



HAL
open science

Les deepfakes : une nouvelle menace pour le journalisme ?

Albane Guichard

► **To cite this version:**

Albane Guichard. Les deepfakes : une nouvelle menace pour le journalisme ?. Sciences de l'information et de la communication. 2020. dumas-03272244

HAL Id: dumas-03272244

<https://dumas.ccsd.cnrs.fr/dumas-03272244>

Submitted on 28 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Mémoire de Master 2

Mention : Information et communication

Spécialité : Journalisme

Option : Journalisme et innovation

Les deepfakes

Une nouvelle menace pour le journalisme ?

Responsable de la mention information et communication
Professeure Karine Berthelot-Guiet

Tuteur universitaire : Valérie Jeanne-Perrier

Nom, prénom : GUICHARD Albane

Promotion : 2019-2020

Soutenu le : 06/11/2020

Mention du mémoire : Très bien

Remerciements

Je tiens à remercier Valérie Jeanne-Perrier et Tristan Mendès-France pour leurs conseils avisés et leur encadrement tout au long de l'avancée de ce mémoire et pendant ma formation.

Je remercie aussi Gerald Holubowicz pour son expertise et son regard journalistique sur le sujet. Un grand merci également à Yanis (alias French Faker) d'avoir accepté de répondre à mes questions et d'avoir eu la gentillesse de réaliser un deepfake pour illustrer ce mémoire.

Enfin, merci à mes camarades de classe du CELSA, mes collègues de *Business Insider France* et mes proches pour leurs encouragements et leur soutien tout au long de cette dernière année de Master de journalisme en alternance.

SOMMAIRE

- **Remerciements** ----- p. 2
- **Sommaire** ----- p. 3
- **Introduction** ----- p. 4
- I. **D'un OVNI à un objet journalistique** ----- p. 7
 - Origines et fonctionnement des deepfakes
 - La vague de deepfakes pornographiques attire l'attention des médias
 - Les deepfakes : arme de désinformation par excellence
- II. **Les deepfakes face à l'information, comment les combattre ?** - p. 18
 - Une menace grandissante dans un contexte propice à la désinformation
 - Les réseaux sociaux et la technologie à l'assaut des deepfakes
 - Le rôle crucial des journalistes et de l'éducation aux médias
- **Conclusion** ----- p. 27
- **Bibliographie et sources** ----- p. 29
- **Annexes (entretiens)** ----- p. 32
- **Résumé et mots-clés** ----- p. 50

Introduction

Le proverbe populaire "*Il ne faut pas toujours croire ce que l'on voit*" — que l'on attribue au peintre du XVIII^e siècle Carmontelle — n'a jamais eu autant de sens qu'aujourd'hui. Photoshop, filtres modifiant les visages sur Snapchat et Instagram, réalité virtuelle et augmentée... au fil des avancées technologiques, la frontière entre perception et réalité se fait de plus en plus floue. Un phénomène qui n'est pas sans conséquence pour l'information.

Sur les réseaux sociaux se mélangent désormais articles de presse et publications d'anonymes, si bien qu'une partie grandissante de la population mondiale ne fait plus le distinguo entre des faits avérés et vérifiés par des journalistes dont c'est le métier, et des analyses personnelles sans fondement, avancées par simple conviction. Certains acteurs malintentionnés profitent des mécanismes des réseaux sociaux, qui exploitent nos biais cognitifs, et de la méfiance grandissante envers les médias pour manipuler l'information, et avec elle l'opinion publique.

Dans cette ère des fake news, le président de la première puissance économique mondiale peut inventer des chiffres et contredire ceux journalistiquement et scientifiquement prouvés, sans que ses supporters ne remettent ses dires en question. Une publication Facebook relayant une théorie du complot, ou une image truquée sur Photoshop, peuvent enflammer la toile en quelques minutes et répandre leur fumée de fausses informations pendant des mois, voire des années. Pour les médias, il est souvent impossible d'éteindre l'incendie à temps et les journalistes se retrouvent à devoir vérifier et expliquer fake news après fake news. C'est dans ce contexte déjà explosif qu'est apparue une nouvelle arme de désinformation : les deepfakes.

Le dictionnaire américain Merriam Webster définit un deepfake comme "*une image ou un enregistrement qui a été modifié et manipulé de manière convaincante pour donner l'impression que quelqu'un a fait ou dit quelque chose qui n'a pas réellement été fait ou dit.*"¹

deepfake noun

 Save Word

deep·fake | \ 'dēp-,fāk  \

plural deepfakes

Definition of deepfake

: an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said

¹ Merriam Webster (s.d.) — Deepfake. Dans *Merriam Webster Dictionary*.

L'étymologie de ce mot-valise ne laisse pas de doute sur sa signification. "deep", qui signifie "profond" en anglais, fait référence au "deep learning" ("apprentissage profond" en français), une méthode poussée d'apprentissage automatique des algorithmes. "fake" désigne lui bien évidemment quelque chose de "faux", au sens d'imitation qui a la volonté de tromper (à distinguer de "false" qui signifie faux au sens d'erroné), comme dans le terme "fake news". Le mot "deepfake" allie donc la technologie à la tromperie, l'intelligence artificielle au mensonge. Les Canadiens francophones emploient d'ailleurs eux le terme d'"hypertrucage"², qui souligne bien cette notion de supercherie.

En France, on utilise très largement l'anglicisme "deepfake", même si le Journal Officiel de la République Française a publié un équivalent français pour les textes de loi. Sous l'entrée "*deep fake, deepfake*", on trouve ainsi la traduction "*infox vidéo, vidéotox*" et la définition : "*Infox qui se présente sous la forme d'une vidéo falsifiée grâce aux techniques de l'intelligence artificielle, en particulier à celles de l'apprentissage profond.*"³ Une note précise que "*La production d'infox vidéo fait notamment appel à l'analyse de l'expression faciale, à la synthèse vocale et à la synchronisation labiale.*"

L'anglais transformant souvent des adjectifs en noms, le terme "deepfake" est aujourd'hui surtout employé pour désigner une vidéo deepfake, et non pas la technologie en elle-même, bien qu'il existe également des deepfakes de voix, et même de texte. Pour des raisons de compréhension générale, j'utiliserai dans ce mémoire le terme "deepfake" dans son sens courant de "vidéo deepfake", et ferai la distinction lorsque j'évoquerai plutôt la technologie elle-même ou d'autres formats de deepfakes.

Avant d'entrer dans le vif du sujet — les deepfakes face à l'information —, il convient de différencier les deepfakes des "cheapfakes". Ces derniers sont également des vidéos ou des enregistrements audio manipulés, mais contrairement aux deepfakes, ils n'ont pas été créés à l'aide d'un algorithme d'intelligence artificielle. Les "cheapfakes" sont réalisés avec des logiciels accessibles à tous et ne nécessitent pas de compétences informatiques particulières.⁴ Les vidéos sont modifiées avec des techniques de montage plus simples et traditionnelles, comme l'accélération ou le ralentissement de la vitesse ou encore le découpage en séquences. Les cheapfakes sont donc moins réalistes et plus grossiers que les deepfakes, d'où leur nom (en anglais, "cheap" signifie peu cher, de mauvaise qualité), et le trucage est souvent visible. Mais attention, ces contenus n'en sont pas moins dangereux. Une photo non truquée, avec une légende qui l'attribue à un autre contexte que celui dans lequel elle a été prise, peut propager une fausse information autant qu'une vidéo réalisée par intelligence artificielle.

² Office québécois de la langue française (2019) — Hypertrucage. Dans *Le grand dictionnaire terminologique*

³ Journal Officiel de la République Française — *Vocabulaire de la culture : édition, médias et mode (liste de termes, expressions et définitions adoptés)*, n°0125 texte n°97, 23 mai 2020

⁴ Gaîté Lyrique (mis à jour le 02.09.2020) — Cheapfake. Dans *Lexique de la Gaîté Lyrique*

Comme les cheapfakes, les deepfakes sont un énième outil de désinformation, plus high-tech et convaincants, même si leur technologie n'a pas été créée dans ce but. Alors qu'il est déjà difficile de limiter la propagation de fausses informations grâce au fact-checking, comment combattre des vidéos qui, en apparence, montrent des choses bien réelles ? Encore faudrait-il être en mesure de les détecter, car malgré un esprit critique aiguisé, n'importe qui peut tomber dans le panneau des deepfakes. Ces "armes d'illusion massives"⁵ trompent le plus important de nos cinq sens, sur lequel les humains se reposent le plus⁶ : la vue.

Face au développement de cette technologie et dans le contexte actuel des fake news, en quoi les deepfakes représentent-ils une nouvelle menace pour le journalisme ?

Cette question nous mènera à une réflexion en deux grandes parties. Dans un premier temps, nous analyserons comment les deepfakes ont fait leur apparition dans notre société, passant du statut d'OVNI technologique sur Internet à un véritable sujet d'inquiétude que les journalistes ne peuvent pas ignorer. Puis, dans une deuxième partie, nous tâcherons de comprendre les dangers des deepfakes comme armes de désinformation et étudierons les solutions pour les combattre.

⁵ Gérald Holubowicz. (16 octobre 2020). Les deepfakes, une « arme d'illusion massive » ?. *Institut national de l'audiovisuel, La revue des médias*

⁶ Fiorenza Gracci. (7 octobre 2017). Comment fonctionne la vision ? *Science & Vie QR n°16* « Nos cinq sens & leurs mystères »

I – D'un OVNI à un objet journalistique

I. 1 – Origines et fonctionnement des deepfakes

La première apparition connue du terme "deepfake" remonte à l'automne 2017, dans les tréfonds d'Internet. Sur Reddit, un forum en ligne où les utilisateurs peuvent partager tout et n'importe quoi (questions, liens, vidéos, memes etc), un utilisateur anonyme publie une série de vidéos pornographiques mettant en scène des célébrités comme les actrices Gal Gadot, Daisy Ridley, Scarlett Johansson ou encore la chanteuse Taylor Swift. Son pseudonyme ? "Deepfakes".

Mais si les stars sont bien reconnaissables, il ne s'agit pas vraiment d'elles. Leur visage a été ajouté sur le corps d'actrices de films de X à l'aide d'un algorithme. Ce dernier aurait été créé à l'aide de TensorFlow, l'outil d'apprentissage automatique de Google Brain, disponible en open source (i.e. le code source est gratuitement accessible pour qui souhaite s'en servir et/ou le modifier). Grâce à cette facilité d'accès au machine learning, "n'importe qui armé de quelques connaissances informatiques – et d'un (bon) ordinateur portable – peut développer et éduquer une IA dans sa chambre"⁷, et donc potentiellement fabriquer des deepfakes pornographiques.



Captures d'écrans de deepfakes pornographiques avec Gal Gadot (en haut à gauche), Daisy Ridley (en haut à droite), Katy Perry (en bas à gauche) et Emma Watson (en bas à droite).

Sources : *Vice*, *Le Figaro*, *Dazed*

Les vidéos deviennent rapidement virales sur Reddit. En novembre 2017, le subreddit "deepfakes" (sorte de sous-section dédiée à un thème spécifique) est créé et rassemble des dizaines d'autres vidéos pornographiques avec les visages de femmes célèbres, comme Emma Watson et Katy Perry. La machine des deepfakes est lancée.

⁷ Pierre Schneidermann. (14 février 2018). TensorFlow, l'outil qui a favorisé l'émergence des deepfakes. *Konbini*

Avant d'aller plus loin dans l'évolution des deepfakes, et même si ce mémoire se concentre sur des questions journalistiques plutôt que techniques, il m'a semblé important de comprendre le fonctionnement de la technologie. N'étant pas familière avec le code, j'ai interrogé un créateur professionnel de deepfakes : Yanis, alias "French Faker", qui n'a pas souhaité dévoiler son nom de famille ni son âge.

Ce jeune homme de moins de trente ans a appris à réaliser des deepfakes en autodidacte à l'été 2019. Après avoir "*commencé à faire des deepfakes pour faire rire [ses] amis*"⁸, il a fini par en faire son métier. Aujourd'hui, avec sa société individuelle French Faker, il réalise des deepfakes pour l'émission télévisée *C'est Canteloup* sur TF1. Yanis a bien voulu m'expliquer le processus de création d'un deepfake dans les grandes lignes : "*C'est relativement simple. Pour réaliser un deepfake il y a différentes étapes :*

1. *Récupérer des vidéos sources de la personne que l'on veut deepfaker. Les images doivent être en très bonne qualité, avec une palette d'expressions si possible fournie et différentes luminosités. Au niveau de la durée, plus c'est long mieux c'est. Ça donne du grain à moudre à la machine. Je dirais qu'il faut 10-15 minutes de vidéos sources, mais bien sélectionnées. C'est une étape qui est très chronophage.*
2. *Avoir une vidéo cible, où l'on voit le visage de la personne, en bonne qualité et avec une bonne luminosité.*
3. *Ensuite, on découpe ces vidéos image par image, on extrait les têtes sur les images sources et les images cibles.*
4. *Puis pendant quelques jours on fait tourner un algorithme qui va "apprendre" les visages de ces personnes. On va lui montrer plusieurs images, plusieurs fois par seconde, idéalement en batch.*
5. *L'algorithme va apprendre à calquer un visage sur l'autre. Il va voir qu'à un moment une personne ouvre la bouche et il va calculer la moyenne qu'il a déjà de l'autre personne pour lui faire ouvrir la bouche aussi.*
6. *Puis on fusionne tout simplement. Et ça donne un deepfake.*

Après il y a plein de paramètres à rentrer, comme la netteté, le type de luminosité, le type de masque. Les deepfakes ont commencé avec des masques qui n'allaient que du front au menton, maintenant on peut faire tout le visage voire toute la tête."⁹

Pour Yanis, ce n'est pas un hasard si les premiers deepfakes viraux étaient des vidéos à caractère pornographique : "*Malheureusement, l'industrie du porno est souvent en avance sur les innovations technologiques. D'ailleurs certains deepfakes pornographiques sont exceptionnels de technicité, même si les algorithmes n'ont pas été créés pour ça.*"¹⁰

⁸ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

⁹ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

¹⁰ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

Suite à la publication des premiers deepfakes pornographiques sur Reddit, il n'aura fallu que peu de temps avant que des applications de création de deepfakes voient le jour, et démocratisent la pratique en la rendant plus facile. FakeApp sera la première du genre à être créée en janvier 2018, par un utilisateur de Reddit connu sous le pseudonyme "deepfakeapp". Ce dernier publie gratuitement son application en ligne dans le but de "*mettre la technologie "deepfakes" à la disposition des personnes sans formation technique ni expérience en programmation.*"¹¹ Avec FakeApp, plus besoin de savoir coder ou d'avoir installé des programmes informatiques comme Tensorflow ou Python. Il suffit d'envoyer à l'application une ou deux vidéos de bonne qualité pour qu'elle réalise un deepfake en 12 heures environ.

Rapidement, les deepfakes se multiplient sur Internet au début de l'année 2018. En janvier, le subreddit "deepfakes" cumule plus de 15 000 abonnés¹². L'écrasante majorité des contenus synthétiques créés reste des vidéos pornographiques ciblant des célébrités féminines, même si quelques internautes commencent à saisir le potentiel artistique et humoristique de la technologie. Certains s'amuse notamment à deepfaker le visage de l'acteur Nicolas Cage dans tout un tas de films dans lequel il n'a pas réellement joué.¹³

Les deepfakes perdront leur étiquette "pornographie" l'année suivante avec le succès d'une autre application, Zao, lancée le 30 août 2019. La facilité d'utilisation de son interface lui permet de toucher le grand public¹⁴ et son nombre de téléchargements explose en quelques jours. Même si à ses débuts, l'application de deepfake requiert d'avoir un numéro de téléphone chinois, elle connaît un succès fulgurant dans le monde entier. Début septembre 2019, Zao devient l'application la plus téléchargée sur iPhone en Chine.

C'est à cette même période que j'ai dû choisir un sujet traitant à la fois de journalisme et d'innovation, pour mon mémoire professionnel de Master 2 au CELSA. Ayant toujours eu une appétence pour les sujets qui mêlent technologie et société, j'avais suivi la montée en popularité des deepfakes, et sentais que cette innovation ne tarderait pas à sortir du contexte pornographique et humoristique. Pourtant, pendant un certain temps, les médias français comme étrangers n'ont traité les deepfakes que sous l'angle des dangers de ces vidéos pornographiques manipulées.

¹¹ Samantha Cole. (24 janvier 2018). We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. *Vice*

¹² Samantha Cole. (24 janvier 2018). We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. *Vice*

¹³ Rob Price. (27 janvier 2018). People are using creepy, cutting-edge AI technology to splice Nic Cage into every movie they can think of. *Business Insider*

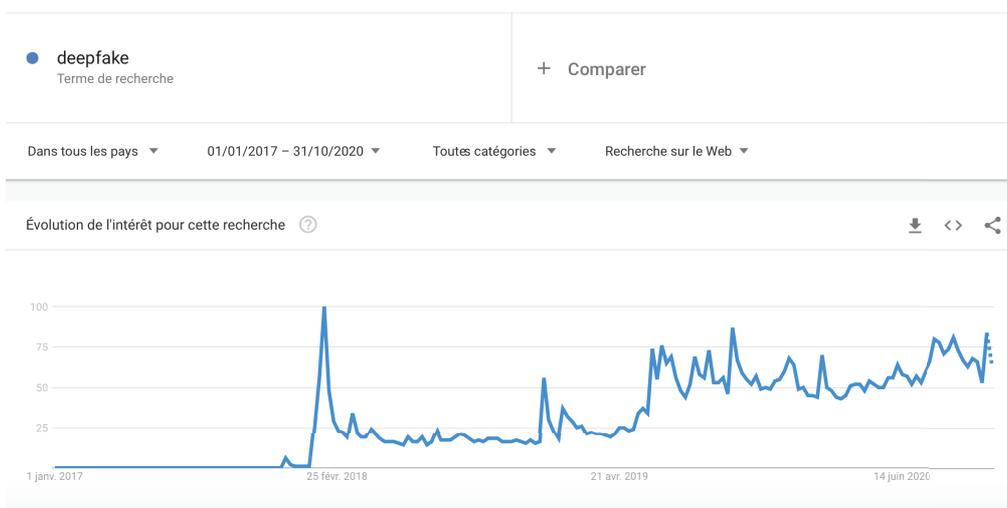
¹⁴ Morgane Tual. (3 septembre 2019). A peine lancée, l'application de vidéos « deepfake » Zao suscite des inquiétudes. *Le Monde*

II. 2 – La vague de deepfakes pornographiques attire l'attention des médias

Comme souvent lorsque qu'il s'agit de tech, un(e) journaliste repère une innovation ou une tendance sur Internet, et sent qu'il y a un sujet à traiter. Dans le cas des deepfakes, c'est un article du site américain *Vice*, plus précisément de sa rubrique tech *Motherboard*, qui va ouvrir le bal médiatique sur les deepfakes. L'article "AI-Assisted Fake Porn Is Here and We're All Fucked"¹⁵ (qu'on pourrait traduire par "La fausse pornographique créée par intelligence artificielle arrive et on est tous foutus") est publié le 11 décembre 2017 sur *Vice*. Dès le lendemain, Google Trends enregistre les premières recherches des mots "deepfakes" et "deepfake" sur le moteur de recherche. Avant le 12 décembre 2017, ces termes n'apparaissent pas dans l'historique de recherche.



Nombre de recherches du terme "deepfakes" sur le Web dans le monde entier entre le 1er janvier 2017 et le 31 octobre 2020. Source : Google Trends



Nombre de recherches du terme "deepfake" sur le Web dans le monde entier entre le 1er janvier 2017 et le 31 octobre 2020. Source : Google Trends

¹⁵ Samantha Cole. (11 décembre 2017). AI-Assisted Fake Porn Is Here and We're All Fucked. *Vice*

D'autres médias anglo-saxons emboîtent le pas à *Vice* et couvrent le sujet des deepfakes pornographiques, dénonçant l'utilisation des visages de femmes, célèbres ou non, sans leur consentement. L'inquiétude de l'usage de cette technologie dans le cadre de *revenge porn* — la publication de contenu sexuellement explicite sans le consentement de la ou des personnes concernées — est évoquée.

Google enregistre un pic de recherches des termes "deepfakes" et "deepfake" en février 2018, lorsque Reddit annonce interdire la publication de deepfakes sur son site et fait le ménage pour supprimer toutes les vidéos manipulées et applications de création déjà en ligne. Dans la foulée, le géant du streaming pornographique Pornhub annonce également bannir les deepfakes. Des internautes continuent cependant de partager des deepfakes pornographiques et des algorithmes pour en créer sur d'autres plateformes comme Github.

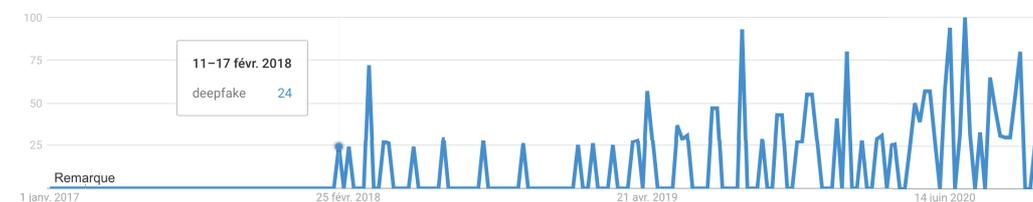
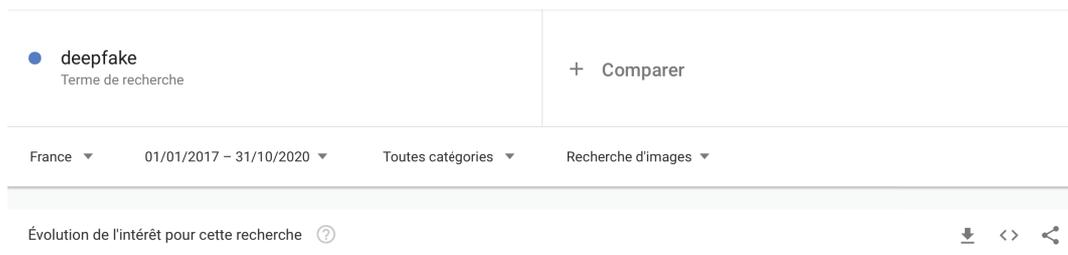
En France, les médias observent le phénomène de loin. Quelques articles y sont consacrés dans les colonnes tech des journaux, mais l'engouement pour le sujet est plus modéré qu'Outre-Atlantique. A l'époque, Gerald Holubowicz travaille pour *Libération* et suit en parallèle un Executive Master à Sciences Po, pour lequel il prépare un mémoire professionnel sur les deepfakes. Ce journaliste de formation, désormais chef de produit spécialisé en innovation éditoriale et nouveaux formats à Condé Nast, a donc suivi de près l'émergence des deepfakes et leur évolution. Il partage aujourd'hui ses analyses et recherches sur les deepfakes sur son site personnel Journalism.design¹⁶.

Comme c'est habituellement le cas avec les sujets qui traitent d'innovation technologique, il a remarqué une frilosité de la part des journalistes français que n'avaient pas leurs confrères anglo-saxons, comme il me l'a raconté lors de notre entretien : *"Quand j'ai commencé à travailler sur le sujet, en février 2018, j'étais à Libération, et je suis allé voir Cédric Mathiot de Checknews [la rubrique de fact-checking en ligne de Libération, ndlr] pour l'interroger sur les deepfakes. Globalement, lui il ne voyait pas l'intérêt, parce que déjà il n'est pas tech du tout, et par ailleurs leur positionnement à Checknews n'est pas là dessus. Mais effectivement en France, je ne sentais pas une fébrilité monstrueuse sur le sujet par rapport à ce qu'on voyait dans les pays anglo-saxons."*¹⁷

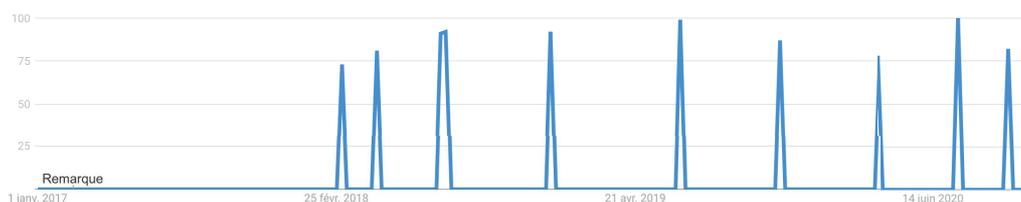
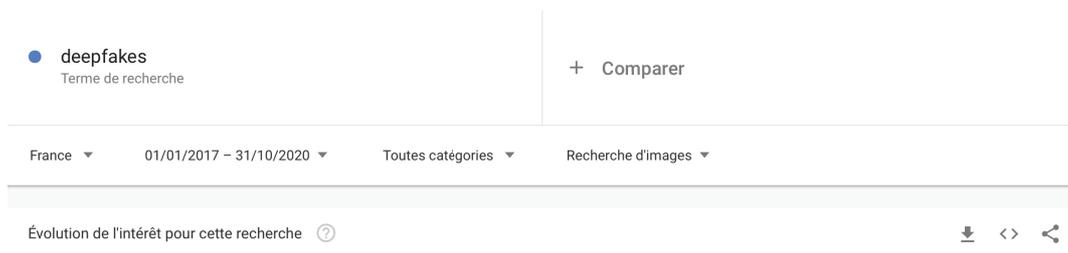
La comparaison entre les recherches Google des termes "deepfake" et "deepfakes" en France et dans le monde entier montre également cette différence de comportement face à cette technologie. Dans l'Hexagone, il y a des pics de recherche à des dates très précises, car les médias français ne traitent des deepfakes que sous le prisme de l'actualité "chaude", alors que les recherches Google tous pays confondus montrent un intérêt plus constant, et moins dépendant de l'actualité.

¹⁶ "Deepfake.media". *Journalism.design* [site Internet]. Gérald Holubowicz. s.d.

¹⁷ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41



Nombre de recherches du terme "deepfake" sur le Web en France entre le 1er janvier 2017 et le 31 octobre 2020. Source : Google Trends



Nombre de recherches du terme "deepfakes" sur le Web en France entre le 1er janvier 2017 et le 31 octobre 2020. Source : Google Trends

Gerald Holubowicz partage cette analyse : *"En journalisme, on traite de l'immédiat, pas du long terme. C'est pour ça que les journalistes ne s'intéressent pas aux deepfakes en France. Ceux que j'ai interrogés m'ont tous dit 'pour l'instant il n'y a pas de deepfake en France donc comme il n'y en a pas, on ne traite pas le sujet.' Aujourd'hui encore, ils restent vigilants, mais ça ne les intéresse pas plus que ça."*¹⁸

Même si le degré d'importance accordé aux deepfakes n'est pas le même en France qu'aux Etats-Unis ou au Royaume-Uni, un événement a cristallisé l'attention des médias du monde entier, car il a apporté une dimension plus personnelle aux deepfakes pour les journalistes. En avril 2018, suite au viol d'une fillette qui a secoué l'opinion publique en Inde, Rana Ayyub, une journaliste d'investigation indienne, est invitée sur les

¹⁸ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

chaînes télévisées BBC et Al Jazeera. Elle y dénonce comment le parti nationaliste Bharatiya Janata (BJP) défend les agresseurs sexuels au lieu de protéger les enfants victimes de viol. Une campagne de harcèlement contre la journaliste débute sur Twitter, mais le pire reste à venir.

Après sa prise de parole, Rana Ayyub reçoit d'une source fiable une vidéo pornographique d'elle-même. Il s'agit évidemment d'un deepfake, fabriqué de toute part, mais il circule dans les hautes sphères gouvernementales avant d'être massivement partagé sur les réseaux sociaux. La journaliste supprime son compte Facebook, disparaît des médias, et finit par être hospitalisée à cause de son anxiété¹⁹. L'organisation Reporters sans frontière en appelle même les autorités indiennes à protéger la journaliste face à la campagne de haine dont elle est la cible²⁰ mais le mal est déjà fait. Les deepfakes ne se contentent plus de servir aux adeptes du *celebrity porn* ou *revenge porn*, ils apparaissent désormais comme un puissant outil de chantage pour faire pression sur des individus, notamment des journalistes.

Des mois plus tard, en novembre 2018, Rana Ayyub témoignera dans le *HuffPost* britannique : *"Pendant longtemps je n'en ai pas parlé parce que je craignais que les gens n'aient pas d'empathie ou de sympathie pour moi, et qu'ils cherchent à explorer [les deepfakes] davantage. Je ne voulais pas que les deepfakes obtiennent ce genre de popularité. Mais malheureusement, ces dernières semaines, j'ai vu de nombreuses vidéos deepfakes de très grandes stars féminines du cinéma, alors j'ai l'impression qu'il est trop tard pour les empêcher. C'est un outil très, très dangereux et je ne sais pas vers quoi nous nous dirigeons avec."*²¹

Selon Gerald Holubowicz, en France, *"la vague du deepfake pornographique a été intéressante mais pas très exploitée."* A l'époque, il avait contacté des associations féministes françaises, connues pour leurs actions contre le cyber harcèlement, pour aborder le sujet des deepfakes, mais il n'a pas eu de réponse de leur part. *"Je pensais que la question du revenge porn allait plus interpeler. Il y a eu quelques grosses histoires en Australie et en Inde, et des journalistes ont été cibles de deepfakes pornos, mais à part Rana Ayyub, on ne connaît pas leurs noms."* D'après ses informations, il y aurait eu cinq ou six autres cas de journalistes cibles de chantage au deepfake pornographique, *"mais les médias n'ont pas voulu nourrir la bête en en parlant."*²²

¹⁹ Rana Ayyub. (21 novembre 2018). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. *HuffPost UK*

²⁰ Reporters Sans Frontières. (27 avril 2018). RSF appelle les autorités indiennes à protéger la journaliste Rana Ayyub. *Reporters Sans Frontières*.

²¹ Rana Ayyub. (21 novembre 2018). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. *HuffPost UK*

²² Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

Si le traitement médiatique des deepfakes semble s'être limité à l'angle de la pornographie pendant longtemps, c'est sûrement aussi parce que les vidéos pornographiques représentent l'immense majorité des deepfakes qui circulent. En septembre 2019, un rapport de Deeptrace²³ — une plateforme de recherche sur les deepfakes et leur détection, aujourd'hui renommée Sensity — révélait ainsi que 96 % des deepfakes en ligne étaient des vidéos pornographiques non consenties. En comparaison, le potentiel artistique de la technologie deepfake, pour le cinéma par exemple, mais aussi humoristique, comme pour les vidéos parodiques, attire forcément moins l'attention.

Mais à mesure que la technologie deepfake fait parler d'elle, et que ses possibilités d'utilisation s'étendent à différents domaines, un autre usage commence à être envisagé, et inquiète sérieusement les journalistes : et si les deepfakes servaient à propager des fake news à l'apparence plus vraie que nature ?

III. 3 – Les deepfakes : l'arme de désinformation par excellence

Lorsque j'ai commencé à travailler ce mémoire à l'automne 2019, on parlait encore assez peu des deepfakes en France, hormis l'angle de la pornographie précédemment évoqué. Je ne disposais pas de beaucoup de ressources en français sur le lien entre deepfake et fake news, mais je pressentais que l'approche de l'élection présidentielle américaine en novembre 2020 allait changer la donne, et que la menace des deepfakes politiques allait se faire de plus en plus pesante. Ça n'a pas manqué. A partir de la fin de l'année 2019, jusqu'à aujourd'hui, le nombre d'articles sur les dangers des deepfakes pour l'information, en français comme en anglais, n'a cessé d'augmenter, preuve que les journalistes se sont emparés du sujet. En 2018, quelques médias avaient déjà touché du doigt le sujet de la désinformation et des deepfakes.

*BuzzFeed*²⁴ a notamment réalisé une vidéo de prévention sous la forme d'un deepfake en avril 2018. On y voit l'ancien président américain Barack Obama s'adresser au spectateur pour lui dire : "*Nous entrons dans une ère où nos ennemis peuvent faire croire que n'importe qui dit n'importe quoi à n'importe quel moment.*"²⁵ Mais au bout de quelques phrases, le langage présidentiel disparaît, et Barack Obama attaque son successeur "*Le président Trump est un total et absolu abruti.*" C'est en réalité l'acteur et réalisateur américain Jordan Peel (qui a reçu l'Oscar du meilleur scénario original pour *Get Out*) qui prononce ces mots, comme la vidéo le révèle peu à peu, démontrant la

²³ "The State of Deepfakes — Landscape, Threats, and Impact" [rapport de recherche—. DeepTrace. Septembre 2019

²⁴ David Mack. (17 avril 2018). This PSA About Fake News From Barack Obama Is Not What It Appears. *BuzzFeed*

²⁵ BuzzFeedVideo. (17 avril 2018). *You Won't Believe What Obama Says In This Video!* 🤪 [Vidéo]. Youtube

puissance de la technologie deepfake, et ses dangers si elle est utilisée à mauvais escient, surtout en politique.



Une vidéo deepfake de prévention contre les deepfakes, mettant en scène le visage de Barack Obama avec la voix de l'acteur Jordan Peel.
Source : *BuzzFeed Video*

La vidéo devient virale et est visionnée près de 3 millions de fois²⁶ en quelques jours. *"Très vite aux Etats-Unis c'est monté d'un coup, en épingle. Le combo intelligence artificielle, information et Trump était explosif, les médias étaient paniqués"*²⁷, se rappelle Gerald Holubowicz. Dans le contexte du moment — un président américain qui qualifie de *"fake news"* chaque média qui ne va pas dans son sens, une crise de confiance envers la presse, la majorité de la population qui s'informe via les réseaux sociaux —, les deepfakes apparaissent en effet comme une bombe de désinformation extrêmement dangereuse. Certains médias comme *BuzzFeed* comprennent rapidement l'intérêt d'informer les gens sur cette technologie, pour mieux les protéger contre ses dangers.

Selon Yanis, alias French Faker, utiliser la technologie des deepfakes pour faire de la prévention sur les deepfakes est une méthode efficace. Il a lui-même imaginé un équivalent français de la vidéo de *BuzzFeed* : *"Mon rêve absolu serait de réaliser un spot télévisé, je le ferais gratuitement pas de problème, dans lequel je ferais dire n'importe quoi à Emmanuel Macron, comme des propos anti-masque par exemple, et à la fin je lui ferais dire quelque chose comme 'ce n'est pas parce que ça va dans votre sens que c'est la vérité'"*²⁸. Il estime que les créateurs ont aussi un rôle à jouer dans l'éducation aux

²⁶ Elisa Braun. (20 avril 2018). La viralité d'une fausse vidéo d'Obama met en lumière le phénomène du "deep fake". *Le Figaro*

²⁷ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

²⁸ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

deepfakes, même si dans la sphère des deepfakeurs professionnels, rares sont ceux qui s'intéressent à la prévention.

Dans la Silicon Valley, des experts s'inquiètent également du combo deepfakes et fausses informations et certains tirent même la sonnette d'alarme. Aviv Ovadya, qui avait déjà anticipé la crise des fake news juste avant l'élection de Donald Trump en 2016, présage l'arrivée imminente d'une "infocalypse", une apocalypse de l'information. En janvier 2018, il explique ainsi à *BuzzFeed* comment il envisage les deux prochaines décennies, qui seront marquées par les fausses informations, les deepfakes et la propagande : *"Nous sommes tellement foutus que cela dépasse ce que la plupart d'entre nous peuvent imaginer. Nous étions déjà complètement foutus il y a un an et demi et nous le sommes encore plus maintenant. Et selon comment on voit l'avenir, ça ne fait qu'empirer."*²⁹

Les deepfakes politiques vont en effet se faire plus nombreux. La technologie est utilisée pour la première fois officiellement dans une campagne électorale en février 2020. Le Bharatiya Janata Party (BJP), l'un des deux principaux partis politiques de l'Inde, diffuse des vidéos trafiquées de son président Manoj Tiwari pour sa campagne à Delhi. Afin de toucher un maximum d'électeurs, la vidéo originale, dans laquelle le candidat prononce son discours en Hindi, a été passée dans un algorithme de deepfake afin d'obtenir la même vidéo, mais dans laquelle il semble parler d'autres langues (anglais et Haryanvi, une langue principalement parlée dans la région de Haryana en Inde). Ces deepfakes ont été envoyés sur 5 800 groupes WhatsApp dans le pays et ont atteint 15 millions de personnes.³⁰

Mais la technologie n'est pas toujours utilisée dans une démarche "positive", comme celle voulue par le BJP. En juin 2020, une vidéo publiée sur Facebook fait le buzz : on peut y voir et entendre l'ambassadeur de France au Cameroun Christophe Guilhou tenir des propos choquants comme *"La République française, c'est la puissance de tutelle qui a colonisé le Cameroun"* ou encore *"Mes ancêtres ont conquis ce territoire par la force et la ruse et le droit international"*. Il s'agit là encore d'un deepfake, réalisé à l'aide d'une application, comme le démontrera le service de fact-checking de *France 24*³¹. La vidéo truquée de Christophe Guilhou est à ce jour le seul deepfake politique français qui a été publié dans une tentative de propager une fausse information.

²⁹ Charlie Warzel. (11 février 2018). He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse. *BuzzFeed*

³⁰ Nilesh Christopher. (18 février 2020). We've Just Seen the First Use of Deepfakes in an Indian Election Campaign, *Vice*

³¹ Les Observateurs. (29 juin 2020). Attention, cette vidéo de l'ambassadeur français au Cameroun est un "deepfake". *France24*

Aux Etats-Unis, notamment pendant la campagne électorale pour l'élection présidentielle, les deepfakes ont été bien plus utilisés qu'en France. Sans surprise, les partisans de Donald Trump et le président lui-même ont partagé des vidéos manipulées pour nuire à ses opposants politiques. En juin 2019 déjà, une vidéo de la présidente de la Chambre des représentants, Nancy Pelosi — fervente démocrate et adversaire de Donald Trump — devient virale³². On la voit avoir du mal à prononcer son discours au Centre Américain pour le Progrès, elle parle lentement et n'articule pas suffisamment. Des membres du parti Républicain partagent la vidéo en accusant Nancy Pelosi d'être ivre pour la décrédibiliser. En réalité, la vitesse de la vidéo a été ralentie pour donner l'illusion qu'elle est ivre ou droguée, et le son modifié afin que le ton de sa voix ne soit pas altéré. Il ne s'agit pas d'un deepfake, mais plutôt d'un cheapfake, car la vidéo n'a pas été modifiée à l'aide d'une intelligence artificielle, mais les médias américains comme français couvrent l'information en faisant le lien entre les deux types de vidéos manipulées et leurs dangers.

La mécanique se répétera à plusieurs reprises tout au long de la campagne présidentielle de Donald Trump, ses supporters républicains étant très prompts à partager des vidéos trompeuses, et souvent dégradantes pour ses adversaires. L'exemple le plus célèbre en date : une vidéo d'une interview de Joe Biden, l'adversaire de Donald Trump dans la course à la Maison Blanche, dans laquelle il semble s'être endormi. Il a les yeux fermés, ne répond plus aux questions de la journaliste et ronfle même. Les partisans de Donald Trump, qui surnomme Joe Biden "Sleepy Joe" (Joe l'Endormi), partagent la vidéo en masse sur les réseaux sociaux. Pourtant le candidat démocrate ne s'est jamais assoupi en pleine interview, il s'agit d'un montage cheapfake : un extrait vidéo où Joe Biden a les yeux fermés a été ajouté au décor d'une interview avec une journaliste, qui n'a en réalité jamais interrogé le candidat, et les ronflements ont eux aussi été rajoutés.³³

Les deepfakes et cheapfakes sont systématiquement fact-checkés et débunkés par des journalistes, mais seulement après avoir été partagés massivement sur les réseaux sociaux. Dès lors qu'une vidéo manipulée devient virale, il est trop tard pour effacer entièrement ses conséquences. En 2020, il est ainsi clairement apparu que les deepfakes, comme les cheapfakes, représentent une menace pour l'information, et plus largement pour la démocratie. Et cette menace ne va aller qu'en s'aggravant si l'on ne trouve pas une solution pour la combattre.

³² Harold Grand. (24 mai 2019). «Deepfake»: une vidéo trafiquée de Nancy Pelosi relayée par des proches de Trump. *Le Figaro*

³³ Mathilde Cousin. (31 août 2020). Présidentielle américaine : Non, Joe Biden ne s'est pas endormi pendant cette interview, qu'il n'a d'ailleurs même pas donnée. *20 Minutes*

II – Les deepfakes face à l'information, comment les combattre ?

I. 1 – Une menace grandissante dans un contexte propice à la désinformation

La menace des deepfakes va grandir proportionnellement à l'importance de la place des réseaux sociaux dans la façon dont les gens s'informent, et le scénario le plus probable penche en faveur d'une aggravation. Tout dans le fonctionnement des réseaux sociaux encourage la propagation des deepfakes, et plus largement des fausses informations. Les boutons de "like" et de partage incitent à la viralité des publications, renforçant le biais cognitif qui nous pousse déjà à réagir plus fortement face à un contenu polémique que face à une information neutre.

Les réseaux sont aussi, et surtout, un terreau fertile à la désinformation par les deepfakes à cause des chambres d'écho et des bulles de filtre, deux phénomènes qui poussent les utilisateurs à n'être exposés qu'à des opinions, des articles et même des personnes avec lesquels ils sont d'accord et qui partagent leur idéologie politique ou sociale.

Le terme "chambre d'écho" désigne plus spécifiquement le comportement humain qui consiste à interagir davantage avec les sujets et les personnes qui nous font plaisir, avec lesquels nous sommes d'accord, ce qui peut fausser notre perception de la réalité. Les bulles de filtres ne viennent, elles, pas directement des utilisateurs, mais des réseaux sociaux, dont les algorithmes filtrent ce qui sera montré à l'utilisateur selon ses goûts et ses opinions, l'enfermant peu à peu dans des chambres d'écho.³⁴

Pour French Faker, ces phénomènes représentent une menace plus lourde pour l'information que la technologie : *"Les deepfakes sont un problème, mais je pense que le plus gros problème ce sont les chambres d'écho sur les réseaux sociaux. C'est facile de rester dans ces chambres d'écho parce qu'on y est bien, tout le monde pense pareil que nous. Mais la vérité, c'est des nuances de gris plutôt que du tout noir ou blanc."*³⁵ Les réseaux sociaux ne semblant pas perdre de leur puissance avec le temps, le contexte devrait continuer d'être propice à la prolifération des deepfakes de désinformation.

Un autre phénomène sociologique est également à l'action et aggrave les dangers des deepfakes. Il s'agit du "Liar's dividend", qu'on peut littéralement traduire par "le bénéfique du menteur". Ce concept a été théorisé par Robert Chesney et Danielle Citron,

³⁴ Dr Richard Fletcher. The truth behind filter bubbles: Bursting some myths. *Reuters Institute*. s.d.

³⁵ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

deux professeurs américains de droit, et Hany Farid, professeur et chercheur en informatique, qui se sont penchés sur l'impact des deepfakes dans notre société :

"La prise de conscience de la menace des deepfakes est elle-même potentiellement néfaste. Cela augmente les chances que les gens soient victimes d'un phénomène que deux d'entre nous (Chesney et Citron) appelons le "bénéfice du menteur". Au lieu d'être "dupés" par les deepfakes, les gens pourraient en venir à se méfier de toutes les vidéos et enregistrements audio. Le déclin de la vérité est une bénédiction pour les personnes moralement corrompues. Les menteurs peuvent échapper à la responsabilité de leurs fautes et rejeter les preuves réelles de leurs méfaits en disant que ce n'est 'qu'un deepfake'"³⁶

Gerald Holubowicz m'a décrit le même phénomène qui l'inquiète : *"Ce qu'il n'y avait pas jusque là, c'est la connaissance chez les gens que les deepfakes pouvaient exister, il manquait cette plausibilité. Maintenant on sait que ça peut être fait. Et c'est là que ça devient dangereux, c'est de voir l'infusion de cette idée là dans la société."*³⁷

Pour lui, la réalisation qu'une telle technologie existe est plus dangereuse que les vidéos deepfakes en elles-mêmes : *"L'objet deepfake ne causera pas de problème, c'est l'idée des deepfakes qui va en poser. La preuve : Winnie Heartstrong [candidate républicaine au poste de représentante à la Chambre, ndlr] a publié un document qui affirme que le meurtre de George Floyd est un deepfake. On a quand même quelqu'un qui a pris le temps d'écrire ces 23 pages complètement absurdes et qui a eu une plateforme pour en parler, notamment dans les réseaux alt-right, et une partie de cette audience là l'a crue."*³⁸ Dans le discours aberrant de Winnie Heartstrong, les deepfakes ne sont ni compris ni maîtrisés, mais servent *"d'écran de fumée rhétorique pour convaincre une audience crédule"*³⁹ : c'est le "liar's dividend" en action.

Le concept du bénéfice du menteur a également déjà été utilisé par Donald Trump, suite à la publication de la vidéo notoire dans laquelle il explique qu'être une célébrité lui permet de faire *"tout ce qu'[il] veut"* avec les femmes, comme *"les attraper par la chatte"*. Après le scandale, le président américain a tenté de semer le doute en avançant que la vidéo devait être fausse ou truquée. Le raisonnement de défense de Donald Trump est simple, et efficace : puisque les médias nous disent qu'il ne faut plus croire tout ce que l'on voit, il ne faut pas croire tout ce que nous montrent les médias.

³⁶ Robert Chesney, Danielle Citron et Hany Farid. (11 mai 2020). All's Clear for Deepfakes: Think Again. *Lawfare*

³⁷ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

³⁸ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

³⁹ Gerald Holubowicz. (27 juin 2020). George Floyd et les deepfakes: la folle théorie de Winnie Heartstrong, *Journalism.design*

Malheureusement, tous ces phénomènes propices à la désinformation ne semblent pas ralentir sur les réseaux sociaux, bien au contraire. Et en plus d'un contexte idéal à la propagation des deepfakes, des facteurs techniques aggravent aussi leur menace.



Capture d'écran de Reface.
Source : Reface/Google Play

Premièrement, il va être de plus en plus facile de réaliser des deepfakes réalistes, sans aucune connaissance en informatique. On a déjà pu observer la création de deepfakes à grande échelle dès que des applications clés en main ont été mises au point. Après le succès de Zao en 2019, la "mode" des deepfakes a de nouveau explosé en septembre 2020 avec Reface. Cette application créée par des ukrainiens permet de prendre un selfie et de deepfaker son visage sur des vidéos ou des gifs en quelques clics. Les utilisateurs partagent ensuite souvent le résultat sur Instagram et TikTok.⁴⁰

*"Quand j'ai commencé, c'était bien plus difficile de réaliser des deepfakes, mais ça va devenir de plus en plus facile et demander de moins en moins de technique, prédit Yanis, alias French Faker. N'importe qui peut faire un deepfake basique maintenant avec l'application Reface, qui est le Zao européen."*⁴¹

En plus de devenir de plus en plus accessible, la technologie des deepfakes est en passe de devenir plus précise, et pourrait bientôt donner naissance à des deepfakes encore plus réalistes et donc potentiellement plus dangereux. La prochaine étape qui pourrait marquer un changement important est la concrétisation des deepfakes vocaux. Dans les deepfakes comme celui de Barack Obama réalisée par *Buzzfeed*, ou les sketches de *C'est Canteloup* produits par French Faker, l'algorithme n'a servi que pour la vidéo, la voix est celle d'un acteur ou d'un imitateur. Sans imitateur, un deepfake ne réussira pas à convaincre les foules que la personne qui s'exprime est celle que l'on voit à l'écran. *"Pour l'instant il manque un élément essentiel, c'est la convergence voix/vidéo"*⁴², estime Gerald Holubowicz.

Mais il est bien plus facile de tromper l'œil que l'oreille. *"Dans chaque voix, il y a une complexité d'intonations, de tons, de profondeur, de tensions, que l'humain va percevoir, mais l'ordinateur en est incapable. Quand on aura la capacité de faire la voix de*

⁴⁰ Aaron Holmes. (26 septembre 2020). Popular deepfake apps are making it easier than ever to make AI-powered manipulated videos — spawning new memes, and an increased potential for abuse. *Business Insider*

⁴¹ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

⁴² Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

n'importe qui en deepfake, à partir d'un sample de 5 secondes, là il y aura un changement, et un plus grand danger"⁴³, annonce Gerald Holubowicz.

Le procédé de création d'un deepfake vocal est le même que celui d'un deepfake vidéo, mais il demande plus de technique et le résultat n'est pas encore au rendez-vous. *"C'est un algorithme un peu plus poussé parce que c'est beaucoup plus évident de détecter quand une voix est fausse, donc ça prend un peu plus de temps, explique Yanis, le deepfakeur professionnel. En plus pour les deepfakes de voix, l'algorithme doit apprendre à parler, c'est-à-dire que chaque syllabe et voyelle, il doit apprendre à les dire. Et c'est aussi plus compliqué de créer une voix en deepfake parce qu'il ne suffit pas de 10 ou 15 minutes de fichiers sources comme pour la vidéo, il faut des heures et des heures d'enregistrement, et leurs retranscriptions."*⁴⁴

Des algorithmes ont déjà été entraînés pour réaliser des deepfakes vocaux en anglais. Les voix obtenues sont encore un peu robotiques, et on entend que quelque chose est étrange, mais les deepfakes vocaux devraient s'améliorer très prochainement et devenir plus convaincants. Et pour que la langue de Molière ne soit pas en reste, Yanis, alias French Faker, travaille en ce moment sur du deepfake de voix en français.

D'ici quelques années, voire peut-être même quelques mois, les deepfakes vocaux devraient donc être au point. Il suffira alors d'ajouter un deepfake de voix sur un deepfake vidéo pour que l'illusion soit quasi parfaite, et imperceptible par nos yeux et nos oreilles. S'il devient impossible pour l'humain de détecter les deepfakes, faudra-t-il se fier aux machines pour le faire ? Un algorithme peut-il nous aider à combattre un autre algorithme ? Des développeurs et experts se penchent déjà sur la question depuis un moment, et les réseaux sociaux semblent se diriger vers cette solution pour limiter la propagation de deepfakes sur leurs plateformes.

II. 2 – Les réseaux sociaux et la technologie à l'assaut des deepfakes

Paradoxalement, les premières mesures coercitives prises à l'encontre des deepfakes viennent des réseaux sociaux, alors même qu'ils sont en grande partie responsable de leur prolifération. Ce n'est pas dans l'intérêt de Facebook, Twitter et autres de limiter les contenus viraux sur leur plateforme, car leur modèle économique repose avant tout sur la collecte de données personnelles. Mais ces dernières années, les réseaux sociaux ont été bien obligés de prendre leurs responsabilités, notamment après que de nombreux experts ont prouvé que leur gestion des fausses informations avait joué un rôle dans l'élection de Donald Trump.

⁴³ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

⁴⁴ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

Dès le début de 2020, l'année de l'élection présidentielle américaine, les réseaux sociaux sont passés à l'action contre les deepfakes. En janvier, Facebook bannit les vidéos modifiées ou fabriquées *"de façon à ce que l'utilisateur moyen ne le voit pas et qui risqueraient d'induire quelqu'un en erreur"* et qui sont *"créées par intelligence artificielle ou par des algorithmes de machine learning"*, à l'exception des vidéos parodiques ou satiriques. Mais les cheapfakes n'étant pas créés par intelligence artificielle ou par un algorithme, ils restent autorisés. Lors de la publication du cheapfake de Nancy Pelosi, Facebook avait d'ailleurs refusé de supprimer la vidéo, en se justifiant : *"Nous n'avons pas de politique qui stipule que les informations que les gens postent sur Facebook doivent être vraies."*

Le 3 février 2020, c'est au tour de Youtube de publier ses nouvelles règles en prévention de l'élection. La plateforme supprima *"tout contenu qui a été manipulé ou truqué de manière à induire les utilisateurs en erreur et qui peut présenter un risque sérieux de nuire gravement à quelqu'un"*. La politique de Youtube exclut cependant *"les extraits de vidéos sortis de leur contexte"*. Youtube ne fait aucune mention de la technologie, sa politique concerne donc aussi bien les deepfakes que les cheapfakes.

Le 4 février, Twitter passe aussi à l'action contre les deepfakes. Le réseau annonce interdire les photos, vidéos et autres contenus manipulés qui sont *"partagés de façon trompeuse"* et qui présentent un risque sérieux de nuire à quelqu'un. Mais Twitter ne supprimera pas tous les deepfakes pour autant. S'ils ne présentent pas de risque pour la sécurité, ils pourront être publiés. Twitter pourra cependant y ajouter une étiquette *"contenu fabriqué ou modifié"* à l'attention des utilisateurs.

Le dernier réseau social en date à avoir pris des mesures contre les deepfakes est TikTok. L'application de vidéos virales a ajouté une nouvelle politique à ses conditions d'utilisation en août 2020, quelques mois seulement avant l'élection présidentielle américaine. Désormais, TikTok *"interdit les contenus fabriqués ou modifiés qui induisent les utilisateurs en erreur en déformant la vérité des faits d'une façon qui pourrait constituer un danger."* Toutes ces mesures préventives en vue de l'élection américaine soulignent bien que les géants de la tech ont conscience que les deepfakes (et surtout les deepfakes politiques) représentent un danger pour la démocratie et l'information.

Pour détecter et supprimer les vidéos interdites, les réseaux sociaux comptent sur le signalement des utilisateurs, mais aussi sur des algorithmes de détection des deepfakes, qui ne sont pas encore parfaitement au point. Yanis, qui publie ses créations deepfakes humoristiques sur les pages Facebook et Twitter de French Faker ainsi que sur sa chaîne Youtube, a senti l'impact de ces algorithmes : *"Sur Facebook, ils ont détecté plusieurs fois que les vidéos que je publiais sur la page French Faker étaient fausses, et du coup ils les ont enlevées. Et j'ai l'impression aussi que, même si je n'ai pas énormément de likes, avant j'en avais plus"*⁴⁵, explique-t-il. Pourtant, ses vidéos relèvent de la parodie

⁴⁵ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

et ne devraient pas être interdites par Facebook, mais les algorithmes de détection des deepfakes ne font pas toujours la distinction.

Afin d'entraîner les systèmes d'intelligence artificielle à mieux détecter les deepfakes, Facebook a mis en ligne une base de données de 100 000 vidéos deepfakes en juin 2020, la plus grosse existante à ce jour. Elle a été réalisée grâce à l'aide des participants au concours "Deepfake Detection Challenge" de Facebook, en collaboration avec le MIT et Microsoft. Mais même avec cette vaste base de données pour s'entraîner, le système de détection des deepfakes n'était efficace qu'à 65 % en juin dernier (test réalisé sur un dataset complexe de 10 000 deepfakes, que l'algorithme n'avait jamais "vus" avant.)⁴⁶

En septembre 2020, Microsoft a lancé à son tour son outil de détection des deepfakes, appelé "Video Authenticator", dans le cadre de son programme de lutte contre la désinformation. Le système peut analyser une image ou une vidéo, et donner un "score de confiance" indiquant le pourcentage de chances que le média ait été manipulé. Tout comme celui de Facebook, l'algorithme de Microsoft a encore besoin d'"apprendre" en passant en revue des deepfakes, et de s'entraîner pour se perfectionner.

Pour Gerald Holubowicz, même si l'on arrivait à créer un algorithme de détection des deepfakes efficace à 100 %, ce ne serait pas pour autant une solution envisageable : *"C'est impossible à mettre en place de façon réaliste. Aujourd'hui, si on impose aux plateformes de réseaux sociaux d'avoir un système de détection des deepfakes, ça voudrait dire que tous les contenus uploadés devraient passer dans l'algorithme. Le volume de data géré en upload devrait être récupéré par des serveurs, traité en direct, avec une certaine puissance, puis mis sur serveur à disponibilité, dans des délais raisonnables pour que le service ne soit pas dégradé, parce que si une vidéo met trois heures à être publiée ça n'a aucun intérêt. Ce n'est pas une solution immédiate."*⁴⁷

Et en plus des problèmes évidents de mise en application de ces algorithmes de détection, faire confiance à la technologie pour nous aider pose d'autres problèmes. *"Est-ce qu'un algorithme est capable de faire la compétition avec un autre algorithme ?, s'interroge Gerald Holubowicz. Oui, mais c'est toujours le jeu du chat et de la souris, il y en aura toujours un qui aura une longueur d'avance sur l'autre."* Alors, si les humains ne peuvent se fier ni à ce qu'ils voient et entendent, ni à des algorithmes pour leur servir de sixième sens, quelle solution reste-il face aux deepfakes ? *"Je pense qu'on essaie d'avoir une solution courte mais on n'en a peut-être pas. C'est un peu ça qui fait flipper, on n'a pas de solution technologique. Il faudrait de l'éducation"*⁴⁸. Et c'est justement là que les journalistes entrent en jeu.

⁴⁶ Will Douglas Heaven. (12 juin 2020). Facebook just released a database of 100,000 deepfakes to teach AI how to spot them. *MIT Technology Review*

⁴⁷ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

⁴⁸ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

III. 3 – Le rôle crucial des journalistes et de l'éducation aux médias

Tout comme pour les fake news, les journalistes ont un rôle central à jouer dans la lutte contre les deepfakes. Leur première mission en tant que garants de la fiabilité de l'information est de vérifier cette dernière. Ce n'est pas parce qu'une information est présentée sous la forme d'une vidéo ou d'un enregistrement audio qu'elle est avérée, l'étape du fact-checking reste primordiale. La technologie donne cependant du fil à retordre aux journalistes. Désormais, s'ils reçoivent une vidéo d'un discours de Barack Obama, ils ne pourront pas se fier au logo de la Maison Blanche qui apparaît, ou à la présentation du journaliste de CNN qui lui précède, car ces informations ne sont peut-être qu'illusions.

Pour autant, la rapidité avec laquelle les journalistes sont amenés à traiter l'information n'est pas compatible avec la possibilité de vérifier chaque seconde de chaque vidéo et enregistrement audio. Comment faire alors pour savoir auxquels il faut prêter plus d'attention ? Lorsque que l'information paraît trop improbable ? Lorsque la source est douteuse ? Il n'y a pas encore de ligne de conduite définie à adopter face aux deepfakes lorsqu'on est journaliste, seulement des précautions supplémentaires à prendre. Certains médias ont commencé à former leur journalistes, c'est le cas de *Reuteurs*, qui a publié une formation en ligne sur les médias manipulés. Le chapitre 2 est consacré aux deepfakes : le cours explique comment ils sont créés et comment les identifier tout en donnant des exemples concrets pour s'entraîner. Car tout comme les algorithmes de détection des deepfakes, les journalistes doivent eux-aussi entraîner leurs yeux, et bientôt leurs oreilles, à détecter ces médias manipulés.

Mais comme même le plus aguerri des journalistes peut être trompé par un deepfake, des médias envisagent également de se faire aider par des systèmes d'intelligence artificielle, comme le font déjà les réseaux sociaux. Dans l'article "Que doivent faire les rédaction au sujet des deepfakes ? Ces trois choses, pour commencer" du *Nieman Lab*⁴⁹, la première piste proposée par les trois chercheurs est d'ailleurs de donner aux médias les moyens techniques de détecter rapidement les médias synthétiques ou altérés. En France, l'*AFP* teste régulièrement les algorithmes de détection de deepfakes qu'on lui soumet, mais pour le moment, selon les informations de Gerald Holubowicz, "*la plupart tombe à côté de la plaque.*"⁵⁰

Malheureusement, s'il ne faut pas relâcher les efforts de fact-checking, les journalistes se heurtent à un problème de rapidité face à la technologie. Lorsqu'un deepfake est publié sur les réseaux sociaux, tout peut aller très vite. Se pose alors le problème du délai entre le temps qu'il faudra à la vidéo pour devenir virale et propager une fausse information à un maximum de monde, et le temps nécessaire aux journalistes pour vérifier et expliquer l'information. La deuxième piste des chercheurs du *Nieman Lab*

⁴⁹ John Bowers, Tim Hwang et Jonathan Zittrain. What should newsrooms do about deepfakes? These three things, for starters. *Nieman Lab*. 20 novembre 2019.

⁵⁰ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

à destination des rédactions pourrait aider à résoudre ce problème : au lieu d'attendre d'avoir un résultat concluant, les journalistes pourraient dans l'immédiat "*préférer rendre compte de leur processus*"⁵¹, c'est-à-dire expliquer leur démarche de fact-checking et pourquoi elle nécessite un certain temps. En bref : faire preuve de transparence et de pédagogie sur le métier de journaliste. Mais c'est difficilement imaginable dans les conditions de travail actuelles des rédactions.

Ce temps de latence est ensuite aggravé par un autre problème : une partie de l'audience de la vidéo n'ira jamais lire l'article de fact-checking qui révèle qu'il s'agissait d'un deepfake, par manque de confiance envers les médias, ou par simple omission. Une réalité dont les journalistes n'ont pas suffisamment conscience aujourd'hui selon Gerald Holubowicz : "*On a une idée surdimensionnée de notre importance alors que la portée que l'on a est très faible. La plupart des gens passe des jours voire des semaines sans regarder les infos.*"⁵²

La viralité sur les réseaux sociaux et la trop faible force de frappe des médias traditionnels compliquent donc la lutte contre les deepfakes une fois que ces derniers sont en ligne. Mais comme le dit le vieux dicton, mieux vaut prévenir que guérir. Avant de se jeter à corps perdu dans le fact-checking, il faut déjà sensibiliser la population sur la question des deepfakes. La prévention par la pédagogie apparaît comme une étape essentielle si l'on veut combattre efficacement les médias manipulés et les fausses informations.

Les journalistes ne peuvent pas pour autant assurer seuls ce rôle. Comme évoqué précédemment, les réseaux sociaux doivent travailler main dans la main avec eux s'ils veulent limiter la propagation de deepfakes sur leurs plateformes, et ils ont déjà commencé à le faire. Et pour Yanis, alias French Faker, les créateurs de deepfakes ont aussi un rôle à jouer dans cet écosystème. Il souhaite qu'une charte éthique du deepfake soit créée, à la manière du code de déontologie journalistique. Et il se dit prêt à agir pour faire de la prévention : "*Ça va être compliqué mais il faut faire de la pédagogie, des formations. Moi je serais prêt à en parler, à faire des formations aux collégiens, lycéens ou personnes âgées. J'essaie déjà de faire de la prévention certaines de mes vidéos.*"⁵³

Aujourd'hui, à l'heure où je rédige ce mémoire, la prévention contre les deepfakes est loin d'être suffisante pour nous prémunir collectivement de leurs dangers. Encore trop peu de rédactions s'emparent de la question et s'attellent à la lourde tâche de l'éducation aux médias. "*Je pense que le journaliste a un rôle d'éducateur qu'il ne veut pas endosser. Il y a une sorte de résistance comme quoi informer ce n'est pas éduquer. C'est ce qui*

⁵¹ John Bowers, Tim Hwang et Jonathan Zittrain. What should newsrooms do about deepfakes? These three things, for starters. *Nieman Lab*. 20 novembre 2019.

⁵² Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

⁵³ Entretien avec Yanis alias French Faker, réalisé le 23 octobre 2020, annexe p. 42 à 49

*empêche les rédactions d'accéder à l'intérêt des deepfakes, et de se dire qu'on a un rôle à jouer dans cette prévention*⁵⁴, estime Gerald Holubowicz.

Pourtant, les deepfakes ne sont pas qu'un sujet réservé aux journalistes tech. Leur potentiels effets sur l'information et la manipulation de l'opinion concernent toute la profession, comme le résume Gerald Holubowicz : *"Prendre du recul et réfléchir à cette question les aiderait également dans la gestion des fake news, parce que ce sont les mêmes mécaniques, et qui peut le plus peut le moins"*⁵⁵

⁵⁴ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

⁵⁵ Entretien avec Gerald Holubowicz, réalisé le 2 octobre 2020, annexe p. 32 à 41

Conclusion

Ces dernières années, la technologie des deepfakes s'est perfectionnée au point de brouiller définitivement la frontière entre perception et réalité. Des contenus pornographiques amateurs publiés sur Reddit aux vidéos parodiques professionnelles diffusés à une heure de grande écoute sur TF1, les deepfakes se sont définitivement imposés dans notre société. Utilisés à mauvais escient, ils peuvent se transformer tour à tour en outil de chantage, de propagande, et surtout de désinformation. Leur récente accessibilité et facilité de réalisation, propulsée par des applications populaires, aggravent leurs dangers. N'importe qui peut aujourd'hui réaliser un deepfake, sans avoir la moindre connaissance informatique. Mais n'importe qui n'est pas capable de reconnaître un deepfake quand on lui met sous les yeux.

Couplée à la viralité des réseaux sociaux, les deepfakes représentent donc une menace grandissante pour le journalisme, et plus largement la démocratie. Mais même si elle la favorise, la technologie n'est pas responsable de la désinformation. C'est bien le contexte et l'intention avec lesquels les deepfakes sont créés qui les rendent potentiellement dangereux. Au bout du compte, il ne s'agit que d'un énième outil de désinformation, certes plus puissant, car il trompe deux de nos sens — notre vue et bientôt notre ouïe —, mais la menace des deepfakes s'inscrit dans un problème plus profond : celui des fake news, et de la crise de confiance envers les médias.

Face aux deepfakes utilisés comme armes de désinformation, il ne suffira donc pas de se reposer uniquement sur des solutions technologiques, pour le moment irréalistes à mettre en place, ou sur les réseaux sociaux, qui tentent de réguler les deepfakes tout en laissant leurs algorithmes favoriser la polarisation de l'opinion et la propagation de fausses informations. Les journalistes ont eux aussi un rôle crucial à jouer dans ce combat, et le fact-checking traditionnel ne suffira pas.

Pour combattre les deepfakes, il faut déjà s'y intéresser, et comprendre les mécanismes de la désinformation. Il n'y a ni solution immédiate, ni recette miracle face à cette nouvelle menace pour le journalisme, mais une solution qui nécessite beaucoup de temps et d'énergie, et qui n'apportera pas la satisfaction d'un résultat immédiat : la prévention à travers l'éducation aux médias. Car les deepfakes ne vont ni disparaître, ni être interdits. Il va donc falloir apprendre à vivre avec, d'où l'importance d'apprendre aux gens à vérifier et comparer leurs sources, lire une information en détectant les biais d'opinion et développer leur esprit critique.

D'autant plus que le plus gros danger des deepfakes reste peut-être la simple idée de leur existence, plus que les contenus manipulés en eux-mêmes. Si l'on se met à confondre le faux pour du vrai, n'importe qui peut alors affirmer que ce qui est vrai est faux. Le "liar's dividend" n'a sûrement pas fini de servir aux plus mal intentionnés. Il est donc plus important que jamais pour les journalistes de faire preuve de transparence et

de pédagogie, afin de rétablir la confiance brisée envers les médias, et de retrouver leur étiquette de garants de la véracité de l'information.

Une approche pédagogique des deepfakes permettrait également de ne pas limiter cette technologie à ses dangers, et d'aborder plus sereinement son potentiel pour l'art, l'humour et tous les métiers de l'audiovisuel. La technologie des deepfakes peut d'ailleurs être mise au service du journalisme. Elle permet par exemple de protéger des sources et des témoins souhaitant rester anonymes, tout en leur donnant un visage, comme l'a fait le journaliste américain David France dans son film documentaire *Welcome to Chechnya*, ou encore de doubler une interview en langue étrangère sans dénaturer les intonations naturelles de la personne interrogée ni recourir aux sous-titres.

La technologie des deepfakes représente sûrement autant d'opportunités à explorer que de dangers à combattre, voilà pourquoi il est primordial de suivre son évolution de près. D'autant plus qu'une autre menace se dessine pour les journalistes : l'émergence des textes générés par intelligence artificielle, parfois surnommés "fake texts" ou "readfakes". Des algorithmes — comme celui d'Open AI, financé par Elon Musk — sont aujourd'hui capables d'écrire un article complet, à partir d'un ou deux mots donnés comme sujet, en quelques secondes à peine. Non seulement ces "fake texts" représentent des armes de désinformation plus dangereuses que les deepfakes vidéos et audios⁵⁶ — car ils nécessitent beaucoup moins de temps et de moyens à produire —, mais en plus, ils pourraient menacer directement le métier de journaliste. Pour réduire leurs coûts, certaines rédactions⁵⁷ ont en effet déjà remplacé des journalistes par ces algorithmes d'intelligence artificielle qui savent rédiger des articles en un temps record. Si cette technologie permet bien sûr de soulager les médias dans certaines de leurs tâches répétitives, l'automatisation du métier de journaliste a de quoi inquiéter.

La révolution de l'intelligence artificielle qui se dessine à l'horizon va-t-elle mettre en danger la profession ? L'avenir nous le dira. Heureusement, pour le moment, aucun algorithme n'est capable de remplacer l'esprit critique et l'intuition d'un(e) journaliste en chair et en os.

⁵⁶ Renee Diresta. (31 juillet 2020). AI-Generated Text Is the Scariest Deepfake of All. *Wired*

⁵⁷ Lucia Moses. (14 septembre 2017). *The Washington Post's* robot reporter has published 850 articles in the past year. *Digiday*

Sources

Publications officielles et dictionnaires :

Journal Officiel de la République Française — Vocabulaire de la culture : édition, médias et mode (liste de termes, expressions et définitions adoptés), n°0125 texte n°97, 23 mai 2020

Merriam Webster (s.d.) — Deepfake. Dans *Merriam Webster Dictionary*.

Gaîté Lyrique (mis à jour le 02.09.2020) — Cheapfake. Dans *Lexique de la Gaîté Lyrique*

Travaux de recherche et formations en ligne :

"Chapitre 2 : Identifier les deepfakes", Les contenus médiatiques manipulés [formation en ligne]. *Reuters*. Décembre 2019

"Deepfake.media". *Journalism.design* [site Internet]. Gérald Holubowicz. s.d.

John Bowers, Tim Hwang et Jonathan Zittrain. What should newsrooms do about deepfakes? These three things, for starters. *Nieman Lab*. 20 novembre 2019.

"Deepfakes and cheap fakes, The Manipulation of Audio and Visual Evidence" [rapport de recherche]. *Data & Society*. 18 septembre 2019

"The State of Deepfakes — Landscape, Threats, and Impact" [rapport de recherche]. *DeepTrace*. Septembre 2019

Dr Richard Fletcher. The truth behind filter bubbles: Bursting some myths. *Reuters Institute*. s.d.

Articles :

Gérald Holubowicz. (16 octobre 2020). Les deepfakes, une « arme d'illusion massive » ?. *Institut national de l'audiovisuel, La revue des médias*

Aaron Holmes. (26 septembre 2020). Popular deepfake apps are making it easier than ever to make AI-powered manipulated videos — spawning new memes, and an increased potential for abuse. *Business Insider*

Mathilde Cousin. (31 août 2020). Présidentielle américaine : Non, Joe Biden ne s'est pas endormi pendant cette interview, qu'il n'a d'ailleurs même pas donnée. *20 Minutes*

Renee Diresta. (31 juillet 2020). AI-Generated Text Is the Scariest Deepfake of All. *Wired*

Les Observateurs. (29 juin 2020). Attention, cette vidéo de l'ambassadeur français au Cameroun est un "deepfake". *France24*

Gerald Holubowicz. (27 juin 2020). George Floyd et les deepfakes: la folle théorie de Winnie Heartstrong, *Journalism.design*

Will Douglas Heaven. (12 juin 2020). Facebook just released a database of 100,000 deepfakes to teach AI how to spot them. *MIT Technology Review*

Robert Chesney, Danielle Citron et Hany Farid. (11 mai 2020). All's Clear for Deepfakes: Think Again. *Lawfare*

Mehdi Chouiten. (30 mars 2020). Deepfake : Menace Imminente Ou Outil D'Aide À La Production ?. *Forbes*

Nilesh Christopher. (18 février 2020). We've Just Seen the First Use of Deepfakes in an Indian Election Campaign, *Vice*

Ian Sample. (13 janvier 2020). What are deepfakes – and how can you spot them?. *The Guardian*

Usbek&Rica. (15 décembre 2019). "Avec les deepfakes, n'importe qui peut devenir un démon de la manipulation". *Usbek&Rica*

Morgane Tual. (24 novembre 2019). On a essayé de fabriquer un deepfake (et on est passé à autre chose). *Le Monde*

Morgane Tual. (3 septembre 2019). A peine lancée, l'application de vidéos « deepfake » Zao suscite des inquiétudes. *Le Monde*

Harold Grand. (24 mai 2019). «Deepfake»: une vidéo trafiquée de Nancy Pelosi relayée par des proches de Trump. *Le Figaro*

Solange Ghernaouti. (7 avril 2019). Contrer Fake news et les Deep fakes par des mesures pragmatiques. *Le Temps, Cybersécurité | Le blog de Solange Ghernaouti*

Rana Ayyub. (21 novembre 2018). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. *HuffPost UK*

Simon Chandle. (4 octobre 2018). Deepfakes 2.0: The terrifying future of AI and fake news. *Daily Dot*

Reporters Sans Frontières. (27 avril 2018). RSF appelle les autorités indiennes à protéger la journaliste Rana Ayyub. *Reporters Sans Frontières*.

Elisa Braun. (20 avril 2018). La viralité d'une fausse vidéo d'Obama met en lumière le phénomène du "deep fake". *Le Figaro*

David Mack. (17 avril 2018). This PSA About Fake News From Barack Obama Is Not What It Appears. *BuzzFeed*

Pierre Schneidermann. (14 février 2018). TensorFlow, l'outil qui a favorisé l'émergence des deepfakes. *Konbini*

Charlie Warzel. (11 février 2018). He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse. *BuzzFeed*

Rob Price. (27 janvier 2018). People are using creepy, cutting-edge AI technology to splice Nic Cage into every movie they can think of. *Business Insider*

Samantha Cole. (24 janvier 2018). We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. *Vice*

Samantha Cole. (11 décembre 2017). AI-Assisted Fake Porn Is Here and We're All Fucked. *Vice*

Fiorenza Gracci. (7 octobre 2017). Comment fonctionne la vision ? *Science & Vie QR n°16* « Nos cinq sens & leurs mystères »

Lucia Moses. (14 septembre 2017). *The Washington Post's* robot reporter has published 850 articles in the past year. *Digiday*

Vidéos, films et émissions de radio :

David France (journaliste réalisateur), (2020), *Welcome to Chechnya* [film documentaire], HBO Films

Ted. (8 octobre 2019). *How deepfakes undermine truth and threaten democracy* | *Danielle Citron* [Vidéo]. Youtube
<https://www.youtube.com/watch?v=pg5WtBjox-Y>

The New York Times. (14 août 2019). *Deepfakes: Is This Video Even Real?* | *NYT Opinion* [Vidéo]. Youtube
https://www.youtube.com/watch?v=1OqFY_2JE1c

Nicolas Martin, *Deepfake : faut-il le voir pour le croire ?*, La méthode scientifique, France culture, 26 juin 2019
Radio-Canada Info. (28 janvier 2019). *Enquête | Deepfake : le vrai du faux* [Vidéo]. Youtube
<https://www.youtube.com/watch?v=1mCwmxW6Xhk>

BuzzFeedVideo. (17 avril 2018). *You Won't Believe What Obama Says In This Video!* 😊 [Vidéo]. Youtube
<https://www.youtube.com/watch?v=cQ54GDm1eL0&t=1s>

Annexes

Entretien avec :

Gerald Holubowicz

Chef de produit numérique chez Conde Nast,
intervenant à l'EDJ Sciences Po et spécialiste des deepfakes

Le 02/10/2020

Comment as-tu commencé à t'intéresser aux deepfakes ?

Un peu comme toi en fait. Je faisais un Executive Master à Sciences Po et j'avais un mémoire professionnel à faire. J'ai découvert les deepfakes quelques semaines avant ma sélection de sujet et je trouvais ça intéressant, à la croisée des chemins entre la tech et le journalisme, mais on peut aussi aller vers la philosophie, ça touche à plein de domaines. Quand j'ai commencé à travailler sur le sujet, en février 2018, j'étais à *Libération*, et je suis allé voir Cédric Mathiot de Checknews pour l'interroger. Globalement lui il ne voyait pas l'intérêt, parce que déjà il n'est pas tech du tout, et par ailleurs leur positionnement à Checknews n'est pas là dessus. Ils fact checkent un peu mais ils sont plutôt dans l'explication et la réponse aux questions qui leur sont envoyées. Donc ce n'était pas étonnant que les deepfakes ne l'intéressent pas, mais effectivement en France je ne sentais pas une fébrilité monstrueuse sur le sujet par rapport à ce qu'on voyait dans les pays anglo-saxons. Après quand je lui ai posé la question, c'était encore vraiment émergent. Il y avait eu un papier sur Motherboard (Vice), il y avait eu quelques trucs mais ce n'était pas encore un sujet très exploré. C'était vraiment le tout début de la réflexion. Dans un premier temps, la réflexion générale était "ça va changer notre point de vue", et ensuite "oh mon dieu ça va être la catastrophe". Je continue à croire aujourd'hui que les deepfakes vont changer notre point de vue, mais que ça ne va pas être la catastrophe.

Toi qui a suivi de près l'émergence des deepfakes dans la sphère journalistique, qu'as-tu observé dans la réception médiatique de cette innovation ?

Il y a eu plusieurs courants. D'abord, il y a eu les gens qui ont bien cerné l'affaire. Ils ont compris que c'était compliqué, que ça n'allait peut-être pas sortir tout de suite mais qu'il fallait commencer à se former sur la question. Ça c'était plutôt une réflexion que je rejoignais. Les journalistes qui réfléchissaient à ça ne voulait pas que ça refasse comme avec les fake news, où on s'était laissé déborder. Donc il fallait anticiper, pour moi c'était plutôt de la prévention. Et puis très vite aux Etats-Unis c'est monté d'un coup, en épingle. Le combo intelligence artificielle, information et Trump était explosif, les médias étaient paniqués. Donc il y a eu cet emballement sur les deepfakes, ensuite on est passés sur quelque chose de plus personnel avec le revenge porn, puis l'inquiétude un peu sécuritaire sur les entreprises qui pourraient se faire leurrer, et après tout le côté politique et attaque à la démocratie. Mais personne n'a parlé du cinéma.

Les deepfakes pornographiques ont beaucoup fait parler d'eux, peut-être plus que les deepfakes politiques. Que penses-tu du traitement qui en a été fait ?

La vague du deepfake pornographique a été intéressante mais je trouve que ça n'a pas été très exploité. Les mouvements féministes, que ma femme surveille, n'en ont pas plus parlé que ça. J'ai même contacté des associations féministes en France qui font des actions contre le cyber harcèlement, et je n'ai pas eu de réponse. Je pensais que la question du revenge porn allait plus interpeler. Il y a eu quelques grosses histoires en Australie et en Inde, et des journalistes ont été cibles de deepfakes pornos, mais à part Rana Ayyub qu'on connaît, on n'a pas les noms. En tout cas il y aurait eu cinq ou six autres cas comme elle, mais les médias n'ont pas voulu nourrir la bête en en parlant.

On parle beaucoup plus des dangers des deepfakes que des opportunités de cette technologie. Est-ce qu'on diabolise trop les deep fakes ?

En fait c'est un débat qu'on n'a pas. L'émergence de cette technologie rejoint l'émergence de la reconnaissance faciale, de l'intelligence artificielle, la 5G et le reste. Ces innovations sont avancées comme une sorte de fatalité, un rouleau compresseur qui doit arriver, auquel on doit se faire et c'est comme ça. Et il n'y a pas de débat de société pour se demander si c'est réellement quelque chose dont on a besoin, si c'est vraiment une évolution souhaitable pour tout le monde. Même pour le cinéma les deepfakes ce n'est pas forcément une bonne chose. Pour l'information c'est aujourd'hui relativement inoffensif même s'il y a certaines choses qui émergent.

En journalisme, l'intérêt le plus évident de la technologie des deepfakes, c'est la protection des sources lorsqu'on veut faire témoigner des gens

Oui, à ce sujet il y a le film de David France *Welcome to Chechnya* qui est vraiment très bien. Et pour faire un avant/après l'existence de la technologie, il faut aussi voir *Z 32* de Avi Mograbi. C'est l'histoire d'un ancien soldat israélien qui a participé à une opération de vengeance dans laquelle deux policiers palestiniens ont été assassinés. Le film est fait sous forme de confession, l'ex soldat raconte sa culpabilité à sa petite amie, et puis il y a aussi une rencontre avec le réalisateur. Et le réalisateur se demande comment faire témoigner quelqu'un qui a commis de tels actes sans lui causer de problème, tout en créant l'empathie pour le spectateur. Il a travaillé avec un créateur d'effets spéciaux qui a mis des masques sur les visages, c'était avant l'arrivée des deepfakes [le film est sorti en 2009, ndlr]. C'est bien fait parce que ça commence avec des masques très *obvious*, et puis progressivement ça termine avec des masques qu'on finit par oublier, même si on ne les oublie pas autant que dans *Welcome to Chechnya*. Ça amène toute une réflexion sur le visage. Pour moi les deux films sont à voir en avant/après, parce que ça pose la question de ce qu'apporte la technologie aujourd'hui, et finalement elle n'apporte pas grand chose.

Est-ce qu'on peut imaginer d'autres usages journalistiques ?

En radio c'est la même chose, pour faire témoigner des gens en les protégeant, car la voix, comme le visage, est reconnaissable, c'est un des éléments qui caractérisent fortement les personnes. Donc ça permettrait de remplacer la voix de la personne pour qu'on ne puisse pas la retrouver. Mais ce ne serait pas une histoire de modulation ou de fréquence différente dans le spectre. Là il s'agirait de substituer la voix par un deepfake de voix. Le deuxième usage possible, c'est la traduction. On pourrait faire parler des gens dans la langue des spectateurs, donc si c'est un Américain, il parlerait français, mais en conservant son ton de voix. Donc en traduction et doublage, c'est intéressant, ça permet d'avoir un rendu un peu plus naturel. C'est ce que permet la technologie des deepfakes, même si c'est presque des artifices.

Tu penses que ça pourrait être amené à se démocratiser dans le journalisme, notamment dans les documentaires télévisés ?

Un journaliste de l'AFP me disait que globalement si ça existe, pourquoi pas, mais en vrai, un bon contre jour, ça fonctionne aussi. La question est de savoir comment ça peut être interprété par les gens. Moi ce qui m'intéresse, c'est de me dire qu'au début on pense qu'un bon contre jour ça suffit, puis un jour un média essaye les deepfakes et puis on finit par en voir un peu partout, dans un contexte qui va être de plus en plus manipulé par ça. Donc l'acceptation de ce phénomène va être de plus en plus facile. Déjà, il y a de moins en moins de gens qui regardent la télévision, donc quand ça va apparaître, ça fera moins de vagues. Si ça apparaît sur d'autres plateformes de distribution, ce ne sera pas forcément vu comme un truc horrible parce qu'aujourd'hui il n'y a plus beaucoup d'innovation qui choquent le public. Donc je pense que ça va se distiller un peu partout, même s'il y aura une latence, et les gens vont s'habituer peu à peu. Et ça c'est le vrai sujet : à quel moment tu acceptes que ce que tu vois n'est pas forcément la réalité, et est-ce que tu te rends compte de ce moment, et est-ce qu'il est encore possible de faire demi-tour quand tu t'en rends compte, c'est à dire probablement très longtemps après ? Et ça, ça veut dire quelque chose. Avec Photoshop, il n'est pas rare qu'on enlève un ou deux défauts avant de publier une photo dans un magazine, mais on n'enlève pas une information, l'image va se vendre comme ça quoi qu'il arrive. Si c'est de la news, on ne le fait pas, si c'est une image de tapis rouge, on le fait. Mais si par exemple l'agence à laquelle tu envoies la photo applique un filtre automatique généré par une intelligence artificielle, qui va modifier je ne sais quoi. A la fin, c'est quoi que tu vois ?

On a le même phénomène avec les filtres sur les réseaux sociaux, on ne voit plus vraiment le vrai visage des gens.

Oui et le contexte actuel est aussi intéressant. On voit de moins en moins les gens en vrai. Quand on était confinés, on ne voyait plus les gens que par l'entremise de la vidéo.

Et maintenant que tout le monde se masque, la seule façon de voir les visages en entier, c'est par la vidéo. Il ne faut pas croire que nous les humains soyons si évolués que ça au point de passer cette transition aussi facilement. Malgré les milliers d'années d'évolution qui nous ont permis d'en arriver là aujourd'hui, notre cerveau est conçu d'une certaine manière, il perçoit des choses. Aujourd'hui on se retrouve dans une situation où on modifie en l'espace de quelques mois l'*output* de ce qu'on va voir. Notre cerveau n'est pas du tout adapté à ça, et dans 10 000 ans notre cerveau sera toujours le même. Peut-être qu'il évoluera si c'est une question de survie, mais pour l'instant il n'a pas évolué. Donc on n'est pas équipés pour se prémunir de ce phénomène. Pour moi c'est ça le plus grand danger, on va considérer qu'on a assimilé ce phénomène, mais on va créer un flou cognitif sur tous les médias qu'on va voir.

Est-ce que c'est le rôle des journalistes de faire de la prévention sur les deepfakes ?

Il y a tellement de problèmes dans les médias aujourd'hui que déjà si on arrivait à se concentrer sur le reste, ça s'améliorerait. Le vrai problème, c'est un problème de culture des journalistes. On a une façon de percevoir le monde qui est intrinsèquement la nôtre et on arrive très peu à se mettre dans les godasses des gens hors du microcosme journalistique. Déjà le fait d'être exposé en permanence aux news crée un espèce de filtre. Moi je regarde souvent des reportages, des photos d'actualité parfois dures, violentes, avec des morts, et ma femme passe à côté et elle est choquée, alors que je ne le suis pas. Et ensuite, en journalisme, on traite de l'immédiat, pas du long terme. C'est pour ça que les journalistes ne s'intéressent pas aux deepfakes en France. Ceux que j'ai interrogés m'ont tous dit "pour l'instant il n'y a pas de deepfake en France", à part celle de l'ambassadeur français Christophe Guillhou au Cameroun il n'y en a pas eu. Donc ils m'ont dit "comme il n'y en a pas, on ne traite pas le sujet." Ils restent vigilants, mais ça ne les intéresse pas plus que ça.

Mais si on se mettait à avoir plein de deepfakes d'un coup en France, est-ce que les journalistes auraient le temps de rebondir puisqu'ils ne l'ont pas anticipé ?

C'est la question que je leur ai posée. Et ils m'ont dit que oui, parce qu'ils parient sur le fait que s'il y a une deepfake sur un sujet, leur enquête sera suffisamment costaud, reposera sur des critères journalistiques rigoureux, qu'ils recouperont l'information etc. Mais bon, s'il faut une, deux ou trois semaines pour détecter le deepfake et le fact checker, si la vidéo est partagée trois jours avant une élection, c'est cuit.

Même sans le risque d'une élection, l'information aujourd'hui circule tellement vite sur les réseaux sociaux que le temps que des journalistes debunk un deepfake, la vidéo aura déjà tournée et des gens y croiront. Et pour dix personnes qui auront vu la vidéo, peut-être que deux personnes iront ensuite lire l'enquête de Checknews ou de l'AFP Factuel, mais les huit autres ne le feront pas, et elles ne sauront pas que

l'information a été démentie. Cette question du délai de réaction des journalistes face aux deepfakes pose problème non ?

Oui et puis on a aussi une idée surdimensionnée de notre importance. Le maximum du tirage d'un journal en France c'est 50 000 exemplaires aujourd'hui. Sur les réseaux ou sur un site de média, c'est pareil. En fait, la portée que l'on a est très faible. La plupart des gens passe des jours voire des semaines sans regarder les infos donc s'ils tombent sur une vidéo ou un son partagé sur Instagram, TikTok, WhatsApp ou peut importe, ils vont être persuadés que ça va être vrai. Donc pour en revenir à la question "est-ce que le journaliste a un rôle à jouer là dessus ?", ma réponse est oui, mais il ne le fera pas parce qu'il n'en est pas capable. Je pense que le journaliste a un rôle d'éducateur qu'il ne veut pas endosser. Je ne sais pas pourquoi, il considère que l'information ce n'est pas de l'éducation, qu'on ne doit pas faire de pédagogie, qu'on doit rester sur les mêmes canaux et la même méthodologie depuis des millions d'années. C'est idiot, c'est absurde, mais c'est comme ça. Même dans certaines écoles de journalisme, des étudiants ont ce discours là. Fondamentalement il y a une sorte de résistance comme quoi informer ce n'est pas éduquer. Parce qu'il y aurait quelque chose de peut-être moins noble ou moins objectif, je ne sais pas. Pourtant, c'est valorisant quand la personne en face de toi comprend et te suit, même si elle n'a pas intellectuellement les mêmes bases que toi. Mais aujourd'hui, on n'est pas dans une démarche différentielle, on traite grosso modo toute l'information ensemble et on demande aux gens de se débrouiller avec ça. C'est ce qui empêche les rédactions d'accéder à l'intérêt des deepfakes, et de se dire qu'on a un rôle à jouer dans cette prévention. Pourtant, prendre du recul et réfléchir à cette question les aiderait également dans la gestion des fake news, parce que ce sont les mêmes mécaniques, et qui peut le plus peut le moins.

Certains médias, plutôt anglo-saxons, s'intéressent quand même au sujet et tentent de faire de la prévention, notamment avec des formations pour sensibiliser les journalistes aux deepfakes, comme celle de Reuters. Il n'y en a pas en France ?

Il faut distinguer le cadre de l'expérimentation dans une entreprise, typiquement l'AFP MediaLab, qui est très cool mais c'est du lab, et puis la rédaction. En général, le lab c'est un mec —à l'AFP c'est Denis Teyssou, à France TV c'est Eric Scherer et il y en a d'autres —, ce sont des gens qui ont atteint un certain âge, qui ont toujours eu un appétit technologique, de prospective, donc un peu les brebis galeuses de la rédaction, suffisamment intéressés par la tech, mais c'est vraiment des inclassables. Et ces mecs là ont créé leur job, ils ont monté leur projet, ils ont fait de l'intrapreneuriat. Mais le reste de la rédaction s'en fiche. Ce n'est pas un truc qui importe au quotidien. Ils peuvent faire de la sensibilisation, faire des ateliers, mais c'est minime.

Si les journalistes décidaient de faire de la prévention contre les deepfakes, comment devraient-ils s'y prendre selon toi ?

Pour moi ce que devrait faire la presse c'est avoir un pôle journalisme scientifique vulgarisateur de la technologie, en plus du journalisme scientifique déjà existant, pour accompagner les gens dans leur usage de la technologie, avoir un aspect serviciel à ce niveau là. Et là on pourrait commencer à expliquer les choses, à expliquer l'écosystème dans lequel les gens se trouvent, les réseaux sociaux, les mécaniques, en faisant en sorte que ces explications ne soient pas uniquement calquées sur l'actualité, en les reliant entre elles, en faisant des corpus. C'est un gros boulot, mais est-ce que ce n'est pas ça l'information finalement ? Aujourd'hui les rubriques tech des médias, c'est de l'actualité des constructeurs, ce n'est pas de l'info. C'est du passage de communiqué de presse sur les nouveaux jeux et téléphones qui sortent, des comparaisons de performance, ça n'a rien à voir avec de l'information. Alors qu'expliquer comment fonctionne Facebook, ce qu'il y a derrière, comment ça impacte notre vie quotidienne, ça c'est autre chose. Donc il faudrait pouvoir structurer ça autour d'un média ou d'une rubrique, faire du serviciel, et à partir de là on pourrait commencer à expliquer les deepfakes aux gens, engager un débat sur l'acceptation de ce procédé à grande échelle. Mais le problème c'est que les journalistes sont souvent technophobes parce qu'ils viennent de sphères intellectuelles et littéraires et pensent que ce n'est pas compatible avec la tech. On les forme à écrire avec des stylos, ils voient la technologie un peu de loin, dans les écoles on forme des journalistes politiques en veux-tu en voilà. C'est pour ça que le débat sur les deepfakes n'existe pas aujourd'hui dans la société. On parle beaucoup de politique, partout en permanence, mais on ne parle pas de ça, parce qu'on n'a pas les gens en capacité d'émettre un message cohérent, intelligent, censé et structuré sur le sujet. C'est très compliqué à mettre en place, mais pour moi le rôle du journaliste est central dans ce débat.

Au vu du contexte médiatique actuel, est-ce que les deepfakes t'inquiètent ?

Pour l'instant ça ne m'inquiète pas spécialement parce qu'il y en a assez peu. Au tout début je me suis inquiété comme tout le monde, puis en fait je me suis dit que ça allait. Déjà ça ne m'inquiète pas sur la partie personnelle, c'est évidemment très embêtant pour les femmes qui sont victimes de deepfakes de revenge porn, mais le volume de vidéos deepfakes qui sont produites n'est pas si important si on compare au volume de vidéos pornographiques classiques qui existent. En terme d'info, la situation des deepfakes pourrait devenir compliquée, mais pour l'instant il manque un élément essentiel, c'est la convergence voix/vidéo. Pour le moment, on a des vidéos qui sont très bien faites, mais il n'y a pas la voix deepfake qui va avec. C'est un problème technique, il y a des gens qui travaillent dessus et c'est plutôt en bonne progression.

Pourquoi la voix est si importante pour qu'une vidéo deepfake soit la plus réaliste possible ?

Dans chaque voix, il y a une complexité d'intonations, de tons, de profondeur, de tensions, que l'humain va percevoir. Entre deux "bonjour" quasiment identiques, toi tu peux percevoir que sur le premier j'étais un peu plus stressé. Mais l'ordinateur en est

incapable, il ne saura pas mettre une intention derrière. On trompe l'œil bien plus facilement que l'oreille. D'un point de vue évolutif on est faits comme ça. L'œil n'est pas si intelligent que ça. Dès qu'on commence à voir des lignes verticales on ne sait plus où on va, dès qu'on a voir des monochromies on perd les pédales. S'il manque des lettres dans un texte, on saura le lire quand même. Tous les trucs qui sont censés nous sauver la vie nous trompent. Donc l'œil n'est pas aussi fiable que la voix, parce que l'évolution ne nous a pas fait nous reposer sur le système auditif et la voix avec la même importance que sur l'œil. On voit très bien les mouvements, donc s'il y a un décalage entre l'image et le son, on va le voir, mais si c'est très bien fait on ne l'entendra pas. Quand on aura la capacité de faire la voix de n'importe qui en deepfake, à partir d'un *sample* de 5 secondes, là il y aura un changement, et un plus grand danger.

Donc pour le moment, les vidéos deepfakes qui circulent sont perfectibles ?

Oui, par exemple Represent US ils ont fait une campagne deepfake que les télés aux Etats-Unis ont refusé de diffuser. Ces deux vidéos là sont super, en 4K, mais la voix n'est pas la bonne, parce que c'est un imitateur, et ça se sent. C'est comme les vidéos deepfakes d'Emmanuel Macron doublées par Nicolas Canteloup, on s'en rend compte. Il y a peu de vrais bons imitateurs qui arrivent à refaire parfaitement la même voix. Donald Trump, tout le monde a essayé de l'imiter, mais la voix ne se place jamais vraiment parfaitement. Les imitateurs ont la gestuelle et tout le reste, mais dans le ton, il y a toujours une différence. Le jour où quelqu'un arrivera à faire un deepfake de voix réussi, il aura gagné, parce qu'après rien qu'avec l'audio ce sera suffisant. On pourra faire du chantage en faisant croire qu'on détient des enregistrements de conversations téléphoniques par exemple. Quand j'ai écouté l'enregistrement audio, qui est un vrai document, de Barack Obama qui parle de Donald Trump assez librement pour la première fois, j'ai d'ailleurs eu peur que ce soit un deepfake.

Est-ce que les deepfakes vont devenir encore plus dangereux dans les années à venir ?

Une fois qu'on saura faire des deepfake de voix, si on le rajoute sur une vidéo mais que la vidéo est trop propre, typiquement si les images viennent d'une conférence de presse, ce sera facile à débunker. Même si elle est diffusée sur les réseaux, il y aura un petit délai mais très vite les journalistes ou autres vont débunker le deepfake. Par contre, ça peut davantage valoir le coup de se servir des deepfakes pour les entreprises je pense. Certains pourraient créer des deepfakes de fausses déclarations pour faire chuter le cours de la Bourse et ramasser l'argent par exemple.

Et dans le domaine politique ?

En fait, dans tous les domaines, il y a forcément un intérêt pour faire pression sur quelqu'un. C'est vraiment ça les deepfakes, c'est l'arme d'illusion massive. Tu n'as pas

besoin de t'en servir pour que ça serve. Tu n'as pas besoin de faire un deepfake ou de le diffuser à grande échelle pour que ce soit utile. Rana Ayyub, quand elle a reçu des deepfakes pornographiques avec son visage, ça lui a fait suffisamment de mal, il n'y avait pas besoin de les diffuser pour la faire chanter. Si tu menaces les gens en leur disant "si tu ne fais pas ce qu'on te demande, on te fait dire ça", c'est déjà suffisant.

Techniquement parlant, on peut déjà le faire depuis plusieurs années, pas besoin que la convergence voix/vidéo soit parfaite pour faire pression sur quelqu'un. Pourtant il n'y a pas eu tant de chantage aux deepfakes que ça.

Oui, seulement ce qu'il n'y avait pas jusque là, c'est la connaissance chez les gens que les deepfakes pouvaient exister, il manquait la plausibilité. Maintenant on sait que ça peut être fait. Et c'est là que ça devient dangereux, c'est de voir l'infusion de cette idée là dans la société. C'est évidemment sexy d'aller vers ce genre de technologies, et en même temps ça créé la confusion chez tout le monde. A un moment, même le plus vigilant va se tromper, retweeter un truc et rentrer dans le jeu.

Justement, comme l'humain peut se tromper, des chercheurs sont en train de mettre au point des algorithmes qui seraient capables de détecter les deepfakes. Est-ce que c'est ça la bonne solution, se fier aux machines puisque l'humain est faillible ?

Oui et non. A l'AFP ils ont testé un certain nombre d'algorithmes qu'on leur soumet régulièrement, et la plupart tombe à côté de la plaque. Pour le moment les algorithmes de détection ne fonctionnent pas à 100%. Et après il faut voir autre chose : c'est impossible à mettre en place de façon réaliste. C'est corrélé avec les législations. Aujourd'hui, si on renforce les législations en mode préventif, qu'on impose aux plateformes de réseaux sociaux d'avoir un système de détection des deepfakes, ça voudrait dire que tous les contenus uploadés devraient passer dans l'algorithme. Le volume de data géré en upload devrait être récupéré par des serveurs, traité en direct, avec une certaine puissance, puis mis sur serveur à disposition, dans des délais raisonnables pour que le service ne soit pas dégradé, qu'il reste comme aujourd'hui pour l'utilisateur, parce que si une vidéo met trois heures à être publiée ça n'a aucun intérêt. Ce n'est pas réaliste. Pour Facebook peut-être à la limite, parce qu'ils ont les moyens, et que la vidéo n'est pas le cœur de leur service. Mais pour Youtube avec son volume de vidéos, ou les plateformes comme Vimeo qui ont beaucoup moins de moyens financiers, c'est impossible. Ce n'est pas une solution immédiate. Et ensuite se pose la question de est-ce qu'un algorithme est capable de faire la compétition avec un autre algorithme ? Oui, mais c'est toujours le jeu du chat et de la souris, il y en aura toujours un qui aura une longueur d'avance sur l'autre.

Quelles solutions sont envisageables pour combattre les deepfakes alors ?

Il y a eu plusieurs méthodes testées, les chercheurs ont regardé les battements du coeur qui se voyaient en vidéo, le clignement des yeux, mais tout ça a été balayé. A chaque fois qu'on pense trouver un truc, ça ne marche pas. Il y a une autre piste, celle de la blockchain. Mais c'est pareil, ça pose deux contraintes. La première c'est qu'il faudrait que toute la chaîne de production soit cryptée, dès la sortie d'usine, tous les appareils devraient utiliser un système de cryptage façon blockchain pour ensuite délivrer du contenu qui soit traçable. Et deuxième contrainte, outre le fait que la blockchain consomme une énergie de dingue, ça voudrait dire que tout serait traçable. Légalement, en terme de vie privée, ce serait un gros problème.

Si la technologie n'est pas la solution contre les deepfakes, quelle serait la bonne solution alors ?

Je pense qu'on essaie d'avoir une solution courte mais on n'en a peut-être pas. C'est un peu ça qui fait flipper, on n'a pas de solution technologique. Il faudrait de l'éducation. Sauf que la notion d'éducation aux médias est aussi un problème, ça revient à dire aux gens qu'ils sont trop bêtes pour pouvoir utiliser les médias dans une certaine mesure, c'est hyper condescendant. Comme ça vient toujours de la part des médias, ça passe très mal forcément. Et si ça ne passe pas par les médias, c'est l'Education nationale, donc c'est très compliqué aussi. En plus il y a des études qui montrent que même en étant régulièrement exposé à cette éducation aux médias, quand on sort du lycée, vers 21-22 ans, on revient à la case zéro. Pour moi, on en revient au rôle crucial des journalistes, qui est peut-être un peu hypothétique et imaginaire, mais c'est l'instauration d'un débat sur la place de la technologie dans nos sociétés, comment elle entre dans nos maisons, dans nos vies et comment on accepte ça. Il faudrait qu'on ait une capacité à réfléchir là-dessus et à se dire qu'en fait, on n'est pas obligés d'accepter, ou qu'on peut l'accepter sous certaines conditions, ou d'avoir conscience que si on accepte des technologies qu'on ne maîtrise absolument pas, il y aura des conséquences. Déjà si on arrive à faire ça, si tout le monde développe cette conscience là et est capable d'articuler des idées, pas forcément de façon très intellectuelle, ce serait déjà le début de quelque chose.

Au quotidien, c'est quand même mentalement fatigant de remettre absolument tout ce que l'on voit en question.

Oui surtout que c'est partout, dans tous les médias, la télé, Youtube, en permanence. Typiquement au cinéma, on peut rajeunir les gens, les vieillir, ou les refaire vivre.

La différence c'est qu'au cinéma, on sait que c'est de la fiction, les gens auront donc moins tendance à confondre les deepfakes avec la réalité non ?

Tu penses ça parce que le phénomène est encore récent. Mais culturellement ça va finir par être accepté. Si dans 20 ans tu vois Samuel L. Jackson dans un film qui vient de

sortir, il n'aura pas la même tête qu'aujourd'hui, mais en même temps si parce qu'il aura déjà été numérisé. Dans le film *The Irishman*, on voit Robert De Niro jeune et vieux, c'est très convaincant. Mais si je regarde la version deepfake du film, je ne sais pas laquelle est la vraie. Imagine un *Star Wars* avec Harrison Ford, et puis tu vas voir un film fait par des fans où l'acteur sera complètement remplacé en deepfake, et ce film là va être tellement bon qu'il va éclipser l'autre, à la fin. Au bout de deux générations, il n'y a plus personne qui sait dire quel est le vrai film. Donc c'est très compliqué. A partir du moment où tu ne réfléchis pas à la technologie, tu autorises la technologie à prendre le dessus et à faire de toi ce qu'elle veut. Ce n'est pas forcément mal au début, mais au bout d'un moment ça devient de l'illusion complète.

Quels sont les risques de ce monde d'illusion cette illusion sur le journalisme ?

On peut se poser la question des conséquences à court terme des deepfakes sur le journalisme. Il n'y en aura aucune en temps que telle. Un deepfake de Donald Trump sera débunké, ça prendra un peu de temps mais c'est tout. L'objet deepfake ne causera pas de problème, c'est l'idée des deepfakes qui va en poser. La preuve : Winnie Heartstrong a publié un document qui affirme que le meurtre de George Floyd est un deepfake. On a quand même quelqu'un qui a pris le temps d'écrire 23 pages complètement absurdes. C'est à la fois drôle mais elle a quand même eu une plateforme pour en parler, notamment dans les réseaux alt-right, et une partie de cette audience l'a cru. Quand Donald Trump dit à ses supporters "allez voir dans les bureaux de vote comment ça se passe", en sous-entendant qu'il va y avoir des problèmes, il souffle sur la braise. Il parle à des gens qui sont déjà mobilisés et motivés et qui le 3 novembre vont aller dans les bureaux de vote et s'attendre à voir tout et n'importe quoi. Et il va y avoir des incidents. Pas parce qu'il doit forcément avoir des incidents, mais parce que ces gens là pensent qu'il va se passer quelque chose, et ce sont eux qui vont créer les incidents. Tu créés ex nihilo un truc qui n'existe pas, c'est comme le Pizza Gate. Et ça, fondamentalement, c'est un problème.

Donc pour toi c'est vraiment le phénomène de la perception de la réalité qui va poser problème, plus que les deepfakes eux-mêmes ?

Oui je pense que c'est plus ça le vrai danger. Et le journaliste aura toujours son rôle de fact checker à jouer bien sûr. Mais je suis aussi pour la récupération technologique des médias, pour le moment on est juste dépendants. Aujourd'hui les médias ne peuvent pas retirer un doigt de Google ou de Facebook sans que tout s'écroule. Il faut que les rédactions se remobilisent autour de ces questions de compréhension de la technologie, de comment ça les affecte elles en premier, pour ensuite savoir comment faire passer l'information aux gens et leur dire que comme pour la politique, l'environnement, l'agriculture, on a le droit d'avoir un débat sur la technologie et de se poser la question de l'impact qu'elle a sur nos vies.

Entretien avec :
Yanis, alias "French Faker"
Créateur de deepfakes professionnel
Le 23/10/2020

Tu es connu sur les réseaux sous le pseudonyme French Faker. Est-ce que tu peux te présenter plus en détails (nom, âge, métier) ?

Sans rentrer dans les détails, je m'appelle Yanis, j'ai moins de 30 ans, et j'ai ma société spécialisée dans le deepfake, French Faker. Mes revenus principaux viennent aujourd'hui des deepfakes, je travaille tous les jours avec l'émission C'est Canteloup. TF1 est mon contrat principal mais j'ai aussi d'autres contrats vidéos, musicaux, et des documentaires qui arrivent mais je ne peux pas en dire plus. Pour le moment je suis seul dans ma société, mais je devrai recruter très rapidement si les contrats continuent à arriver.

Comment est-ce que tu as appris à maîtriser la technologie deepfake et pourquoi tu t'y es intéressé ?

J'ai toujours été passionné de technologie, j'ai toujours aimé les gadgets, les innovations. Je me suis penché sur chaque grosse innovation technologique, comme les crypto monnaies, la 3D, toujours en autodidacte. Ce qu'il faut savoir aussi c'est que j'ai un parcours IT, j'ai toujours été dans la vente de produits technologiques et informatiques, et sur mon temps libre j'ai appris à coder, j'ai toujours bidouillé. Je suis venu à m'intéresser aux deepfakes un peu comme tout le monde, avec l'application Zao au printemps 2019. Zao est arrivée de Chine et j'ai trouvé ça vraiment hallucinant, du coup je voulais tester, mais on ne pouvait pas parce qu'il fallait absolument chinois. J'ai réussi à me procurer un numéro chinois, téléchargé l'application, j'ai testé et j'ai trouvé ça génial mais limité. Donc j'ai cherché sur le web ce qui existait de similaire, des solutions à la base Open Source, et de fil en aiguille je m'y suis mis. J'ai commencé à faire des deepfakes pour faire rire mes amis, la vidéo de Marine Le Pen avec le voile c'était au départ juste pour faire rire mes collègues. De fil en aiguille je suis arrivé à faire de la vidéo mon métier, mais c'est un accident.

Toi qui connaît bien l'aspect technique, est-ce que tu dirais que la réalisation de deepfakes est accessible à tous ?

Oui, produire un deepfake c'est à la portée de plus ou moins de tout le monde. C'est comme tout, dans l'absolu si on a envie on y arrive. Quand j'ai commencé, c'était bien plus difficile, mais ça va devenir de plus en plus facile et demander de moins en moins de technique. A l'heure actuelle pour faire un deepfake réussi, auquel on peut croire, ça demande quand même de la technique. Mais n'importe qui peut faire un deepfake basique maintenant avec l'application Reface, qui est le Zao européen.

Beaucoup de deepfakes sont en réalité des cheapfakes, des vidéos réalisées plus grossièrement, sans intelligence artificielle. Quelles sont tes exigences de qualité ?

Moi je me considère comme un artisan du deepfake, je vais prendre plusieurs jours de mon temps à faire quelque chose que j'estime très réussi. Après ça dépend des clients, certains clients me demandent entre sept et neuf deepfakes par jour, donc là c'est de l'industrialisation, mais de bonne qualité quand même.

Est ce que tu peux m'expliquer le processus de production d'un deepfake vidéo ?

C'est relativement simple, il y a une technicité mais plus sur la réalisation qu'il y a derrière, sur la post production. Pour réaliser un deepfake il y a différentes étapes :

- 1) Il faut récupérer des vidéos sources de la personne qu'on veut deepfaker, en très bonne qualité, avec une palette d'expressions si possible fournie (avec des sourires, où on voit ses dents etc), avec différentes luminosités. Au niveau de la durée, pour faire des deepfakes de qualité, je trouve que plus c'est long mieux c'est. Ça donne du grain à moudre à la machine. Je dirais qu'il faut 10-15 minutes de vidéos sources, mais bien sélectionnées, avec plusieurs profils, expressions, luminosités. C'est une étape qui est très chronophage.
- 2) Il faut avoir une vidéo cible, où on voit le visage de la personne bien entendu, en bonne qualité et avec une bonne luminosité.
- 3) Ensuite, on découpe ces vidéos image par image, on extrait les têtes sur les images sources et les images cibles.
- 4) Puis on fait tourner un algorithme qui va « apprendre » les visages de ces personnes, et les recréer lui même. C'est l'algorithme qui fait l'alignement, il place des points clés sur les visages. On va lui montrer plusieurs images, plusieurs fois par seconde idéalement, en batch, c'est à dire qu'on va lui montrer 4 ou 5 images d'un coup et lui demander de faire une moyenne. On fait tourner ça quelques jours, selon la qualité de la vidéo. Moi j'en fais plusieurs par jour donc j'ai modifié l'algorithme Open Source pour régler certains paramètres, pour que ça soit le plus en HD possible, c'est pour ça que j'ai des gros PC.
- 5) Donc l'algorithme va apprendre les visages et va apprendre à calquer un visage sur l'autre. Il va voir qu'à un moment une personne ouvre la bouche, il va calculer la moyenne qu'il a déjà de l'autre personne pour lui faire ouvrir la bouche aussi.
- 6) Puis on fusionne tout simplement. Et ça donne un deepfake.

Après il y a plein de paramètres à rentrer, comme la netteté, le type de luminosité, le type de masque. Les deepfakes ont commencé avec des masques qui n'allaient que du front au menton, maintenant on peut faire tout le visage voire toute la tête.

Et une fois qu'on a un deepfake vidéo, que faut-il faire pour la voix ?

Pour l'émission C'est Canteloup, c'est eux qui rajoutent la voix de Nicolas Canteloup. Rajouter une voix d'imitateur ça se fait très facilement sur Adobe Première, ce n'est pas compliqué. Après pour créer une voix en deepfake, c'est le même procédé que pour les deepfakes vidéos, sauf qu'à la place des images on donne à l'algorithme des voix. Mais c'est un algorithme un peu plus poussé parce que c'est beaucoup plus évident de détecter quand une voix est fausse, donc ça prend un peu plus de temps. En plus de la voix, l'algorithme doit apprendre à parler, c'est à dire que chaque syllabe et voyelle, il doit apprendre à les dire, pour qu'après il puisse parler. Toute cette partie là est passionnante. Pour créer une voix en deepfake c'est aussi plus compliqué parce qu'il ne suffit pas de 10 ou 15 minutes de fichiers sources comme en vidéo, il faut des heures et des heures d'enregistrement et leurs retranscriptions. Pour des discours d'Emmanuel Macron c'est relativement simple, parce que je peux les trouver déjà retranscrits, mais si on veut une personnalité publique ou autre il faut tout retranscrire soi-même.

Sur tes réseaux sociaux, tu publies plutôt des deepfakes humoristiques, des scènes de films. Est-ce une volonté de ta part de rester sur des sujets légers en se servant de cette technologie ?

Oui, ma chaîne YouTube c'est vraiment pour rigoler, j'ai envie de faire du divertissement plus qu'autre chose. Mais je peux faire un peu ce que je veux dessus, j'ai déjà fait des vidéos où je fais une fausse théorie du complot et des vidéos un peu préventives. J'ai pour idée d'en faire une très qualitative bientôt. Donc oui mon domaine principal c'est le divertissement, mais aussi la prévention. C'est pour ça aussi que je fais des scènes qui sont peu probables. Marine Le Pen avec le voile, personne ne va y croire. Quand on est deep faker professionnel, je pense qu'il y a une éthique à avoir, celle de ne pas tromper les spectateurs.

Est-ce pour cette raison que dans l'émission C'est Canteloup, on voit toujours son crâne et que le deepfake ne semble pas abouti ? Est-ce une demande de la production pour s'assurer que les téléspectateurs comprennent tout de suite que c'est une vidéo truquée ?

Oui totalement. Moi en tant que deep faker professionnel, ça me gêne parce que j'ai envie de produire quelque chose de plus qualitatif, mais c'est tout à fait louable de la part de la production de vouloir garder le crâne chauve apparent. Au moins, ça reste humoristique et on sait tous qu'il s'agit d'un deepfake. Mais j'arrive aussi à prendre un discours d'Emmanuel Macron et à lui faire dire n'importe quoi, ce n'est pas encore parfait, mais une partie de la population pourrait y croire.

Est-ce que tu as forcément besoin de la voix d'un imitateur comme Nicolas Canteloup ?

Pour le moment oui, mais je suis en train de travailler sur du deepfake de voix, ça devrait arriver très prochainement. Ça existe déjà en anglais, avec des qualités variables forcément, mais je veux absolument faire du français donc je suis en train d'y travailler.

A partir du moment où on pourra aussi créer la voix, ça rendra les deepfakes encore plus dangereux non ?

Oui, totalement. J'y vois une certaine dualité. D'un côté ça me fait peur, c'est aussi pour ça que je m'y intéresse, pour mieux comprendre l'envers du décor et pourquoi pas pouvoir mieux l'expliquer à d'autres personnes. Mais une autre partie de moi se dit que finalement avec Photoshop on peut aussi faire ce que l'on veut et ce n'est pas pour autant qu'on est noyés de fake news à cause de Photoshop. Ça arrive, c'est une dérive de Photoshop, mais Photoshop est utilisé pour bien d'autres aspects. La technologie des deepfakes va devenir artistique je pense. Il y aura forcément des dérives, et malheureusement pour le moment les dérives sont plus importantes que l'apport artistique. Mais je pense que c'est amener à changer.

C'est vrai qu'on parle beaucoup moins du potentiel artistique de cette technologie, les médias ont tendance à plus parler de ses dangers. Est-ce que tu trouves qu'on diabolise trop les deepfakes ?

Oui totalement. On le diabolise, et je trouve qu'on pense que l'humain est foncièrement mauvais. Sans rentrer dans de la psychologie de comptoir, c'est sûr que certaines personnes l'utilisent déjà pour des mauvaises choses, mais ça peut aussi faire des choses extraordinaires. On peut faire revivre des acteurs, quand on arrivera à mieux deep faker la voix ça va être extraordinaire. Par contre encore une fois, artistiquement on avance mais éthiquement peut être qu'on recule. Est-ce que c'est normal de faire revivre un acteur, de refaire sa voix ? A qui appartiennent les droits ? Qu'en pense la famille, le public ? C'est des questions qui se posent, mais je trouve que c'est passionnant.

Ces questions n'ont pas encore de réponses légales, il n'existe pas de loi relative aux deepfakes. Tu penses qu'il faudrait légiférer ?

Oui, il faut légiférer. Justement pour que les créateurs aient une charte éthique et puissent faire du deep fake de qualité, il faudrait légiférer.

Il n'y a pas encore de législation gouvernementale mais des entreprises ont déjà créé la leur. Facebook, Twitter et d'autres réseaux sociaux ont récemment interdit la publication de vidéos deepfakes sur leur plateforme. Est-ce que ça t'a déjà posé problème ?

Sur Facebook oui, ils ont détecté plusieurs fois que les vidéos que je publiais sur la page French Faker étaient fausses, et du coup ils les ont enlevées. Et j'ai l'impression aussi que, même si je n'ai pas énormément de likes, avant j'en avais plus. Donc j'ai l'impression que quelque chose s'est passé dans l'algorithme. Ce n'est pas un problème pour moi, je fais ça, pour rigoler pas pour les likes. Et puis ce n'est pas forcément une mauvaise chose qu'ils suppriment les vidéos deepfakes.

Tu estimes qu'il vaut mieux que les plateformes suppriment tous les deepfakes, quitte à enlever les vidéos humoristiques comme les tiennes, plutôt que d'autoriser toutes les vidéos, dont celles potentiellement dangereuses ?

C'est difficile à dire. D'un côté je me dis « laissons tout et les utilisateurs pourront trier eux-mêmes ». Mais il y a des personnes qui sont plus crédules que d'autres, je pense notamment aux parents, aux générations qui n'ont pas eu accès à Internet dès leur plus jeune âge et qui croient déjà un peu à tout ce qu'ils lisent sur les réseaux sociaux. Si on leur rajoute une couche de vidéo ça va être pire. Donc supprimer tous les deepfakes c'est peut être la solution pour le moment. Mais peut-être que juste mettre un petit encart qui précise que cette vidéo est une fausse news suffirait.

En parlant de crédulité, est-ce aux gens d'être individuellement plus responsables et de questionner ce qu'ils voient, ou est-ce aux créateurs de deepfakes, aux réseaux sociaux, et même aux médias et à l'Education nationale de faire de la prévention ?

Mon rêve absolu serait de réaliser un spot télévisé, je le ferais gratuitement pas de problème, dans lequel je ferais dire n'importe quoi à Emmanuel Macron, comme des propos anti masque par exemple, et à la fin lui faire dire « ce n'est pas parce que ça va dans votre sens que c'est la vérité ». Les créateurs ont un rôle à jouer dans l'éducation aux deepfakes, en tout cas c'est mon avis, je sais que d'autres s'en fichent totalement. Moi en tout cas je trouve ça intéressant de faire de la prévention. Et c'est aussi aux médias de dédramatiser les deepfakes, d'expliquer sans entrer dans les détails comment ça marche, que même dans les vidéos il peut y avoir des fakes news.

C'est difficile de répondre à cette question, parce que si on amène trop de craintes chez les gens, on ne va plus croire non plus aux vraies informations, et on va arriver dans un monde où l'on se méfierait de tout et où l'on ne saurait plus quoi croire.

Il faudrait dès l'école, et ça je suis prêt à militer pour, expliquer au delà des deepfakes, comment vérifier des sources quand on lit un article, expliquer le principe des chambres d'écho sur les réseaux sociaux. Quand on a plein d'informations qui vont dans notre sens sur notre fil Twitter par exemple, ce n'est pas forcément que ce qui est dit est vrai. C'est facile de rester dans ces chambres d'écho parce qu'on y est bien, tout le monde pense pareil que nous, mais finalement ce n'est peut être pas cela la vérité. La vérité c'est des nuances de gris plutôt que du tout noir ou blanc, et ça il faut l'inculquer dès l'école.

Pour rebondir sur ton idée de spot télévisé, Solidarité Sida avait fait quelque chose de similaire en octobre 2019. Dans une vidéo, on voit Donald Trump dire qu'il a éradiqué le sida. Au bout de 36 secondes de discours, un bandeau affiche « ceci est une fake news, la première fake news qui peut devenir vraie ». Cette campagne de sensibilisation a reçue de nombreuses critiques car la vidéo était longue, et le moment où l'on apprend que c'est un deepfake arrive tard, alors qu'on sait bien que beaucoup de gens ne vont pas jusqu'au bout des vidéos. Est-ce que tu penses que dès le début il aurait fallu mettre un encart précisant qu'il s'agissait d'un deepfake ?

Le spot était très bien fait. Mais oui c'était un peu maladroit. Je pense qu'il ont voulu faire un peu la même chose que mon idée, c'est à dire faire dire à Donald Trump un truc trop gros pour qu'on y croit, mais on y a cru, parce qu'on a tellement envie d'y croire au fait que le VIH soit vaincu. Je pense que c'est sur cet aspect là qu'ils se sont trompés. S'ils avaient dit quelque chose de vraiment impossible comme « on a ressuscité un mort », personne n'y aurait cru. Donc c'était maladroit, mais je ne pense pas que cela partait de mauvaises intentions. Et je ne pense pas que le problème vienne du fait que ça ne soit révélé qu'à la fin. On peut ne dire que c'est un deepfake qu'à la fin, mais il faut que le message soit plus gros, plus improbable que ça.

La majorité des vidéos deepfakes qui existent aujourd'hui sont des vidéos pornographiques, réalisées avec les mêmes algorithmes Open Source que ceux que tu utilises sûrement. Est-ce que dans le cercle des deepfakeurs professionnel, il t'est déjà arrivé de tomber sur ce genre de vidéos et de créateurs ?

A la base, la technologie n'a pas été créée pour ça, mais elle a été largement utilisée pour ces vidéos pornographiques. Déjà moi, il faut le noter, je n'en ai jamais fait, je n'en ferai jamais, je trouve ça totalement abjecte, j'assimile ça à du viol d'une certaine manière. Mais malheureusement, l'industrie du porno est souvent en avance sur les innovations technologiques. D'ailleurs, certains deepfakes pornographiques sont exceptionnels de technicité même si les algorithmes n'ont pas été créés pour ça. J'ai été dans des groupes de créateurs de deepfakes où certains faisaient des deepfakes pornographiques, moi je ne disais rien à ce sujet, j'écoutais juste pour en apprendre plus sur la technicité, parce que certains étaient vraiment des machines de guerre du deepfake. Ils étaient vraiment en avance. Et c'est un business qui est très lucratif, bien plus lucratif je pense que moi qui travaille pour TF1, c'est dire. Ils inondent les plateformes classiques, dont ça fait de l'appel au clic, par exemple en disant que c'est Emma Watson dans une vidéo porno, alors que ce n'est pas elle. Et après il y a des requêtes directes.

C'est à dire que des gens lambdas peuvent demander à un créateur de produire un deepfake pornographique avec le visage de n'importe qui ?

Oui, ça peut être une requête d'une vidéo de votre voisine, c'est horrible. Mais juste avec des vidéos qu'ils trouvent sur Facebook ou ailleurs, les créateurs peuvent en faire un deepfake porno. Sur les forums ou discussions Telegram entre professionnels du deepfake, j'ai déjà vu ce genre de requête. Et je pense qu'après ils se font rémunérer en crypto monnaie pour qu'on ne puisse pas remonter la trace.

C'est pour ça qu'il faut faire très attention à ce qu'on met sur les réseaux, moi je ne montre jamais mon visage. Ces deepfakes pornographiques peuvent après être utilisés pour faire du chantage, surtout à des personnalités politiques ou des célébrités. C'est un des aspects des deepfakes qui me font très peur et qu'il faut combattre. Mais ça se faisait déjà avant avec Photoshop, et à un moment ça se faisait aussi beaucoup avec l'application Nudify. Grâce à l'IA, on pouvait envoyer une photo de quelqu'un habillé à l'application et l'algorithme imaginait le corps de la personne nu. Avant les deepfakes, ça a commencé comme ça.

Tu évoquais tout à l'heure l'idée d'une charte éthique. Est-ce que tu dirais que la majorité des créateurs de deepfakes sont dans ton cas et réfléchissent aux questions d'éthique ?

Non. D'après ce que j'ai lu et ce j'ai vu, non. Il y a des chiffres qui sont aberrants, je crois que plus de 90% des deepfakes qui circulent sont des contenus pornographiques, c'est gravissime. En plus je ne pense pas que ce soit très difficile à combattre, parce qu'il n'y a que quelques nids, il suffirait de les éradiquer ou de les déréférencer. Cela suffirait amplement, au pire ils resteraient entre eux dans leur groupe de 20 personnes.

Les contenus pornographiques représentent la majeure partie des deepfakes, mais un autre type de vidéos inquiète encore plus : les deepfakes politiques. Publiées sur les réseaux sociaux, ces vidéos peuvent devenir virales avant que des journalistes aient le temps de les fact checker. Que penses-tu de ce phénomène ?

Tu as tout à fait raison, surtout avec l'ère Trump, on a d'abord une news qui apparaît et ensuite elle est fact checkée mais ça peut faire très très mal entre temps. Donald Trump a déjà partagé des vidéos qui n'étaient pas des deepfakes, mais des vidéos modifiées, de Joe Biden et de Nancy Pelosi. Donc oui ça va malheureusement arriver. Et je ne comprends pas qu'il n'y ait aucune répercussion suite au partage de vidéos truquées. En plus c'est facilement vérifiable, surtout si ça a été fait sciemment, et pour Donald Trump on a aucun doute la dessus. Donc ça me fait très peur. Après il faut compter sur l'intelligence de chacun, parce que dès qu'une vidéo a été partagée, on peut voir dans les commentaires des démentis. Encore une fois, les deepfakes sont un problème, mais je pense que le plus gros problème ce sont les chambres d'écho sur les réseaux sociaux.

Effectivement il n'y a eu aucune répercussion suite au partage de ces vidéos par Donald Trump, parce qu'aucun pays n'a mis en place de loi propre aux deepfakes. Est-ce que ça veut dire qu'il n'y a jamais eu de condamnation ?

Non il n'y a pas de loi spécifique mais je crois que ça peut tomber sous le coup des loi contre la diffamation. Après ça dépend si c'est un deepfake humoristique ou pas. Mais je sais qu'en Australie, un homme a été condamné après avoir fait un deepfake de revenge porn sur une femme qui n'avait rien demandé et à qui il a pourri la vie. Et au Japon aussi, plusieurs personnes ont été attrapées, mais pour l'instant de ce que j'ai lu c'est toujours pour du revenge porn. Parce que le mal qui est fait avec ce genre de vidéo est facilement quantifiable, donc condamnable.

Même si tu es un deep faker professionnel et pas un journaliste, penses-tu que la technologie des deepfakes représente une menace pour les journalistes ?

Oui je pense que c'est clairement une menace pour le journalisme, mais que ce sera une opportunité grandiose pour le domaine artistique. On va peut-être apprendre à vivre avec, comme Photoshop. Ça va être compliqué mais il faut faire de la pédagogie, des formations. Moi je serais prêt à en parler, à faire des formations aux collégiens, lycéens ou personnes âgées. J'essaie déjà de faire de la prévention dans mes vidéos. J'ai fait un deepfake de Donald Trump qui montre des graphiques sur le Covid-19. Les voix ont été faites avec de l'intelligence artificielle. C'est clairement un exemple de fausse information qu'on peut faire en deepfake, et à la fin j'explique comment j'ai fait, pour faire de la prévention.

Et pour d'autres raisons, ce qui menace aussi les journalistes, c'est le deepfake texte. Ça existe déjà ?

Oui, tout existe. Par exemple avec l'Open AI d'Elon Musk, il a créé une sorte de méga machine à absorber le web et il fait du machine learning avec. Il prend des téraoctets entier de web, et il la fait tourner, tourner. L'algorithme peut même rédiger des articles, sans écrire des phrases, juste en lui demandant un sujet en disant « attaque Iran » par exemple. Je pense qu'il va y avoir une révolution de l'intelligence artificielle, ça va chambouler pas mal de métiers. C'est pour ça que c'est important d'être à jour et de se renseigner sur ce qui se fait.

Résumé

A l'ère des fake news, les deepfakes représentent une sérieuse menace pour le journalisme. Ces vidéos ou enregistrements manipulés par intelligence artificielle brouillent la frontière entre perception et réalité. Avec cette technologie, il est possible de faire dire n'importe quoi à n'importe qui, et l'illusion est quasi parfaite. Utilisés à mauvais escient, les deepfakes sont donc une puissante arme de désinformation. Comment se sont-ils peu à peu fait une place dans notre société ? Et comment combattre cette nouvelle menace pour le journalisme ? C'est ce que nous tâcherons de comprendre dans ce mémoire professionnel.

Mots-clés

Deepfake — Hypertrucage — Infox vidéo — Videotox — Cheapfake — Fake news — Désinformation — Intelligence artificielle