Note de synthèse:
Usages des outils d'Intelligence Artificielle générative dans le domaine de la recherche - Points de vigilance en matière de déontologie et d'intégrité scientifique

Céline Blitz-Frayret (UMR Eco&Sols)
Estelle Jaligot (Délégation à la déontologie et à l'intégrité scientifique)
Colline Orsini (Délégation aux affaires juridiques et à la conformité)

La présente note se veut une tentative de résumer les principales questions soulevées par les nouveaux outils d'Intelligence Artificielle (IA) générative et leurs applications dans les activités de recherche scientifique, dans le but d'attirer l'attention sur les biais et risques potentiels associés avec leur usage. Dans la mesure où i) ces outils sont relativement nouveaux et ii) ils présentent des spécificités techniques et modes de fonctionnement différents, ce document ne peut prétendre aborder toutes leurs applications possibles dans le cadre de la recherche ni la totalité des biais et risques associés. Néanmoins, nous tentons ici de donner ici quelques repères afin de permettre aux utilisateurs potentiels d'employer les IA génératives de manière responsable et éclairée.
Les ressources sur lesquelles s'appuie cette synthèse sont récapitulées en fin de document. Certaines d'entre elles vous fourniront également des informations plus générales sur les questions éthiques soulevées dans le cadre de la conception, du développement, du déploiement et de l'utilisation des outils basés sur l'IA, ainsi que des pistes pour leur gestion dans le cadre des projets de recherche.

Ce document est amené à évoluer à mesure que de nouveaux usages des IA génératives sont testés et leurs limites analysées. Dans cet esprit, toute contribution permettant de l'enrichir est bienvenue.

De quoi s'agit-il?

Récemment, de nouveaux outils d'IA reposant sur de grands modèles de language (large language models, LLMs) ont été mis à disposition du grand public à grand renfort de publicité: ChatGPT (proposé par OpenAI) est le plus connu d'entre eux, mais il existe aussi Bard (Google), Llama (Meta), ainsi que de nombreux autres.
Ces différents outils ont en commun d'être des agents conversationnels (chatbots), réagissant à une requête (prompt) émanant de l'utilisateur pour produire une réponse crédible, sinon correcte. Ils fonctionnent sur la base d'un algorithme dont les performances ont été progressivement affinées et validées sur des jeux de données d'entraînement (dites aussi d'apprentissage). La réponse fournie se présente comme une synthèse d'informations préexistantes, fournies par l'utilisateur en complément du prompt ou puisées par l'IA dans les ressources auxquelles elle a accès.
Dans la présente note, nous avons choisi de mettre en lumière les problèmes posés par l'utilisation des IA génératives dédiées au texte, du fait de leur potentiel d'application très large dans un contexte de recherche. Il existe également des outils d'IA générative dédiés à la génération d'images (par exemple Dall-E, Midjourney, Firefly) ou de sons (AudioCraft, AIVA, SOUNDRAW, etc.), qui partagent une partie de ces problèmes.

Problèmes connus

Biais de représentation

Fonctionnant sur la reconnaissance des patrons (patterns), les IA génératives tendent à reproduire préférentiellement les enchaînements de mots et de phrases en fonction de leur plus forte représentation au sein de leurs données d'entraînement et/ou au sein des données fournies par l'utilisateur. Elles auront donc tendance à fournir des réponses basées sur des éléments (faits, opinions) majoritaires au sein de ces données, et à éluder les éléments minoritaires. Ce mode de fonctionnement peut les conduire à produire des réponses stéréotypées et qui omettent des nuances importantes, voire à générer des contresens.

Ce biais peut s'avérer d'autant plus critique que le développement de l'algorithme ainsi que les jeux de données d'entraînement peuvent eux-même être biaisés. L'IA fournira alors des réponses qui reproduiront, voire amplifieront ces biais. Une décision prise sur la base de réponses fournies par une IA peut donc conduire à des situations inéquitables, voire des discriminations.

Enfin, la formulation du prompt initial est susceptible de faire varier sensiblement la réponse: dans le cadre d'un usage en recherche, cela soulève des problèmes de répétabilité et de reproductibilité des résultats.

Production ou propagation de réponses erronées ou tronquées

Les IA génératives produisent des réponses sur la base de corrélations statistiques (établies en fonction de la fréquence d'association entre des chaînes de caractères), qui peut n'avoir aucun rapport avec leur sens, ni avec leur validité scientifique. L'outil n'a en effet pas la capacité d'effectuer une analyse critique des informations traitées ou produites, et ne sait donc pas les interpréter, ni les hiérarchiser, ou leur attribuer un niveau de confiance qui soit totalement fiable. De plus, les jeux de données utilisés pour l'entraînement de l'IA peuvent aussi être incomplets, de mauvaise qualité ou erronés, ce qui augmente la probabilité de produire des réponses inadéquates.

Dans le contexte scientifique, les IA génératives peuvent notamment produire des textes dits "hallucinés", sans rapport avec des informations issues du monde réel (par exemple des références bibliographiques) en réarrangeant voire en reformulant des éléments préexistants sans liens entre eux. La capacité des IA génératives à produire des combinaisons d'assertions vraies et fausses sur un sujet donné est d'ailleurs reconnue par leurs concepteurs[1]. Ainsi, l'IA produira presque toujours une réponse, si inexacte ou inadaptée qu'elle soit par rapport au prompt de départ: rares sont en effet les IA qui mentionnent leur incapacité à répondre.

Si elles peuvent accélérer l'écriture de lignes de code informatique, les IA génératives sont susceptibles de produire des erreurs conduisant à des pertes de données voire à des failles de sécurité.

Les IA génératives devant "apprendre" à partir de données qui leur sont accessibles avant de pouvoir fournir une réponse, elles sont donc "aveugles" aux informations les plus récentes: ainsi, la version gratuite actuelle de ChatGPT (version 3.5) ne peut accéder qu'aux données antérieures à janvier 2022. Un l'état de l'art produit par une IA générative nécessitera donc, de la part de l'utilisateur, une actualisation.

Problèmes associés à l'utilisation des sources par l'IA

[1] Ainsi la mention systématique figurant en bas de la fenêtre lors de l'utilisation de ChatGPT : « ChatGPT [date] Version. ChatGPT may produce inaccurate information about people, places, or facts ».

Les IA génératives ne créent aucun matériel intrinsèquement nouveau, les réponses produites sont constituées du réarrangement d'éléments préexistants. Dans la mesure où le fonctionnement des IA génératives peut reposer en partie sur du matériel protégé par des droits d'auteur (qu'il s'agisse des données d'apprentissage ou des données fournies par l'utilisateur), les réponses produites constituent une violation de ces droits. Dans le cas de la production d'écrits ou d'illustrations scientifiques, l'utilisation d'une IA générative peut ainsi être assimilable à un plagiat voire à de la contrefaçon.
Par ailleurs, de par le fait que l'IA ne mentionne pas l'appartenance des œuvres d'origine, ses productions ne respectent pas les principes du droit d'auteur (droit moral, droits patrimoniaux). Par extension, l'utilisateur commet une faute en les utilisant.

Problèmes associés à la réutilisation par l'IA des données fournies par l'utilisateur
Les données fournies par l'utilisateur à une IA générative abondent ensuite au fonctionnement de celle-ci, et sont donc susceptibles d'être réutilisées par n'importe quel autre utilisateur dans les futures réponses données par l'outil. L'utilisation d'une IA peut donc entraîner une fuite de données, dès lors que les données fournies ne sont pas d'accès public: c'est par exemple le cas de documents internes à l'établissement.
Dans un contexte de recherche, l'utilisation d'une IA générative sur des documents confidentiels (évaluation de projets, de publications, etc) constitue de ce fait une violation du devoir de confidentialité. Si cette obligation de confidentialité a été incluse dans un contrat, cette divulgation va à l'encontre des engagements contractuels, ce qui peut engager la responsabilité de l'utilisateur ou celle de l'établissement.
Selon la même logique, l'utilisation d'une IA générative sur des documents contenant des données personnelles (listes de contacts, données d'enquête, etc) contrevient à l'obligation d'assurer la protection des données personnelles et/ou de la vie privée, conformément au réglement général pour la protection des données (RGPD)[2].

Manque de transparence pour l'utilisateur
L'élaboration des outils d'IA générative les plus populaire par des compagnies privées s'assortit d'une grande opacité quant à leurs modalités de fonctionnement et aux données d'apprentissage utilisées. L'utilisateur n'a donc en général pas les moyens de se prémunir contre les problèmes mentionnés ci-dessus et d'exercer un réel contrôle qualité. Dans le contexte d'un usage en recherche, ce manque de transparence prive l'utilisateur de la possibilité de garantir pleinement l'intégrité des données produite et la rigueur de la démarche. Cette obligation faisant partie des responsabilités de toute personne acceptant le rôle d'auteur d'une publication scientifique, certains journaux scientifiques conditionnent strictement l'usage des IA génératives à la possibilité de déclarer en détail les modalités de leur fonctionnement et l'étendue de leur utilisation[3]. Selon les principes de traçabilité, l'accès à la chaîne de traitement d'un résultat produit par l'IA générative (incluant les codes sources de l'IA générative, les données, etc .) devrait être permis afin de pouvoir reproduire ce résultat en cas d'audit ou de litige. Cependant, le modèle d'IA ne produit aucune référence aux sources, rendant difficile la vérification. Au-delà de la vérification de la véracité des réponses produites, cette opacité de la conception et du

[2] Règlement (UE) 2016/679 modifié du Parlement européen et du Conseil du 27 avril 2016 : https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:32016R0679.
[3] Voir à ce sujet le focus de l'Office français de l'intégrité scientifique en mai 2023 ainsi que les "ChatGPT, Generative Artificial Intelligence and Natural Large Language Models for Accountable Reporting and Use Guidelines" (CANGARU), initiative multi-acteurs visant à proposer des lignes directrices homogénéisées.

fonctionnement de l'outil empêche l'utilisateur de satisfaire aux principes FAIR de bonne gestion des données[4].

En résumé: bonnes pratiques dans l'utilisation des IA génératives en recherche

Les IA génératives constituent des outils puissants pour synthétiser et reformuler rapidement des quantités importantes d'information, pour un coût faible ou nul pour l'utilisateur. De ce point de vue, elles peuvent s'avérer utiles pour accélérer certaines tâches. Néanmoins, le Comité National Pilote d'Ethique du Numérique (CNPEN) indique que les systèmes d'IA générative mis sur le marché peuvent être à haut risque. En cas de système d'IA générative en libre accès sous licence non-commerciale, une obligation de transparence et de test par les concepteurs est requise.

Au niveau de l'utilisateur, il convient de:

- Toujours déclarer l'utilisation d'un outil d'IA (générative ou non), et d'en détailler les modalités de manière transparente. L'utilisation non déclarée constitue en effet un manquement à l'intégrité scientifique[5].
- Ne jamais utiliser une IA générative pour produire tout ou partie d'un matériel présenté comme un travail original engageant la responsabilité de l'auteur ou, a fortiori, celle de l'établissement, que ce soit dans une publication, un rapport d'expertise, un policy brief, un projet soumis pour financement, une évaluation, etc. (en-dehors d'usages limités tels que décrits ci-dessus).
- Ne jamais les alimenter une IA générative à l'aide de données qui ne sont pas destinées à être rendues publiques, et notamment des données personnelles ou confidentielles ou des données protégées par un droit de propriété intellectuelle.
- Dans le cadre d'une publication scientifique, prendre connaissance de la politique du journal concernant les IA génératives et la respecter. Il est fortement recommandé de limiter strictement cette utilisation aux produits sur lesquels l'utilisateur possède les connaissances et compétences nécessaires pour effectuer un contrôle de la qualité et de la validité de la réponse. Il s'agit par exemple de la production d'un résumé (sur la base de documents auxquels il a accès), de la reformulation de la traduction (entre deux langues qu'il maîtrise raisonnablement bien) ou de la validation d'un programme informatique (dans un langage et pour des fonctions avec lesquels il est familier).
- Traiter les réponses fournies par un outil d'IA comme un élément d'aide à la décision parmi d'autres, et les confronter systématiquement à des éléments obtenus sans l'appui d'outils d'IA. Dans tous les cas, l'ensemble des résultats doit être soumis à une ou plusieurs analyses critiques humaines avant toute décision.
- En cas d'utilisation d'une IA générative, opter pour un logiciel dont le code source est ouvert et indiquer la version utilisée si possible. De même, expliciter et rendre accessibles les stratégies d'interaction avec l'IA générative ainsi que les connaissances utilisées, tout en respectant le cadre juridique et éthique de ces dernières.
- Dans un cadre de gestion d'un collectif, demeurer vigilant face à la tentation de recourir à l'utilisation des IA génératives comme un substitut à des compétences ou à des collaborations absentes ou insuffisantes, et aux risques de dérives qui en découlent.

[4] Findable, Accessible, Interoperable, Reusable (soit Trouvable, Accessible, Interopérable et Réutilisable en français): voir Ouvrir la Science.

[5] Parmi les comportements listés dans la section **Research Misconduct and other unacceptable practices:** "Hiding the use of AI or automated tools in the creation of content or drafting of publications" (European Code of Conduct for reseach integrity, version actualisée en 2023).

Pour en savoir plus:

- *LINC - Laboratoire d'innovation numérique de la CNIL:*

Dossier "IA générative"

- *Comité national pilote d'éthique du numérique (CNPEN):*

Avis n°3: Agents conversationnels: enjeux d'éthique, 15 septembre 2021.
Avis n°7: Systèmes d'intelligence artificielle générative : enjeux d'éthique, 30 juin 2023.

- *Commission Européenne (CE):*

DG Communications Networks, Content and Technology. Ethics guidelines for trustworthy AI, 2019.
DG Research & Innovation. Ethics By Design and Ethics of Use Approaches for Artificial Intelligence, version 1.0 du 25 novembre 2021.

- *Office français de l'intégrité scientifique (OFIS):*

Systèmes d'intelligence artificielle générative : quelques points de vigilance, février 2024.

- *UK Research integrity office (UKRIO):*

AI in research (mis à jour en janvier 2024).