Les Echos
N° 24335
mardi 12 novembre 2024
Édition(s) : Edition principale
Page 15
1208 mots - 5 min

Face à des IA trop humaines, le défi de l'identification

Alors que les bots génèrent plus de la moitié du trafic en ligne, identifier les contenus créés par l'IA est devenu un enjeu commercial, sécuritaire et éthique majeur.

Blade Runner », le film culte de Ridley Scott sorti en 1982, comporte un test de Turing qui tourne mal : Leon Kowalski, l'androïde sur la sellette, assassine le policier qui tente de déterminer s'il est un humain ou un robot en l'interrogeant et en scrutant son iris. Quarante-deux ans plus tard, « *il est devenu impossible de distinguer à l'oeil ou à l'oreille les contenus générés par certaines Intelligences Artificielles* », s'inquiète Erwan Le Merrer, chercheur au sein de l'équipe Artishau (ARTificial Intelligence, Security, trutHfulness and Audit) d'Inria, à Rennes.

« *Certaines IA ont déjà passé mieux que des humains les tests de Turing, y compris les Captchas, utilisés sur Internet pour vérifier qu'un visiteur n'est pas un robot* », ajoute Steven Smith, vice-président ingénierie chez TFH (Tools for Humanity), une entreprise de San Francisco qui propose une technologie permettant de s'assurer que la personne avec qui l'on discute sur Telegram ou Zoom n'est pas un bot. Il suffit de se faire enregistrer comme humain auprès de TFH en faisant photographier son iris par l'Orb, sorte de gros oeil mécanique et bourré d'IA, programmé pour repérer les robots.

L'ombre des bots

Cynisme ? Intuition de ce que sera notre avenir ? Un des deux cofondateurs de TFH n'est autre que Sam Altman, également patron d'OpenAI, l'éditeur de ChatGPT. Cette GenAI (Intelligence Artificielle générative) qui fêtera bientôt ses deux ans est à l'origine du grand carnaval actuel, où l'on ne peut plus distinguer un humain d'une IA. « *Les bots, ces logiciels qui se font passer pour des humains, représentent entre 50 et 60 % du trafic sur Internet* », rappelle Christophe Lebrun, data scientist et adjoint scientifique à la Haute Ecole de Gestion de Genève où il s'occupe, entre autres, de plagiat. Sur le site de Harvard, Latanya Sweeney, professeur dans cette université, estime qu'à l'avenir, sur Internet, « *90 % du contenu ne sera plus généré par des humains mais par des robots* ». La start-up américaine NewsGuard a déjà identifié plus de 1.110 sites d'information, y compris en français, non fiables car entièrement rédigés par une IA.

« *Si nous ne faisons rien, les activités malveillantes permises par l'IA risquent de définitivement polluer le Web* », avertissent trente-deux chercheurs d'OpenAI, Microsoft, Harvard, Berkeley, du MIT etc., dans un article paru en août. Ces juristes, informaticiens et spécialistes de l'éthique appartiennent à une nouvelle discipline, la « *sécurité de l'IA* ». « *Ce domaine scientifique, qui explore de nouvelles méthodes pour s'assurer qu'un contenu n'a pas été généré par une GenAI, est à la jonction de la cybersécurité, l'algorithmique, la statistique* », détaille Erwan Le Merrer.

Tests de Turing

Même si ses travaux peuvent également protéger les humains, la sécurité de l'IA tente surtout de garantir sa propre intégrité : l'utilisation de data synthétiques pour l'entraînement des prochaines générations de ces algorithmes provoquera un effondrement de leurs performances. « *Dans ce cas, il pourrait y avoir à terme une baisse de la richesse linguistique des nouvelles données produites* », prévient Chloé Clavel, directrice de recherche en IA d'Inria Paris. « *Il faut donc une sorte de test de Turing inversé, afin de vérifier si les données présentes sur le web sont le fait d'humains ou,*

au contraire, de GenAI », insiste Erwan Le Merrer.

Même s'il s'agit des deux versants d'un même problème, prouver que l'on est bien un humain lors d'une visioconférence ou de l'ouverture d'un compte bancaire en ligne est une chose ; certifier l'origine d'un contenu en est une autre, beaucoup plus difficile car facilement contournable. Pour s'assurer de la réalité humaine de leurs clients, les sites en ligne recourent à des techniques de KYC (Know Your Customer) qui, jusqu'à présent, reposaient souvent sur des selfies. Pour distinguer le vrai (humain) du faux (généré par une IA), il faudrait pouvoir se livrer à une analyse spectrale de la photo envoyée et comparer les résultats à ceux d'une base de données. Des chercheurs du Huawei Noah's Ark Lab, à Montréal, au Canada, spécialisé dans l'IA, viennent de bâtir une photothèque comportant 1,3 million de vrais clichés et autant de faux.

L'identification biométrique, comme celle proposée par TFH, constitue également une solution fiable mais très longue à mettre en oeuvre à l'échelle planétaire : TFH n'a scanné que 7 millions d'iris, essentiellement en Amérique du Sud ; en Europe, elle fait l'objet d'une enquête des CNIL locales. En attendant, les 32 chercheurs d'OpenAI, Microsoft, Harvard, Berkeley, du MIT etc.,proposent de mettre en place des PHC (PersonHood Credentials), *« des certificats numériques qui permettent aux utilisateurs de prouver qu'ils sont des personnes réelles aux services en ligne auxquels ils souhaitent s'inscrire sans divulguer d'informations personnelles »*. Des tiers indépendants seront chargés de vérifier les preuves apportées.

Au problème de la certification de l'origine des contenus, il n'existe pour l'instant pas de solution. Jusqu'ici, la validation des textes, photos ou vidéos mis en ligne reposait soit sur la détection d'un tatouage, un filigrane invisible (suites de mots, pixels…) volontairement introduit au moment de leur génération, soit sur leur analyse par une autre IA, tentant d'y déceler le style typique d'une GenAI.

Exemple de tatouage : SynthID-text, partagé en open source fin octobre par des chercheurs de DeepMind, la filiale de Google spécialisée dans l'IA. *« SynthID-text introduit des informations supplémentaires au moment de la génération du texte en modulant la probabilité que certains morceaux de phrase soient présents, sans compromettre la qualité du texte »*, explique, par mail, Pushmeet Kohli, en charge de la recherche chez DeepMind. Mais il suffit le plus souvent de demander à une seconde IA de réécrire le texte rédigé par une première GenAI pour faire disparaître toute trace de filigrane ou tout style particulier… ·

Jacques Henno ■

par Jacques Henno

La chasse au plagiat

La copie d'un contenu existant concerne aussi bien l'enseignement que les places de marché éditoriales. DCM Swiss a collaboré avec Christophe Lebrun, de la Haute Ecole de gestion de Genève, pour élaborer une méthode de validation des textes de ses contributeurs. *« Nous convertissons le texte en vecteurs qui permettent d'identifier des mots possédant le même sens,* explique-t-il. *Cela permet de comparer le texte à une base de données et de voir s'il y a plagiat ou pas. »*
L'université d'Aix-Marseille traque les « faussaires »dans les publications scientifiques. « Certaines revues prédatrices acceptent des articles générés par une IA, constate Ismail Badache, maître de conférences en informatique à l'Inspé de cette université. Une liste noire de ces supports sans scrupule existe, mais un article bidonné peut aussi se glisser dans une revue sérieuse. »

Les chiffres clés

100 milliards de mots
sont produits chaque jour par ChatGPT et les outils d'OpenAI, selon Sam Altman, son dirigeant fondateur.
1.110 sites d'information
sont déjà rédigés par des IA à travers le monde selon Newsguard.

Les chiffres clés

100 milliards de mots
sont produits chaque jour par ChatGPT et les outils d'OpenAI, selon Sam Altman, son dirigeant fondateur.
1.110 sites d'information
sont déjà rédigés par des IA à travers le monde selon Newsguard.

Parution : Quotidienne
Diffusion : 135089 ex. (Diff. payée Fr.) - © ACPM PV 2020-2021
Audience : 676000 lect. (LNM) - © ACPM One 2021 V4