mediapart.fr

Les détecteurs d'intelligence artificielle, une solution séduisante mais imparfaite

Charlotte Mauger

10–12 minutes

Thélyson Orélien a bien utilisé de l'intelligence artificielle (IA) dans le processus d'écriture de *C'était ça ou mourir*. Le 21 octobre, les éditions du Boréal (l'éditeur québécois du roman) ont dit avoir découvert *« des traces de balises d'IA dans les métadonnées »* de certains documents de travail de l'auteur.

Aucun détail supplémentaire sur ces *« traces »* n'est communiqué. Peut-être s'agit-il de données inscrites dans les propriétés du texte lorsqu'on le télécharge en format Word ou PDF. Auprès de *Libération*, l'auteur reconnaît désormais un recours à l'IA, mais seulement pour des recherches.

Cela signe la fin d'un déchaînement médiatique au cœur duquel se trouvaient le roman et son auteur. Les soupçons sont nés le 21 septembre 2026, à la suite d'une publication sur X où les auteurs du compte « Balance ton Claude » assuraient que l'ouvrage était *« rédigé presque intégralement par une intelligence artificielle »*, après l'avoir fait analyser par l'outil Pangram, un détecteur censé identifier l'IA générative.

© Photo illustration Justine Vernier / Mediapart

Aussi paradoxal que cela puisse paraître, nous avons

désormais besoin de systèmes d'IA pour débusquer d'éventuels contenus générés par… IA. Une brèche dans laquelle se glissent plusieurs entreprises qui commercialisent déjà ce type de solution. Vidéo, photo, musique ou texte : la chasse aux sorcières est désormais ouverte. Pourtant, alertent les scientifiques, ces méthodes sont encore loin d'être totalement fiables.

Les signatures d'une photo ou d'une musique générée

Les modèles d'IA générative sont entraînés à imiter des contenus authentiques. Sauf qu'en réalité, l'intelligence artificielle et l'être humain ne fabriquent pas une œuvre de la même manière.

Par exemple, *« un appareil photo laisse des traces caractéristiques, notamment propres aux capteurs ou à l'interpolation des couleurs* [le calcul de la couleur des pixels – ndlr] *»*, explique Tina Nikoukhah, vice-présidente recherche chez GetReal Security, une entreprise de détection de deepfakes, et autrice de l'ouvrage *Dans les coulisses des images* (Eyrolles, 2026). Des signatures invisibles à nos yeux, mais présentes dans l'image finale.

« Le générateur d'images par IA ne reproduit pas ces traces. Son objectif à lui, c'est seulement de créer une image plaisante à l'œil humain », dit Patrick Bas, directeur de recherche CNRS au Cristal, le Centre de recherche en informatique, signal et automatique de Lille.

En outre, ces outils génératifs d'images produisent leurs propres signatures. *« Générer une image coûte cher. Par conséquent, ces modèles créent souvent une petite image, puis*

l'agrandissent et ajoutent des détails au dernier moment. Mais cet agrandissement laisse des traces », décrit Teddy Furon, chercheur en informatique à l'Institut national de recherche en informatique et en automatique (Inrial) de Rennes.

Des modèles d'IA sont ainsi entraînés à distinguer les images authentiques de celles créées par IA selon la présence de ces signatures. *« On guide le modèle en lui demandant de regarder à ces endroits. Car même si on sait quelles anomalies chercher, ces outils identifient des motifs très compliqués, qui échapperaient peut-être à l'analyse humaine »*, détaille Marina Gardella, postdoctorante à l'École normale supérieure (ENS) Paris-Saclay, spécialiste en analyse d'images.

Un constat similaire peut être fait pour la génération de musiques. *« Chaque modèle laisse ses propres artefacts, des choses qu'on n'entend pas à l'oreille »*, explique Aurélien Hérault, directeur de l'innovation chez Deezer. Ils apparaissent lors d'une analyse du signal : l'IA laisse des pics anormaux à certaines fréquences. Mais *« il faut rester prudent : un pic ne signe pas à lui seul une génération par IA. D'autres traitements en laissent aussi, comme la compression, le rééchantillonnage ou certains filtrages »*, précise Tina Nikoukhah.

Depuis le début 2025, la plateforme a déployé un détecteur de musiques générées qui repère ces anomalies. *« Si toute la piste a été générée, notre détecteur va s'allumer, car cela dépend de la proportion d'utilisation de l'IA »*, détaille Aurélien Hérault.

Désormais, en moyenne, 90 000 musiques générées par IA sont révélées chaque jour. Deezer ne les exclut généralement pas, mais ajoute une pastille pour prévenir l'utilisateur ou l'utilisatrice. *« Nous ne sommes pas contre l'IA mais nous souhaitons de la transparence »*, souligne-t-il.

Le problème des textes

Cette recherche guidée de traces fonctionne moins bien pour les textes. Ici, il n'y a plus de pixels ou de signaux à analyser. Les chercheurs et chercheuses laissent alors le modèle trouver par lui-même les motifs qui font « humains » ou « artificiels » à partir de nombreux exemples. *« Qu'apprend-il exactement ? On ne sait pas. Il n'y a aucune interprétabilité de ses décisions »*, éclaire Teddy Furon. Au mieux peut-on modifier légèrement le texte et voir si cela affecte le résultat. *« Mais c'est limité »*, poursuit-il.

Pour éviter que la décision repose sur des éléments sans lien avec le caractère artificiel, il faut sélectionner les données d'entraînement avec soin : fournir des textes similaires côté IA et côté humain, et limiter les trous dans la raquette (comme des genres littéraires moins courants).

« Les détecteurs d'IA ont fait de véritables progrès, contrairement à ce que beaucoup imaginent encore », insiste Marzena Karpinska, chercheuse à l'université Simon-Fraser, au Canada. Une évolution notamment due à un meilleur choix des textes.

« On ne prend plus seulement des données humaines et générées, mais aussi des données "miroirs" : des textes générés sur le même sujet, dans le même format et de la même longueur qu'un texte humain », explique-t-elle. C'est notamment pourquoi l'outil Pangram (utilisé pour analyser *C'était ça ou mourir*) affiche aujourd'hui de bonnes performances.

Cependant, ce type de détecteur appuie ses réponses sur la forme, délaissant le fond. Pourtant, c'est peut-être cet autre aspect qui nous intéresse…

Une idée originale, ou celle de l'IA ?

Marzena Karpinska et ses collègues ont justement voulu s'attaquer à la question *« Qui a eu cette idée ? »*. Dans un article (pas encore validé par les pairs) d'octobre 2026, ils et elles introduisent le modèle IdeaLens et en proposent une démonstration en ligne.

« Avant l'analyse, tout est paraphrasé par un système d'IA pour que la détection ne soit pas polluée par le style d'écriture », précise Marzena Karpinska. Le modèle s'attarde ainsi sur les personnages, les événements, les informations ou l'organisation : les détails paraissent-ils trop universels ? Les chiffres sont-ils contextualisés ?

« Nous avons analysé les deux premiers chapitres de C'était ça ou mourir*, notre outil aussi a vu de l'IA »*, ajoute-t-elle. Prudence néanmoins : dans sa configuration par défaut, l'outil se trompe 1 fois sur 100. Tous ces détecteurs d'IA peuvent doublement faillir : ne pas reconnaître l'IA (faux négatif) ou, plus embêtant, confondre l'humain avec la machine (faux positif).

S'intéresser à leur fiabilité est donc important. Par exemple, le taux de faux positifs de Pangram est de 0,0026 % pour les textes en français, ce qui veut dire que vingt-six écrits humains sur un million seront détectés comme artificiels. *« Il faut donc utiliser un détecteur comme un test médical : un résultat positif déclenche des examens complémentaires, il ne pose pas le diagnostic à lui seul »*, suggère Tina Nikoukhah.

Il faut aussi recontextualiser ces statistiques, qui correspondent au taux d'erreur d'un modèle sur un corpus de données. *« Cette valeur n'est correcte que sur ce corpus »*, souligne Teddy Furon. Pour qu'un taux de faux positifs soit fiable, il faut que ces données soient variées pour recouvrir la diversité de ce qui est

écrit, composé, photographié. *« Dès que les données diffèrent de ce qui a été vu en entraînement, les performances sont moins bonnes »*, prévient Marina Gardella.

Une frontière de plus en plus floue

Depuis cet été, le Code de bonnes pratiques sur la transparence des contenus générés par l'IA de l'Union européenne (UE) impose aux fournisseurs d'IA générative de « tatouer » leurs contenus. *« C'est-à-dire de les modifier pour y ajouter une information imperceptible et indélébile »*, développe Patrick Bas, qui travaille sur ce sujet depuis les années 2000.

Par exemple, en changeant la valeur de certains pixels d'une image ou en biaisant la probabilité d'apparition de certains mots dans les modèles de langage. Ces détecteurs, beaucoup plus fiables, ne seront pas (pour l'heure) à la disposition du public – pour ne pas rendre trop facile l'ajout ou le retrait du tatouage.

Mais même cette solution ne règle pas tout. Où placer la frontière entre une œuvre humaine et une autre artificielle ? L'assistance de l'IA s'immisce partout et notre processus créatif tend à s'hybrider : l'IA améliore nos photos ou reformule nos écrits. Elle nous influence même.

« Ces systèmes proposent des formulations simples, un langage facile à penser, qu'on finit par adopter », partage Nicolas Spatola, chercheur en sciences cognitives et sociales chez Artimon Perspectives, et qui explore nos relations avec les systèmes d'IA dans son ouvrage *Ce que l'IA fait de nous* (Fyp éditions, 2026). Ainsi, comme l'a constaté une équipe en 2026, certaines expressions caractéristiques de l'IA, comme « approfondir » ou « méticuleux », explosent dans nos usages.

Une œuvre n'est plus seulement humaine ou artificielle, mais

parfois un entre-deux. Quelques modèles explorent déjà une analyse moins binaire, comme GPTZero, qui précise la part de prose humaine et artificielle d'un texte *(voir notre boîte noire)*.

Reste à savoir si le public adopte ces œuvres hybrides. L'affaire Thélyson Orélien montre à quel point nous aimons savoir qui est vraiment l'auteur ou l'autrice derrière le roman pour pouvoir en apprécier le contenu. *« La valeur qu'on associe à une œuvre diminue si nous savons qu'elle a été produite par un système d'IA »*, estime Nicolas Spatola.

Alors, et presque ironiquement, on se tourne vers une machine faillible pour « connaître la vérité » sur l'origine d'une œuvre, quitte à croire ses statistiques plutôt que les promesses d'un humain.