Ce mot-valise, fusion de *deep learning* (apprentissage profond, technique d'intelligence artificielle) et *fake* (faux), désigne un contenu truqué et sciemment trompeur. Un *deepfake* est un faux, quelle que soit la nature du contenu – vidéo, photo, audio ou texte –, conçu grâce à l'intelligence artificielle. Aux États-Unis, la manipulation malveillante d'images vidéo affecte déjà la sphère politique.

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. Le réalisme des images recomposées artificiellement fait d'un *deepfake* un instrument potentiellement efficace pour manipuler l'opinion. Les progrès de l'intelligence artificielle promettent la reproduction de plus en plus précise des caractéristiques propres aux humains. Dans un avenir proche, il est donc envisageable qu'un appel vidéo sur internet soit parfaitement manipulé par un *deepfake*.

Deepfakes est également le nom du développeur qui le premier a posté, fin 2017, des vidéos truquées à caractère pornographique sur le forum Reddit. Pour contrefaire ces vidéos avec le visage d'actrices célèbres, des logiciels, comme Deepfake et FakeApp, étaient disponibles gratuitement sur internet dès le début de l'année 2018. Développés à partir de Tensor Flow, outil d'apprentissage machine open source de Google, ces logiciels, capables de composer l'image 3D d'un visage à partir de milliers de photos récupérées sur internet, ont permis la propagation rapide des premiers *deepfakes* appliqués à la pornographie. La plupart des sites concernés comme YouPorn ou Pornhub, les forums Reddit ou Discord, le site de partage de vidéo Gfycat et Twitter, n'ont pas tardé à réagir, entre la fin janvier et le début février 2018, en interdisant ces *deepfakes* pornographiques. En octobre 2019, l'entreprise néerlandaise Deeptrace, spécialisée dans les risques en ligne, comptabilisait 15 000 vidéos *fake porn* en circulation sur internet au cours des sept derniers mois.

Le degré de perfection atteint est tel en matière de *fake*, qu'il convient de distinguer les *deepfakes*, qui sont le fruit de l'intelligence artificielle, des autres *fakes* qui ne le sont pas. Baptisées *cheapfakes* ou *shallow fakes*, les vidéos qui relèvent de cette dernière catégorie sont peu retravaillées, le truquage consistant simplement à ralentir les images d'une séquence. Les *cheapfakes* ne sont pas pour autant inoffensifs. Aux États-Unis, ils ont servi avec succès à attaquer un adversaire politique. En mai 2019, une vidéo a circulé sur internet, notamment relayée par le Parti républicain, l'Alt-right et même par le président Donald Trump, montrant Nancy Pelosi « comme ivre » lors d'un discours. L'enquête menée par le *Washington Post* a révélé que le débit de paroles de la présidente (démocrate) de la Chambre des représentants avait été sciemment très ralenti sur la vidéo afin de faire croire que l'intéressée souffrait d'une addiction à l'alcool. Vu plus de deux millions de fois, ce *cheapfake* a été retiré sur Youtube mais maintenu accessible sur Twitter. Facebook l'a simplement déclassé, en y ajoutant un avertissement à l'attention de ses membres, répondant aux critiques que *« nous n'avons pas de règles stipulant que l'information publiée sur Facebook doit forcément être vraie ».* Quelques mois auparavant, en novembre 2018, un autre *cheapfake*, montrant une soi-disant dispute entre une collaboratrice du

président Trump et un journaliste de CNN, a permis l'éviction de ce dernier des conférences de presse à la Maison Blanche.

À l'origine des *deepfakes*, il y a une technique inventée en 2014 par le chercheur Ian Goodfellow : les GAN (*Generative Adversarial Networks* ou réseaux antagonistes génératifs). Selon cette technologie, deux algorithmes s'entraînent mutuellement, l'un œuvrant à fabriquer une image tandis que l'autre cherche, en parallèle, à déterminer si cette image est fausse, entraînant de cette façon le premier à faire mieux. Cette technologie a permis d'engendrer « le portrait d'Edmond de Belamy », tableau conçu par une intelligence artificielle et vendu aux enchères chez Christie's en octobre 2018. Les images créées par les GAN ont atteint un niveau de crédibilité très satisfaisant, la frontière entre le réel et le virtuel n'étant plus toujours perceptible. En 2018, le spécialiste américain des processeurs graphiques Nvidia a présenté un programme, fondé sur la technologie des GAN, capable d'inventer des visages au réalisme sidérant. Les préventions contre le logiciel de retouche d'images Photoshop s'en trouvent très largement dépassées… En effet, les résultats du jeu *Which Face Is Real ?*, qui consiste à identifier un vrai visage d'un faux à partir d'une paire de photos, 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.

Dès avril 2018, le site Buzzfeed entend dénoncer la gravité de cette pratique du faux en réalisant une vidéo à partir du logiciel FakeApp, un *deepfake* qui fait dire à l'ancien président Barack Obama – avec la participation du comédien Jordan Peele – tout le mal qu'il pense de son successeur à la Maison Blanche. En juin 2019, c'est une démarche plus incisive qui guide les deux artistes britanniques Bill Posters et Daniel Howe, en collaboration avec l'agence de communication Canny AI, lorsqu'ils postent sur Instagram une vidéo *deepfake* faisant dire à Mark Zuckerberg que *« quiconque contrôle les données contrôle l'avenir ».* Avec leur vidéo clairement estampillée #deepfake, les artistes invitent à imaginer *« la prochaine étape de notre évolution numérique où chacun pourrait éventuellement avoir une copie numérique, un humain universel éternel. Cela changera notre façon de partager et de raconter des histoires, de nous souvenir de nos proches et de créer du contenu. »* En toute logique, ce *deepfake* du patron n'a pas été supprimé sur le réseau social.

Un faux pourrait également servir à mieux communiquer. Une vidéo truquée d'un discours du président américain Donald Trump, qui s'enorgueillit d'avoir éradiqué le sida dans le monde, a été utilisée comme instrument de campagne par l'association Solidarité Sida à l'occasion de la sixième conférence de reconstitution des ressources du Fonds mondial de lutte contre le sida, la tuberculose et le paludisme en octobre 2019 à Lyon. Totalisant plus de 3,5 millions de vues quelques heures seulement après son lancement, le 6 octobre 2019, ce *deepfake* de cinquante secondes a rempli son objectif selon Eric Tong Cuong, fondateur de La Chose, agence de publicité qui l'a réalisé : *« Être efficace »* au regard de la génération internet pour laquelle le sida est *« un non-sujet ».* 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 ». 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 »*.

Les outils de manipulation de l'audio et de la vidéo, pour faire dire n'importe quoi à n'importe qui, se perfectionnent et se multiplient. Des chercheurs de l'université de Washington ont mis au point un algorithme basé sur la technologie VDR (remplacement de dialogue vidéo), qui permet la synchronisation trompeuse d'une voix sur des images. À Stanford, le programme Face2Face s'appuie sur la technique Facial Reenactment pour recomposer des expressions faciales en temps réel. Du côté du géant coréen de l'électronique Samsung, les scientifiques spécialistes en intelligence artificielle ont créé, à partir d'une seule photo, une copie humaine de synthèse douée de mouvements et de la parole. En Europe, des chercheurs allemands travaillent à un logiciel de « marionnettisation » servant à calquer sur le visage d'une personne des expressions et des paroles factices, ce qui permet d'imaginer un jour une conférence de presse en direct mais fausse et totalement détournée. La propagation des outils permettant de fabriquer des *deepfakes* serait donc une menace pour tous. En août 2019, le *Wall Street Journal* relatait qu'une entreprise avait été victime d'une tentative d'extorsion de fonds par des escrocs ayant utilisé une voix artificielle, imitant celle du PDG. En Belgique, un *deepfake*, qui montre le président Trump inciter le pays à sortir de l'Accord de Paris sur le climat, a été utilisé par le Parti socialiste flamand mais le message avertissant de la supercherie n'a pas été entendu par de nombreux internautes.

Vraisemblables quoique fausses, ces vidéos de personnes connues dans des situations fictives viennent grossir le flux quotidien de la désinformation, entendue comme la communication volontaire d'une information fausse dans le but de nuire, selon la définition proposée par l'OCDE (voir *La rem* n°45, p.62). Avec les *deepfakes*, l'information est de plus plus relativisée et la vérité est devenue de plus en plus précaire. Centres de recherche universitaires, start-up, gouvernements et médias se mobilisent autour d'un même objectif : identifier les *deepfakes*, mais avec la même crainte de se lancer dans un *« jeu du chat et de la souris »*, comme l'exprime Francesco Marconi, responsable R&D du *Wall Street Journal*. Cette bataille se présente en effet comme celle que se livrent, sans fin, les pirates informatiques et les professionnels de la cyber-sécurité, les trouvailles des uns servant de tests de fiabilité aux autres. On sait déjà que le niveau de crédibilité de certains *deepfakes* est testé par leurs auteurs grâce à la technologie des GAN.

L'agence de la recherche du ministère américain de la défense, la Darpa, finance depuis 2016 à travers un programme baptisé MediFlor, pour Media Forensics, plusieurs projets pour *« vérifier l'authenticité et établir l'intégrité des médias visuels »*, associant des organismes américains à d'autres équipes de recherche à l'étranger, notamment en Europe. Pour combattre l'invasion des *deepfakes*, les pistes empruntées par les

chercheurs sont diverses. Pour certains, la priorité serait de constituer une immense base de données d'images afin de pouvoir comparer les images originales avec celles qui sont truquées. D'autres chercheurs s'attellent plutôt à entraîner leurs algorithmes à repérer les truquages par le biais de leurs défauts ; alors qu'une autre approche comportementale est envisagée par ceux qui ne croient pas à la pérennité des solutions techniques face à la sagacité des pirates. Selon Walter Quattrociocchi, directeur du laboratoire de science de la donnée et de la complexité à l'université Ca'Foscari de Venise, la solution passerait par l'identification des sujets qui suscitent la polémique sur internet, car *« ce sont souvent les plus populaires et ceux qui suscitent le plus d'engagement, mais aussi ceux qui attireront le plus de* deepfakes », le but étant de pouvoir alerter les internautes engagés dans ces discussions.

Souvent décriés pour ne pas lutter assez activement contre la désinformation, les géants internet ont annoncé mettre à la disposition des chercheurs une grande quantité de contenus contrefaits afin d'entraîner leurs algorithmes à les détecter. En septembre 2019, Google a rendu publiques plus de 3 000 vidéos vraies et fausses réalisées avec des acteurs issues de sa *DeepFake Detection Dataset*. Par ailleurs, les groupes américains de la tech et des institutions universitaires ont lancé un *« deepfake challenge »* en septembre 2019. Financée à hauteur de 10 millions de dollars par Facebook, cette initiative, qui vise à trouver des solutions *anti-deepfakes*, regroupe notamment les géants Apple, Amazon, Microsoft, IBM, ainsi que le MIT et l'université d'Oxford.

Les médias sont également des acteurs majeurs dans la détection des faux contenus. *« La sensibilisation de l'ensemble de la rédaction à la question des* deepfakes *est cruciale,* déclare Francesco Marconi du *Wall Street Journal. Les processus et les normes du journalisme ne changent pas : bien qu'il s'agisse d'une technologie de pointe, le fondement du métier ne change pas – vérifier l'origine et la fiabilité des sources, faire des recherches sur le contexte, comparer des informations, etc. »*. Depuis 2018, le *Wall Street Journal* dispose d'un Media Forensics Committee composé de 21 journalistes et rédacteurs en chef de tous les services (rédaction, photo, vidéo, produit, R&D, audience et analyse, normes et éthique). *« Chacun d'entre eux est de garde pour répondre aux questions des journalistes sur la manipulation d'un élément de contenu,* explique Francesco Marconi. *Après chaque question d'un journaliste et l'analyse subséquente, les membres rédigent un rapport contenant les détails de ce qu'ils ont appris. »* Si la mobilisation des grands médias américains est générale, de l'agence Reuters au *New York Times* en passant par le *Washington Post*, elle marque néanmoins la fragilité des plus petites entreprises de presse qui ne disposent pas des mêmes moyens.

En Europe, dans le cadre d'un projet de recherche soutenu par l'Union européenne, dix pays, dont la France à travers l'AFP, ont mis au point, entre 2016 et 2018, la plateforme InVID (In Video Veritas – Vérification du contenu vidéo des réseaux sociaux pour l'industrie de l'information), afin d'aider les journalistes à détecter les vidéos truquées. Un module installé sur n'importe quel navigateur internet leur permet de savoir si la vidéo a déjà été diffusée sur le web dans un autre contexte et si elle a fait l'objet d'une manipulation technique. Un autre projet de recherche en cours, baptisé WeVerify, comporte la

création d'une base de données de faux connus.

Avec les élections américaines de 2020 en ligne de mire, le FBI et des représentants du gouvernement ont réuni, en septembre 2019, les équipes responsables de la sécurité de Facebook, Google, Microsoft et Twitter. Les vidéos truquées sont une nouvelle menace qui viendrait amplifier les campagnes de désinformation sur les réseaux sociaux. Le procureur Robert Mueller, après deux années d'enquête sur une possible collusion entre Moscou et l'équipe de campagne de Donald Trump, a signalé un potentiel risque d'interférence en 2020. *« La Russie a sans doute été enhardie à recommencer en 2020, étant donné le peu de réactions des États-Unis face aux révélations de 2016 »*, estime Alex Stamos, professeur à Stanford et ancien responsable de la sécurité chez Facebook. Le Parti démocrate a d'ailleurs demandé expressément aux équipes de campagne des candidats à la primaire de ne pas « jouer » avec FaceApp, la très populaire application russe de retouche photographique pour se voir vieilli. Selon Maurice Turner, spécialiste de la sécurité électorale, le problème reste complexe car une vidéo, même identifiée comme fausse, peut *« renforcer une opinion chez ceux qui veulent y croire, et détourner l'attention des informations réelles »*.

À propos du *cheapfake* s'attaquant à Nancy Pelosi, largement répandu sur les réseaux sociaux alors qu'il était pourtant facile de ne pas se laisser duper, Francesco Marconi explique *« qu'il n'en faut pas beaucoup pour tromper certains internautes. Ce sont les effets de ce que les experts appellent le "biais de confirmation" : lorsqu'une vidéo semble prouver quelque chose à laquelle ils croient déjà, les internautes penseront plus volontiers que la vidéo est réelle et la partageront. »*

En janvier 2019, la chaîne de télévision locale Q13 à Seattle a diffusé un *deepfake* imitant Donald Trump au cours d'un discours tenu par le président des États-Unis à peine quelques minutes auparavant. À la perspective de voir éclore des *deepfakes* en temps réel ou en streaming, Francesco Marconi répond : *« Avec l'informatique quantique et l'expansion de la 5G, nous arriverons sans aucun doute à un point où les simulations seront très proches de la réalité. »*

Sources :

- « L'AFP partenaire du projet européen InVID sur la vérification des vidéos sur le web », AFP, afp.com, 1er février 2016.
- « Du porno aux fausses informations, l'intelligence artificielle manipule désormais la vidéo », Morgane Tual, lemonde.fr, 4 février 2018.
- « Comment lutter contre les "deepfakes" », Jacques Henno, *Les Echos*, 29 janvier 2019.
- « "Deepfake" : une vidéo trafiquée de Nancy Pelosi relayée par des proches de Trump », Harold Grand, lefigaro.fr, 24 mai 2019.
- « Une vidéo "deepfake" de Mark Zuckerberg met à l'épreuve la modération de Facebook », Harold Grand, lefigaro.fr, 12 juin 2019.
- « Are You For Real ? », Tom Simonite, Wired, July-August 2019.
- « Un "deepfake challenge" pour lutter contre la désinformation », AFP, tv5monde.com, 6 septembre 2019.
- « La détection des "deepfakes", une course contre la montre », AFP, tv5monde.com, 11 septembre 2019.
- « Les élections américaines de 2020 menacées par les fausses vidéos et attaques informatiques », Rob Lever, AFP, tv5monde.com, 22 septembre 2019.
- « Pornographie et politique au cœur des "deepfakes" selon une étude », AFP, tv5monde.com, 8 octobre 2019.
- « Fausse vidéo de Trump : pourquoi Solidarité Sida a sorti un "deepfake" pour sa campagne », Morgane Tual, lemonde.fr, 9 octobre 2019.
- « WSJ on DeepFakes: "It's a cat & mouse game". WSJ's R&D Chief, Francesco Marconi, shares lessons learned », Ana Lomtadze, GEN, Global Editors Network, medium.com, Oct 10, 2019.
- « Dépister les deepfakes : le jeu du chat et de la souris », Isabelle Bellin, DAP, dataanalyticspost.com, 15 octobre 2019.