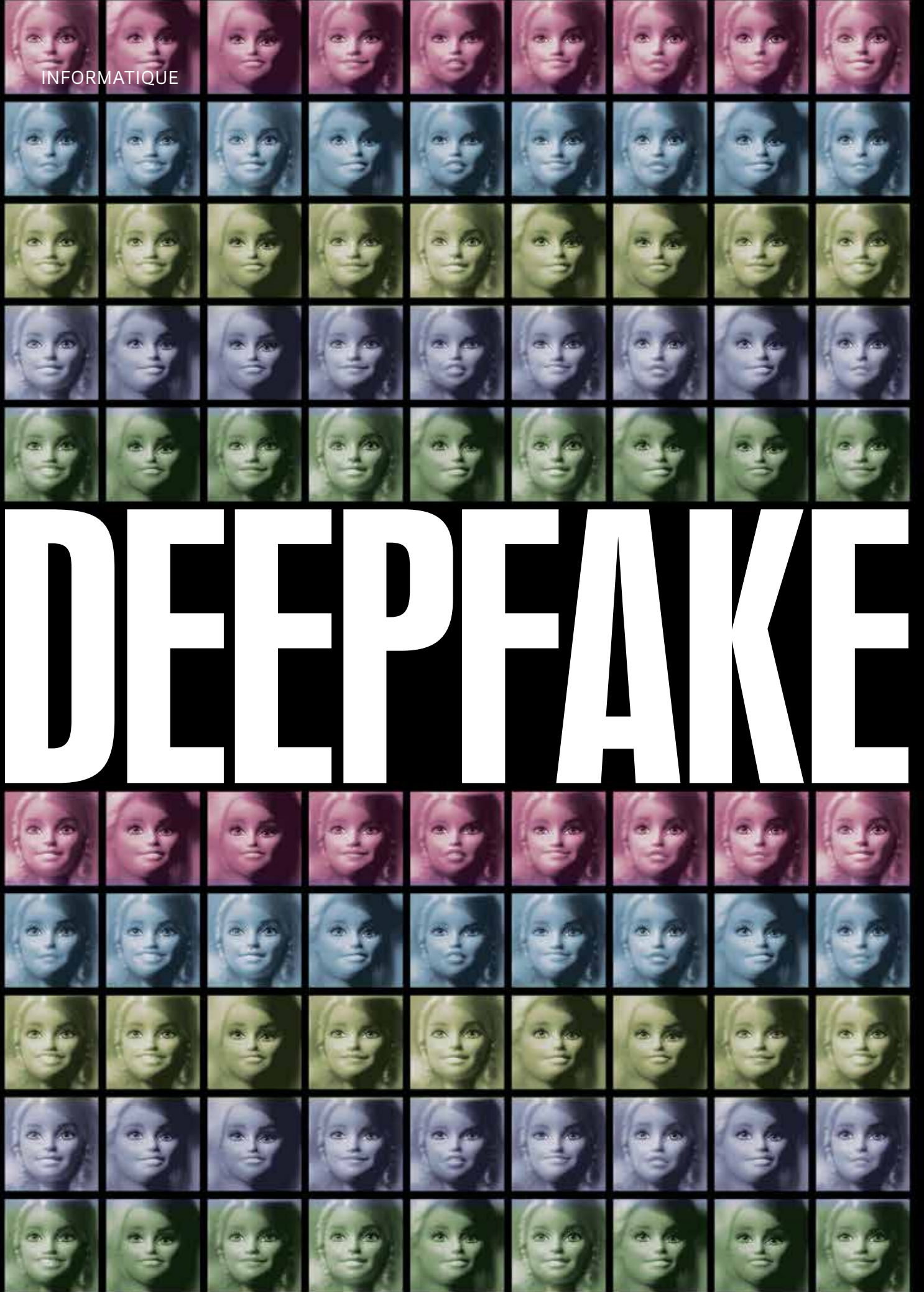
INFORMATIQUE



DEEPFAKE

Un algorithme détecte des images et des vidéos modifiées grâce à l'intelligence artificielle, la nouvelle étape technologique au service de la désinformation

Sarah Schmidt

En septembre, une vidéo contrefaite du *Jornal Nacional* (Journal National), principal programme d'information de la chaine de télévision Globo, a fait le tour des réseaux sociaux. Dans cette vidéo, les présentateurs William Bonner et Renata Vasconcellos présentaient les résultats d'un sondage relatif aux intentions de vote pour la Présidence, mais les données des candidats étaient inversées, tant sur les graphiques que dans les commentaires des présentateurs. Le lendemain, la chaîne d'information a apporté un démenti en signalant que la vidéo avait été utilisée dans le but de désinformer la population et qu'il s'agissait d'un deepfake, une technique qui fait appel à l'intelligence artificielle pour remanier en profondeur un contenu numérique. Il est ainsi possible, par exemple, de modifier numériquement le visage d'une personne ou de reproduire sa voix, en lui faisant faire ce qu'elle n'a pas fait où dire ce qu'elle n'a pas dit.

Au mois d'août, une autre vidéo du *Jornal Nacional* qui manipulait également les résultats d'un sondage pour la Présidence avec un montage similaire, a été publiée sur le réseau social TikTok et a atteint 2,5 millions de vues, selon le projet Comprova, une initiative regroupant des journalistes issus de 43 médias brésiliens en matière de fact-checking. « Il est possible qu'une méthode de deepfake ait été utilisée dans ces vidéos, mais il faudrait procéder à une analyse plus approfondie. Le plus important pour nous est de savoir qu'elles sont fausses », observe l'informaticien Anderson Rocha, directeur de l'Institut Informatique de l'Université Publique de Campinas (Unicamp), où il coordonne le Laboratoire d'Intelligence Artificielle (Recod.ai). Le chercheur étudie des manières permettant d'identifier des falsifications malveillantes de photos et de vidéos, y compris de deepfakes, également appelées médias synthétiques.

Au mois de mars de cette année, peu de temps après le début de la guerre entre la Russie et l'Ukraine, le président ukrainien Volodymyr Zelensky a été victime d'un deepfake. Une vidéo où il demandait aux Ukrainiens de déposer les armes et de rentrer chez eux comme si le pays allait se rendre, a circulé sur les réseaux sociaux, obligeant Facebook et YouTube à la supprimer après avoir constaté qu'il s'agissait d'une contrefaçon. Sur les images, le visage du président apparaissait sur un corps qui bougeait à peine et vêtu d'un T-shirt vert.

Dans certains cas, comme pour les vidéos du *Jornal Nacional*, il est très facile de se rendre compte qu'elles ont été modifiées car les informations originales sont facilement vérifiables. Mais ce n'est pas toujours le cas. Confronté aux médias synthétiques, le dicton « il faut le voir pour le croire » perd tout son sens, et l'intelligence artificielle peut aussi devenir une alliée.

« Généralement, les vidéos synthétiques sont conçues en deux temps, tout d'abord en utilisant une plateforme deepfake pour changer le visage ou synchroniser la bouche, puis en réalisant un montage à l'aide de logiciels d'édition vidéo », explique Anderson Rocha. Les personnes qui savent quoi chercher détectent généralement une faille dans le programme utilisé pour produire le trucage, comme un jeu de lumières différent ou un contraste entre la vidéo originale et le nouveau visage qui a été inséré.

C'est comme si l'on découpait un visage sur une photo et qu'on le plaçait sur une autre où l'incidence de la lumière et la photo capturée par l'appareil photo sont différentes. Ces traces sont autant d'indices détectés grâce aux techniques de l'informatique légale, un domaine de recherche qui a pris de

ILLUSTRATIONS FELIPE BRAZ / REVUE PESQUISA FAPESP **PHOTOGRAPHIES** ALEXA / PIXABAY

l'ampleur ces dernières années et dont Anderson Rocha fait partie.

Le chercheur a mis au point, en collaboration avec des collègues de l'Université de Hong Kong, un algorithme permettant de détecter simultanément une éventuelle manipulation des visages et, le cas échéant, de localiser les régions modifiées. Il peut s'agir, par exemple, du visage entier ou seulement de la bouche, de la région des yeux ou des cheveux. « Le taux de réussite moyen a été de 88% pour les vidéos à faible résolution et de 95% pour les vidéos à haute résolution », explique Anderson Rocha. Pour ce faire, ils ont testé un échantillon de 112 000 visages dont une moitié était réelle et l'autre manipulée à l'aide de quatre logiciels de deepfake. La méthode révèle également si l'image a été créée de toutes pièces, et non à partir d'une photographie existante. Les résultats ont été publiés en avril 2022 dans la revue *Transactions on Information Forensics and Security*.

Selon l'informaticien, les nouveaux algorithmes qui permettent de détecter des traces de manipulation dans les deepfakes se basent principalement sur les indices laissés par des logiciels plus connus et qui correspondent à deux catégories, ceux qui permettent de changer les visages et ceux qui permettent de modifier les expressions faciales. L'un de ces logiciels est réputé pour présenter des défauts dans la synchronisation de la bouche, l'algorithme est alors programmé pour rechercher cette erreur spécifique. « Ceci est problématique car si nous ne connaissons pas le logiciel de deepfake utilisé, ces traces seront difficilement détectables et de nouveaux logiciels apparaissent tous les jours », observe Anderson Rocha.

Anderson Rocha et ses collègues ont donc entraîné leur nouvel algorithme à détecter des traces sans avoir à connaître le logiciel qui produit le deepfake. « Nous partons de l'idée que, quel que soit le programme, il va laisser un bruit, un élément qui ne correspond pas au reste de l'image ». La méthode opère sur deux fronts en recherchant des signatures de bruit, c'est-à-dire, des changements subtils dans le contour du visage, par exemple, et en identifiant la signature sémantique, qui peut être une anomalie au niveau de la couleur, de la texture ou de la forme. « L'algorithme automatise le travail que ferait un spécialiste et qui consiste à détecter des anomalies, comme des contrastes de lumière », dit-il. « La prochaine étape sera de le tester avec des vidéos manipulées par un plus grand nombre de logiciels afin de confirmer son efficacité ».

Ce type d'algorithme de détection peut être utilisé à des fins diverses, telle la lutte contre l'utilisation malveillante de deepfakes. Anderson Rocha fait partie d'un projet international, appelé Semantic Forensics, aux côtés de chercheurs des universités de Sienne, de l'École Polytechnique de Milan, en Italie, et de Notre-Dame, aux États-Unis, avec le soutien du Département Américain de la Défense. L'objectif est de développer des outils automatisés qui détectent ces contrefaçons. « Nous avons déjà vu des cas de vidéos modifiées relatives à des exercices militaires de certains pays qui multipliaient leur nombre de missiles pour étaler leur puissance militaire », dit-il.

Ces algorithmes peuvent également se révéler utiles dans les cas de deepfakes politiques, comme dans l'épisode du président ukrainien, ou même de deepfakes pornographiques. C'est à partir des films pornographiques que cette méthode s'est fait connaître, fin 2017. À l'époque, certains internautes ont commencé à insérer des visages de célébrités dans des scènes de films à caractère sexuel. En septembre 2019, 96% des deepfakes répertoriés sur internet concernaient des contenus pornographiques non consensuels, selon une enquête de DeepTrace Labs, une entreprise de cybersécurité néerlandaise. Les principales victimes étaient des femmes, principalement des actrices, mais il y avait également des personnes qui n'étaient pas connues. La chanteuse Anitta a également été victime d'un deepfake pornographique au mois de juillet de cette année. La vidéo originale avait déjà été utilisée pour produire des images contrefaites utilisant le visage de l'actrice Angelina Jolie.

Dans une vidéo contrefaite, le président ukrainien Volodymyr Zelensky a appelé ses compatriotes à déposer les armes

IL EST POSSIBLE DE CHANGER LE VISAGE D'UNE PERSONNE OU D'IMITER SA VOIX EN LUI FAISANT DIRE CE QU'ELLE N'A PAS DIT

Le Brésil est déjà victime de deepfakes et il est nécessaire de les démentir, selon la journaliste Cristina Tardáguila, directrice de programme du Centre International des Journalistes (ICFJ) et fondatrice de l'Agence Lupa, spécialisée dans la vérification de faits. Les programmes permettant de détecter les médias synthétiques peuvent être des alliés précieux pour les journalistes et les vérificateurs de faits qui travaillent contre la montre. « Quand on a affaire à de fausses informations, il faut être rapide. Il faut donc investir davantage dans l'intelligence artificielle, dans des outils qui permettent de détecter et de répertorier plus rapidement ces faux contenus. Nous parviendrons ainsi à écourter le délai entre la propagation de la contrefaçon et le résultat du fact-checking », explique-t-elle.

« Les deepfakes sont le summum des fake news. Ils peuvent tromper plus facilement, car s'il s'agit d'une vidéo, la personne regarde cette scène », observe la journaliste Magaly Prado, qui suit une formation postdoctorale à l'Institut d'Études Avancées de l'Université de São Paulo (IEA-USP). « Le format audio peut également être créé de manière artificielle », dit-elle. Elle est également l'auteur du livre *Fake news et intelligence artificielle : le pouvoir des algorithmes dans la guerre de la désinformation*, publié en juillet par Edições 70.

Elle estime que les deepfakes strictement audio, même s'ils sont moins connus et moins courants, peuvent facilement se propager sur des plateformes telles que WhatsApp, un média largement utilisé par les Brésiliens. Ils suivent une logique semblable à celle des vidéos, car avec des logiciels chaque fois plus accessibles et performants, il est possible de simuler la voix de quelqu'un. Les principales victimes sont des personnalités publiques, dont la voix est largement disponible sur Internet. Cette technique peut également être utilisée dans le cadre d'escroqueries financières. « Il y a déjà eu des cas comme celui d'un employé d'une entreprise technologique qui avait reçu un message vocal d'un cadre supérieur lui demandant un transfert d'espèces. Il a toutefois eu des soupçons et le message a été ensuite analysé par une entreprise spécialisée. Il s'est alors avéré qu'il s'agissait d'un message élaboré à l'aide d'une intelligence artificielle », explique-t-il.

Le journaliste Bruno Sartori, directeur de la société FaceFactory, explique que produire des deepfakes bien faits, tant audio que vidéo, n'est pas pour autant si simple. Son entreprise produit des médias synthétiques à usage commercial et fournit des contenus pour des émissions humoristiques sur les chaînes de télévision Globo et SBT.

Il a travaillé en 2021 sur une publicité pour Samsung dans laquelle la présentatrice Maísa, déjà adulte, interagissait avec sa version d'elle-même étant enfant. Cette vidéo a été produite avec des techniques de deepfake montrant une petite fille virtuelle dansant, jouant et lançant un ordinateur portable vers le haut. À un autre moment, il a dû insérer le visage d'un acteur sur celui d'un cascadeur. « Pour bien entraîner une intelligence artificielle, il faut avoir une bonne banque d'images et de sons de la personne que l'on souhaite imiter. Les bons logiciels permettant un traitement de haute qualité doivent également disposer de paramètres avancés. Dans le cas contraire, des défauts visibles peuvent apparaître sur le visage ou, dans le cas de l'audio, une voix robotique », explique-t-il.

Selon lui, les vidéos du faux sondage du *Jornal Nacional* n'ont pas été modifiées à l'aide d'une intelligence artificielle. D'après son analyse, elles ont fait l'objet un montage classique, en coupant et en inversant l'ordre des audios. « Il s'agit de *shallowfakes*, mais comme ils sont bien faits, ils peuvent facilement tromper les gens », déclare Bruno Sartori. Il estime que d'ici quelques années, ces programmes seront plus légers, plus intelligents et plus accessibles.

Il y a plusieurs moyens de se protéger de la désinformation numérique. L'un d'entre eux est de se renseigner sur les licences d'utilisation et sur le niveau de confidentialité des applications gratuites les plus diverses utilisées au quotidien, comme celles qui sollicitent un accès aux photos de l'utilisateur pour produire des effets amusants, ou celles qui peuvent enregistrer sa voix. Selon Anderson Rocha, de l'Unicamp, de nombreuses applications stockent une grande quantité de données qui peuvent être partagées à d'autres fins, comme pour l'entraînement de logiciels de deepfakes.

L'éducation aux médias est un autre aspect essentiel. « Bien qu'il y ait de nombreux logiciels qui nous permettent de déceler les contrefaçons, la première chose à faire est de se méfier de ce que l'on reçoit sur les réseaux sociaux. Il faut également vérifier les sources d'information et se renseigner à leur sujet », conclut-il. ■

Projet

Déjà vu : Cohérence temporelle, spatiale et caractérisation des données hétérogènes pour l'analyse et l'interprétation de l'intégrité. (N° 17/12646-3) ; **Modalité** Projet Thématique ; **Chercheur responsable** Anderson Rocha ; **Investissement** 1 912 168,25 reais BRL.

Article scientifique

KONG, C. *et al.* Detect and locate: Exposing face manipulation by semantic- and noise-level telltales. **Transactions on Information Forensics and Security**. v. 17. avr. 2022.

Livre

PRADO, M. **Fake news et intelligence artificielle : Le pouvoir des algorithmes dans la guerre de la désinformation**. São Paulo : Edições 70, 2022.