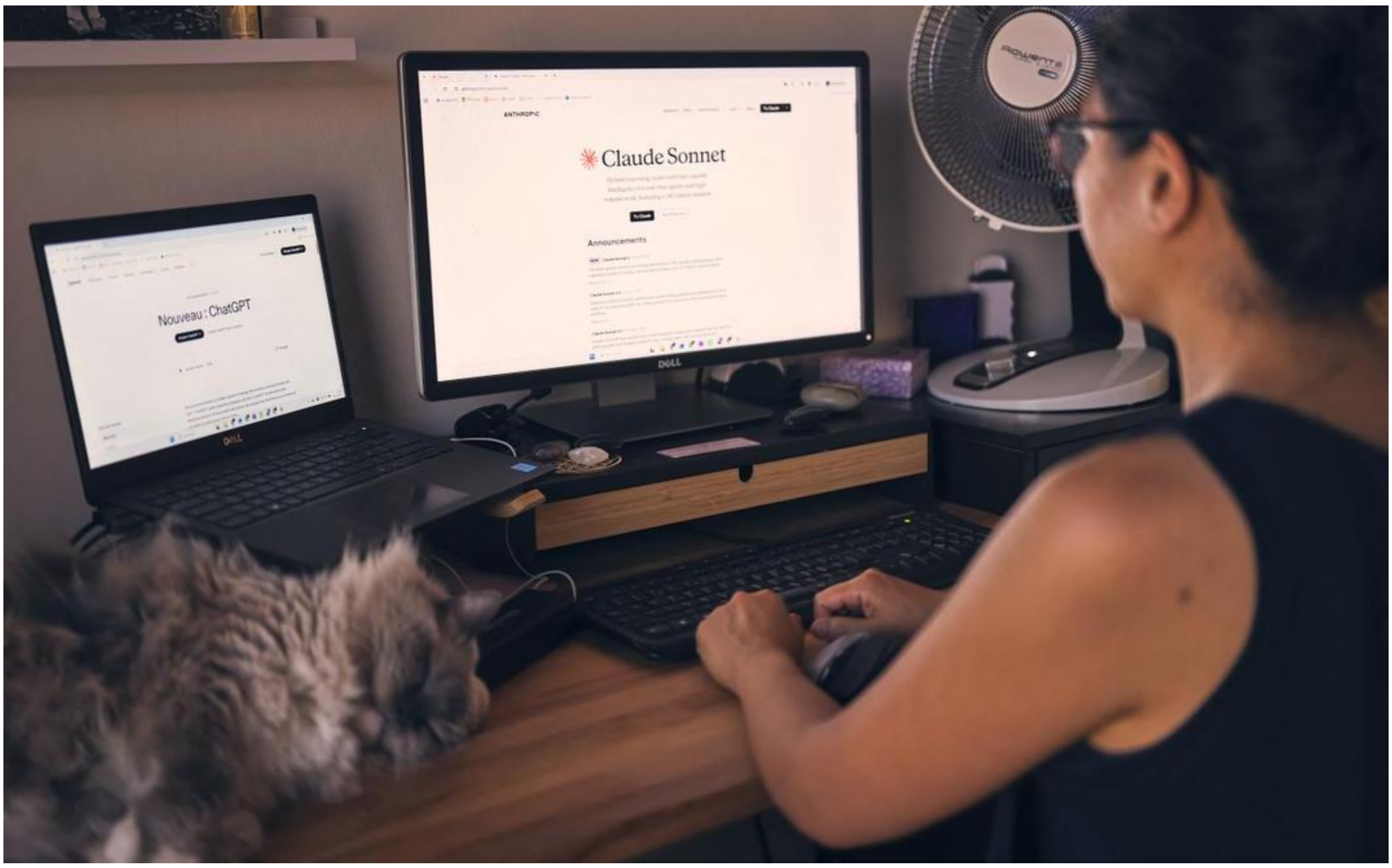
INTELLIGENCE ARTIFICIELLE

Peut-on se fier aux détecteurs d'IA ?



Depuis 2022 et le lancement de ChatGPT, impossible de savoir avec certitude si ce qu'on lit a été écrit par un humain ou créé par une machine. © HANS LUCAS VIA AFP.

Pangram fait office de référence dans la traque aux contenus générés par des machines. Mais, comme les autres détecteurs, il n'est pas infaillible. De toute façon, à l'instar des microplastiques, l'IA est désormais partout, même à très petite dose.

THOMAS CASAVECCHIA

Le Goncourt, le Médicis et le Femina se seraient-ils « fait avoir » ? La semaine dernière, Thélyson Orélien a été accusé d'avoir eu recours à l'intelligence artificielle pour rédiger son roman, *C'était ça ou mourir*, sélectionné par ces prix prestigieux – déselectionné entre-temps par l'académie Goncourt. L'ouvrage, un des phénomènes de la rentrée littéraire, a déjà reçu le prix du roman Fnac.

Repérer les productions de l'IA n'est, certes, pas chose aisée. Depuis 2022 et le lancement de ChatGPT, impossible de savoir avec certitude si ce qu'on lit a été écrit par un humain ou créé par une machine. Sans même parler de toutes les nuances possibles entre ces deux extrêmes. Il y avait clairement un marché pour la détection. Les sociétés promettant de repérer les écrits générés par des robots ont donc poussé aussi rapidement que des champignons. Entraînées sur des millions de textes humains et des millions de tirades pondues par les différents modèles d'IA du marché, ces intelligences artificielles sont censées repérer les formulations et structures produites par leurs pairs.

Ces OriginalityAI, GPTZero, RoBERTa… représentent une arme de choix pour les écoles, les universités, les maisons d'édition et les sélectionneurs de prix littéraires, qui veulent éviter de se laisser « enfumer » par un rédacteur plus à l'aise avec les prompts qu'avec la plume.

Le succès ne fait pas tout

Parmi ces instruments, celui qui fleurit sur toutes les lèvres depuis quelques mois est Pangram. Il l'est encore davantage depuis la polémique Thélyson Orélien des derniers jours. L'outil, créé fin 2023 par deux alumni de l'Université de Stanford, revendique quelque 120.000 utilisateurs mensuels. Sa popularité a décollé cet été quand Substack, la célèbre plateforme de création et d'envoi de newsletters, l'a intégré à son interface. Le fait que l'outil puisse être testé gratuitement a également contribué à sa notoriété. Tester les emails d'un collègue n'a jamais été aussi simple.

Mais c'est aussi son efficacité qui fait son succès : Pangram est considéré comme le meilleur de ces détectives numériques. Ses créateurs assurent que leur logiciel est fiable entre 99,5 % et 99,9 %, selon les modèles d'IA utilisés pour la rédaction de textes. En plus de leurs propres chiffres issus de tests en interne, les créateurs de Pangram mettent en avant des recherches indépendantes. Les chercheurs de l'Université de Chicago estiment ainsi que les taux de faux positifs (lorsqu'un texte est identifié à tort comme une production IA, NDLR) et faux négatifs (quand l'outil se trompe en estimant que le texte est d'origine humaine, NDLR) de Pangram avoisinent zéro.

Autant dire qu'il frôlerait l'infaillibilité. Pourtant, la solidité de ces IA continue de poser question et il convient de faire preuve de prudence. « Les chiffres avancés par les constructeurs ne peuvent pas être des bases auxquelles on peut se fier », assure Yves Deville, professeur à l'Ecole polytechnique de l'UCLouvain et référent de l'université pour les questions de l'utilisation de l'IA. Selon une enquête de l'Université du Maryland, l'outil atteindrait 2,7 % de faux positifs. Soit assez loin de la promesse d'un taux de certitude de 99,9 %.

Il existe bien des raisons de rester prudent. « La performance d'un tel outil est peut-être très bonne dans nombre de ces études indépendantes. Encore faut-il se poser la question du genre de texte qu'elles utilisent. Si l'on fait analyser des articles de presse à la machine, ce n'est pas tout à fait la même chose que lui demander de déterminer l'origine d'un poème en français. Or, traditionnellement, ces modèles de langage sont plus efficaces en anglais que dans d'autres langues », indique par exemple Yves Deville.

Pour le professeur, de tels détecteurs ne peuvent en aucun cas être utilisés comme des preuves. « Le fonctionnement et la méthodologie de ces systèmes restent inconnus. On ne sait rien de leurs critères, de ce qui leur permet d'affirmer que telle ou telle phrase a été générée ou rédigée. » Contrairement à l'ADN, l'argument n'a rien de tangible ; il faut donc éviter de condamner qui que ce soit sur la seule base de ces outils. « Puisque ces systèmes de détection ne sont jamais à 100 % fiables, à l'université, nous avons fait le choix qu'ils ne pourraient jamais être utilisés comme preuve qu'un étudiant a triché », relate l'enseignant.

Un jeu du chat et de la souris

D'autant que la frontière entre ceux qui ont recours à l'IA et les autres est de plus en plus floue. Comme l'a fait remarquer l'écrivaine Virginie Despentes, sur le plateau de *C ce soir* sur France Télévision, les grandes entreprises du numérique ont forcé l'arrivée d'outils d'intelligence artificielle dans nos quotidiens. Dès que l'on ouvre un document Word, Copilot, l'IA de Microsoft débarque et demande comment il peut aider à remplir la page blanche.

Autre exemple : depuis des mois, une simple recherche Google se transforme en prompt générant une réponse IA qui s'affiche sur la moitié de l'écran. Comment reprocher à qui que ce soit d'utiliser ces services, alors qu'ils nous sont, dans les faits, imposés dans nos vies numériques ? Et que dire des correcteurs orthographiques qui s'appuient de plus en plus sur les larges modèles de langage ? Quelle aide de l'IA est-elle considérée comme acceptable quand on écrit un roman ?

Si le débat n'est pas encore tranché dans l'univers de l'art, il semble qu'il soit désormais bel et bien clos dans le monde académique. Le mois dernier, un groupe de travail du MIT a présenté un rapport sur l'éducation, très relayé, recommandant de ne pas employer les outils de détection d'IA, sous peine de risquer de créer un climat de méfiance entre les enseignants et les apprenants. C'est aussi ce que préconise Yves Deville : « Notre rôle n'est pas de faire la police pour vérifier si l'on utilise l'IA ou non. Notre but, c'est de s'assurer que les étudiants ont atteint les objectifs d'apprentissage visés. Si l'IA fait partie des moyens pour y arriver, tant mieux. Dans certaines étapes de l'apprentissage, l'IA peut être très utile. » Elle peut sans doute l'être tout autant pour rédiger un bon roman.

Puisque ces systèmes de détection ne sont jamais à 100 % fiables, à l'université, nous avons fait le choix qu'ils ne pourraient jamais être utilisés comme preuve qu'un étudiant a triché

Yves Deville
Professeur à l'Ecole polytechnique de l'UCLouvain

Autre raison de prendre ses distances : les faux négatifs restent nombreux. Un petit tour sur YouTube suffit à trouver des tonnes de vidéos qui expliquent comment rendre plus « humain » un document généré par IA, à l'aide d'autres intelligences artificielles « humanisantes ».

Les outils gratuits et payants d'« humaniseurs » de textes pullulent sur internet et leurs performances varient grandement. Evidemment, les développeurs des détecteurs sont bien conscients du problème et entraînent leurs modèles sur la base des écrits générés par ces sites pour les détecter, eux aussi. Mais ces derniers continuent d'entraîner leurs propres modèles à duper les détecteurs. Un jeu du chat et de la souris qui laisse généralement une longueur d'avance aux « humaniseurs ».

Une part d'incertitude

L'*AI Act*, la législation européenne sur l'intelligence artificielle, impose aux développeurs de ces logiciels une plus grande transparence afin de permettre aux citoyens de détecter les productions non-humaines. Pour les photos et les vidéos, cela passe souvent par l'ajout d'un « watermark », une signature qui marque le contenu comme étant générée par IA. Cela arrive aussi pour le texte. La société Anthropic, qui a créé le chatbot Claude, a introduit, cet été, un filigrane invisible, développé par Google pour tous les contenus rédigés par son IA.

Concrètement, le robot conversationnel va placer une série de marqueurs (des successions de mots, de signes de ponctuation, etc.) qui montreront de manière invisible que le texte a été généré. Pour repérer cette identification, il faudra disposer de la clé de décryptage secrète.

Le procédé souffre toutefois de quelques limites. Ainsi, si le texte a été lourdement modifié, il y a de fortes chances pour que cette empreinte numérique invisible disparaisse. A l'inverse, l'IA risque bien d'apposer ce filigrane à un texte écrit par un être humain qui l'a fait corriger par une IA. Il faudra vraisemblablement accepter, à l'avenir, une part d'incertitude.