« Deepfakes », la traque aux vidéos truquées est ouverte

Depuis leur apparition en 2018, ces contenus manipulés par intelligence artificielle amusent autant qu'ils inquiètent. Les méthodes de détection, qui reposent elles-mêmes sur des algorithmes, se multiplient.

EN CE MOIS DE MARS 2021, sur son compte Twitter, Darla n'a de cesse de défendre son employeur Amazon et de critiquer ses collègues, en Alabama (États-Unis), qui tentent de monter un syndicat. Mais le voile est vite levé : ce compte est une parodie signée d'un comédien pour se moquer des tentatives du géant américain du commerce en ligne de délégitimer la création du syndicat avec de faux comptes Twitter d'employés. Et la photo de Darla, jeune femme à lunettes, est un *deepfake* (un « hypertrucage »), un visage de synthèse créé par intelligence artificielle à partir du site ThisPersonDoesNotExist (« Cette personne n'existe pas »).

Analyser les reflets sur l'iris des yeux

Coïncidence, au même moment, des chercheurs de l'université de Buffalo dans l'État de New York (États-Unis) dévoilent un moyen de débusquer automatiquement de tels artifices. Il consiste à analyser les reflets de lumière sur l'iris des yeux. Sur un vrai visage, les deux iris sont similaires, par leur forme comme leur couleur. Pas sur des visages artificiels. Les chercheurs ont d'abord entraîné un algorithme à en reconnaître les caractéristiques sur une base de visages non truqués. Après quoi, de faux visages du site ThisPersonDoesNotExist ont été soumis au même algorithme : celui-ci a su repérer l'artifice dans 94 % des cas.

Le procédé ne fonctionne toutefois que sur des visages où les deux reflets apparaissent. Et elle ne permet de s'attaquer pour le moment qu'à des images fixes. « *Dans son principe, cette méthode de détection pourra fonctionner aussi sur des vidéos*, assure Siwei Lyu, spécialiste en « sciences forensiques » et coauteur de ce projet. *Mais les séquences actuelles n'étant pas en haute résolution, les reflets sur les yeux ne sont pas forcément visibles.* »

Les *deepfakes* sont en effet surtout connus — et redoutés — dans leur forme vidéo. Apparus en 2018 pour des contenus pornographiques, ils diffèrent des montages et des modèles 3D créés avec des logiciels d'infographie : ils sont produits automatiquement par des algorithmes appelés réseaux génératifs antagonistes (*lire* S. et A. *n° 862, décembre 2018*), à partir de visages authentiques. Le résultat peut être un visage totalement inventé, ou remplacé par celui d'une autre personne réelle par un procédé appelé *faceswap* (« remplacement de visage ») (1). Il est possible d'aller encore plus loin en appliquant aux lèvres d'une personne un mouvement réel capté sur un autre individu par une caméra (2). On obtient ainsi une véritable marionnette numérique à laquelle on fait dire ce que l'on veut. Début mars, un spécialiste belge en effets spéciaux a posté sur le réseau social TikTok de fausses séquences confondantes avec Tom Cruise *(voir photos ci-contre)*. Il a fait appel à un acteur capable d'imiter la gestuelle de la star américaine et a peaufiné le tout par des retouches d'images, montrant jusqu'où il était possible d'aller.

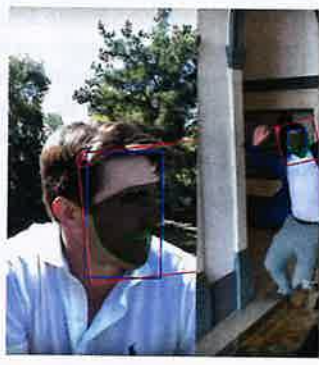
EUROCOM

« Ce n'est pas parce qu'aucun défaut n'est visible par l'humain que la machine ne va rien détecter »

Jean-Luc Dugelay, enseignant chercheur à Eurocom, à Sophia-Antipoplis

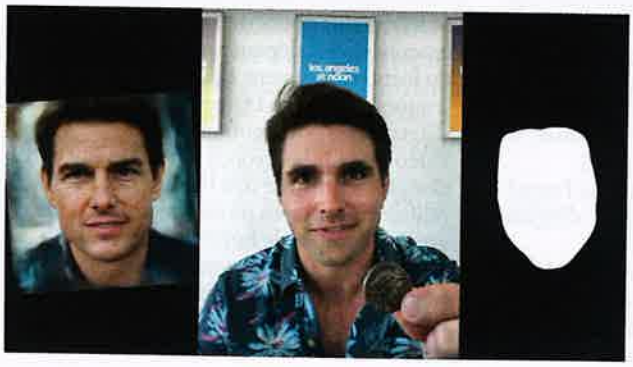
Ces moyens ne sont pas — encore — à la portée de tout le monde et les *deepfakes* restent essentiellement cantonnés à des blagues sur Internet. Pourtant, ils inquiètent. Facebook et Twitter ont annoncé qu'ils supprimeraient ceux cherchant à tromper le public. D'autant que les outils sont de plus en plus faciles d'usage, avec des applications mobiles comme Wombo qui a séduit dix millions d'utilisateurs deux mois après son apparition en février. À l'origine, pour manipuler un visage existant, il fallait fournir aux algorithmes quantité d'images d'une même personne

pour obtenir un visuel satisfaisant. Or, « *on parvient maintenant à faire un* faceswap *avec juste une vidéo et une photo de la personne que l'on veut faire apparaître* », prévient Jean-Luc Dugelay, chercheur à l'école d'ingénieurs Eurecom à Sophia Antipolis (Alpes-Maritimes). Les risques en matière de désinformation sont évidents, de même que pour l'identification en ligne. « *Bientôt, on pourra non seulement générer en temps réel une vidéo de quelqu'un d'autre mais aussi lui faire bouger la tête, cligner des yeux, selon ce que l'interlocuteur, une banque en ligne, un animateur de session Zoom ou Skype par exemple, aura instauré comme critère de vérification* », continue Jean-Luc Dugelay. C'est pourquoi depuis un an, les travaux en matière de détection s'accélèrent. Ils reposent sur des algorithmes d'apprentissage automatique capables de débusquer ▸



PHOTOS : VFXCHRISUME

pour obtenir un visuel satisfaisant. Or, « *on parvient maintenant à faire un* faceswap *avec juste une vidéo et une photo de la personne que l'on veut faire apparaître* », prévient Jean-Luc Dugelay, chercheur à l'école d'ingénieurs Eurecom à Sophia Antipolis (Alpes-Maritimes). Les risques en matière de désinformation sont évidents, de même que pour l'identification en ligne. « *Bientôt, on pourra non seulement générer en temps réel une vidéo de quelqu'un d'autre mais aussi lui faire bouger la tête, cligner des yeux, selon ce que l'interlocuteur, une banque en ligne, un animateur de session Zoom ou Skype par exemple, aura instauré comme critère de vérification* », continue Jean-Luc Dugelay. C'est pourquoi depuis un an, les travaux en matière de détection s'accélèrent. Ils reposent sur des algorithmes d'apprentissage automatique capables de débusquer ▶



◀ ▲ **L'acteur américain Tom Cruise** (à droite) et le comédien Miles Fischer (à gauche). Le spécialiste des « deepfakes » Chris Ume a mis en scène dans de courtes vidéos le second imitant la gestuelle du premier. Il a ensuite remplacé le visage de Miles Fischer par celui de l'acteur américain.

défauts et incohérences. Les reflets sur les iris, mais aussi le clignement des yeux, mal géré par les premiers *deepfakes*, des problèmes de résolution d'image, de couleurs, d'ombres, etc. « *Les erreurs détectables par les humains et celles qui le sont par les machines ne sont pas les mêmes*, explique Jean-Luc Dugelay. *Ce n'est pas parce qu'aucun défaut n'est visible que la machine ne va rien détecter.* »

SHU HUET AL.

En analysant les reflets de la lumière dans les yeux, un algorithme est parvenu dans 94 % des cas à diifférencier les visages truqués (les trois portraits de gauche) des vrais (les trois de droite).

Un code intégré au contenu pour garantir l'authenticité

Le Media Lab du MIT (États-Unis) a mis sur pied en avril 2020 le projet expérimental Detect Fakes. Dans un premier temps, les internautes sont invités à distinguer séquences authentiques et *deepfakes*. « *Nous comptons nous servir des données recueillies pour bâtir de futurs systèmes de détection et comprendre comment mieux sensibiliser les gens* », explique Matt Groh, doctorant au laboratoire. Plusieurs bases de données de *deepfakes* ont été constituées pour entraîner des algorithmes à repérer les manipulations. Mais c'est Facebook qui a frappé le plus grand coup avec un corpus de 104 000 vidéos créées pour le Deepfake Detection Challenge, un concours de programmation qui a mobilisé 2114 équipes de recherche entre décembre 2019 et juin 2020.

Pour l'heure, les approches telle que celle de l'analyse des iris souffrent de la spécificité des éléments considérés : il faut d'autres algorithmes, entraînés différemment, pour traquer des incohérences ailleurs dans l'image. Autre problème : quand un algorithme se montre performant, il ne l'est pas tout le temps selon la manière dont l'image a été obtenue. Ainsi, la solution qui a remporté le Deepfake Detection Challenge détectait les faux à plus de 82 % quand elle était testée sur la base de données sur laquelle elle avait été entraînée. Mais face à des *deepfakes* créés selon d'autres méthodes, le score est tombé à 65,18 %. « *Cette différence est énorme*, commente Jean-Luc Dugelay. *Cela signifie que les réseaux de neurones de détection sont spécialisés dans un type* de deepfakes, *généré d'une certaine façon, avec ses défauts types.* »

Des travaux récents ouvrent néanmoins des perspectives. Une équipe d'universitaires californiens est partie du principe que parler déclenche chez une personne des mouvements musculaires faciaux qui lui sont propres. Elle a créé une empreinte de ces mouvements à partir de vidéos authentiques de personnalités politiques. Cette empreinte est alors comparée à celle issue de *deepfakes* : la détection marche à plus de 90 %, quelle que soit la méthode ayant produit le truquage. Mais, par définition, une empreinte ne marche que sur les vidéos d'une personne en particulier. De quoi en tout cas débusquer celles ciblant des célébrités. La notion d'empreinte, justement, est une direction privilégiée. Microsoft, notamment, a le projet d'intégrer à un contenu, au moment de sa création, un code qui servirait à vérifier si une image a été truquée après coup. À l'université d'État de New York à Binghamton, une équipe s'est concentrée sur un critère biologique : le changement de couleur de la peau en fonction du rythme cardiaque. Or non seulement cette technique s'est révélée efficace à plus 93 % mais les chercheurs ont découvert qu'elle pouvait permettre d'extraire une signature du générateur de *deepfakes* utilisé. Une direction prometteuse. Car savoir comment a été fabriquée une marionnette numérique, c'est déjà entrevoir qui en tire les ficelles. ■ **Arnaud Devillard**

@A_Devila

POUR EN SAVOIR PLUS

- L'interview complète de Jean-Luc Dugelay d'Eurecom : **sciav.fr/892Dugelay**
- Le projet du MIT Detect Fakes : **sciav.fr/892MIT**
- Le site du spécialiste en effets spéciaux Chris Ume avec des deepfakes de Tom Cruise ou Jacques Brel : **sciav.fr/892ChrisUme**

(1) sciav.fr/892Faceswap
(2) http://sciav.fr/892Paroles

AUDIO

Les fausses voix alertent aussi

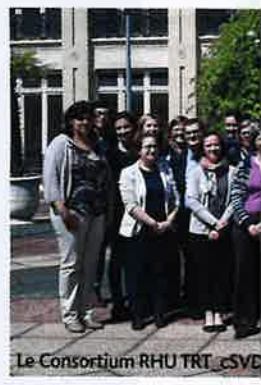
L'an dernier, le P-DG d'une entreprise britannique d'énergie reçoit un appel téléphonique du patron de la maison mère en Allemagne : il faut transférer en urgence 200 000 livres sterling (254 000 euros) vers un fournisseur hongrois. Le P-DG s'exécute... avant de découvrir que l'appel était un faux. En l'occurrence, un *deepfake* de la voix du patron allemand, en reproduisant la diction et l'accent. Générés de la même manière que les images et nécessitant très peu de données de départ, les faux sons animent aussi la recherche. Des données phonétiques, notamment sur les consonnes, dépendantes de l'anatomie du locuteur, peuvent aider à trahir une voix truquée. Le projet Detect Fakes du MIT (États-Unis) porte autant sur des visages en vidéo que sur les propos tenus. De son côté, Google a constitué une base de données de milliers de phrases prononcées par 68 voix artificielles dans le but d'entraîner des algorithmes. Elle est notamment utilisée pour un concours de systèmes de détection, l'ASVspoof Challenge, dont la quatrième édition s'est tenue en avril.

RHU TRT_CSVD

Les fragiles peti de notre c

Basé sur une découverte française majeure, un aml à dévoiler les mécanismes à l'origine des maladies connues, elles s'expriment avec l'âge. Dans le ce hémorragies et autres lésions.

De minuscules vaisseaux sanguins parcourent le cerveau. Ils sont si fins qu'on ne peut les observer directement, même avec des techniques d'imagerie sophistiquées. Leur rôle est pourtant immense : ils fournissent l'oxygène indispensable à la survie et au fonctionnement de chaque neurone. En réponse aux variations de pression artérielle, ils modifient le diamètre de leur section pour adapter finement le débit sanguin et irriguer correctement le tissu cérébral. Quand ils sont malades, les troubles sont multiples : l'humeur, l'équilibre, la mobilité, la mémoire, la cognition peuvent être perturbés. Après 65 ans, plus des deux tiers d'entre nous seraient concernés. Certaines personnes sont frappées plus jeunes. Ces pathologies (SVD, pour small vessel diseases, en anglais) sont à l'origine de 30 % des accidents vasculaires cérébraux, 25 % des infarctus cérébraux et 90 % des hémorragies intracérébrales. En cause : l'hypertension artérielle et le vieillissement, mais pas seulement. Plusieurs facteurs génétiques peuvent être impliqués.



Le Consortium RHU TRT_cSVD

« Après 65 ans, plus tiers d'entre nous se concernés, mais d'a personnes sont frap plus jeunes. »

Depuis une trentaine d'années, un temps très court à l'aune de la connaissance scientifique, on commence à y voir plus clair. Hugues Chabriat, coordonnateur du Centre de référence pour les maladies vasculaires rares du cerveau et de l'œil, à Paris, et Anne Joutel, chercheuse Inserm à l'Institut de psychiatrie et neurosciences de Paris, ont consacré une grande partie de leur carrière aux SVD. Avec Marie-Germaine Bousser et Elizabeth Tournier-Lasserve, ils ont découvert CADASIL, une artériopathie cérébrale héréditaire. En identifiant les mutations génétiques qui en sont à l'origine, ils ont mis la main sur un modèle d'étude des SVD. Jusqu'alors, le déclin cognitif et les troubles moteurs liés au vieillissement étaient le plus le compte de l'hypertensio autre réponse. L'étude de C le rôle joué par d'autres ac mécanismes moléculaires petits vaisseaux du cerveau diagnostics, laisse espérer le de traitements médicamen découverte française a été Brain Prize, le plus grand p en neurosciences. Les deux aussi réuni une grosse ving de recherche et trois ind d'un programme de reche universitaire dédié.

Démarré en 2017, le projet a avancées. Plusieurs autres g