de satire et de critique. Il exclut également les activités commerciales telles que les publicités, pour autant qu'il s'agisse d'informations, d'un documentaire ou d'une parodie* »[146].

Autre grand problème, le dévoiement pornographique des images de célébrités qui peut être un frein à l'exposition publique des artistes. Nous l'avons vu, la Corée du Sud est au cœur du phénomène. En 2024, « *d'importants labels de K-pop, JYP Entertainment et ADOR, qui gèrent notamment les groupes Twice et NewsJeans, ont annoncé vouloir lancer des actions en justice pour protéger leurs artistes. Mais peu d'affaires aboutissent : entre 2021 et juillet 2024, 793 délits liés aux "deepfakes" ont été signalés, mais seulement seize personnes ont été arrêtées et poursuivies, selon les données de la police* »[147]. Et nous ne parlons ici que de la Corée du Sud.

143 La rédaction d'Evangéliques Info, « 665 millions d'évangéliques dans le monde en 2021 », *Evangéliques Point Info*, le 22 janvier 2021, [en ligne :] https://www.evangeliques.info/2021/01/22/665-millions-d-evangeliques-dans-le-monde-en-2021/, consulté le 17 septembre 2024.

144 En 2022, l'organisation Christians United For Israel (Cufi) comptait 10 millions de membres.

145 Lettre ouverte, « Disrupting the Deepfake Supply Chain », *FAccT Conférence*, le 21 février 2024, *op. cit.*

146 David E., « No Fakes Act wants to protect actors and singers from unauthorized AI replicas », *The Verge*, le 12 octobre 2023, [en ligne :] https://www.theverge.com/2023/10/12/23914915/ai-replicas-likeness-law-no-fakes-copyright, consulté le 10 septembre 2024.

147 La rédaction de l'Essentiel, « Les victimes de "deepfakes" porno se sentent démunies », *l'Essentiel*, le 13 septembre 2024 [en ligne :] https://www.lessentiel.lu/fr/story/en-coree-du-sud-les-victimes-de-deepfakes-porno-se-sentent-demunies-103184210, consulté le 17 septembre 2024.

À l'heure actuelle, il est en effet, difficile d'imaginer des poursuites efficaces contre ce genre de crimes, puisque réalisables depuis l'étranger, sans avoir besoin d'énormes moyens matériels et juridiques. On ne peut imaginer que la censure des grandes plateformes pour freiner leur succès, mais il se trouvera toujours un site pour les accueillir.

6. Pistes de solutions

La piste la plus répandue est celle d'un marquage numérique, une sorte de tatouage des images, indiquant qu'elles ont été fabriquées par une IA[148]. Selon la lettre, « *il est possible pour les caméras de générer des sceaux numériques infalsifiables sur des photographies et des vidéos non modifiées du monde réel, en utilisant des techniques de signature cryptographique similaires aux certificats de sites web et aux identifiants de connexion. S'ils étaient utilisés à grande échelle, ces sceaux permettraient à quiconque d'utiliser des applications d'authentification open source pour vérifier l'authenticité d'une photo ou d'une vidéo correctement signée. Les fabricants d'appareils, les développeurs de logiciels et les sociétés de médias devraient collaborer et populariser ces méthodes d'authentification du contenu ou d'autres méthodes similaires* ». Ajoutant « *les lois actuelles ne ciblent et ne limitent pas de manière adéquate la production et la diffusion de deepfakes, et même les exigences imposées aux créateurs – qui sont souvent mineurs – sont inefficaces. L'ensemble de la chaîne d'approvisionnement des "deepfakes" devrait être tenu pour responsable, comme c'est le cas pour les logiciels malveillants et la pédopornographie* »[149]. Cela dit, le marquage numérique des deepfakes est une solution proposée par les géants du secteur mais il se trouvera toujours un site, quelque part dans le monde, pour contourner ces lois. Dans l'état actuel des choses, la Russie pourrait par exemple offrir un tel service aux utilisateurs européens ou américains pour chercher à y envenimer un peu plus les polarisations politiques et/ou idéologiques. C'est pourtant la seule piste réellement développée par les élites actuellement.

VI. COMMENT ENCADRER LE PHÉNOMÈNE DEEPFAKE ?

1. Côté américain

De manière générale, les grandes plateformes semblent accepter les limites à la liberté d'expression en ce qui concerne les deepfakes images, particulièrement dans les domaines politiques et pornographiques.

5. « *OpenAI, la société à l'origine de l'outil de génération d'images DALL-E, a déjà supprimé le contenu explicite de ses données et filtre les demandes de création d'images de célébrités et d'hommes politiques.*
6. *Un autre modèle d'IA populaire, Midjourney, bloque certains mots-clés et encourage les utilisateurs à signaler les images problématiques aux modérateurs.*
7. *TikTok a également imposé que les contenus manipulés soient étiquetés comme faux ou altérés, et a interdit les "deepfakes" de personnalités privées et de jeunes.*

148 Un marquage aux données d'entraînement serait également imposé, pour savoir si des contenus ont été utilisés pour l'apprentissage de modèles de génération d'images, de textes…

149 Lettre ouverte, « Disrupting the Deepfake Supply Chain », *FAccT Conférence*, le 21 février 2024, *op. cit.*

8. *Twitch a averti que la promotion, la création ou le partage intentionnels de "deepfakes" pornographiques entraîneraient un bannissement immédiat.*
9. *Parallèlement, Meta, OnlyFans et Pornhub ont tous commencé à participer à un nouvel outil "Take It Down" permettant aux adolescents de signaler des images et des vidéos explicites d'eux-mêmes sur l'internet*[150]*. Cependant, les visages de vedettes transposées sur des films pornos sont encore légion sur un site comme Pornhub.*
10. *Google a ajusté ses algorithmes pour minimiser l'affichage de contenus explicites et fallacieux parmi les résultats de recherches, en particulier lorsqu'une recherche mentionne des noms spécifiques associés à des deepfakes. La plateforme a aussi simplifié le processus de suppression de contenus, particulièrement pour les victimes de deepfakes, pour qui un filtrage sera mis en place afin de réduire l'exposition de la victime.*
11. *Sensity dispose de son propre outil d'analyse des pixels et de la structure du fichier pour détecter s'il a été modifié.*
12. *Intel a lancé un détecteur de deepfake en temps réel en 2022 qui inspecte la façon dont la lumière interagit avec les vaisseaux sanguins du visage.*
13. *Les plates-formes Meta's étiquetteront bientôt le contenu généré par l'IA pour leurs utilisateurs* »[151].

Côté politique aussi les choses avancent. Le Sénat américain a, par exemple, adopté à l›unanimité un projet de loi fin juillet 2024, le Disrupt Explicit Forged Images and Non-Consensual Edit Act (DEFIANCE) Act, qui permet aux victimes de deepfakes sexuellement explicites de poursuivre au civil ceux qui ont produit ou traité l'image dans l'intention de la distribuer. La Californie, le Texas, le Wisconsin, l'État de Washington, le Minnesota ou encore le Michigan ont adopté une législation conçue pour lutter contre l'IA lors des élections. Les deepporns font également l'objet de lois et de poursuites judiciaires. Le président Joe Biden, lui-même a publié un ordre exécutif, en octobre 2023, chargeant le Département du Commerce de créer des conseils sur le contenu de l'IA « watermarking »[152] pour indiquer clairement que certaines vidéos deepfake n'ont pas été créées par des humains.

Le 16 février 2024, une vingtaine de grandes entreprises du numérique, parmi lesquelles Google, Meta, OpenAI, Microsoft, Amazon, X, IBM, TikTok, Adobe, Snap, ou encore Stability AI, ont signé un accord permettant « *d'aider à empêcher les contenus trompeurs générés par IA d'interférer dans les élections prévues cette année (2024, ndlr) dans le monde* »[153], dans le cadre du forum sur la sécurité de Munich. Elles s'engagent à développer des outils communs, notamment via des sortes de « tatouage numérique », lisibles par les machines. Un travail déjà

150 El Attilah I., « La condamnation à perpétuité des victimes de "deepfake porn" », *Euronews*, le 30 juin 2023, [en ligne :] https://fr.euronews.com/next/2023/06/30/la-condamnation-a-perpetuite-des-victimes-de-deepfake-porn, consulté le 18 septembre 2024.

151 Duboust O., « Deepfakes are spreading in scams and on social media 'faster than expected', experts warn », *Euronews*, le 13 juin 2024, [en ligne :] https://www.euronews.com/next/2024/06/13/deepfakes-are-spreading-in-scams-and-on-social-media-faster-than-expected-experts-warn, consulté le 15 septembre 2024.

152 Sorte de tatouage numérique.

153 La rédaction du Monde, « Les géants du numérique signent un accord contre l'utilisation trompeuse de l'IA dans le cadre d'élections », *Le Monde*, le 16 février 2024, [en ligne :] https://www.lemonde.fr/pixels/article/2024/02/16/les-geants-du-numerique-signent-un-accord-contre-l-utilisation-trompeuse-de-l-ia-dans-le-cadre-d-elections_6216967_4408996.html, consulté le 15 septembre 2024.

en cours, avec, notamment, le standard C2PA[154] [155], pour Coalition for Content Provenance and Authenticity. Mais les termes du contrat manquent de clarté, notamment sur les réponses à apporter et sur le contexte du deepfake qui peut avoir des vertus éducatives, satiriques, politiques, artistiques ou autres. Les entreprises de la Tech ont également insisté sur le fait que cette lutte ne relevait pas de leur seule responsabilité : « *Nous nous engageons à faire notre part en tant qu'entreprises technologiques, tout en affirmant que l'usage trompeur de l'IA ne représente pas seulement un défi technique, mais un problème politique, social et éthique, et nous espérons que d'autres s'engageront également à agir dans le reste de la société* »[156].

2. La Chine, un parti, un récit

Avec le succès de l'application d'échange de visage Zao[157], devenue n° 1 sur la liste des applications de divertissement gratuites dans l'App Store d'Apple dans les deux jours suivant ses débuts[158], le gouvernement chinois a décidé de prendre des dispositions radicales pour éviter toute dérive. Depuis janvier 2020 il est obligatoire de préciser qu'une vidéo a été créée grâce à l'intelligence artificielle et qu'elle rapporte de fausses informations, pour qu'elle soit publiée de manière légale. Si ces mentions n'apparaissent pas, le créateur de la deepfake sera considéré comme un criminel aux yeux des autorités chinoises et donc traité comme tel.

En Chine, on ne badine pas avec la « fausse propagande », ou du moins avec les informations non officielles, ni même avec la satire politique.

3. L'Europe crée des garde-fous

L'Union européenne a, de son côté, déjà voté son AI Act, début 2024, qui devrait entrer en vigueur en 2025. Le texte impose, entre autres, l'étiquetage des deepfakes (watermarking). Il y a également le code de bonnes pratiques 2022, élaboré par « *les principales plateformes en ligne, les plateformes émergentes et spécialisées, les acteurs du secteur de la publicité, les vérificateurs de faits, les organismes de recherche et les organisations de la société civile … contre la désinformation, conformément aux orientations de la Commission de mai 2021 … Les signataires se sont engagés à prendre des mesures dans plusieurs domaines, tels que: démonétiser la diffusion de la désinformation; garantir la transparence de la publicité à caractère politique; donner aux utilisateurs les moyens d'agir; renforcer la coopération avec les vérificateurs de faits; et offrir aux chercheurs un meilleur accès aux données* »[159].

---

154 La norme C2PA a été fondée en février 2021 par des leaders de l'industrie tels qu'Adobe, ARM, Intel, Microsoft, The New York Times et la BBC. Elle fonctionne en intégrant une métadonnée spécifique, considérée comme une signature numérique, dans les fichiers d'images générées par IA. Cette métadonnée fournit des informations clés sur l'origine de l'image et qu'elle a été produite par une IA.

155 La rédaction du Monde, « Les géants du numérique signent un accord contre l'utilisation trompeuse de l'IA dans le cadre d'élections », *op. cit.*

156 *Ibid.*

157 Voir chapitre sur le Face Swapping.

158 Meng Jing, « La Chine publie de nouvelles règles pour limiter les technologies de contrefaçon utilisées pour créer et diffuser de fausses nouvelles », *South China Morning Post*, 29 novembre 2019, [en ligne :] https://www.scmp.com/tech/apps-social/article/3039978/china-issues-new-rules-clamp-down-deepfake-technologies-used, consulté le 14 mai 2020.

159 Commission Européenne, « The 2022 Code of Practice on Disinformation », Commission Européenne, mai 2021, [en ligne :] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation, consulté le 17 septembre 2024.

4. Côté belge

Dans le cas de deepfakes politiques, selon Sandrine Carneroli, avocate spécialiste en droit des médias à Bruxelles, « *si l'intention de son auteur est «de manipuler et de travestir la vérité politique, économique, sociale»*, ce dernier risque des sanctions. «*On peut faire valoir l'atteinte à la vie privée de la personne qui est présentée en disant des choses qui ne correspondent pas à ce qu'elle dirait en temps normal. On arrive dans ce qu'on appelle le délit de presse. On n'est plus sur la voie pénale parce qu'on va devoir analyser les propos* »[160]. Le contexte d'un deepfake reste donc au cœur de tout débat.

Dans le cas de *deepporn*, elle ajoute qu'« *On peut demander que le tribunal donne une injonction à la plateforme pour identifier l'auteur du contenu préjudiciable. Puisque la plateforme connaît les identifiants, a accès à un compte bancaire, etc. La loi permet également de faire une action en urgence, un référé, pour faire supprimer le contenu dans un délai de 6 heures, non seulement à l'auteur, s'il est connu, mais également au diffuseur, qui s'expose à une amende. Autre point intéressant souligné par l'avocate, le RGPD et "la loi belge d'application protègent les données nominatives, et l'image d'une personne est une donnée nominative" Une plainte pourrait ainsi être déposée auprès de l'Autorité de protection des données* ».

Depuis 2020, la loi sur le *revenge porn*, punit également la « diffusion non consensuelle d'images à caractère sexuel ».

5. Attention, tous les fakes ne sont pas égaux

Les deepfakes audio ont cet « avantage » sur les images, qu'ils sont plus difficiles à détecter, meilleurs marchés et plus facile à réaliser. On peut ainsi faire dire un mensonge à un présentateur de journal parlé ou à une personnalité politique en pleine campagne électorale.

Une vraie gageure pour les plateformes qui vont devoir les repérer pour se conformer, par exemple, au Code de pratique de l'UE en matière de désinformation. Renforcé en juin 2022, il interdit les Deepfakes et enjoint les plates-formes à utiliser leurs outils (modération, déplateformisation[161]…) pour s'en assurer.

Ces faux sonores, générées par l'IA, pourraient-ils être carrément mentionnés par une autre IA en réponse à une recherche de citoyen ? Pour l'éviter, il faudrait un marquage numérique à la production des faux sonores également et là-dessus, il n'y a pas encore d'information.

Ces contrôles pourraient être faits, dans une certaine mesure, par des journalistes spécialisés dans le *factchecking*. Mais les journalistes eux-mêmes sont inquiets des évolutions fulgurantes des deepfakes et font notamment appel, via Reporters sans Frontières, au corps juridique, pour la protection des journalistes, afin qu'ils créent un « délit de deepfake » capable de dissuader les manipulateurs[162]. De plus, les journalistes ne pourront vérifier que des infos d'ac-

160 Jacquet T., « Deepfakes pornographiques, politiques, économiques : quelles sont les sanctions prévues par le droit belge contre ces pratiques ? », *op. cit.*

161 Suspension ou bannissement d'une plateforme.

162 Grimonpont A., « Deepfakes : quatre solutions pour éviter la violation du droit à l'information », *RSF*, le 29 février 2024, [en ligne :] https://rsf.org/fr/deepfakes-quatre-solutions-pour-%C3%A9viter-la-violation-du-droit-%C3%A0-l-information, consulté le 18 septembre 2024.

tualités utiles mais il ne sera pas de leur ressort de vérifier les arnaques, utilisant une fausse voix, qui pullulent sur les réseaux. Les plateformes relayant ces faux ont, ici encore, clairement une responsabilité et un rôle à jouer dans leur filtrage.

6. Des outils d'aide se mettent en place

Par ailleurs, des outils en ligne comme Deepware Scanner, GPTZero, SynthID ou Copyleaks AI Détecteur de contenu, peuvent vous aider à repérer le contenu généré par l'IA – qu'il s'agisse de texte, d'images ou de vidéos [163].

Il est certain que d'autres outils de ce type vont se développer. L'un ou l'autre pourrait même supplanter les autres par son universalité et son efficacité à repérer les deepfakes. Mais il y a fort à parier que celui-ci devienne payant et n'aide que les plus nantis. Il devrait pourtant être un outil de service public.

CONCLUSION

Les deepfakes peuvent donc théoriquement avoir des conséquences diverses et entraîner diffamation, violation de propriété intellectuelle, violation de la vie privée et de la protection des données, harcèlement, fraude, chantage, violation de droits de publicitaires, interférence électorale ou encore incitation à la violence et aux troubles sociaux et civils.

Et, contrairement à la prise en main d'un outil comme Photoshop, plutôt complexe, les outils d'IA comme Dall-e et Midjourney, pour fabriquer une image, ou Sora, pour la vidéo, permettent de générer des contenus réalistes à partir de quelques mots clés. Il est donc urgent de porter à la connaissance du plus grand nombre, non seulement leur existence, mais aussi et surtout de leurs possibles et néfastes applications.

L'une des plus choquante est la réalisation de deep nudes, voire de deep porns, au sein d'écoles pour humilier et/ou racketter une fille, et plus rarement un garçon, de son entourage. Selon une enquête réalisée par l'Université d'Anvers, 7 % des Belges entre quinze et vingt-cinq ans ont déjà tenté d'en réaliser un. Si on rapporte ce chiffre à la pyramide des âges de Statbel [164], cela représente près de 100 000 jeunes qui ont créé un deep sexuel. Alors que beaucoup de parents et de professeurs n'ont pas encore vraiment entendu parler du phénomène qui explose, et pas seulement dans les écoles. Peu sont à l'abri car tant qu'une photo de quelqu'un est en ligne, sur un réseau social par exemple, il peut faire l'objet d'un deepfake. Et les gens en mettent par dizaines en ligne. Il est particulièrement urgent de porter ces perfectionnements abjects du harcèlement à la connaissance de tous et de sévir rapidement pour éviter une prolifération accompagnée d'un sentiment d'impunité, particulièrement chez les mineurs. Car aujourd'hui les poursuites restent rares. Le plus souvent, les victimes ne sont pas au courant de l'infraction, ou elles n'ont pas le courage ou les moyens de se confronter aux créateurs de ces vidéos. Certains mineurs ne distinguent même pas spécialement la

163 Carbonaro G., « The AI detector tools that can help you check content for plagiarism, fakes and scams », *Euronews*, le 24 janvier 2024, [en ligne :] https://www.euronews.com/next/2024/01/24/plagiarism-fakes-and-copycats-these-are-the-best-tools-to-spot-ai-generated-content, consulté le 18 septembre 2024.

164 Statistiques Statbel, « Pyramide des âges », *Statbel*, 2024, [en ligne :] https://statbel.fgov.be/fr/figures/pyramide-des-ages, consulté le 20 septembre 2024.

différence entre ajouter des oreilles de lapin à la photo d'une fille de leur école et utiliser son visage pour en faire un deep nude. Pour eux, c'est juste marrant. Beaucoup d'hommes voient même les deep nudes comme une forme de satire. Il y a donc, en parallèle, un gros travail de conscientisation qui doit être fait pour que ces personnes prennent la mesure de la portée de leurs actes.

L'autre grand danger est clairement le perfectionnement des arnaques en ligne. Pourra-t-on encore croire ce qu'on voit et entend sur le net, sachant qu'un deepfake est toujours possible ? Comment sera-t-il possible de communiquer en toute confiance désormais avec son patron ou sa famille ? Via des VPN[165] ? Oui mais les meilleurs sont payants et les personnes ayant le moins de moyens risquent d'être plus exposés que les autres, alors qu'elles sont déjà fragilisées numériquement. L'arrivée de l'IA risque de complexifier encore plus les choses. Quelle que soit la réponse, l'État a le devoir de protéger ses citoyens mais il ne s'en donne pas assez les moyens. Rien qu'en ce qui concerne les cyber-attaques, « *plus de 60 % des professionnels européens de la cybersécurité déclarent que l'équipe de cybersécurité de leur organisation manque de personnel, et plus de la moitié (52 %) pensent que le budget de cybersécurité de leur organisation est insuffisant* ». L'arrivée de l'IA a clairement boosté ces chiffres et le nombre d'appareils connectés se multiplient pour donner toujours plus de données, exploitables par des escrocs du monde entier. Et désormais les deepfakes facilitent les vols d'identité.

Il est certain qu'une concertation internationale devient essentielle pour partager les meilleures pratiques, légiférer et condamner dans le plus de pays possible. Mais il est difficile de croire que tous y participeront. D'ailleurs, si l'influence des deepfakes dans un processus électoral est encore rare et est rapidement dénoncée côté occidental, des pays moins démocratiques comme l'Inde, la Turquie ou le Vénézuela ont montré que l'outil permettait aux pouvoirs en place d'asseoir leur propagande et toucher les citoyens presque directement. C'est un énième outil de propagande, mais il peut être particulièrement efficace sur les populations non averties, en l'absence d'une presse indépendante notamment. Et l'IA ne fera que le perfectionner encore et encore. C'est là l'un des points d'attention pour les deepfakes politiques, punir les producteurs et responsabiliser les diffuseurs et les consommateurs. La justice aura également son rôle à jouer. En Europe, elle bénéficie de lois pour défendre les citoyens face à la désinformation, aux dérives de l'IA ou encore au vol de données. Et ce sera à la justice de déterminer la pertinence ou non d'une satire, du moins si on lui en donne les moyens.

Mais les pouvoirs publics ne doivent pas oublier qu'une des bases du problème est que la consommation de fausses informations est souvent motivée par des sentiments d'opposition envers les institutions et les corps sociaux établis, perçus comme ayant failli dans leur mission. La crise du Covid-19 en a fourni une illustration récente, avec l'émergence rapide de figures très médiatisées, en opposition frontale et systématique avec les mesures proposées, et très soutenues par leurs supporters sur les médias sociaux. En 2023, la campagne présidentielle argentine a été 'pimentée' par quelques deepfakes, comme celui de l'ex-président Massa en train de sniffer de la cocaïne. À la suite de quoi Natalia Zuazo, spécialiste argentine de la politique et des technologies, disait en interview : « *Ces contenus sont bon marché à produire et apportent de l'originalité à la campagne électorale, qui, d'ordinaire, en manque … La fake news n'a pas forcément besoin d'être convaincante pour qu'on la croit. Elle doit juste valider une croyance*

[165] Virtual Private Network pour Réseau Privé Virtuel, permettant une navigation privée.

préexistante chez ce public qui brûle de voir quelque chose d'horrible chez le candidat qu'il rejette » [166]. La diffusion de fausses informations créée ainsi un sentiment d'appartenance et de solidarité au sein de groupes qui s'opposent au pouvoir en place. Mais ces petits jeux peuvent aller loin. Verra-t-on un jour une campagne fondée sur des batailles de deepfakes en lieu et place d'arguments politiques ? Les deepfakes sont un élément de show supplémentaire aux élections polarisées auxquelles nous assistons de plus en plus de par le monde, quitte à flirter avec la calomnie et la diffamation.

Reste que le public devra être mis au courant des nouvelles technologies et de leurs opportunités d'utilisation trompeuse. Mais à l'heure actuelle, cela demande quasi une démarche proactive de la part des citoyens. Ils doivent trouver une formation ou un EPN, dont beaucoup ne connaissent même pas encore l'existence, monter dans le TGV du numérique et se protéger contre des escroqueries boostées à l'IA et pensées par des professionnels. C'est absurde, sachant qu'il y a déjà un gros travail pour réduire la fracture numérique et susciter l'intérêt pour la chose numérique, à l'heure où nombre de gens, toute classe sociale confondue, ne savent pas ce qu'est un cookie et s'en contrefichent.

> *Il faut toujours connaître les limites du possible. Pas pour s'arrêter, mais pour entreprendre l'impossible dans les meilleures conditions.*
>
> **Romain Gary, *Charge d'âme*, 1977**

**

Philippe Courteille est licencié en journalisme et communication de l'ULB. Il a travaillé comme journaliste-réalisateur freelance pour de nombreuses émissions de télévision pendant 25 ans. Il est aujourd'hui responsable de la thématique Médias & Actions citoyennes chez Citoyenneté & Participation.

166 La rédaction du Courrier International, « Élections. "Deepfake", "cheapfake" : l'IA au service de la campagne présidentielle argentine », Courrier International, le 17 novembre 2023, [en ligne :] https://www.courrierinternational.com/article/elections-deepfake-cheapfake-l-ia-au-service-de-la-campagne-presidentielle-argentine, consulté le 20 septembre 2024.

COURTEILLE Philippe, *Deepfakes, le mensonge à l'ère de l'intelligence artificielle*, Bruxelles : Citoyenneté & Participation, Étude n°50, 2024, [en ligne :] http://www.cpcp.be/publications/e50-deepfakes.

Désireux d'en savoir plus !

Animation, conférence, table ronde... n'hésitez pas à nous contacter,
Nous sommes à votre service pour organiser des activités sur cette thématique.

www.cpcp.be

FÉDÉRATION
WALLONIE-BRUXELLES

Avec le soutien du Ministère de la Fédération Wallonie-Bruxelles

CPCP ASBL – Avenue des Arts 50/bte 6 – 1000 Bruxelles | info@cpcp.be | www.cpcp.be
N° d'entreprise : 0409 117 690 | RPM Bruxelles | BE67 3101 6586 0487

Citoyenneté & Participation

Avenue des Arts, 50/6 – 1000 Bruxelles
02 318 44 33 | **info@cpcp.be**
www.cpcp.be | www.facebook.com/CPCPasbl

Toutes nos publications sont disponibles en téléchargement libre :
www.cpcp.be/publications/

E.R. : John Lewis – Avenue des Arts; 50/6 – 1000 Bruxelles