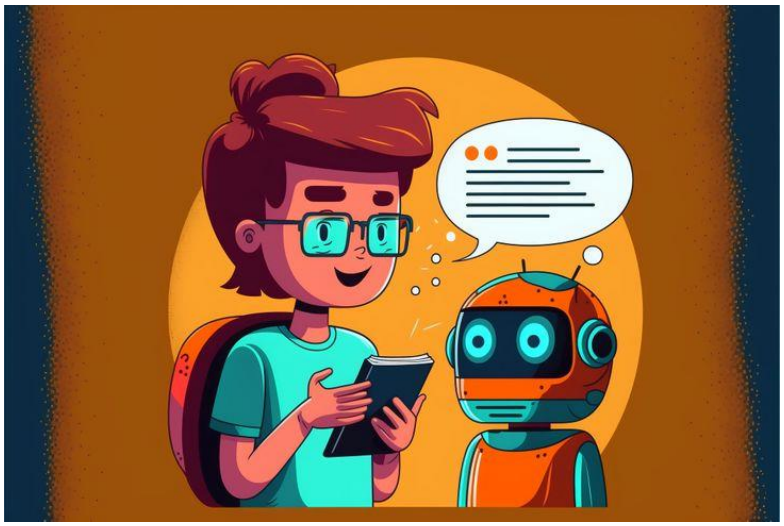
ChatGPT et triche : identifier l’usage de l’IA en mêlant « détection automatique et observation humaine »

Christian Goglin, Enseignant-Chercheur en Intelligence Artificielle à l’ICD Business School

- Christian Goglin,
- le 08/02/2023 à 10:46
- Lecture en 3 min.



Le lancement de ChatGPT, basé sur l’intelligence artificielle, a engendré des cas de triche sur le plan scolaire (photo d’illustration).

ChatGPT est le dernier-né des robots conversationnels, ou chatbot, développé par la société Open AI, fondée notamment par le milliardaire fantasque Elon Musk. Ce chatbot défraie la chronique depuis sa mise à disposition gratuite auprès du grand public au début du mois de décembre dernier, non sans raison, au regard de ses capacités de génération de contenus bluffantes. Le monde de l’enseignement semble pris de court par ce phénomène inédit dont le principal risque se traduit par la généralisation d’une fraude que l’on peut qualifier d’automatique. Dès lors se pose la question de la juste réaction face à ce phénomène.

Un robot performant mais peu créatif

Tout d’abord, précisons les capacités de ChatGPT. Ce robot, dopé à l’intelligence artificielle, peut générer des contenus textuels réalistes dans tous les domaines et dans toutes les langues répertoriées sur l’Internet. La connaissance de ChatGPT se base sur une masse immense de textes, dont tout Wikipédia. Même si sa sophistication est sans commune mesure, le principe de génération de texte est analogue à la complétion de l’éditeur de textos de votre smartphone.
À La Croix, ce sont plus de 100 journalistes qui travaillent à fournir une information de qualité précise et vérifiée.

Concrètement, ce robot combine des mots pour construire de nouvelles phrases vraisemblables par extrapolation. L'approche est statistique, mais la machine n'a aucune compréhension du sens des mots et phrases manipulées. On peut aussi dire que ChatGPT fait « du neuf avec du vieux » ; en conséquence, le texte généré ne peut pas être véritablement innovant, car le robot est comme prisonnier de ses données d'entraînement, ce qui bride ses possibilités créatives.
Pour en revenir à la « fraude automatique », comment considérer le cas de l'élève utilisant telles quelles les réponses fournies par ChatGPT en vue de réaliser un devoir scolaire ? À l'évidence, l'élève s'approprie le texte produit par un autre, en l'espèce un robot. Ce type d'utilisation revient à contourner l'exigence de production d'un travail personnel et original. Ainsi, de la même façon que le plagiat d'œuvres de l'esprit produites par des humains est sanctionné, insérer à l'identique des fragments de texte produits par un robot dans un devoir scolaire devrait, en toute logique, être également puni par l'institution.

La difficile détection de la « fraude automatique »

D'abord, ChatGPT commet peu de fautes d'orthographe, ce qui constitue bien souvent, malheureusement, le signal d'une fraude. Aussi, la phase d'entraînement de ChatGPT étant antérieure à 2022, il ne produira pas d'information récente. Il faut aussi noter que le texte généré est caractéristique, le style est réaliste mais demeure artificiel, il ne « sonne pas » comme celui produit par un humain. Ce style artificiel peut être identifié avec plus ou moins de succès par d'autres IA, capables de déterminer la probabilité qu'un texte ait été généré artificiellement.

Allant plus loin, Open AI prépare une solution plus subtile présentée comme un antidote, sous la forme d'un filigrane, soit une sorte de code caché dans le texte. Il s'agira plus précisément d'une contrainte déterministe appliquée à la génération stochastique (aléatoire) du texte. À titre d'illustration, cela pourrait consister à insérer un mot de 5 lettres tous les 30 mots ou toute autre contrainte imaginable plus sophistiquée, basée sur la cryptographie et non détectable par un être humain.

Cependant, non seulement ces méthodes de détection ou de signature peuvent être, en partie, contournées par d'autres systèmes automatiques de paraphrase mais, en outre, une base de texte générée automatiquement, puis remaniée par un humain, avec son style propre, et son orthographe approximative, sera pratiquement indétectable, au moins sur la forme. Il faudra alors compter avec l'observation humaine ; ainsi, le professeur constatant plusieurs devoirs assez semblables sur le fond pourra légitimement s'interroger.

Du bon usage d'un robot

En somme, la meilleure façon d'identifier un texte produit par ChatGPT consiste à combiner outils de détection automatique et observation humaine. Si apporter la preuve d'une fraude automatique n'est pas simple, une piste de solution existe. Requérant imagination et originalité, elle consiste à demander aux élèves de produire des contenus sur la base de travaux spécifiques étudiés en cours. En effet, dans ce cas, ChatGPT ne pourra produire que des généralités, ne répondant pas précisément à l'exercice.

Il faut tout de même noter qu'une autre utilisation de ChatGPT, positive cette fois, est possible. Elle consiste à considérer ce robot comme une source de documentation parmi d'autres, apportant l'information nécessaire à une réflexion plus personnelle, de la même façon que consulter Wikipédia ou toute autre source

est légitime et même nécessaire. Dans cette optique, il faut toutefois garder à l'esprit que la validité des informations restituées par ChatGPT n'est pas garantie puisque ses réponses sont construites sur la base de ressources d'Internet dont la qualité est très fluctuante.

En somme, puisqu'il est impossible d'empêcher les étudiants d'utiliser ChatGPT à leur domicile, il semble raisonnable d'autoriser son usage, en le bornant à celui d'une simple source d'information. Les institutions du monde de l'enseignement devraient donc s'équiper d'outils de détection de textes générés automatiquement et fixer un seuil limite, de façon analogue au seuil défini pour la lutte anti-plagiat, en ayant à l'esprit que l'enseignant, dûment formé, devra également faire preuve d'originalité et de vigilance.