RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

Usages en recherche de logiciels (dits d'« intelligence artificielle ») de production de texte, code ou images : questions d'intégrité scientifique

Association RESINT – 27 septembre 2024

Catherine Tessier
referent-integrite-ethique@onera.fr

À la une…

Article de franceinfo du 8 août 2024

au sujet d'une publication scientifique australienne

lien

L'article "What happens to our bodies after death", *Cosmos,* July 26, 2024

produit par « IA »

lien

...

L'article "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery",
Ch. Lu, C. Lu, R. Tjarko, J. Foerster, D. Ha. *arXiv,* 15 August 2024

lien

L'article "Artificial intelligence and illusions of understanding in scientific research",
L. Messeri, M.J. Crockett.
Nature 627, 49-58, 6 March 2024

lien

L'article "AI is complicating plagiarism. How should scientists respond?",
D. Kwon.
Nature News Features, 30 July 2024

lien

L'article "Can AI Replace Human Research Participants? These Scientists See Risks",
C. Stokel-Walker.
SciAm, 22 March 2024

lien

« Intelligence artificielle » ?

L'intelligence artificielle

- est une discipline scientifique
- qui recouvre un ensemble de techniques différentes, voir ici par exemple

AfIA
Association française pour l'Intelligence Artificielle

Terme utilisé par abus de langage pour parler :

- de techniques d'apprentissage machine, quelles qu'elles soient
- de ce qui nous concerne ici :

> des agents conversationnels, donc des logiciels,
> fondés sur plusieurs techniques d'apprentissage machine
> qui produisent du texte, du code informatique ou des images,
> en réponse à des requêtes d'un utilisateur

Comment fonctionnent de tels logiciels ?

En *très bref* et *très schématique*, pour le texte (ou le code informatique) :

- L'utilisateur, par ses requêtes* (*prompts*), programme l'outil
- L'outil produit de nouvelles données, par exemple la séquence de mots la plus probable après le *prompt*, à partir de caractéristiques communes « apprises » sur un corpus de données de très grande taille [CNPEN23]

→ corrélations statistiques + filtrage de résultats + annotations humaines

Les grands modèles de langue (LLM - *Large Language Model*) sont entraînés sur un corpus de textes

- généraliste (internet)
- ou spécifique (exemple : *ScopusAI* – Elsevier lien)

* La *Liste relative au vocabulaire de l'intelligence artificielle (termes, expressions et définitions adoptés),* JORF n° 0212 du 6 septembre 2024, préfère le terme d' « instruction générative » - lien

Par nature (1/2)

- Requêtes de l'utilisateur

S'il s'agit d'un outil ouvert (en ligne, ex : ChatGPT)

→ divulgation des informations contenues dans les requêtes

  - sécurité
  - vie privée
  - propriété intellectuelle

- Résultats fournis

→ issus de corpus contenant des éléments originaux

→ respect de la propriété intellectuelle des tiers ?

Par nature (2/2)

C’est un logiciel

→ ne « comprend » pas

→ illusion de « dialogue »

Fonctionnement statistique

→ possibilité de résultats erronés (manque de robustesse)

→ pas de notion de « vérité »

→ reproductibilité ?

Biais induits chez l'utilisateur

- qualité perçue du résultat (bien écrit, bien construit, sans fautes…)

→ peut induire la sensation que ce résultat est scientifiquement valide

- biais d'automatisation (si l'outil le dit, c'est que c'est vrai)

et biais de confirmation (renforcement de ses propres croyances)

→ susceptibles d'être accentués

- risque d'une illusion de compréhension

Usages en recherche

Recherche bibliographique et production automatisée d'état de l'art

Élaboration de la recherche

Élaboration d'articles en vue de publications

Examen par les pairs

… bientôt tout le processus de recherche ??

Recherche bibliographique et production automatisée d'état de l'art [Bouchard24]

- Quel corpus d'entraînement du logiciel ?

risques : comprend revues ou conférences de qualité médiocre, articles corrigés ultérieurement ou rétractés, documents non évalués par les pairs (archives ouvertes, blogs personnels), opinions…

- Représentativité des résultats

fraîcheur, pertinence, complétude (ouvrages ? documents payants ?), existence

– ne pas prendre connaissance des références (citer les références sans les avoir lues ni même consultées) ↑

– plagiat d'un état de l'art ou d'un *survey* existant, plagiat de citations

\+ privilégier des outils de littérature académique, en utiliser plusieurs, croiser les résultats

\+ vérifier les références suggérées et les lire !

Élaboration de la recherche

- Où vont les données des requêtes (*prompts*), comment sont-elles utilisées par les sociétés qui proposent les logiciels ?

+ Vigilance relativement aux données
  - sensibles (données personnelles, données de recherche)
  - confidentielles (hypothèses, données brevetables, données pas encore publiées)
  - protégées (propriété intellectuelle)

– production (très) facilitée de données créées de toutes pièces (fabrication), de données « arrangées » (falsification)

– risque de pression temporelle accrue sur la production de résultats scientifiques, « parce que l'IA fait gagner du temps »

Élaboration d'articles en vue de publications

- Un logiciel n'est pas « co-auteur » d'un document scientifique (consensus)

- Les auteurs d'un document scientifique qui utilisent un tel logiciel :

+ doivent le signaler dans le document : nom de l'outil, version, (requêtes)
+ expliquer les usages

rédaction, production d'images, de graphiques, usages pour le processus de recherche…

+ s'assurer de la correction de ce qui est produit
+ vérifier qu'il n'y a pas atteinte à la propriété intellectuelle d'autrui
  - difficile ?

– plagiat délibéré : facilité et de plus en plus difficile à détecter

plusieurs passes dans le logiciel pour reformuler, adapter le style, changer la langue…

– inflation de faux documents scientifiques

Examen par les pairs

Par définition (sauf si publication de la version soumise)

l'examen par les pairs porte sur des documents non publiés, (articles, thèses, propositions de projets, etc.), donc à ne pas diffuser

– divulgation de connaissances originales par téléchargement dans un logiciel externe, ouvert
– multiplication d'évaluations automatisées (qualité ?)

\+ si usage (mise en forme du texte d'évaluation), le signaler

Références (1/4)

L'Ofis FAIT LE POINT

OFFICE FRANÇAIS DE L'INTÉGRITÉ SCIENTIFIQUE

Février 2024

Principes fondamentaux

Contrôle humain et responsabilité. Il est important de rappeler que les chercheuses et chercheurs ayant recours à des systèmes d'IA générative sont responsables des contenus générés qu'ils reproduisent dans leurs articles, commentaires ou autres productions de recherche - entre autres de leur fiabilité et de leur adéquation avec la règlementation en vigueur. Toute utilisation de ces outils exige le contrôle du résultat final par la personne responsable.

Transparence. Depuis juin 2023, le code de conduite européen pour l'intégrité scientifique[1] recommande la transparence : cacher l'utilisation d'IA ou d'outils automatisés dans la création de contenu ou dans la rédaction de publications y est désormais considéré comme un manquement à l'intégrité scientifique.

lien

Points de vigilance pour l'intégrité scientifique

L'Ofis attire aussi l'attention de ceux et celles qui utilisent ces outils sur quelques points de vigilance. Etant données la vitesse des avancées en IA et l'évolution en cours de la règlementation, cette liste est appelée à évoluer.

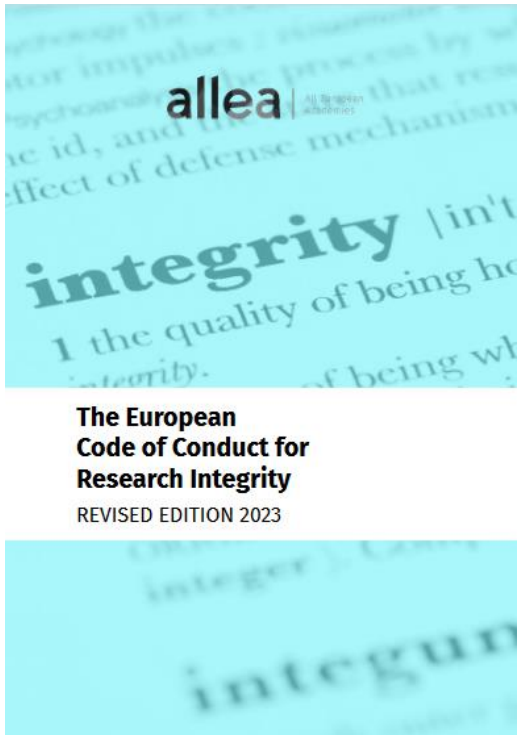
Fiabilité Les systèmes d'IA génératives font parfois des erreurs, et présentent de manière très vraisemblable des informations erronées voire complètement inventées (couramment appeler des "hallucinations").[2] Cela a par exemple été mis en évidence dans la génération d'états de l'art et de références bibliographiques,[3] ou de réponses à des questions scientifiques.[4] Cela expose les chercheuses et chercheurs à un risque de diffuser de fausses informations, voire à de la fabrication et de la falsification.

Propriété Il est possible que le contenu généré par les systèmes d'IA générative soit issu de données d'entrainement protégées par un *copyright*.[5] Il existe dont un risque de plagiat par l'utilisateur qui s'approprie ce contenu. Par ailleurs, un système d'IA générative ne peut pas être reconnu comme auteur d'un article ou d'autres productions de recherche (lire « ChatGPT, auteur de publication scientifique ? »). Les détails de l'utilisation des LLMs doivent être déclarés dans les méthodes et/ou les remerciements.

Confidentialité Les systèmes actuels d'IA générative n'offrent pas une protection suffisante en matière de données personnelles – qu'il s'agisse de la protection des données d'entrainement ou de celles qui proviennent des requêtes des utilisateurs.[2,5] Ces systèmes présentent un risque de violation de plusieurs dispositions du RGPD, telles que la confidentialité, le consentement ou le droit à l'oubli.[2,5] Les chercheuses et chercheurs doivent donc éviter de partager des données personnelles ou confidentielles dans leurs requêtes.

[1] ALLEA, *The European Code of Conduct for Research Integrity*, 2023
[2] CNPEN, Avis n°7, "Systèmes d'intelligence artificielle générative : enjeux d'éthique", 2023
[3] W. H. Walters et E. I. Wilder, « Fabrication and errors in the bibliographic citations generated by ChatGPT », *Sci Rep*, vol. 13, n° 1, Art. n° 1, sept. 2023, doi: 10.1038/s41598-023-41032-5.
[4] H. Zheng et H. Zhan, « ChatGPT in Scientific Writing: A Cautionary Tale », *The American Journal of Medicine*, mars 2023, doi: 10.1016/j.amjmed.2023.02.011.
[5] Novelli, Claudio *et al.*, « Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity », 2024. arXiv. https://doi.org/10.48550/arXiv.2401.07348.

Références (2/4)



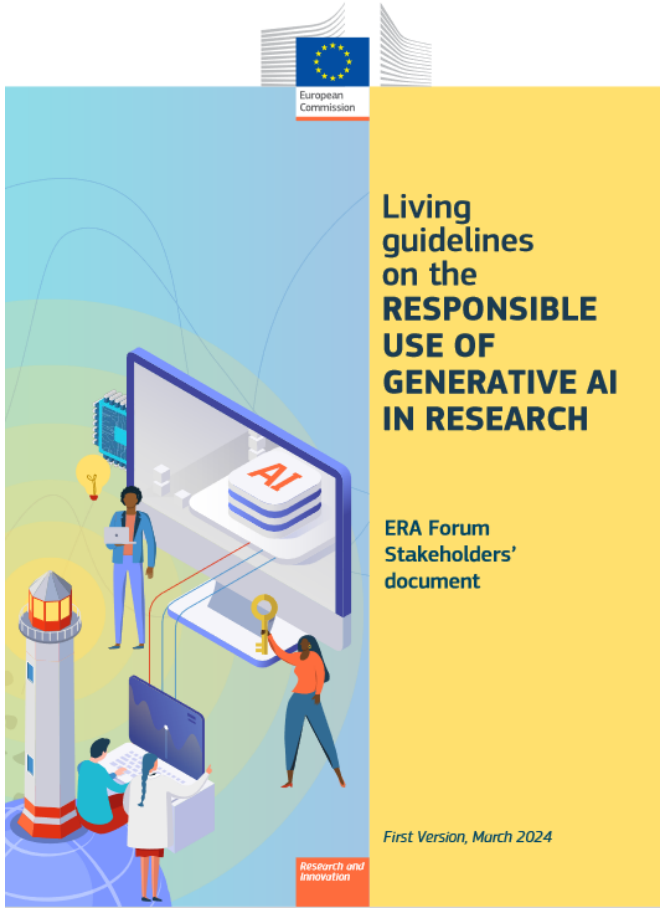
lien

- Researchers report their results and methods, including the use of external services or AI and automated tools, in a way that is compatible with the accepted norms of the discipline and facilitates verification or replication, where applicable.

- Researchers, research institutions, and organisations review and assess submissions for publication, funding, appointment, promotion, or reward in a transparent and justifiable manner, and disclose the use of AI and automated tools.

- Hiding the use of AI or automated tools in the creation of content or drafting of publications.

Références (3/4)



lien

Références (4/4)

[Bouchard 24] Aline Bouchard. *Au-delà de ChatGPT – Recherche d'informations académiques et intelligence artificielle*. URFIST Paris, mai 2024. lien

[CNPEN23] Comité national pilote d'éthique du numérique. Avis 7 – *Systèmes d'intelligence artificielle générative : enjeux d'éthique*, juin 2023. lien