Le défi de détecter les contenus issus des IA

Alexandre Piquard

Des entreprises et des régulateurs veulent marquer les images et les textes créés par l'intelligence artificielle

Faut-il se résoudre à un monde où il est impossible de discerner les contenus générés par des intelligences artificielles (IA) de ceux produits par des humains ? La question est chaque jour plus brûlante : les textes bluffants prolifèrent depuis le lancement, en novembre 2022, du robot conversationnel ChatGPT, et les photos trompeuses comme celle du pape en doudoune blanche sont appelées à se multiplier avec l'essor de logiciels comme Midjourney.

En réaction, certains cherchent des moyens de rendre détectables ces contenus synthétiques. Le défi est complexe, mais d'actualité : mardi 23 mai, le géant du logiciel Microsoft a annoncé des solutions dans ce sens et le ministre de l'économie français, Bruno Le Maire, a évoqué la question à Paris avec Sam Altman, le PDG d'OpenAI, le créateur de ChatGPT.

« [Rendre détectables les contenus créés avec l'IA] *aiderait à lutter contre la triche à l'université, ou contre la génération massive de propagande et de désinformation dans le but, par exemple, d'inonder des blogs de commentaires favorables à l'invasion de l'Ukraine »,* a argumenté, dans une conférence, en novembre 2022, Scott Aaronson, le chercheur chargé de travailler sur cette question chez OpenAI. *« Le maintien des distinctions est un impératif éthique pour des raisons liées aux usages de l'IA, dans l'éducation, la santé ou le droit, mais aussi, au niveau philosophique, pour délimiter ce qui est du ressort humain et ce qui est fait par les machines »,* ajoute Alexei Grinbaum, membre du Comité national pilote d'éthique du numérique et auteur de *Parole de machines* (Humensciences, 192 p., 17,90 euros).

Outils intégrés ou externes

Dans cet esprit, Microsoft a annoncé l'intégration d'un *« filigrane cryptographique invisible »* (ou *watermark*) dans les images créées par ses logiciels Designer et Bing Image Creator : en consultant les métadonnées d'une photo ou vidéo, c'est-à-dire les informations attachées à ce fichier, *« l'utilisateur pourra voir qu'elle a été créée avec une IA »,* dit le groupe.

Disponible *« dans les prochains mois »,* cette indication de la *« provenance »* d'un contenu repose sur un standard appelé C2PA. Celui-ci a été aussi intégré par Adobe dans l'outil de retouche d'image grâce à l'IA disponible dans son célèbre logiciel Photoshop. OpenAI étudie aussi les techniques de filigrane pour Dall E 2, son logiciel de génération d'image à partir d'une description texte. Et son concurrent Midjourney a adopté un système de métadonnées créé par l'IPTC, un organisme de standardisation de l'industrie des médias.

La tendance se généralise : le 10 mai, Google a assuré travailler à créer *« des outils qui permettent d'identifier les contenus générés de façon synthétique quand vous en rencontrez ».* Ces *« métadonnées et filigranes »* devraient

être inclus dans ses logiciels d'IA d'ici la fin de l'année. Ces informations permettront aussi de signaler, dans son moteur de recherche, les contenus synthétiques utilisant le même standard.

Google précise vouloir prendre ce genre de mesures pour ses modèles de génération de son. Tout comme OpenAI. Et ces efforts s'étendent aussi au texte. Celui-ci pose un défi particulier, car les phrases ne sont pas comme les images cantonnées à un fichier, mais peuvent être copiées et collées. *« Le watermark doit donc être un code caché dans le texte lui-même »,* explique M. Grinbaum.

Une version *« ultrasimpliste »* serait de placer la lettre *« e »* tous les 256 caractères. Mais ce système pourrait facilement être rendu inefficace en changeant quelques mots du texte… Depuis août, plusieurs chercheurs, notamment de l'université du Maryland, aux Etats-Unis, cherchent des méthodes plus robustes : elles aussi inspirées du chiffrement, celles-ci introduisent un *« petit biais mathématique dans l'algorithme qui génère les mots »,* dit M. Grinbaum.

Vers des réglementations

« En gros, nous voulons qu'à chaque fois que nos modèles comme ChatGPT créent un long texte, celui-ci contienne un signal secret dans le choix des mots qui puisse ensuite être utilisé pour prouver que celui-ci a été créé avec un de nos logiciels », a expliqué M. Aaronson. Ce système nécessite de connaître le biais introduit par le fabricant de l'IA. D'autres systèmes – comme GPTZero, DetectGPT ou Turnitin – tentent d'identifier les textes synthétiques de l'extérieur, mais leur taux de réussite est limité : celui proposé par OpenAI *« identifie 26 % des textes générés par IA et classifie à tort 9 % des textes écrits par des humains »,* explique l'entreprise.

M. Grinbaum et tous les spécialistes en conviennent : les systèmes de détection ne sont pas infaillibles. Le filigrane d'un texte restera présent si quelques mots sont changés et des phrases reformulées, mais il ne résistera pas si l'on utilise un autre logiciel d'AI pour *« paraphraser »* le texte, explique ainsi un article universitaire paru en mars. *« Une personne déterminée pourra le contourner »,* a prévenu M. Altman dans un entretien en janvier.

Autre défi de taille : certains logiciels de génération de texte ou d'images, par exemple en accès libre en open source, pourraient ne mettre en place aucun filigrane. Et faire perdurer les contenus indétectables. *« Il est impératif de prendre des décisions réglementaires pour imposer le principe de maintien des distinctions »,* en conclut Alexei Grinbaum. Le CNPEN pourrait évoquer ce point dans l'avis sur l'IA qu'il doit rendre au gouvernement fin juin.

Il faudrait aussi, selon M. Grinbaum, imposer un *« standard »* simple, afin d'éviter d'avoir à utiliser plusieurs logiciels de différents fabricants pour détecter l'origine d'un contenu. La traçabilité des contenus synthétiques est également prônée dans la lettre demandant une *« pause »* des recherches en IA, signée, mardi 28 mars, par des milliers de personnalités, dont le dirigeant de Tesla, Elon Musk. La mise en place de filigranes est aussi défendue par l'ONG DAIR.

Une telle obligation n'est pas intégrée dans l'AI Act, le règlement européen en cours de discussion à Bruxelles. Mais Bruno Le Maire a plaidé mardi pour des *« signalements systématiques »* des images générées par une IA, voire pour des *« bandeaux »* barrant les textes. Patrick Kuban, cofondateur de l'association de comédiens Les Voix, espère aussi que l'utilisateur sera *« prévenu »* s'il écoute une voix synthétique dans un livre audio ou un film doublé.

« *Tout ce qui sera généré par l'IA devra obligatoirement être signalé »*, avait plaidé le commissaire européen Thierry Breton, le 3 avril, sur Franceinfo. Les discussions sur l'AI Act promettent toutefois d'être longues. M. Breton espère les voir aboutir d'ici à la fin de l'année, a-t-il dit lors d'une entrevue, mercredi, avec le PDG de Google, Sundar Pichai.