↓ *Dessin de* **Señor Salme,** *Espagne.*

SYNERGY ART

SOURCE

NATURE
Londres, Royaume-Uni
Hebdomadaire
nature.com
Depuis 1869, cette revue scientifique au prestige mérité accueille – après plusieurs mois de vérifications – les comptes rendus des innovations majeures dans tous les domaines : de la biologie à la physique, en passant par l'astronomie. Outre les articles destinés aux chercheurs et aux scientifiques, la revue propose des pages d'actualités, de débats et de dossiers accessibles au grand public.

Discerner le vrai de l'IA

Les images trompeuses entièrement créées par intelligence artificielle générative se multiplient. Elles menacent la démocratie, estiment certains. Pour limiter les dégâts, des chercheurs se sont lancés dans une course au développement d'outils capables de les détecter. — *Nature* [Londres]

En juin dernier, dans le cadre des joutes politiques en amont des primaires de la présidentielle américaine de 2024, plusieurs images ont été diffusées montrant Donald Trump embrassant l'un de ses anciens conseillers à la santé, Anthony Fauci. Sur certains clichés, on voit Trump bécoter gauchement le visage de Fauci, le "monsieur santé" vilipendé par certains conservateurs américains pour avoir défendu le port du masque et les vaccins pendant la pandémie de Covid-19.

"*Ça sautait aux yeux*" qu'ils étaient faux, se souvient Hany Farid, informaticien à l'université de Californie à Berkeley et l'un des nombreux spécialistes à avoir examiné ces clichés. En scrutant trois des photos, on s'aperçoit que les cheveux de Trump sont flous, curieusement, que les inscriptions n'ont aucun sens à l'arrière-plan, que les bras et les mains ont des positions peu naturelles et que l'oreille visible de Trump défie les lois de l'anatomie. Autant d'éléments qui trahissent – pour l'heure – l'utilisation d'une intelligence artificielle générative (IAG).

Les images et vidéos de ce type, appelées *deepfakes*, produites par des générateurs *text-to-image* [qui créent des images à partir d'une indication textuelle] en faisant appel au *deep learning* ["apprentissage profond"], sont désormais monnaie courante. Si les tricheurs recourent depuis belle lurette à la tromperie pour extorquer de l'argent, influencer l'opinion, voire déclencher des guerres, la vitesse et la facilité avec lesquelles des volumes considérables de contenus fallacieux (mais extrêmement convaincants) peuvent être dorénavant produits et diffusés – associées à un manque de sensibilisation du grand public – constituent une menace grandissante.

"*Les gens connaissent encore mal la technologie générative. Ce n'est pas comme si elle était apparue progressivement. Ç'a été fulgurant : tout à coup, elle était là, ce qui nous empêche d'avoir le recul nécessaire*", observe Cynthia Rudin, informaticienne spécialiste de l'IA à l'université Duke de Durham, en Caroline du Nord.

Des dizaines d'outils sont désormais disponibles, permettant à l'internaute lambda de générer artificiellement n'importe quel contenu ou presque, à n'importe quelle fin, qu'il s'agisse de créer une fausse vidéo de Tom Cruise sur TikTok pour s'amuser, de faire revivre une victime d'une fusillade dans une école dans une vidéo réclamant un durcissement de la réglementation sur les armes à feu ou de simuler un appel à l'aide d'un proche pour délester une victime de quelques dizaines de milliers de dollars. Les *deepfakes* vidéo peuvent même être générés en temps réel, au cours d'une visioconférence, par exemple. Au début de l'année, Jerome Powell, le patron de la Réserve fédérale américaine, a ainsi discuté en visio avec quelqu'un qu'il pensait être le président

TECHNOLOGIE

ukrainien, Volodymyr Zelensky… sauf que ce n'était pas lui.

La quantité de contenus générés par IA est inconnue, mais elle serait en train d'exploser. Les universitaires citent souvent une estimation selon laquelle 90 % des contenus d'Internet pourraient être créés artificiellement d'ici quelques années. *"Tout le reste serait alors noyé dans le bruit ambiant"*, prévient Cynthia Rudin, ce qui compliquerait la tâche de celles et ceux qui sont à la recherche de contenus utiles ou authentiques. Les moteurs de recherche et les réseaux sociaux ne feront qu'amplifier la désinformation, ajoute-t-elle. *"Jusqu'à présent, on recommandait et on faisait circuler toutes ces bêtises. Maintenant, on va les générer."*

Si les contenus de synthèse ont souvent pour but premier de divertir et d'amuser, comme l'image devenue virale du pape François portant une doudoune de créateur, certains obéissent à une stratégie partisane et parfois malveillante – par exemple via la diffusion tous azimuts de contenus pornographiques dans lesquels le visage d'une personne est associé au corps d'une autre sans son consentement. Même un seul contenu de synthèse peut être lourd de conséquences : une image générée par l'IA d'une explosion au Pentagone s'est ainsi propagée comme une traînée de poudre en mai dernier, créant un trou d'air sur le marché boursier. L'existence de ces contenus de synthèse permet également à des individus de rejeter les preuves bien réelles d'un délit en les qualifiant de "fausses".

"Les gens savent de moins en moins à quel saint se vouer. Et c'est un vrai problème pour la démocratie, analyse Sophie Nightingale, psychologue à l'université de Lancaster (Royaume-Uni), qui étudie les effets de l'IA générative. *C'est un point sur lequel il faut réagir, et vite. La menace est d'ores et déjà considérable."* Elle précise que la question va se poser avec acuité dans les deux ans à venir, avec les élections majeures qui vont se tenir aux États-Unis, en Russie et au Royaume-Uni.

Pour l'instant, certains contenus de synthèse contiennent encore des indices révélateurs, tels que des mains à six doigts sur les photos. Mais l'IAG s'améliore de jour en jour. *"C'est l'affaire de quelques mois"* avant que les gens ne soient plus capables de faire la différence, pointe Wael Abd-Almageed, informaticien et ingénieur informatique à l'université de Californie du Sud à Los Angeles.

Dans ce contexte, les chercheurs mettent les bouchées doubles pour tenter de tirer le meilleur parti des *deepfakes* tout en concevant des garde-fous contre les abus. Il existe à ce jour deux parades technologiques : le marquage des contenus authentiques ou de synthèse au moment de leur création et l'utilisation de détecteurs d'IA pour repérer les faux après publication. Aucune de ces deux solutions n'est parfaite, mais l'une et l'autre compliquent la falsification, assure Shyam Sundar, psychologue et fondateur d'un laboratoire de recherche sur les effets des contenus à l'université d'État de Pennsylvanie. *"Si vous êtes malveillant et que vous avez de la suite dans les idées, vous pouvez à l'évidence aller assez loin. Le but de la manœuvre, c'est de vous mettre des bâtons dans les roues"*, explique-t-il.

À court terme, affirme Sophie Nightingale, la technologie jouera le premier rôle, mais *"à plus long terme peut-être mettra-t-on davantage l'accent sur la sensibilisation et la réglementation"*. L'Union européenne (UE) ouvre la voie à l'échelle internationale avec sa loi sur l'IA, qui a été adoptée par le Parlement en juin dernier et qui attend l'aval des deux autres instances dirigeantes de l'Union. *"On va certainement en tirer des leçons importantes,* juge la psychologue, *que leurs efforts soient couronnés de succès ou non."*

↑ Une fausse vidéo mettant en scène l'ancien président américain Barack Obama. Les logiciels de reconnaissance faciale, comme celui à l'œuvre ici, peuvent être utilisés pour faire dire à quelqu'un quelque chose qu'il n'a jamais dit, notamment en période électorale. *Photo AP/Sipa*

Pour les chercheurs, l'IA générative est un outil très utile. Elle est par exemple utilisée pour créer des jeux de données médicales, en contournant la question de la confidentialité, afin de concevoir des molécules ou encore d'améliorer les articles et les logiciels scientifiques. On envisage également d'utiliser les *deepfakes* pour anonymiser les participants aux thérapies de groupe en visio, créer des avatars personnalisés de médecins ou d'enseignants qui soient plus attrayants pour le public, ou améliorer l'encadrement des études dans le domaine des sciences sociales. Shyam Sundar assure : *"Je suis plus optimiste qu'inquiet. Je pense que c'est révolutionnaire comme technologie."*

Face au risque d'abus généralisés, les chercheurs et les éthiciens n'en tentent pas moins de poser des règles encadrant l'usage de l'IA, comme la Déclaration de Montréal de 2018 et la Recommandation sur l'intelligence artificielle de 2019. Le Partnership on AI ["Partenariat sur l'IA"], une organisation à but non lucratif qui réunit des poids lourds du secteur, promeut le dialogue sur les bonnes pratiques, même si certains observateurs et participants doutent de son efficacité réelle.

Tous défendent les principes de transparence et de signalement des contenus de synthèse. Les entreprises s'emparent du sujet : en mars, par exemple, TikTok a mis à jour les règles de sa communauté en obligeant les créateurs de contenus à dire s'ils utilisaient l'IA dans les scènes d'apparence réaliste. En juillet, sept géants de la tech, dont Meta, Microsoft, Google, OpenAI et Amazon, se sont engagés auprès de la Maison-Blanche à signaler les contenus générés par l'IA. Et, en septembre, Google a fait savoir qu'à compter de la mi-novembre tout contenu généré par l'IA utilisé dans des publicités à caractère politique devra être signalé comme tel sur ses plateformes, y compris YouTube.

Une des manières de marquer les images de synthèse est de leur appliquer un filigrane en modifiant des pixels. La retouche est imperceptible à l'œil nu mais décelable à l'analyse. En modifiant certains pixels de sorte que leur valeur colorimétrique soit un nombre pair, par exemple, on crée un filigrane – mais un filigrane rudimentaire qui disparaît à la moindre manipulation de l'image ou presque, comme l'application d'un filtre de couleur. Certains filigranes ont ainsi été jugés trop faciles à supprimer. Mais d'autres plus complexes peuvent, par exemple, prendre la forme d'un dégradé d'un bord à l'autre de l'image, superposé à plusieurs autres motifs de même type, de sorte qu'il soit impossible de les effacer en manipulant les pixels individuellement. Ces filigranes peuvent être difficiles (mais pas impossibles) à supprimer, explique Hany Farid. En août dernier, Google a dévoilé un filigrane destiné aux images de synthèse, appelé SynthID, sans donner plus de détails sur son fonctionnement ; reste à savoir s'il est vraiment efficace, commente l'informaticien.

Dans le même ordre d'idées, une autre option consiste à taguer les métadonnées d'un fichier avec des informations sécurisées sur sa provenance. Pour la photographie, cette méthode commence à la prise de vue avec un logiciel intégré à l'appareil photo qui vérifie que les coordonnées GPS et l'horodatage de l'image sont valides et que le cliché n'est pas la photo d'une photo, par exemple. Les compagnies d'assurances se servent de ce type d'outils pour authentifier les images de biens et de dégâts, et l'agence de presse Reuters a testé cette technologie d'authentification pour valider des photos de la guerre en Ukraine.

La Coalition sur la provenance et l'authenticité des contenus (C2PA), qui réunit des géants de la tech et de l'édition, a publié en 2022 une première version d'un document technique sur le traçage des données de provenance, à la fois pour les images réelles et de synthèse. De nombreux outils conformes à la C2PA permettant

d'intégrer, d'identifier et de vérifier les données de provenance sont désormais disponibles, et de nombreuses entreprises, dont Microsoft, s'engagent à suivre les lignes directrices de la C2PA. *"La C2PA va être très importante, elle va nous aider"*, se félicite Anderson Rocha, chercheur en IA à l'université de Campinas, au Brésil.

Les outils d'identification de la provenance des images devraient permettre de réduire le nombre de fichiers douteux, estime Hany Farid, qui siège au comité directeur de la C2PA et propose ses services de consultant à Truepic, une société de San Diego (Californie) qui vend des logiciels d'authentification des photos et des vidéos. Encore faut-il que les *"acteurs bien intentionnés"* adhèrent à un programme comme la C2PA. D'autant que *"certains contenus passeront entre les mailles du filet"*, observe-t-il. D'où l'intérêt des détecteurs d'IA, qui peuvent être un bon outil complémentaire.

"Je suis plus optimiste qu'inquiet. Je pense que c'est révolutionnaire comme technologie."

Shyam Sundar,
PSYCHOLOGUE ET CHERCHEUR AMÉRICAIN

Les laboratoires universitaires et les entreprises ont créé une multitude d'outils de classification faisant appel à l'IA. Ces derniers reposent sur des modèles qui permettent de distinguer les contenus générés artificiellement par IA des photos réelles. La plupart de ces outils seraient capables de repérer plus de 90 % des faux et ne prendraient les images réelles pour des faux que dans 1 % des cas. Reste qu'il est facile de les mettre en échec. Un individu malintentionné peut falsifier des images de sorte d'induire le détecteur en erreur, prévient Hany Farid.

Les outils faisant appel à l'IA peuvent être associés à d'autres techniques qui reposent sur l'intuition humaine pour démêler le faux du vrai. Hany Farid recherche des indices comme des lignes de perspective qui défient les lois de la physique. D'autres signes sont plus subtils. Avec ses confrères, il a découvert que les visages créés par les générateurs StyleGAN avaient notamment tendance à placer les yeux toujours au même endroit sur les photos, ce qui permettait de repérer ceux qui avaient été générés par IA. Les détecteurs d'IA peuvent faire appel à des algorithmes sophistiqués qui peuvent, par exemple, lire une horloge située sur la photo et vérifier que la luminosité de l'image correspond à l'heure qu'elle donne. Le logiciel FakeCatcher, de la société Intel, analyse les vidéos à la recherche des changements de couleurs attendus sur le visage en fonction des fluctuations de la circulation sanguine. Certains détecteurs d'IA, poursuit Anderson Rocha, recherchent le bruit distinctif créé par les capteurs de luminosité des appareils photo, que l'IA a encore du mal à reproduire.

Faussaires et débusqueurs de faux se livrent une guerre féroce. Hany Farid se souvient ainsi d'un article rédigé par un de ses anciens étudiants, Siwei Lyu, devenu informaticien à l'université de Buffalo (État de New York), qui mettait en évidence que, sur certaines vidéos générées par IA, les gens clignaient des yeux à des cadences différentes. Les générateurs avaient réglé le problème en quelques semaines. C'est la raison pour laquelle, même si le laboratoire de Farid publie la grande majorité de ses travaux, il ne diffuse le code qu'au cas par cas aux universitaires qui en font la demande. Wael Abd-Almageed pratique la même politique. Il redoute : *"Si on met notre outil à la disposition du public, les gens vont perfectionner encore leurs méthodes de génération."*

Plusieurs services de détection proposant des interfaces utilisateurs publiques ont vu le jour et beaucoup de laboratoires universitaires sont également sur le coup, notamment le projet DeFake du Rochester Institute of Technology à New York et le projet DeepFake-o-meter de l'université de Buffalo. L'agence américaine Darpa (Agence de recherche avancée sur la défense) a lancé en 2021 un projet baptisé SemaFor ayant pour mission de découvrir le qui, le quoi, le pourquoi et le comment derrière chaque fichier de synthèse. Dans le cadre du projet, une équipe d'une centaine d'universitaires et de chercheurs privés a travaillé main dans la main pour créer plus de 150 outils d'analyse, explique le responsable du projet, Wil Corvey. La plupart sont des algorithmes de détection qui peuvent être utilisés ensemble ou séparément.

En raison du très grand nombre de générateurs et de détecteurs d'IA, et de la spécificité de chaque cas, le taux de précision est très variable. D'autant que la situation évolue constamment du fait de la course aux armements entre ces deux types d'outils. Pour de nombreux types de contenus, les taux de réussite sont plutôt faibles à l'heure où l'on parle. Pour un texte généré par l'IA, une étude menée cette année sur 14 outils de détection a révélé que tous avaient en commun de n'être *"ni précis ni fiables"*. Pour les vidéos, un concours très médiatisé, organisé en 2020, a été remporté par un outil dont le taux de précision n'était que de 65 %. Pour les images, Anderson Rocha explique que, si l'outil de génération est connu, les détecteurs peuvent afficher un taux de réussite supérieur à 95 %. Mais si le générateur est de type nouveau ou inconnu, ce taux s'effondre dans la plupart des cas. L'utilisation de plusieurs détecteurs d'IA pour une même image peut alors augmenter le taux de réussite, précise Wil Corvey.

L'expert ajoute que la détection des contenus de synthèse ne constitue qu'une pièce du puzzle : à l'heure où de plus en plus d'internautes font appel à l'IA pour modifier leurs productions, la question n'est pas de savoir dans quelle mesure un contenu est artificiel, mais plutôt dans quel but il a été créé. C'est la raison pour laquelle une bonne partie du travail de SemaFor consiste à déterminer l'intention qui se cache derrière les faux, en attribuant le contenu à un créateur et en caractérisant son objectif. Un projet parallèle de la Darpa, l'Incas (Programme de sensibilisation et d'interprétation du sens), planche actuellement sur des outils automatisés capables de détecter les indices des campagnes de désinformation de masse faisant appel à l'IA.

SemaFor en est actuellement à la troisième et dernière étape de son projet, qui voit Wil Corvey se concentrer sur la recherche d'utilisateurs potentiels, comme les réseaux sociaux. *"On a contacté un certain nombre de sociétés, dont Google. À notre connaissance, aucune ne se sert de nos algorithmes de manière régulière"*, dit-il. Meta a collaboré avec des chercheurs de l'université d'État du Michigan, à East Lansing, sur les détecteurs d'IA mais n'a rien dit sur l'usage qu'il prévoyait d'en faire. Hany Farid collabore avec la plateforme professionnelle LinkedIn, qui se sert de détecteurs basés sur l'IA pour repérer et éliminer les visages de synthèse apparaissant sur les comptes frauduleux.

Wael Abd-Almageed est favorable à ce que les réseaux sociaux utilisent des détecteurs pour scanner toutes les images apparaissant sur leurs sites, éventuellement en mettant un bandeau d'avertissement sur les images soupçonnées d'avoir été créées artificiellement. Mais il s'est cassé le nez lorsqu'il l'a proposé voilà quelques années à une société qu'il ne nommera pas. Wael Abd-Almageed soupire : *"J'ai dit à un réseau social :* 'Prenez mon logiciel et utilisez-le gratuitement.' *Ils m'ont répondu :* 'Si vous n'êtes pas capable de nous montrer comment faire de l'argent avec, ça ne nous intéresse pas.'"

Hany Farid estime pour sa part que les détecteurs automatiques ne sont pas adaptés à ce type d'usage : même un outil fiable à 99 % se tromperait encore une fois sur cent, ce qui suffirait selon lui à saper totalement la confiance du public. Pour lui, la détection doit se concentrer sur des enquêtes approfondies, pilotées par des humains, sur des cas précis, plutôt que de vouloir faire la police sur l'ensemble de la Toile.

Beaucoup soutiennent que les entreprises comme les éditeurs et les réseaux sociaux auront besoin d'un cadre réglementaire pour les inciter à adopter un comportement responsable. En juin dernier, le Parlement européen a adopté un projet de loi visant à encadrer les utilisations à haut risque de l'IA et à imposer la transparence pour les contenus générés par ce type d'outils. *"Le monde regarde, parce que l'UE a pris de l'avance dans le domaine"*, commente Sophie Nightingale. Reste que les experts sont loin d'être d'accord sur les mérites d'un tel encadrement et redoutent qu'il ne bride l'innovation. Aux États-Unis, quelques projets de loi sur l'IA sont en cours d'examen, dont l'un visant à empêcher les *deepfakes* d'images intimes et un autre sur l'utilisation de l'IA dans la publicité à caractère politique, mais aucun des deux n'est encore sûr de passer.

Les experts s'accordent au moins sur une chose : l'amélioration de la culture technologique contribuera à empêcher la société et la démocratie de se retrouver noyées sous les faux. *"Il faut faire passer le message pour que les gens prennent conscience de ce qui est en train de se passer*, met en garde Anderson Rocha. *Quand ils sont informés, ils ont des leviers. Ils peuvent réclamer que ça figure par exemple dans les programmes scolaires."*

Même avec tous les outils technologiques et sociaux à notre disposition, estime Hany Farid, c'est un combat perdu d'avance que de vouloir enrayer ou repérer toutes les supercheries. *"Mais ce n'est pas si grave, parce que, même si ça échoue, j'aurais au moins enlevé [cette technologie] des mains du commun des mortels"*, dit-il. Ainsi, comme avec la fausse monnaie, il sera toujours possible de duper le monde avec des faux générés artificiellement par l'IA, mais ce sera beaucoup plus dur.

—Nicola Jones,
publié le 27 septembre

SUR NOTRE SITE

courrierinternational.com

Dernières innovations, débats, analyses, reportages et témoignages : suivez sur notre site l'actualité des intelligences artificielles, en pleine effervescence.