OFFICE FRANÇAIS
DE L'INTÉGRITÉ SCIENTIFIQUE

ENTRETIEN

« IA : LES FILIGRANES AU SERVICE DE L'INTÉGRITÉ SCIENTIFIQUE »

Février 2024

Contexte

Pour encadrer le développement et les usages des outils d'intelligence artificielle (IA) générative, des réglementations se mettent en place partout dans le monde : l'AI Act en Europe, un décret aux Etats-Unis, un autre en Chine. Dans ce cadre général en pleine évolution, l'utilisation accrue de ces outils par les chercheuses et chercheurs force également à revoir les codes de conduite qui leur sont spécifiques. Alexei Grinbaum, spécialiste des questions d'éthique du numérique –il participe au projet européen iRECS sur le sujet- revient sur des enjeux d'intégrité scientifique, notamment liés aux mécanismes d'identification des contenus générés par de l'IA.

ALEXEI GRINBAUM

Directeur de recherche et président du Comité opérationnel pilote d'éthique du numérique du CEA

Depuis l'arrivée en force de ChatGPT en novembre 2022, le monde de la recherche, qui n'échappe pas à une utilisation accrue et massive d'outils d'IA générative, voit ses différents acteurs se doter de nouvelles règles – maisons d'édition (voir infolettre N°5), établissements de recherche, agences de financement (eg NSF), etc. En juin dernier, la dernière révision du code de conduite européen pour l'intégrité scientifique a également intégré cette nouvelle dimension : que retenir de ces lignes directrices ?

Alexei Grinbaum : Le message essentiel de l'ensemble de ces codes de conduite est le principe de transparence : en soi, utiliser des outils d'intelligence artificielle générative quand on est chercheur ou chercheuse n'est pas un problème, mais il faut le déclarer et expliciter l'usage qui en est fait.
Les questions portent en effet plutôt sur la façon d'utiliser ces outils et sur l'objectif de leur utilisation. Certains usages posent problème, d'autres pas, d'où la nécessité d'être transparent. Prenons un exemple : Alice écrit un article de manière tout à fait classique mais le résumé ne la satisfait pas. Elle demande alors à ChatGPT de réécrire ce résumé de manière plus percutante, à partir de cette première version. Puis elle le relit, le modifie au besoin, pour assumer la pleine responsabilité du résultat : si elle déclare l'avoir écrit avec l'assistance d'IA, ce n'est pas problématique. En revanche, si elle copie directement un texte de ChatGPT sans le relire et qu'en plus elle ne le déclare pas, c'est inadmissible.
Dans le cadre du projet européen Techethos, j'ai participé aux discussions pour la révision du code de conduite européen (lire l'encadré). Outre le principe de transparence, l'idée d'assistance est cruciale : l'IA est une aide pour produire des contenus scientifiques mais le résultat final nécessite toujours un contrôle humain par la personne qui en endosse la responsabilité. La prochaine version du code doit aller encore plus loin mais, pour l'heure, et c'est important, il indique déjà que ne pas déclarer un recours à un système d'IA constitue un manquement à l'intégrité scientifique.

Quelles évolutions envisagez-vous ?

A. G. : La limite de ces déclarations est qu'aujourd'hui, elles ne sont pas vérifiables. Aucun logiciel n'est en mesure de détecter de manière fiable l'usage de l'IA générative. Mais les avancées sur les filigranes, inscrits en amont dans le contenu généré pour permettre de l'identifier, vont changer les choses. Le travail sur les filigranes avance très vite : les recherches[1] ont démarré en automne 2022 ; un an plus tard, les pays membres du G7 signaient un accord dans le cadre du processus d'Hiroshima qui introduit l'obligation d'insérer les filigranes. Ce même principe a aussi été inscrit dans l'article 52 de la loi européenne AI Act, qui devrait être votée au parlement européen en avril. Aux Etats-Unis, le président Biden a aussi signé le 30 octobre un décret dans le même sens.

En 2025, l'inscription des filigranes sera donc devenue obligatoire ?

A. G. : Oui, et tous les développeurs des modèles d'IA génératives auront à l'appliquer. Grâce à des logiciels qui liront ces filigranes, tout utilisateur pourra authentifier les contenus générés par la machine. La détection ne sera jamais garantie à 100 %, mais des niveaux de confiance assez élevés pourront être atteints. Il reste que, techniquement, créer des filigranes dans des textes est encore très compliqué, bien plus que dans les images, les vidéos ou les audios, pour lesquels différentes solutions existent. L'une des difficultés principales est aussi de fabriquer des filigranes à la fois suffisamment robustes et interopérables. C'est à dire qui fonctionnent pour tous les contenus, qu'ils soient fabriqués par les modèles de Google, Facebook, OpenAI, Mistral ou autre. Aujourd'hui, la solution miracle n'existe pas, mais puisque les filigranes relèvent désormais d'une obligation légale, les développeurs vont y travailler. La nécessité de standards interopérables mondiaux pose d'autres questions : qui va en décider? Qui va contrôler que les centaines voire les milliers de modèles d'IA générative adoptent bien ces standards ?

La Commission européenne a annoncé ce 24 janvier la création du bureau européen de l'intelligence artificielle à Bruxelles, effective à la fin février, ce bureau jouera-t-il ce rôle ?

A. G. : Ce nouveau bureau se met tout juste en place. Je pense qu'il jouera effectivement un rôle dans la création de ces standards et dans le contrôle de leur bonne application. Il faut être conscient qu'un filigrane incassable, qui résiste à toute attaque, n'existe pas. Se développeront certainement des modèles qui vont insérer des filigranes suffisamment élaborés et d'autres qui tenteront de les détecter pour les supprimer. Des sanctions seront à prendre contre les seconds. Cela fera normalement partie des missions de ce nouvel office.

Le code européen s'adresse aussi aux évaluateurs des articles et des projets de recherche : y voyez-vous un enjeu spécifique ?

A. G. : Oui, car l'examen par les pairs, socle de l'évaluation des résultats de recherche, rencontre aujourd'hui un énorme problème avec la croissance massive du nombre de publications. Trouver des scientifiques qui acceptent de réviser les articles sérieusement, devient extrêmement compliqué. Force est de constater que face à ces difficultés, les maisons d'édition sont moins regardantes. Or, l'utilisation de ChatGPT pour évaluer un article, sans l'avoir lu in extenso, est une pratique qui se répand. La question de l'avenir de ce modèle d'évaluation n'est pas nouvelle, mais il faut désormais la poser à l'aune de ces nouveaux outils. Dans l'hypothèse d'un système dans la continuité de l'existant, on pourrait imaginer que les éditeurs s'équipent d'un modèle génératif qui, avant d'envoyer un article pour revue, en ferait un résumé, extrairait les sections les plus importantes, les arguments essentiels, les innovations, pour cibler le travail du *reviewer* et économiser son temps, en ayant identi-

[1] A. Grinbaum et L. Adomaitis, "The Ethical Need for Watermarks in Machine-Generated Language", 2022 arXiv - CS - Computation and Language https://doi.org/10.48550/arXiv.2209.03118

fié les contenus générés par de l'IA et vérifié qu'ils ont été déclarés comme tels. Le tout avec toujours un contrôle humain. Ce n'est qu'une piste, mais il est certain que le système va devoir évoluer.

Une autre question émerge dans plusieurs domaines : quand l'IA est créatrice de solutions inédites, à qui revient la découverte ?

A. G. : C'est une question très intéressante, à laquelle on ne peut répondre en quelques phrases, mais il faut la soulever. Il est clair que ces modèles génératifs fournissent aujourd'hui des résultats jamais envisagés auparavant. C'est le cas pour les travaux sur les mécanismes de repliement des protéines ou encore pour le déchiffrage d'inscriptions antiques, comme les rouleaux d'Herculanum par exemple. Pour autant, la découverte revient à mon sens à celui qui a formulé le bon prompt -ou requête- et à ceux qui ont fabriqué le logiciel, un collectif d'auteurs regroupant les personnes ayant élaboré le corpus d'entraînement, fabriqué le modèle, etc... Il faut bien garder en tête que si un modèle d'IA propose de multiples solutions, impossibles à imaginer sans elle, il est incapable de discerner celle qui a le plus de sens. Il faut des spécialistes pour cela, de biologie moléculaire pour les repliements de protéines ou des papyrologues pour les rouleaux. C'est une nouvelle façon de faire de la science mais elle exige toujours une vérification par l'être humain. Celui-ci doit s'habituer à le faire car un système génératif n'évalue ni le vrai, ni le faux et ne sait pas qu'il se trompe. Lui déléguer une responsabilité qu'il n'est pas en mesure d'endosser serait une faute.

Un autre débat porte sur la transparence même des modèles d'IA génératives, en particulier sur l'ouverture des bases de données sur lesquelles ils sont entrainés. Qu'en pensez-vous ?

A. G. : L'ouverture dans ce domaine est effectivement un autre grand enjeu. La notion même de modèle d'IA générative « ouvert » fait débat. Est-ce que cela signifie ouvrir le corpus d'entrainement ? Les paramètres d'apprentissage ? Les filtres du fine-tuning ? Pour certains, il faut tout ouvrir. Pour d'autres, comme OpenAI ou Google, il est dangereux d'ouvrir complètement ces grands modèles, en raison de l'usage dual qui peut en être fait[2]. Est-ce une raison suffisante pour ne pas l'ouvrir ? Bien sûr, ces outils permettent de fabriquer des publications frauduleuses, mais ils peuvent aussi être utilisés pour les détecter. Je ne partage pas la posture de fermeture, mais je reconnais qu'il s'agit d'un vrai dilemme. On n'ouvre pas par exemple la façon de fabriquer des virus très dangereux. Si l'on suit la ligne « aussi ouvert que possible, aussi fermé que nécessaire », cela pose une autre question : qui décide du nécessaire ?
Par ailleurs, il faut savoir qu'il existe un frein à l'ouverture d'une autre nature : aujourd'hui ces modèles ont besoin de corpus géants qui ne sont pas uniquement composés de données de qualité. Toute la Bibliothèque nationale de France ne représente qu'une infime fraction de la quantité nécessaire. Dans ces conditions, le développeur qui ouvre son corpus d'apprentissage expose sa réputation : il est quasiment certain que quelqu'un pourra y trouver des données répréhensibles qui feraient la une des journaux. La discipline qui s'occupe des données d'apprentissage s'appelle la *data curation*, c'est-à-dire le nettoyage des données pour éliminer toutes sortes de données toxiques C'est un énorme enjeu.

Propos recueillis par Hélène Le Meur

Lire l'encadré >>

[2] A. Grinbaum et L. Adomaitis, "Dual Use Concerns of Generative AI and Large Language Models", Journal of Responsible Innovation 11:1, 2304381, 2024 https://doi.org/10.48550/arXiv.2305.07882

CC BY

« IA : LES FILIGRANES AU SERVICE DE L'INTÉGRITÉ SCIENTIFIQUE »

Février 2024

L'IA fait son entrée dans le code de conduite européen

L'exigence de transparence quant à l'utilisation d'outils d'intelligence artificielle a été introduite à trois endroits dans la nouvelle version de juin 2023. Selon ce code, dissimuler un tel usage entre désormais dans la catégorie des mauvaises pratiques.

> 2.3 Procédures de recherche
> Les chercheuses et chercheurs rendent compte de leurs résultats et de leurs méthodes, y compris l'utilisation de services externes ou d'outils d'intelligence artificielle automatisés, d'une manière qui soit compatible avec les normes acceptées dans la discipline et qui facilite la vérification ou la réplication, le cas échéant.
>
> 2. Révision et évaluation
> Les chercheuses et chercheurs, les institutions et les organismes de recherche examinent et évaluent les demandes de publication, de financement, de nomination, de promotion ou de récompense de manière transparente et justifiable, et divulguent l'utilisation de l'IA et d'outils automatisés.
>
> 3.1 Manquement à l'intégrité scientifique et autres pratiques inacceptables
> Cacher l'utilisation de l'IA ou d'outils automatisés dans la création de contenu ou la rédaction de publications.

Code de conduite européen pour l'intégrité en recherche, 2023.