Anton Zeilinger
Prix Nobel de physique
LES MYSTÈRES DE L'INTRICATION QUANTIQUE
p.122

Corinne Le Quéré
Climate Change Committee
INSÉRER LE CLIMAT DANS L'ENSEMBLE DES POLITIQUES
p.6

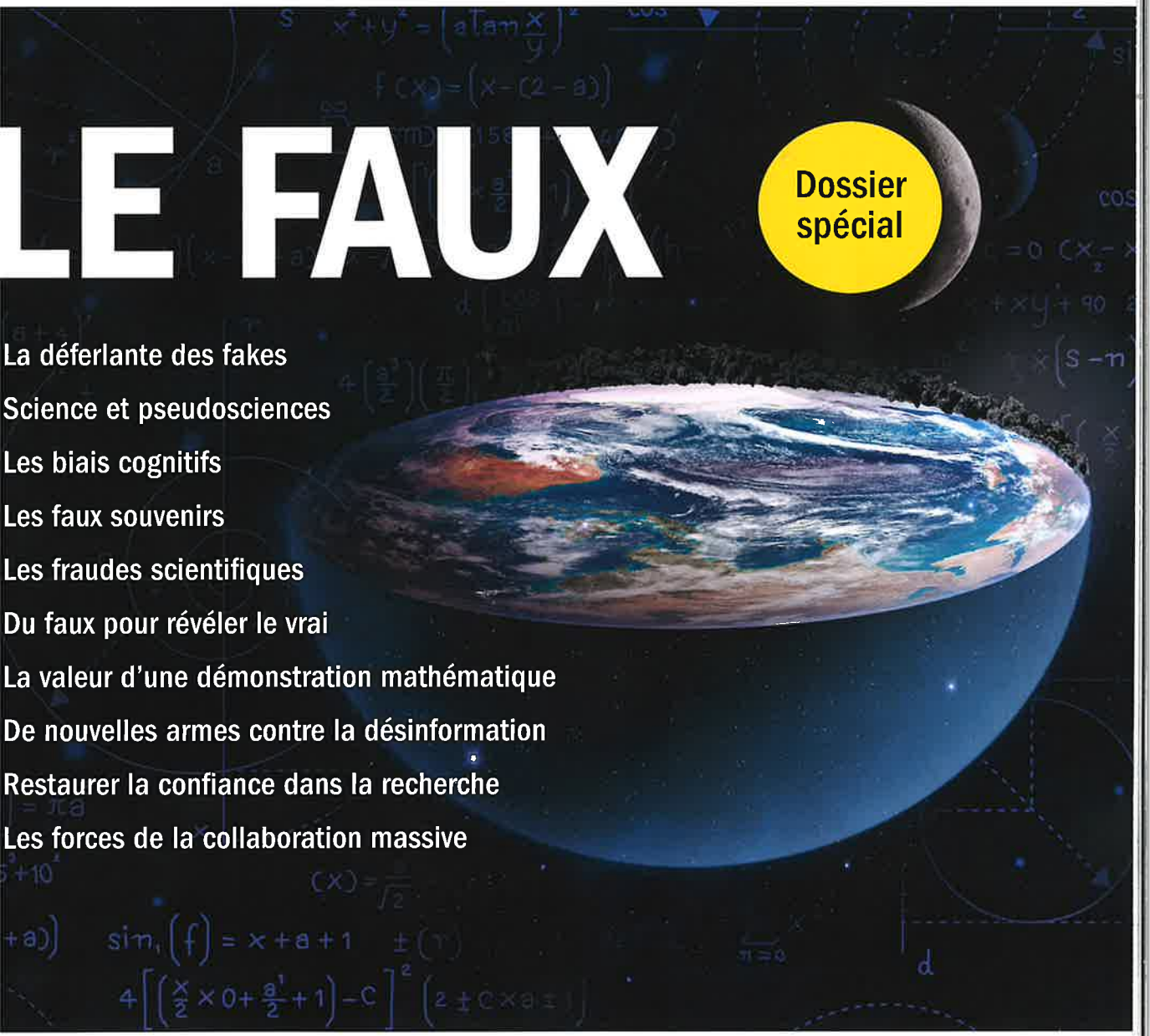
La Recherche

LE MAGAZINE DE RÉFÉRENCE SCIENTIFIQUE - JUILLET / SEPTEMBRE 2025 - 9€90

LE FAUX

Dossier spécial

La déferlante des fakes
Science et pseudosciences
Les biais cognitifs
Les faux souvenirs
Les fraudes scientifiques
Du faux pour révéler le vrai
La valeur d'une démonstration mathématique
De nouvelles armes contre la désinformation
Restaurer la confiance dans la recherche
Les forces de la collaboration massive



rosciences
NEURONES
.A SATIÉTÉ
p.82

Planétologie
LA SURFACE DE MARS REVISITÉE
p.108

Informatique affective
LES ROBOTS LISENT-ILS NOS PENSÉES...
p.102

Paléoécologie
IL PLEUT DANS LE SAHARA
p.96

Éthologie
LES GORILLES DES PLAINES
p.76

Français sont les plus nombreux en Europe à avoir déserté ce réseau : 2,7 millions d'utilisateurs en sont partis, soit plus de 25 %. Cela signifie que des milliers de citoyens ont fait le choix d'explorer d'autres réseaux, même si la problématique d'une appropriation plus large d'espaces numériques transparents et ouverts reste d'actualité.

S'EMPARER DES OUTILS DE L'ÉMANCIPATION NUMÉRIQUE

Cette expérience révèle aussi une bascule entre ces deux paradigmes du numérique mentionnés plus haut : celui des plateformes fermées, centralisées, basées sur la captation des données et la rétention des utilisateurs ; et le second, émergent, reposant sur des protocoles ouverts, des logiciels libres et la portabilité des comptes. Mastodon et BlueSky ne sont pas des clones de X, mais les premiers éléments d'un nouvel écosystème. Comme le Web à ses débuts, ces protocoles permettront demain des usages encore insoupçonnés, bien au-delà du simple microblogging.

Certes, la tentation de rester sur X, par confort, par habitude ou pour préserver son audience, est forte. Mais c'est une illusion. Ces audiences sont gonflées par les comptes inactifs ou les *bots* et, surtout, elles ne sont pas à l'abri de pratiques de censure ciblées et arbitraires, comme l'a montré la fermeture de comptes d'opposants au président Recep Tayyip Erdogan par Musk, en pleine campagne électorale en Turquie. Nous devons sortir de cette dépendance. D'autres solutions existent et elles s'améliorent chaque jour. Si chacun fait évoluer ses pratiques, nous pourrons reconstruire un espace public plus sain, plus résilient, et moins manipulable.

Quelles leçons tirer de ces attaques ? Les affronter n'a été ni agréable ni facile, mais elles ont confirmé l'importance de notre démarche. Car elles illustrent à quel point les initiatives citoyennes, même scientifiques, sont devenues des cibles lorsqu'elles remettent en cause les rentes informationnelles. La meilleure réponse, c'est la transparence. Raconter ce que mon équipe a vécu, documenter les attaques, analyser leurs logiques sont une manière de résister. Et d'inviter chacun à ne plus se contenter d'un Internet dégradé et captif, mais à s'emparer des outils de l'émancipation numérique. ■

Les vertiges d'une réalité artificielle

Apparus il y a dix ans avec les premières intelligences artificielles génératives, les *deepfakes* sont désormais un problème de société. Photos, vidéos ou voix sorties d'un algorithme mais capables de passer pour authentiques, ces contenus posent aussi le défi technique de leur détection automatique.

▲ *Ce deepfake devenu célèbre, montrant le pape François, a été généré par le programme Midjourney et diffusé à l'échelle mondiale en 2023.*

DR

nçais sont les plus nombreux en Europe à
oir déserté ce réseau : 2,7 millions d'utilisateurs
sont partis, soit plus de 25 %. Cela signifie que
s milliers de citoyens ont fait le choix d'explo-
d'autres réseaux, même si la problématique
ne appropriation plus large d'espaces numé-
ues transparents et ouverts reste d'actualité.

EMPARER DES OUTILS DE EMANCIPATION NUMÉRIQUE

tte expérience révèle aussi une bascule entre
s deux paradigmes du numérique mentionnés
us haut : celui des plateformes fermées, centra-
ées, basées sur la captation des données et la
ention des utilisateurs ; et le second, émergent,
posant sur des protocoles ouverts, des logiciels
res et la portabilité des comptes. Mastodon et
ueSky ne sont pas des clones de X, mais les pre-
ers éléments d'un nouvel écosystème. Comme
Web à ses débuts, ces protocoles permettront
main des usages encore insoupçonnés, bien
-delà du simple microblogging.
ertes, la tentation de rester sur X, par confort,
r habitude ou pour préserver son audience,
t forte. Mais c'est une illusion. Ces audiences
nt gonflées par les comptes inactifs ou les
ts et, surtout, elles ne sont pas à l'abri de pra-
ques de censure ciblées et arbitraires, comme
montré la fermeture de comptes d'opposants
président Recep Tayyip Erdogan par Musk, en
eine campagne électorale en Turquie. Nous
evons sortir de cette dépendance. D'autres
lutions existent et elles s'améliorent chaque
ur. Si chacun fait évoluer ses pratiques, nous
ourrons reconstruire un espace public plus
in, plus résilient, et moins manipulable.
uelles leçons tirer de ces attaques ? Les affron-
r n'a été ni agréable ni facile, mais elles ont
onfirmé l'importance de notre démarche.
ar elles illustrent à quel point les initiatives
toyennes, même scientifiques, sont devenues
es cibles lorsqu'elles remettent en cause les
ntes informationnelles. La meilleure réponse,
est la transparence. Raconter ce que mon
quipe a vécu, documenter les attaques, ana-
ser leurs logiques sont une manière de résister.
t d'inviter chacun à ne plus se contenter d'un
ternet dégradé et captif, mais à s'emparer des
utils de l'émancipation numérique. ■

Les vertiges d'une réalité artificielle

Apparus il y a dix ans avec les premières intelligences artificielles génératives, les *deepfakes* sont désormais un problème de société. Photos, vidéos ou voix sorties d'un algorithme mais capables de passer pour authentiques, ces contenus posent aussi le défi technique de leur détection automatique.

▲ Ce deepfake *devenu célèbre, montrant le pape François, a été généré par le programme Midjourney et diffusé à l'échelle mondiale en 2023.*

DR

Gérôme Billois est expert cybersécurité au cabinet Wavestone. Ce 23 janvier 2025, il intervient en visioconférence dans une salle du Campus Cyber à la Défense, dans le cadre de la présentation annuelle des risques cyber faite par le Clusif, association de professionnels du domaine. Et d'emblée, ses propos détonnent : *« Nous avons analysé en groupe de travail toutes ces évolutions en profondeur, et j'ai une annonce importante : il n'y a aucun risque lié à l'intelligence artificielle. Cette technologie est parfaite, il n'y a pas de faille. »* Provocation ? Méconnaissance du sujet ? Ni l'un ni l'autre : un *deepfake* ! C'est-à-dire un contenu généré par intelligence artificielle (IA). La voix et le visage de Gérôme Billois ont été recréés par un algorithme d'apprentissage, puis substitués et synchronisés au flux audiovisuel d'un autre intervenant basé à Nantes. Manière d'illustrer les dangers bien réels de l'IA.

Les premiers *deepfakes* étaient produits avec une méthode inventée en 2014 (le terme, lui, date de 2017) par un doctorant en informatique de l'université de Montréal (Canada) : les réseaux génératifs antagonistes, ou GAN. De nombreux travaux académiques, comme le pionnier Face2Face (Institut Max-Planck d'informatique, universités Friedrich-Alexander d'Erlangen-Nuremberg, en Allemagne, et Stanford, aux États-Unis), ont suivi. En 2019, le projet *Do as I Do* d'une équipe de l'université de Californie à Berkeley altérait le comportement du corps entier d'un individu à partir d'une autre vidéo.

Dès l'origine, le potentiel malveillant n'a échappé à personne. Il n'a fait que s'amplifier avec le raffinement des techniques, la disponibilité massive de contenus sur Internet pour entraîner les algorithmes et l'apparition, après les GAN, des modèles de diffusion sur lesquels reposent les IA génératives actuelles telles que Dall-E, Stable Diffusion, Imagen de Google DeepMind ou le système Make-A-Video de Meta. *« Les GAN ont l'avantage d'être plus rapides, alors que les modèles de diffusion reposent sur un processus itératif lourd consistant à débruiter une image au fur et à mesure,* note Alexandre Libourel, doctorant spécialisé en sécurité numérique à l'école d'ingénieur Eurecom. *Mais ces derniers ont une grande variabilité sur les contenus générés, tandis que les GAN restent proches de leurs données d'entraînement. »*

Il y a dix ans, il fallait des dizaines voire des centaines d'images pour entraîner un algorithme à générer le visage d'une personne. Le faux Gérôme Billois de 2025, lui, n'en a nécessité que quelques-unes, ainsi que trois minutes d'extrait de voix. *« Sur des salons, nous faisons même, de temps en temps, une démonstration en temps réel,* renchérit Alexandre Libourel. *Nous filmons quelqu'un et au même moment, sur grand écran, on fait apparaître à la place de son visage le deepfake d'une personnalité. »* À l'automne 2024, Microsoft Research Asie a dévoilé le modèle Vasa, capable, à partir du fichier audio d'une voix, d'appliquer automatiquement les mouvements faciaux et de tête adéquats à l'image fixe d'un visage.

DES IMPLICATIONS GÉOPOLITIQUES

Photos et vidéos photoréalistes, voix, texte, etc. : tout, aujourd'hui, peut devenir un *deepfake* exploitable, au travers d'outils accessibles, sans être un informaticien chevronné. Début 2024, un employé du groupe britannique Arup, à Hong Kong, a transféré à des escrocs un total de 25,6 millions de dollars, trompé par les *deepfakes* de membres de sa hiérarchie lors d'une visioconférence. Six mois plus tard, la société KnowBe4 de sensibilisation à la cybersécurité embauchait un ingénieur logiciel américain au terme d'un processus incluant quatre entretiens par visioconférence et diverses vérifications. La recrue était en fait un cybercriminel nord-coréen, caché derrière une fausse identité.

Le phénomène a des implications géopolitiques, suivies en France par l'agence Viginum, chargée de la surveillance des ingérences numériques étrangères. Mi-février 2025, ce service du secrétariat général de la Défense et de la Sécurité nationale a présenté deux métadétecteurs de contenus générés par IA, l'un pour les images, l'autre pour les textes. Développés avec le Pôle d'expertise de la régulation numérique (PEReN), ils évaluent les outils de détection existants, notamment au regard de leur circulation en ligne. *« Les images sont continuellement transformées au fil des partages sur les réseaux sociaux, par l'appli-*

FaceSwap versus *reenactment*

Il existe deux grandes techniques pour altérer la vidéo d'un visage par IA. Le FaceSwap consiste à substituer celui d'une personne par celui d'une autre. Le second va adopter automatiquement les poses et mouvements du premier, comme un moulage numérique. Dans le *reenactment*, une personne se filme en train de jouer une scène et transfère les mouvements, la gestuelle sur une personne cible, laquelle devient comme une marionnette numérique.

cation de filtres, des recadrages, l'ajout de texte, ce qui altère l'efficacité des détecteurs», explique Gaspard Defréville, expert en science des données, au PEReN. La célèbre fausse photo du pape François en doudoune blanche est par exemple bien reconnue comme un *deepfake*, mais il suffit de modifier la couleur du vêtement pour que les mêmes détecteurs jugent l'image authentique.
Les analystes voient également passer de plus en plus de textes manipulateurs contenant des formules-types de modèles de langage (LLM) comme «*Bien sûr. Voici un article…*», «*En conclusion…*» ou «*En tant que modèle de langue…*». Or, les détecteurs testés par Viginum ne les classent pas comme artificiels. L'agence a aussi trouvé des textes anglophones incluant des idéogrammes chinois, preuve d'une «hallucination» du LLM chinois Qwen, mais qui échappe à certains outils de détection. Là encore, ces derniers sont entraînés sur des corpus académiques trop «propres» par rapport à ce qui se partage en ligne.

La biométrie comportementale, un moyen inédit de détection

La détection s'est initialement concentrée sur les défauts des *deepfakes*, notamment dans les images (répartition des pixels, incohérence, etc.). Or, non seulement les techniques de génération ne cessent de s'améliorer mais, en plus, toutes ne produisent pas les mêmes artefacts. Et un détecteur ne peut pas être entraîné sur des contenus issus de tous les générateurs existants. D'où l'approche d'une équipe du laboratoire CortAIx de Thales : un métadétecteur qui se base sur les interactions de cinq détecteurs du monde académique. Dans une même image, chacun analyse des choses différentes pour évaluer la probabilité qu'elle soit artificielle. «*Si le modèle 1 et le modèle 2 donnent le même verdict sur une image et que, dans la quasi-totalité de ces cas, ils ont raison, notre métamodèle va apprendre comme règle de faire confiance à ces deux modèles quand ils sont d'accord*», explique Rodolphe Lampe, qui travaille au sein de l'équipe. D'autres combinaisons de ce type permettront au métadétecteur de multiplier les critères afin de gagner en efficacité, sans avoir été lui-même entraîné sur des *deepfakes*.

UN AVANTAGE STRATÉGIQUE

Quant au projet DeTox, mené actuellement dans le cadre du programme Astrid de l'Agence de l'innovation de défense, il opte pour un axe inédit en matière de *deepfake* : celui de la biométrie comportementale d'une personnalité spécifique, politique, militaire, économique. L'idée étant de parer aux manipulations à vocation de stratégie géopolitique. «*On s'intéresse à la dynamique faciale de la personne*, explique Jean-Luc Dugelay, enseignant-chercheur à Eurecom et porteur du projet. *Chacun a une façon bien à lui de froncer les sourcils, de hocher ou osciller la tête, et nous allons exploiter ces informations pour détecter si, dans une vidéo, les mouvements sont bien ceux de la personne.*» Cela passe par la création d'un réseau de neurones entraîné à reconnaître un genre de signature comportementale d'une personne quand elle s'exprime. Avec la contrainte de devoir réentraîner le modèle pour chaque personnalité, mais sans avoir à se soucier du générateur qui a été utilisé.
DeTox bénéficie de l'Institut de recherche et coordination acoustique/musique en synthèse sonore pour la création de *deepfakes* audiovisuels les plus réalistes possibles. Ils sont destinés à être soumis aux détecteurs d'Eurecom, et donc à entraîner au mieux ces derniers. Ces faux contenus sont engendrés à partir des mêmes données publiques accessibles aux créateurs de *deepfakes* (chaînes YouTube, sites officiels), mais aussi avec des données propres, captées spécialement pour cette recherche et de bien meilleure qualité que ce qui circule en ligne. «*Une ancienne ministre est venue enregistrer dans nos studios devant trois caméras couvrant l'intégralité des poses pour la modélisation*, raconte Nicolas Obin, spécialiste en synthèse de sons, *et la voix a été enregistrée avec des techniques professionnelles. Cela nous donne un avantage stratégique sur l'attaquant.*»
Ce qui n'est pas si fréquent dans un contexte où, comme en cybercriminalité, ce dernier a toujours un temps d'avance. ■ **Arnaud Devillard**

débat et les heures qui suivent, chaque camp va remettre puis enlever le 3. Au total, une cinquantaine de modifications seront effectuées en à peine plus de deux heures ! La direction de Wikipédia finira par intervenir dans cette guerre d'édition stérile, qu'elle nomme « vandalisme », en gelant la page dans sa version avant le raid de manipulation.

LE LONG TRAVAIL DE SAPE DE LA FACHOSPHÈRE AMÉRICAINE

Wikipédia est aussi devenue une cible de choix dans la guerre culturelle qui se déroule aux États-Unis. Elle est notamment entrée dans le viseur d'Elon Musk et de ses soutiens, qui ont appelé à boycotter le financement de l'encyclopédie à but non lucratif (elle vit de dons) en l'accusant d'être biaisée et de servir d'outil de « propagande » au wokisme. Par l'usage d'un jeu de mots stigmatisant, « wokepédia », ils lui reprochent d'être trop progressiste et d'introduire des descriptifs critiques vis-à-vis des théories conspirationnistes et des opérations de désinformation auxquelles le magnat trumpiste adhère et qu'il encourage, notamment sur le réseau social X dont il est le propriétaire.

Ces attaques participent d'un mouvement plus ancien lancé par la fachosphère américaine ; dès 2020, celle-ci qualifiait Wikipédia, dans un tweet, de *« groupe politique de gauchistes fanatiques »*, juste parce qu'elle reste attachée à la science, aux faits, aux preuves à apporter pour justifier ce qui est publié, et à la notion de consensus.

Après avoir mené un long travail de sape contre les journalistes et les médias, et contre les scientifiques – notamment les climatologues –, les réactionnaires ont donc pris pour nouvelle cible l'encyclopédie citoyenne et participative, dans l'espoir qu'elle cesse d'être un point de repère de la connaissance dans un océan de mensonges.

Face à ces attaques, préserver le principe de fonctionnement de Wikipédia, reposant sur la collaboration (massive), l'utilisation du consensus pour arbitrer les différends ou encore la vérification et le contrôle des contenus en temps réel – très similaire à la manière dont fonctionne la recherche scientifique –, est crucial. C'est ce principe qui fait de l'encyclopédie un outil précieux dans la lutte contre la désinformation. ■

Le marquage numérique en devenir

Distinguer le faux du vrai devient difficile avec l'intelligence artificielle. L'idée de marquer numériquement les contenus créés est une piste privilégiée pour garantir la transparence. Plusieurs méthodes existent, mais elles sont encore loin d'avoir fait leurs preuves.

Première loi européenne sur l'intelligence artificielle, l'*IA Act* est entrée en vigueur début août 2024, avec pour enjeu de se prémunir contre les risques et dérives inhérents à ces technologies. Au rang desquels, les contenus générés, source de confusion et de manipulation. Pour cela, le préambule de la loi avance explicitement une solution : *« Le marquage dans un format lisible par machine et la détection du fait que les sorties ont été générées ou manipulées par un système d'IA, et non par un être humain. »*

Le texte mentionne plusieurs méthodes possibles, comme les filigranes (ou *watermarks*), les métadonnées ou la cryptographie. La démarche n'a rien de nouveau, elle servait déjà aux premiers temps de la lutte contre le piratage de musique et de films sur Internet. Mais, à l'aune des *deepfakes* (lire p. 32), elle nourrit nombre de recherches, sachant que la diversité des contenus à marquer (texte, son, image) complique la tâche.

En février 2021, le *New York Times*, la BBC, Microsoft, Intel et Adobe ont créé le consortium C2PA autour d'un standard de marquage des images (qui ne concernait pas spécifiquement l'IA générative au départ). Ce genre de signature numérique consiste en l'ajout automa-

eures qui suivent, chaque camp
uis enlever le 3. Au total, une
de modifications seront effec-
ne plus de deux heures ! La direc-
édia finira par intervenir dans
l'édition stérile, qu'elle nomme
», en gelant la page dans sa ver-
aid de manipulation.

AVAIL DE SAPE
OSPHÈRE AMÉRICAINE

aussi devenue une cible de choix
e culturelle qui se déroule aux
le est notamment entrée dans le
Musk et de ses soutiens, qui ont
cotter le financement de l'ency-
ut non lucratif (elle vit de dons)
d'être biaisée et de servir d'ou-
gande » au wokisme. Par l'usage
nots stigmatisant, « wokepédia »,
ent d'être trop progressiste et d'in-
descriptifs critiques vis-à-vis des
spirationnistes et des opérations
ation auxquelles le magnat trum-
et qu'il encourage, notamment sur
al X dont il est le propriétaire.
participent d'un mouvement plus
par la fachosphère américaine ; dès
qualifiait Wikipédia, dans un tweet,
olitique de gauchistes fanatiques »,
u'elle reste attachée à la science,
preuves à apporter pour justifier
blié, et à la notion de consensus.
nené un long travail de sape contre
es et les médias, et contre les scien-
tamment les climatologues –, les
es ont donc pris pour nouvelle cible
lie citoyenne et participative, dans
le cesse d'être un point de repère de
nce dans un océan de mensonges.
ttaques, préserver le principe de
nent de Wikipédia, reposant sur la
n (massive), l'utilisation du consen-
itrer les différends ou encore la véri-
e contrôle des contenus en temps
nilaire à la manière dont fonctionne
e scientifique –, est crucial. C'est ce
fait de l'encyclopédie un outil pré-
a lutte contre la désinformation. ■

Le marquage numérique en devenir

Distinguer le faux du vrai devient difficile avec l'intelligence artificielle. L'idée de marquer numériquement les contenus créés est une piste privilégiée pour garantir la transparence. Plusieurs méthodes existent, mais elles sont encore loin d'avoir fait leurs preuves.

Première loi européenne sur l'intelligence artificielle, l'*IA Act* est entrée en vigueur début août 2024, avec pour enjeu de se prémunir contre les risques et dérives inhérents à ces technologies. Au rang desquels, les contenus générés, source de confusion et de manipulation. Pour cela, le préambule de la loi avance explicitement une solution : *« Le marquage dans un format lisible par machine et la détection du fait que les sorties ont été générées ou manipulées par un système d'IA, et non par un être humain. »*

Le texte mentionne plusieurs méthodes possibles, comme les filigranes (ou *watermarks*), les métadonnées ou la cryptographie. La démarche n'a rien de nouveau, elle servait déjà aux premiers temps de la lutte contre le piratage de musique et de films sur Internet. Mais, à l'aune des *deepfakes* (lire p. 32), elle nourrit nombre de recherches, sachant que la diversité des contenus à marquer (texte, son, image) complique la tâche.

En février 2021, le *New York Times*, la BBC, Microsoft, Intel et Adobe ont créé le consortium C2PA autour d'un standard de marquage des images (qui ne concernait pas spécifiquement l'IA générative au départ). Ce genre de signature numérique consiste en l'ajout automatique de métadonnées, les *content credentials*, à une image, renseignant à la fois sur l'origine de cette dernière mais aussi sur toutes les altérations qu'elle a subies à mesure des publications et autres partages sur les réseaux. En 2024, Google, Meta, OpenAI, Amazon ont rejoint ce projet en vue d'intégrer ce standard à leurs outils d'IA. La même année, le réseau social LinkedIn l'a adopté pour permettre à ses utilisateurs de vérifier si une image, fixe ou vidéo (pas le texte ni la voix), a été créée artificiellement. Une initiative d'envergure donc, mais qui a ses limites. Notamment celle de nécessiter une interopérabilité avec d'autres solutions adoptées hors consortium. Sans oublier que les créateurs de *deepfakes* vont privilégier des générateurs sans marquage, quitte à les développer eux-mêmes. *« Il faudrait que les attaquants n'aient accès qu'à des générateurs qui marquent automatiquement, et qu'en plus, les logiciels de lecture refusent de lire une vidéo qui n'aurait pas été marquée »*, résume Jean-Luc Dugelay, enseignant-chercheur à l'École d'ingénieurs Eurecom à Sophia Antipolis et spécialiste de *deepfakes*.

Quant au *watermarking*, il consiste à utiliser le processus de génération lui-même d'une manière telle que l'on puisse reconnaître un mécanisme de création artificielle, sans nuire à la lisibilité du contenu. On sait le faire depuis longtemps pour des images, mais c'est plus difficile pour le texte.

(*) **Les *tokens*** sont les groupes de lettres et de caractères appris par un modèle de langage et qui forment les mots et les phrases.

En mai 2024, une équipe d'experts en apprentissage automatique et sécurité informatique de l'université du Maryland (États-Unis) a toutefois publié une technique innovante consistant à introduire un biais dans la génération de *tokens*(*). Un tel algorithme produit du texte en opérant une sélection dans une liste de *tokens* probables sur le plan statistique. Mais, là, la méthode privilégie des *tokens* verts , autrement dit, elle force la probabilité que ceux-ci soient générés et baisse celle d'autres (les *tokens* rouges).

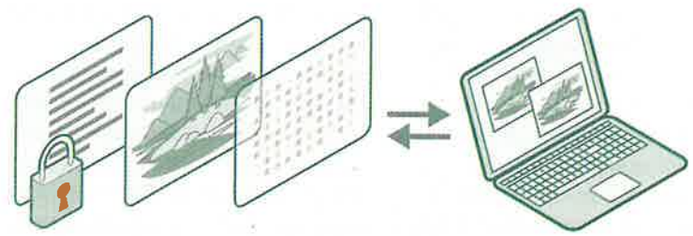
UN MANQUE DE ROBUSTESSE

On peut aussi faire varier ce biais au fil de la génération, selon divers critères. C'est non seulement indécelable par l'œil humain et sans impact sur l'intelligibilité du résultat mais, de surcroît, produire un texte artificiel débarrassé de ce *watermark* pour échapper aux détecteurs exige de connaître la manière dont le LLM utilisé « tokenise ». Car chaque modèle a sa propre manière de le faire. C'est aussi la limite, là encore, de cette solution : elle n'est pas interopérable, il faut tout redévelopper pour chaque LLM et l'adapter à sa « tokenisation ». Pourtant, quelques mois après cette publication, des chercheurs des universités Harvard, George Mason (États-Unis) et *La Sapienza* à Rome (Italie) montraient qu'il était possible de retirer ce *watermark*. Fatalistes, ils concluent même leur article par *« l'impossibilité d'un marquage robuste dans les modèles génératifs »*. Ce qui n'empêche pas la communauté scientifique de continuer d'explorer de nouvelles pistes.

Le défi technique se double aussi de considérations pratiques. Comment faire le distinguo entre un texte dont on a juste amélioré la syntaxe avec un LLM, et un autre complètement artificiel rempli de fausses informations, de biais volontaires ou de plagiats ? Un *watermark* risque en effet de pénaliser injustement toute une frange d'utilisateurs. C'est l'un des arguments qui retient OpenAI d'intégrer à ChatGPT une technique de *watermark* prête depuis l'an dernier. Il existe toutefois une autre raison, révélée par le *Wall Street Journal* en août 2024 : la crainte que les utilisateurs se détournent de l'interface pour privilégier des concurrents ne pratiquant pas le marquage numérique… ■

Arnaud Devillard

PROTECTION EN TROIS ÉTAPES : EMPREINTE, MARQUAGE ET VÉRIFICATION CERTIFIÉE



BRUNO BOURGEOIS

1 REGISTRE INVIOLABLE

Un registre enregistre avec le fichier une empreinte numérique unique et des métadonnées de provenance afin de créer une preuve vérifiable de l'intégrité, impossible à altérer.

2 FILIGRANES (*WATERMARKS*)

On intègre au fichier un filigrane imperceptible (*watermarking*) qui reste détectable, même si le fichier est modifié ou si ses métadonnées sont altérées ou supprimées.

3 SITE DE VÉRIFICATION

Un site permet de vérifier l'authenticité d'un fichier reçu en le comparant à la version stockée ; on peut aussi vérifier son intégrité grâce aux empreintes et aux filigranes.