Analyse des risques de mésinformation liés à l'utilisation des IAG pour la recherche d'information scientifique

Analysis of Misinformation Risks Related to the Use of AI Tools for Scientific Information Retrieval

Robert VISEUR

UMONS

robert.viseur@umons.ac.be

Résumé

Cette étude explore les risques pour la rigueur scientifique inhérents à l'utilisation de moteurs de recherche génératifs par les chercheurs pour accéder à des informations scientifiques. Une revue de la littérature met en évidence des politiques fréquentes de blocage des robots collectant des données pour l'entraînement des intelligences artificielles génératives (IAG). En s'appuyant sur l'analyse des fichiers *robots.txt* (protocole d'exclusion des robots), la recherche examine la constitution des jeux de données scientifiques et les stratégies de blocage mises en œuvre par les éditeurs de revues. Les résultats révèlent un lien entre le niveau d'autorité d'une revue et l'intensité des restrictions imposées, particulièrement chez les éditeurs commerciaux dominants. Cette étude met en évidence un biais de validation résultant de la possible surreprésentation des revues de moindre qualité dans les jeux de données d'entraînement, compromettant potentiellement la fiabilité des réponses générées par les *chatbots*. Une analyse approfondie des limitations des moteurs de recherche génératifs (régurgitations, hallucinations, biais) est menée, accompagnée d'une discussion sur les

implications dans quatre scénarios d'utilisation typiques. Enfin, un ensemble de recommandations pratiques est proposé pour les fournisseurs et les utilisateurs de ces technologies émergentes, afin de promouvoir une utilisation responsable de ces systèmes.

Mots-clefs : intelligence artificielle générative, agent conversationnel, recherche d'information, botshit, recherche scientifique.

Abstract

This study explores the risks to scientific rigour inherent in the use of generative search engines by researchers seeking scientific information. A review of the literature highlights frequent policies blocking bots from collecting data used to train generative artificial intelligences (GAI). Drawing on the analysis of *robots.txt* files (robot exclusion protocol), the research examines the composition of scientific datasets and the blocking strategies implemented by journal publishers. The findings reveal a correlation between the level of a journal's authority and the intensity of imposed restrictions, particularly among dominant commercial publishers. This study highlights a validation bias resulting from the possible overrepresentation of lower quality journals in training datasets, potentially undermining the reliability of chatbot-generated responses. A thorough analysis of the limitations of generative search engines (regurgitations, hallucinations, biases) is conducted, along with a discussion of the implications across four typical usage scenarios. Finally, a set of practical recommendations is proposed for providers and users of these emerging technologies, aimed at fostering responsible use of such systems.

Keywords: generative artificial intelligence, chatbot, information retrieval, botshit, scientific research.

1. Introduction

Introduit auprès d'un large public en novembre 2022, ChatGPT a connu une croissance rapide de son nombre d'utilisateurs, dépassant en quelques mois les cent millions d'utilisateurs actifs mensuellement (Hu, 2023). ChatGPT est une intelligence artificielle générative (IAG). Il prend la forme d'un agent conversationnel (*chatbot*) généraliste basé sur un grand modèle de langage (LLM, *Large Langage Model*) (Hannigan et al., 2024). Ce dernier, nommé GPT (*Generative Pre-trained Transformer*), est capable d'interpréter des directives formulées en langage naturel, appelées « *prompts* », et de générer des réponses textuelles cohérentes en adéquation avec ces directives (Floridi & Chiriatti, 2020 ; Hutson, 2024). Le moteur de recherche Google se trouve dès lors concurrencé par des moteurs de réponse, comme ChatGPT, et par des moteurs de recherche assistés par IAG, comme Perplexity. L'insatisfaction suscitée par les moteurs de recherche classiques ainsi que l'interactivité apportée par les intelligences artificielles génératives encouragent la migration vers ces nouveaux outils (Zhou & Li, 2024). L'attrait des IAG dans le cadre d'activités de recherche scientifique s'est ainsi rapidement développé parmi les étudiants mais aussi parmi les chercheurs. Les fonctionnalités jugées d'intérêt concernent la recherche d'informations mais aussi la synthèse d'articles et l'aide à la rédaction, voire la co-rédaction (Hutson, 2024). Des limitations de l'outil ont cependant rapidement été identifiées sur le plan de la fiabilité des réponses incluant les faits rapportés et les références suggérées (de Corbière et al., 2023). Cette irruption dans le contexte académique, et les activités de la recherche scientifique en particulier, intervient alors que l'on constate un niveau de littératie numérique inégal parmi les étudiants du supérieur incluant les étudiants en bachelier (licence dans le système français), en master et en doctorat (Soung & Dumouchel, 2019).

L'intelligence artificielle (IA) contribue au phénomène de mésinformation (Bontridder & Poullet, 2021). La mésinformation « *désigne une information fausse, inexacte ou trompeuse partagée sans*

intention de tromper », contrairement à la désinformation, qui « *est une information fausse, inexacte ou trompeuse qui est diffusée dans l'intention de tromper le destinataire* » (Bontridder & Poullet, 2021 ; p. e32-2). Au-delà des opportunités de création délibérée de fausses informations et de l'assistance à la diffusion de celles-ci vers des audiences ciblées, l'intelligence artificielle, et en particulier les IA génératives (IAG) accessibles au grand public depuis 2023 (ChatGPT, Gemini, Claude, Copilot, Le Chat...), peut aussi contribuer à la mésinformation dès lors que les réponses aux *prompts* de l'utilisateur contiennent elles-mêmes des erreurs. Ce risque est identifié et a été qualifié d'« *hallucination* » (Ye et al., 2023). L'utilisation sans recul critique de ces contenus erronés conduit à un risque épistémique que Hannigan, McCarthy et Spicer (2024) ont baptisé « *botshit* » (par analogie au « *bullshit* »).

Nous analysons donc dans cette recherche les risques de mésinformation en matière d'information scientifique par les IA génératives, et plus précisément ce que la littérature nomme, selon les auteurs, « *moteur de recherche génératifs* » ou « *systèmes de recherche conversationnels* » (Sharma et al., 2024 ; Zhou & Li, 2024). Les producteurs (p. ex. OpenAI) d'IA génératives (p. ex. ChatGPT) ont besoin de contenus de qualité, en quantité, pour entraîner leurs grands modèles de langage (p. ex. GPT) (Floridi & Chiriatti, 2020). Pour ce faire, ils utilisent notamment les contenus de qualité publiés par la presse en ligne et par les éditeurs scientifiques (Dodge et al., 2021). Cependant, les sources de données utilisées ne sont pas toujours constituées d'articles validés par les pairs (p. ex. *preprints*). Par ailleurs, comme Viseur et Delcoucq (2024) le montrent dans le secteur de la presse en ligne, les éditeurs tendent à bloquer les robots chargés de collecter des données d'entraînement. De plus, les politiques de blocage ne sont pas nécessairement homogènes. Dans le cas de l'édition scientifique, cela pourrait entraîner non seulement des biais mais aussi des problèmes importants de qualité dans l'information scientifique incluse dans les données d'entraînement des *chatbots*. C'est donc ce point que nous explorons dans ce papier : les robots d'exploration utilisés par les producteurs d'intelligences artificielles génératives disposent-ils d'un

accès homogène à des sources de qualité ou les producteurs doivent-ils se contenter d'un entraînement sur des recherches de moindre qualité voire frauduleuses ? La réponse à cette question nous servira de base pour, dans un second temps, nous appuyant sur les travaux de Hannigan et al. (2024), analyser globalement le risque de mésinformation scientifique en matière d'information scientifique par les IA génératives.

Cet article est organisé en trois parties. La première comporte un état de l'art relatif aux *datasets* dédiés aux contenus scientifiques permettant d'entraîner les modèles d'intelligence artificielle générative. La seconde décrit la méthodologie d'analyse ainsi que les cinq hypothèses testées puis les résultats pour chaque hypothèse. La troisième, précédant la conclusion, discute les implications des résultats sur le plan de la fiabilité des réponses aux *prompts*, des risques associés aux différents usages et aux stratégies de mitigation envisageables.

2. Revue de la littérature

Les producteurs d'intelligences artificielles génératives collectent de vastes ensembles de données en parcourant le Web (Viseur & Delcoucq, 2024). Cette politique, généralement non négociée, considérée comme une forme de prédation par certains éditeurs de contenus, conduit à des politiques de blocage, passif ou actif, par les propriétaires des sites présentant des contenus originaux (Viseur & Delcoucq, 2024 ; Dinzinger & Granitzer, 2024). En particulier, Viseur et Delcoucq (2024) ont analysé le comportement des éditeurs de presse face aux producteurs d'IA génératives. Ils montrent que les blocages des robots d'exploration alimentant les jeux de données (*datasets*) sont fréquents, basés sur le protocole d'exclusion des robots, et que cela occasionne de nombreux biais, notamment linguistiques, culturels et idéologiques (Ferrara, 2023). Le même type de dispositif de protection de la propriété intellectuelle est-il mis en place par les éditeurs scientifiques ? Quatre robots sont d'un usage courant : GPTbot (utilisé pour collecter des données

d'entraînement), ChatGPT-User (utilisé pour les actions réalisées par l'utilisateur dans ChatGPT[1]), Google-Extended (utilisé par Google) et CCbot (associé au Common Crawl, un jeu de données disponible en ligne). Fin 2024, OpenAI a rajouté OAI-SearchBot, associé à son moteur de recherche. Des listes plus complètes existent[2]. Cependant, elles se répercutent actuellement peu dans les fichiers robots analysés (Viseur & Delcoucq, 2024).

Les robots d'exploration des intelligences artificielles génératives voient donc l'accès aux contenus scientifiques conditionné à l'absence d'interdiction exprimée au travers du protocole d'exclusion des robots. Or, l'édition scientifique est devenue un marché lucratif caractérisé par des marges élevées (Larivière et al., 2015). Les articles sont fréquemment publiés derrière des *paywalls*. Aussi les grands éditeurs (Elsevier, Springer Nature, Wiley Blackwell, Taylor & Francis...) tendent à défendre la propriété des contenus qu'ils publient (Chawla, 2017). Leur position sur le marché les autorise par ailleurs à régulièrement augmenter leurs tarifs. Cette situation a suscité plusieurs réactions. D'une part, les articles derrière *paywall* se retrouvent publiés sur des plateformes alternatives. Par exemple, Sci-Hub est une base de données gratuite, riche de plusieurs dizaines de millions d'articles, souvent toujours couverts par droit d'auteur, dès lors considérée comme illégale par les éditeurs scientifiques (Banks, 2016). Compte tenu de son caractère peu ou prou légal, il ne s'agit pas d'un jeu de données exploitables par les producteurs d'IA génératives. D'autre part, le monde académique a encouragé la création de nouveaux journaux publiés en *open access* (Gershenson et al. 2020). La publication des résultats de recherche dans de tels journaux s'est d'ailleurs trouvée encouragée par certains organismes de financement (p. ex. Plan S). Le développement des journaux en *open access* (OA) s'est malheureusement accompagné de la prolifération de revues pratiquant un marketing agressif et offrant des taux d'acceptation élevé (Richtig et al., 2018). Ces revues acceptent des articles sans processus rigoureux de révision par les pairs, dans un but de profit (Xia et al., 2015 ; Richtig et al., 2018). Le phénomène a notamment été

1 Voir https://platform.openai.com/docs/bots.

2 Voir par exemple https://github.com/ai-robots-txt/ai.robots.txt.

étudié par Jeffrey Beall. Ce dernier a désigné ces journaux comme « *prédateurs* » et maintenu une liste pour sensibiliser la communauté académique aux pratiques de publication malhonnêtes (Beall, 2010). Les producteurs d'IAG voient donc l'accès facilité à ces revues en *open access*, sans cependant que la qualité des publications soit garantie.

La littérature existante donne quelques éclairages sur les *datasets* proposant de l'information scientifique (Brown et al., 2020 ; Gao et al., 2020 ; Dodge et al., 2021) : Common Crawl, Colossal Clear Crawled Corpus C4 et The Pile. Le Common Crawl est un jeu de données constitué par une exploration à large échelle du Web. Il est notamment utilisé par OpenAI (Brown et al., 2020). Il est aussi utilisé comme *dataset* de base pour la constitution de *datasets* de meilleure qualité après l'application de règles de filtrage. C'est notamment le cas du Colossal Clear Crawled Corpus C4 (Dodge et al., 2021). Ce dernier s'appuie substantiellement sur les éditeurs de presse (New York Times, LA Times, Washington Post…) et les éditeurs scientifiques (PLOS One, Frontiers…), en plus de Google Patent pour l'accès aux connaissances scientifiques. The Pile est un jeu de données de haute qualité incluant 22 sous-*datasets* (Gao et al., 2020). Parmi les jeux de données, deux sont de nature scientifique : ArXiv (8,96 % du poids total) et PubMed Central (14,40 % du total). Le premier est un serveur de *preprints*, le second, un répertoire de documents issus de la recherche médicale. Aucun des deux ne propose donc une information soumise à un processus strict de *peer reviewing*. Quant au traitement des sites des revues scientifiques proprement dites, il est peu ou prou documenté. Enfin, les *datasets* intègrent classiquement des données issues de Wikipédia (Dodge et al., 2021). Or, il apparaît que Wikipédia est un bon relais pour l'information scientifique publiée, d'une part, dans des journaux en *open access*, d'autre part, dans des journaux à facteur d'impact élevé, éventuellement protégés par *paywall* (Teplitskiy et al., 2017).

Tous les contenus scientifiques n'ont en effet pas la même valeur. Premièrement, un contenu scientifique peut avoir fait ou non l'objet d'une révision par les pairs. Un contenu publié dans une

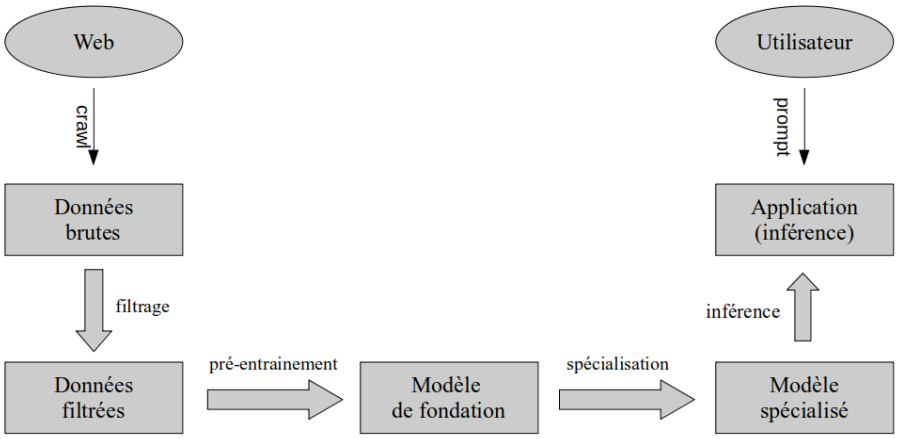
conférence ou une revue à comité de lecture bénéficiera donc d'un niveau de validation supérieur à un article en *preprint*, même si la qualité de ces derniers peut se révèler globalement correcte. Deuxièmement, à l'intérieur même des conférences ou des journaux scientifiques, une hiérarchie existe, que les articles soient ou non publiés en *open access*. Par exemple, l'indicateur SCImago Journal Rank (SJR) offre un classement des revues scientifiques contenues dans la base de données Scopus. Il est basé sur une mesure, inspirée du Pagerank de Google, qui tient compte à la fois du nombre de citations reçues par une revue et du prestige des revues d'où proviennent les citations. En France, la FNEGE « *publie tous les 3 ans le classement des revues scientifiques en sciences de gestion* »[3]. Pour les producteurs d'IAG, il ressort, d'une part, que les contenus scientifiques les plus accessibles ne sont pas nécessairement les meilleurs, d'autre part, qu'il existe un risque que les comportements protecteurs des éditeurs soient d'autant plus forts qu'une revue scientifique est réputée pour sa haute qualité. Troisièmement, l'injonction à publier a conduit à la publication d'articles frauduleux faisant aujourd'hui l'objet de détection à large échelle puis, avec cependant un délai, d'une rétractation (Cabanac, 2024).

Les intelligences artificielles génératives (IAG) souffrent de plusieurs limitations susceptibles d'affecter la qualité des recherches académiques. Premièrement, elles souffrent du phénomène de « *mémorisation* » (de Wynter et al., 2023), aussi appelé régurgitation. Les réponses incluent alors des séquences de mots, que l'on peut directement retrouver dans les données d'entraînement, occasionnant un problème de plagiat. Les régurgitations sont notamment favorisées par l'existence de contenus dupliqués dans les données d'entraînement (de Wynter et al., 2023). Deuxièmement, elles peuvent générer des informations erronées. Ce phénomène est désigné par le terme « *hallucination* » (Hannigan et al., 2024 ; Ye et al., 2023). Ces dernières peuvent être atténuées en fournissant un texte en entrée mais ne peuvent jamais être totalement supprimées (Ye et al., 2023). Troisièmement, les réponses peuvent souffrir de différents biais induits par les jeux de données d'entraînement (Chu et al., 2024 ; Maleki et al., 2024 ; Ferrara, 2023). Ferrara (2023) identifie sept

3 Voir https://fnege.org/classement-des-revues-scientifiques-en-sciences-de-gestion/.

types de biais affectant les réponses de ChatGPT : les biais démographiques, les biais culturels, les biais linguistiques, les biais temporels, les biais de confirmation ainsi que les biais idéologiques et politiques. La qualité des données influence ces défectuosités (Chu et al., 2024). Navigli et Conia (2023) introduisent ainsi, en complément des biais sociaux (sexisme, âgisme, racisme...), le « *biais de sélection de données* », qu'ils définissent comme « *le biais causé par le choix des textes qui composent un corpus d'entraînement* ». Ce biais se produit lorsque « *les textes sont identifiés, ou lorsque les données sont filtrées et nettoyées* ». Au-delà des données primaires, ces biais peuvent également être causés, d'une part, par des jeux de données utilisés à des fins de *finetuning* (spécialisation) et, d'autre part, par le texte utilisé au sein des *prompts* (Maleki et al., 2024 ; Chu et al., 2024). D'autres facteurs conduisent à des biais : les algorithmes d'apprentissage logiciel et le design des applications (Ferrara, 2023). Navigli et Cornia (2023) identifient d'ailleurs le « *biais statistique* », à savoir « *la tendance d'un modèle statistique à sur- ou sous-estimer certaines informations* », et le biais inductif, à savoir « *l'ensemble des hypothèses faites par le créateur du modèle de machine learning* ».

Figure 1. Chaîne de production d'une intelligence artificielle générative



C'est donc la totalité de la chaîne de production d'une intelligence artificielle générative (IAG), et en particulier tout en amont les jeux de données brutes, qui influence la qualité des réponses aux *prompts* (cf. Figure 1). Les grands modèles de langage, entraînés sur de vastes ensembles de

données, sont capables de produire « *un charabia technique basé sur des motifs de mots dans les données d'entraînement, qui sont elles-mêmes une boîte noire* » (Hannigan et al., 2024). En résulte le phénomène de « *botshit* » par analogie au « *bullshit* », soit le risque épistémique que représentent les réponses erronées ou fabriquées par les *chatbots*, mais dont l'apparence est cohérente et techniquement plausible, reprises sans recul critique par les utilisateurs. En contexte académique, cette problématique est particulièrement critique. Notre principale contribution portera donc sur l'analyse des limitations en matière d'information scientifique induites par les restrictions supposées d'accès aux articles publiés en ligne.

Notre revue de littérature nous permet de formuler les hypothèses suivantes, qui vont être testées dans la suite de l'article : (H1) les robots des IA génératives sont davantage bloqués que les robots des moteurs de recherche ; (H2) le robot GPTbot est davantage bloqué que les robots d'autres IA génératives ; (H3) les éditeurs scientifiques commerciaux dominants bloquent davantage les robots d'IA génératives que les autres éditeurs ; (H4) les revues prédatrices bloquent moins les robots d'IA génératives que les revues non prédatrices ; et (H5) mieux une revue scientifique est classée et plus elle bloque les robots d'IA génératives.

3. Méthodologie et résultats

Deux jeux de données sont utilisés. Le premier est constitué de la liste de revues prédatrices publiée par Beall, et disponibles sur le site Beall's List. Le second est constitué des revues évaluées dans le Norwegian Register for Scientific Journals, Series and Publishers. Les revues y sont classées sur trois niveaux (level 0, level 1, level 2). Le niveau 1 intègre des revues scientifiques respectant les critères de qualité académique élémentaires tandis que le niveau 2 rassemble les meilleurs canaux de publication. Le niveau 0 contient des revues non scientifiques, dédiées par exemple à la vulgarisation, mais aussi parfois des revues prédatrices, comme « Progress in Physics », également

incluse dans la liste de Beall. L'appartenance éventuelle au Directory of Open Access Journals (DOAJ) y est indiquée. Parmi les classements publiés, celui de 2024 a été utilisé. Parmi ces listes, certains sites sont cependant injoignables, inaccessibles ou associés à des redirections (avec un envoi, correct ou non, du code HTTP correspondant). De plus, de nombreux doublons existent car plusieurs revues peuvent être publiées sur le même site (cas des grands éditeurs par exemple). Aussi un filtrage des URL (suppression des sites injoignables, calcul des redirections…) est réalisé à l'aide d'un script codé en Python. Au final ont été considérés 1153 sites de revues prédatrices, 4129 de niveau 0, 7276 de niveau 1 et 542 de niveau 2 (ces valeurs peuvent être légèrement inférieures lors de l'analyse proprement dites, en cas d'inaccessibilité du site par exemple).

Les bases de données obtenues en sortie sont ensuite utilisées pour collecter automatiquement les fichiers *robots.txt*. Ces fichiers sont sauvés localement (mise en cache) puis explorés. Cette analyse nous permet d'obtenir la liste des robots d'exploration cités, dont est décliné un Top 10 des robots les plus fréquemment cités, d'identifier les pratiques de blocage (trois cas : pas de fichier robots.txt, liste blanche ou liste noire) et de déterminer la politique de blocage adoptée pour chaque robot. Sur cette base, il est possible de calculer le taux de blocage ainsi que le biais global par robot. Le biais global, introduit initialement par Sun et ses co-auteurs (2007), est une valeur comprise entre -1 et 1 qui « *représente le pourcentage, en valeur absolue, de sites (de l'échantillon) qui favorisent (signe positif) ou défavorisent (signe négatif) le robot* » (Viseur & Delcoucq, 2024). Il est ainsi possible d'estimer la discrimination des robots d'IAG comparativement au robot universel (cas général) puis aux robots des moteurs de recherche, qui sont globalement bien acceptés par les éditeurs de sites web. Le calcul du biais global utilise la procédure documentée par Viseur et Delcoucq (2024). Les résultats sont enregistrés dans un fichier journal. Ces fichiers peuvent ensuite être ingérés par ChatGPT pour l'assistance à la production de tableaux de synthèse spécifiques.

H1 : Les robots des IA génératives sont davantage bloqués que les robots des moteurs de recherche.

Les robots les plus fréquemment cités par les revues non prédatrices sont, juste après le robot universel (« * »), GPTbot, CCbot, Google-Extended, Googlebot et ChatGPT-User. Les trois premiers robots sont les robots d'exploration utilisés pour la création et la mise à jour des jeux de données.

Tableau 1. Taux de blocage des robots d'exploration

Robot	Citations	Blocages	Taux de blocage (si robot cité)	Taux de blocage (tous les sites)
googlebot	689	9	1,31 %	0,08 %
bingbot	367	26	7,08 %	0,23 %
ccbot	763	666	87,29 %	5,96 %
gptbot	921	834	90,55 %	7,46 %
chatgpt-user	679	603	88,21 %	5,39 %
google-extended	720	657	91,25 %	5,88 %

Le blocage des robots d'exploration des deux moteurs de recherche dominants (Google et Bing) par les revues non prédatrices apparaît sensiblement moindre que celui des robots d'exploration des producteurs d'IA génératives (cf. Tableau 1). Même ChatGPT-User, le robot associé aux actions réalisées au sein de ChatGPT (p. ex. résumé d'un texte) fait l'objet d'un blocage fréquent. Surtout, dès lors que le robot est cité, c'est dans l'immense majorité des cas pour être finalement bloqué.

H2 : Le robot GPTbot est davantage bloqué que les robots d'autres IA génératives.

Le robot GPTbot fait l'objet d'un blocage par les revues non prédatrices dans 90,55 % (cf. Tableau 1) des cas où il est mentionné dans le fichier *robots.txt* (81,68 % des sites configurent un tel fichier). Au final, 7,46 % des sites web interdisent l'accès aux pages de contenu. Cette valeur est légèrement

plus élevée (9,44 %) pour les sites des revues estampillées DOAJ. Les autres robots d'exploration des IA génératives font l'objet d'un taux de blocage légèrement inférieur même si l'ordre de grandeur est équivalent. Le biais global pour GPTbot, soit -0,0743, est le plus élevé, et traduit une discrimination du robot comparativement à d'autres robots poursuivant ou non les mêmes objectifs de collecte.

H3 : Les éditeurs scientifiques commerciaux dominants bloquent davantage les robots d'IA génératives que les autres éditeurs.

Les éditeurs internationaux comme Sciencedirect ou Springer ont un taux de blocage très sensiblement plus élevé que les revues prédatrices ou même que la moyenne des revues non prédatrices. Ce blocage accru conduit à un biais global (cf. Tableau 2) sensiblement plus élevé, en particulier pour le robot d'exploration GPTbot.

Tableau 2. Biais global (revues prédatrices vs Top 50)

Robot	Revues prédatrices	Top 50
googlebot	-0,0141	0,0323
bingbot	-0,0247	0,0323
ccbot	0	-0,1290
gptbot	-0,0018	-0,4194
chatgpt-user	0	-0,1613
google-extended	0	-0,2581

Les politiques de blocage des robots peuvent prendre des allures parfois radicales à l'image de Sciencedirect qui renvoie une erreur HTTP 403 (« *forbidden* ») lors de la lecture avec un script du fichier *robots.txt*. En pratique, le fichier existe (un hyperlien non fonctionnel déclencherait une

erreur 404) mais son accès est activement bloqué après détection du robot. Ce dernier précise d'emblée la politique : « *# go away ? tell all others not in the list below to stay out!* ». Ce dispositif pourrait s'expliquer par la volonté de freiner l'exploration des sites à large échelle et de limiter l'identification de ressources protégées par le droit d'auteur.

H4 : Les revues prédatrices bloquent moins les robots d'IA génératives que les revues non prédatrices.

Les revues prédatrices se distinguent par, d'une part, le plus faible pourcentage de sites disposant d'un fichier *robots.txt*, d'autre part, le très faible biais global associé aux robots d'IA générative (cf. Tableau 3). Le biais global augmente sensiblement pour les revues de niveau 2.

Tableau 3. Biais global par type de revue

Robot	Revue prédatrice	Revue (niveau 0)	Revue (niveau 1)	Revue (niveau 2)
googlebot	-0,0141	-0,0027	0,0082	0,0112
bingbot	-0,0247	-0,0012	0,0082	0,0112
ccbot	0	-0,0194	-0,0551	-0,1453
gptbot	-0,0018	-0,0334	-0,0700	-0,2682
chatgpt-user	0	-0,0147	-0,0467	-0,1415
google-extended	0	-0,0167	-0,0529	-0,1750

H5 : Mieux une revue scientifique est classée et plus elle bloque les robots d'IA génératives.

Les revues non prédatrices ont une politique de régulation des robots d'exploration d'autant plus systématique que la revue est d'un niveau plus élevé. Cela se marque par l'utilisation plus systématique d'un fichier *robots.txt* (cf. Tableau 4).

Tableau 4. Utilisation du protocole d'exclusion des robots

Robot	Nombre de sites	Nombre de sites avec *robots.txt*
Revues prédatrices	1134	852 (75,1 %)
Revues de niveau 0	4070	3262 (80,1 %)
Revues de niveau 1	7169	5845 (81,5 %)
Revues de niveau 2	537	486 (90,5 %)

Le taux de blocage augmente avec le niveau de la revue, légèrement jusqu'au niveau 1 puis plus brutalement pour les revues de niveau 2 (cf. Tableau 5).

Tableau 5. Taux de blocage en fonction du niveau

Robot	Taux de blocage (ccbot)	Taux de blocage (google-extended)	Taux de blocage (gptbot)
Revues prédatrices	0 %	0 %	0,18 %
Revues de niveau 0	1,89 %	1,77 %	3,39 %
Revues de niveau 1	5,36 %	5,33 %	7,04 %
Revues de niveau 2	15,08 %	17,88 %	27,37 %
Top 50	19,35 %	32,26 %	48,39 %

De plus, plus la revue est bien classée et plus le biais global présente une valeur négative élevée (cf. Tableau 3). Les revues de niveau 2 se distinguent particulièrement des revues de niveau 0 ou 1.

Les résultats observés à partir des taux de blocage et de biais globaux calculés sont cohérents avec les hypothèses H1, H2, H3, H4 et H5.

4. Discussion

Hannigan et ses co-auteurs (2024) soulignent que les *chatbots* génératifs, comme ceux basés sur des grands modèles de langage (LLM, *Large Language Models*), sont conçus pour prédire des contenus plutôt que pour en comprendre la signification profonde. En particulier, les grands modèles de langage ne disposent pas de la capacité de dégager un consensus scientifique par la compréhension d'un corpus de documents, ni de vérifier la véracité des informations qu'ils relaient. Par ailleurs, la qualité des prédictions est fortement tributaire de celle des données d'entraînement. Cependant, les barrières d'accès posées par les éditeurs scientifiques en situation d'oligopole, comme Springer ou ScienceDirect, réduisent la disponibilité de données de haute qualité. En contrepartie, les revues prédatrices, qui ne bloquent pas les robots d'exploration, se révèlent être beaucoup plus accessibles. Ces revues, dépourvues de processus rigoureux de validation par les pairs, publient des articles souvent incorrects ou non vérifiés, et leur contenu risque de dégrader les résultats générés par les *chatbots*. Cela entraîne une diminution globale de la qualité des informations scientifiques disponibles pour entraîner ces outils. Les utilisateurs pourraient ainsi recevoir des réponses obsolètes, incomplètes ou incorrectes, sapant la fiabilité des *chatbots*. En lien avec le biais de sélection de données (Navigli & Conia, 2023), nous enrichissons la typologie de Ferrara (2023) par l'ajout d'un biais de validation, que nous définissons comme la surreprésentation parmi le corpus d'entraînement de données faiblement validées sur un plan scientifique.

Les effets induits par le degré variable de validation des données scientifiques n'est sans doute homogène dans l'ensemble des disciplines scientifiques. Larivière et al. (2015) mettent ainsi en évidence les différences de dépendance aux éditeurs scientifiques commerciaux en fonction des disciplines. Les sciences sociales apparaissent par exemple beaucoup plus affectées que la physique dès lors que cette dernière bénéficie du support de puissantes sociétés savantes qui conservent davantage de contrôle sur la diffusion de la production scientifique. Par ailleurs, la diffusion des connaissances sur Wikipédia, qui peut servir de proxy pour l'accès à la connaissance scientifique derrière *paywall*, n'est pas homogène non plus pour tous les domaines (Teplitskiy et al., 2016). Il en résulte que les risques de mésinformation scientifique au sein des IAG varient probablement en fonction de la discipline.

Les politiques différenciées de blocage par les éditeurs scientifiques peuvent-ils engendrer d'autres biais que le biais de validation précédemment discuté ? Les biais temporels paraissent les plus évidents. Les LLM souffre en effet d'un temps de création élevé. D'une part, les jeux de données nécessitent du temps pour être constitués puis traités (Navigli & Cornia, 2023). Ces délais peuvent être accrus par l'existence d'étapes de traitement manuelles, qui encouragent par ailleurs la réutilisation de jeux de données plus anciens. D'autre part, l'entraînement de l'IAG est lourd et prend donc lui-même du temps. Au 30 octobre 2024, les données d'entraînement du modèle GPT 4o n'allaient pas au-delà d'octobre 2023[4]. Les articles publiés au cours de l'année écoulée sont dès lors inconnus pour ChatGPT. Par ailleurs, les articles plus anciens ne sont pas nécessairement numérisés. Par exemple, certains *datasets* sont récents, comme arXiv qui ne remonte pas au-delà de 1991[5].

Les revues prédatrices modifient également sensiblement la provenance géographique des publications scientifiques. L'analyse de la localisation des serveurs hébergeant les revues

4 Voir https://platform.openai.com/docs/models/gpt-4o.

5 Voir https://en.wikipedia.org/wiki/ArXiv.

prédatrices et non prédatrices permet ainsi de mettre en évidence des disparités entre ces deux types de revue. La localisation des sites web a été déterminée avec un script Python basé sur la solution GeoLite2 de MaxMind. Les 10 localisations les plus fréquentes pour les revues prédatrices ont été conservées puis comparées aux revues listées (niveaux 0, 1 et 2). Cette localisation met en évidence une surreprésentation des revues indiennes parmi les revues prédatrices, ce qui est également constaté par Xia et al. (2017). De plus, les revues non prédatrices ressortent comme globalement moins concentrées sur quelques pays, soit les États-Unis d'Amérique et l'Inde (plus de 50 % des revues prédatrices). Dans les deux cas, le déséquilibre géographique est source de biais démographiques et de biais culturels, sans que l'impact soit facilement évaluable.

Si l'on prend plus spécifiquement ChatGPT, plusieurs modes d'interaction sont possibles en matière de recherche d'informations, profanes ou scientifiques : comme moteur de réponse (mode par défaut), comme moteur de recherche génératif (« Rechercher sur le Web ») ou comme moteur de réponse personnalisé (« custom » GPT). L'utilisation comme moteur de réponse permet des usages plus ou moins élaborés. Signalons que l'interrogation du *chatbot* peut faire appel à des stratégies d'interrogation plus ou moins avancées. L'usage le plus courant, celui du *direct* (ou *I/O*) *prompt*, consiste en une instruction simple, tandis que le c*hain-of-thought (CoT) prompt* permet d'expliciter ou d'imposer un raisonnement (Yao et al., 2024). Ainsi, un usage simple va consister à explorer une thématique à l'aide d'une succession de *direct prompts* (p. ex. question posée en langage naturel de manière à identifier le vocabulaire du domaine, exploitable dans un second temps dans un moteur de recherche d'information profane, p. ex. Google, ou scientifique, p. ex. Google Scholar ou Semantic Scholar). L'utilisation de la réponse est possible mais expose l'utilisateur, d'une part, aux hallucinations (de Corbière et al., 2023 ; Ye et al., 2023), d'autre part, au plagiat assisté par IA (Hutson, 2024). Un usage plus avancé consiste à imposer un processus de recherche plus élaboré. Par exemple, une recherche initiale peut être suivie de l'extraction de concepts dans les articles trouvés puis de documents associés à ces concepts, et enfin par l'exportation des références

bibliographiques aux formats APA ou Bibtex. L'utilisation comme moteur de recherche[6] permet de déployer les mêmes stratégies de *prompt* mais permet en plus d'obtenir des réponses associées à des références listées au sein d'un panneau latéral dédié. L'utilisation comme GPT personnalisé[7] permet de s'appuyer sur une spécialisation de ChatGPT à l'aide d'un *prompt system* spécifique (p. ex. affectation d'un rôle à l'IA) et d'une base de connaissances composée de documents maîtrisés.

Hannigan et ses coauteurs (2024) proposent une typologie pour comprendre ces limites, croisant deux dimensions : la vérifiabilité de la réponse et l'importance de sa véracité. Ils identifient ainsi quatre modalités d'utilisation des *chatbots* (authentifié, autonome, automatique, augmenté), associées à des risques spécifiques (ignorance, mauvaise calibration, routinisation et boîte noire). Voyons maintenant plus précisément quels sont les risques associés aux modes d'interaction avec ChatGPT dans le cas particulier de la recherche d'informations scientifiques, en nous appuyant sur les risques identifiés par Hannigan et ses coauteurs (2024) :

1. Le mode augmenté, où le *chatbot* est utilisé pour générer des idées ou des concepts puis les transformer, inclut l'utilisation du *chatbot* en recherche d'information pour identifier le vocabulaire du domaine et des noms d'auteurs. Ces informations extraites peuvent ensuite être utilisées dans un moteur de recherche scientifique tel que Google Scholar ou Semantic Scholar. La vérification a priori de la réponse n'est pas importante puisqu'elle sert de matière première pour la recherche d'information dans le moteur de recherche spécialisé. Par contre, l'utilisateur risque, dans ce cas d'utilisation, de passer à côté de certaines opportunités offertes par l'outil.

2. Le mode authentifié, où l'utilisateur soumet des tâches et vérifie ensuite méticuleusement les réponses, inclut l'utilisation du *chatbot* comme moteur de réponse, avec ou sans la suggestion de références. Le risque de mauvaise calibration concerne notamment la surestimation ou la sous-

6 Voir https://openai.com/index/introducing-chatgpt-search/.

7 Voir https://openai.com/index/introducing-gpts/.

estimation du phénomène d'hallucination, dès lors un usage trop timoré ou au contraire trop optimiste de l'outil, cette seconde situation conduisant au *botshit*. Dans le cas d'un *direct prompt* sans accès au Web, il est par ailleurs connu que la réponse souffrira de biais présents dans les données d'entraînement mais aussi que la formulation du *prompt* orientera les réponses (Sharma et al., 2024). Ce sont les problèmes du biais de sélection des données (Hannigan et al., 2024) et du biais de confirmation (Sharma et al., 2024). Dans le cas d'un *direct prompt* avec accès au Web, ces problèmes subsistent mais s'y ajoute l'effet du blocage des robots associés à OpenAI. En particulier, comme nous l'avons montré, de nombreuses plateformes de publication d'articles scientifiques bloquent l'accès aux robots d'OpenAI. ChatGPT va dès lors se baser sur des plateformes plus ouvertes telles que ResearchGate. Cela conduit à un biais de couverture important ainsi qu'à un biais de validation. En effet, les articles publiés sur ResearchGate couvrent inégalement les disciplines scientifiques et présentent des niveaux variables de qualité (Thelwall & Kousha, 2017). De plus, la re-publication d'articles derrière *paywall* s'y voit attaquée par les éditeurs commerciaux (Chawla, 2017), ce qui pourrait freiner la publication de recherches finalisées sur cette plateforme. Dans le contexte d'information scientifique, le phénomène d'hallucination couvre non seulement les erreurs de définition mais aussi des erreurs dans les références (de Corbière et al., 2023).

3. Le mode automatique, où l'utilisateur soumet des tâches standards et routinisées au *chatbot*, inclut par exemple le résumé d'un article et l'exécution de questions-réponses sur base de cet article. Le gain de temps produit par la délégation de ce type de tâches, pour lesquelles les intelligences artificielles génératives sont réputées performantes, est important. Cette efficacité peut faire perdre de vue à l'utilisateur le risque résiduel d'hallucination incluant la mauvaise interprétation du contenu ou l'ajout d'informations inventées (Ye et al., 2023).

4. Le mode autonome, où l'utilisateur délègue les tâches au *chatbot* avec un entraînement spécifique, inclut différentes configurations dans le cas de la recherche d'informations scientifiques.

Le premier concerne le codage de procédures de recherche sophistiquées à l'aide de *prompts* élaborés (p. ex. CoT) comme la réalisation d'une recherche en plusieurs étapes. Le second concerne l'utilisation de « customs » GPT, pour l'utilisation comme moteur de réponse spécialisé. Dans ces deux cas de figure, l'utilisateur ne maîtrise pas les tâches réellement réalisées par l'agent conversationnel puisque ce dernier fonctionne comme une boîte noire. Le *prompt* élaboré ne permet pas de réaliser une revue systématique de la littérature tandis que le GPT personnalisé peut malgré tout halluciner, en particulier dans les tâches de résumé (Ye et al., 2023) et d'attribution de citations, même si le risque d'erreur est plus limité. À titre d'illustration, le test d'un « custom » GPT nourri de documents traitant d'éthique utilitariste et d'éthique déontologique, mais pas d'éthique de la sollicitude (*care ethics*), a conduit le *chatbot* à sourcer sa réponse à un *prompt* abordant l'éthique de la sollicitude par un document traitant de l'éthique dans les soins de santé (*ethics*, *care*, *healthcare*), démontrant ainsi les limites du fonctionnement statistique des LLM.

Selon Dignum (2019), la production de systèmes d'intelligence artificielle « *responsables et dignes de confiance* » repose sur trois principes (ART). Premièrement, la redevabilité (« *accountability* ») « *fait référence à l'exigence pour un système d'être en mesure d'expliquer et de justifier ses décisions aux utilisateurs et à d'autres acteurs concernés* » (p.53). Deuxièmement, la responsabilité (« *responsability* ») « *fait référence au rôle des personnes elles-mêmes dans leur relation avec les systèmes d'IA* » (p.53). Elle concerne donc à la fois les producteurs et les utilisateurs des IAG ainsi que les personnes chargées des régulations et des réglementations. Troisièmement, la transparence (« *transparency* ») « *désigne la capacité de décrire, d'inspecter et de reproduire les mécanismes par lesquels les systèmes d'IA prennent des décisions et apprennent à s'adapter à leur environnement, ainsi que la provenance et la dynamique des données utilisées et générées par le système* » (p.54). Elle recouvre donc la documentation des jeux de données brutes et des processus de filtrage. Ces quelques définitions permettent de structurer un ensemble de recommandations à destination des producteurs et des utilisateurs d'IAG. Ces préconisations recouvrent différentes préoccupations

d’ordre éthique incluant notamment l’accessibilité des technologies, la fiabilisation des informations et la responsabilisation des acteurs (Hamet & Michel, 2018).

Premièrement, la recherche met en évidence le besoin de *datasets* composés d’information scientifique de qualité. Cela suppose d’aller plus loin que l’entraînement sur base de *preprints* tel que proposé par The Pile. Cependant, cette recommandation pourrait se révéler difficile à mettre en œuvre dès lors que la publication d’articles scientifiques dans de tels jeux de données relèverait de la reproduction et de la communication d’œuvres protégées par le droit d’auteur et pourrait donc exposer ses initiateurs à des risques légaux. Une telle initiative pourrait néanmoins être facilitée, d’une part, par la tolérance juridique affichée en Europe au titre de l’exception TDM (*Text and Data Mining*) à l’égard des structures non commerciales dédiées à la mise en commun des jeux de données (Goldstein et al., 2024), d’autre part, par la négociation avec les éditeurs de journaux en *open access* reconnus de qualité. En complément, la publication sous licence libre de *pipelines* de traitement de jeux de données scientifiques pourrait réduire ce risque légal tout en garantissant l’audit et l’automatisation de la création de jeux de données scientifiques de qualité. De la sorte, le biais de validation et le biais temporel seraient réduits. Par ailleurs, la formation aux limitations, incluant les problèmes connus des intelligences artificielles génératives (hallucinations, régurgitations, biais...) doit devenir une priorité dans les équipes de développement, de manière en éviter la propagation au travers, soit de nouveaux modèles de langage, soit de modèles existants intégrés dans des applications innovantes.

Deuxièmement, l’usage raisonné des outils d’IAG nécessite également la formation des utilisateurs. Cela passe tout d’abord par une évaluation des risques de *botshit* en fonction des scénarios d’utilisation (Hannigan et al., 2024). Ensuite, il convient de former les utilisateurs aux biais induits par le mode conversationnel lui-même. Sharma et al. (2024) ont ainsi étudié les comportements de recherche basés sur, d’une part, des systèmes de recherche conventionnels, d’autre part, des

systèmes de recherche conversationnels. Les auteurs montrent que les requêtes conversationnelles sont plus verbeuses et expressives. Cela tend à polariser davantage les réponses et à accroître le biais de confirmation. Les auteurs parlent de « *chambres d'écho génératives* ». Notons cependant que cette problématique des biais de confirmation induits par la formulation des requêtes n'est pas un problème neuf. En effet, il s'observe également dans l'utilisation d'outils de recherche plus classiques (Azzopardi, 2021), incluant Google Scholar ou Semantic Scholar pour la recherche d'informations scientifiques (Kacperski et al., 2023). De plus, ils relèvent que les utilisateurs tendent à moins lire les documents sources lorsqu'il y en a. La mésinformation se révèle dès lors liée non seulement aux limites de la technologie mais aussi au comportement d'interaction des utilisateurs avec les applications recourant à des LLM. Ces constats nous permettent d'identifier un besoin de formation des utilisateurs, d'une part, à la formulation textuelle d'un *prompt* (p. ex. neutralité), d'autre part, à la structuration d'un *prompt* (p. ex. type de *prompt* et affectation d'un rôle). En effet, Rawte et ses co-auteurs (2023) montrent notamment comment l'utilisation d'un langage formel et tangible permet de réduire les hallucinations tandis que Shen et al. (2023) démontrent l'impact sous condition de l'affectation d'un rôle à ChatGPT. Les solutions englobent la formation à des outils permettant aux utilisateurs de travailler sur des documents chargés et validés dans l'IA (p. ex. GPT « custom » et NotebookML). Ces recommandations concernent aussi la formation des étudiants en master et en doctorat. Une part substantielle de ces derniers éprouvait en effet des difficultés, avant l'apparition des intelligences artificielles génératives, dans la recherche d'informations scientifiques (Soung & Dumouchel, 2019). Leur penchant pour Google Scholar, du fait de sa convivialité, éventuellement au détriment de base de données plus classiques comme Scopus, devrait par ailleurs se reproduire avec les systèmes de recherche conversationnelle qui, à défaut d'une fiabilité irréprochable, offrent une interface à première vue intuitive.

Soulignons enfin l'apparition de nouveaux moteurs génératifs dédiés à la recherche d'informations scientifiques (Consensus, Elicit, SciSpace...). Ces outils apportent une sécurité à l'utilisateur dans la

mesure où ils réalisent une synthèse des éléments importants en rapport avec le *prompt* sur base d'éléments trouvés dans leur base de données. Cela réduit notamment le phénomène d'hallucination (sans néanmoins totalement le supprimer ; Ye et al., 2023). Cependant, ces services apportent une nouvelle limitation que nous pourrions qualifier de biais de couverture. Leur facilité d'utilisation masque en effet le caractère plus ou moins restreint, plus ou moins transparent, des bases de données d'articles sous-jacentes. Consensus et Elicit s'appuient par exemple sur la base de données Semantic Scholar, avec des écarts non expliqués dans les volumétries communiquées. En réaction, la communauté scientifique tente, à l'image d'Artirev, de développer des outils davantage maîtrisés (Walsh et al., 2022).

Cette recherche souffre de trois limitations. Premièrement, elle ne permet pas de particulariser finement les conclusions par discipline ou par famille de disciplines. Nous avons en effet vu, d'une part, que la dépendance aux éditeurs commerciaux dépendait du domaine de recherche, d'autre part, que des classements, notamment sectoriels, existaient. Le premier élément pourrait faciliter l'accès à des données d'entraînement, mais uniquement dans certaines disciplines, tandis que le second pourrait décourager la création de revues prédatrices dans des disciplines où des logiques de listes blanches prévalent dans l'évaluation des dossiers scientifiques (par exemple : classement français FNEGE[8] en science de gestion). Deuxièmement, le calcul de biais global a actuellement été réalisé sur l'ensemble des sites sans prendre en compte la concentration des revues sur quelques sites de grands éditeurs scientifiques (Springer, Sciencedirect...). Il en résulte une sous-estimation du biais global. Le calcul de ce dernier pour les 50 domaines les plus représentés en base de données donne cependant une indication de la borne supérieure du biais global pour les revues non prédatrices. Troisièmement, la recherche est limitée par l'utilisation de la liste de Beall. D'une part, la transparence de la méthodologie permettant de dresser cette liste a été critiquée (Richtig et al., 2018). D'autre part, cette liste populaire commence à dater puisque sa mise à jour a été stoppée en

8 Voir https://fnege.org/classement-des-revues-scientifiques-en-sciences-de-gestion/.

2017 (Richtig et al., 2018). Aussi serait-il intéressant de recalculer la mesure de biais global sur une liste actualisée de revues prédatrices.

5. Conclusion

L'étude met en lumière l'impact des restrictions d'accès appliquées par les éditeurs scientifiques aux robots d'exploration des intelligences artificielles génératives. D'une part, l'accès aux contenus intégraux est souvent bloqué par *paywall* ; d'autre part, l'accès aux sites des revues, donc des résumés des articles (si ces derniers sont derrière *paywalls*), est interdit par usage du protocole d'exclusion des robots. Nos résultats montrent le risque d'une prépondérance de contenus issus de revues de moindre qualité, voire prédatrices, dans les données d'entraînement des IAG, créant ainsi un biais de validation. Ce biais expose les utilisateurs à une mésinformation scientifique, potentiellement amplifiée dans des domaines sensibles. La recherche souligne l'urgence de développer des stratégies de rééquilibrage des *datasets* en favorisant un accès contrôlé et éthique aux contenus validés par des pairs, afin d'améliorer la qualité et la fiabilité des réponses fournies par les grands modèles de langage. Notre recherche contribue ainsi à la compréhension des biais dans les LLM (Ferrara, 2023 ; Navigli & Conia, 2023), à leur mesure (Chu et al., 2024) et à l'évaluation des risques (Hannigan et al., 2024), dans le cadre de stratégies de recherche d'information scientifique. Les outils d'intelligence artificielle générative offrent donc des perspectives intéressantes en matière d'aide à la recherche d'informations scientifiques. Ils permettent d'identifier plus rapidement le vocabulaire associé à un domaine, de découvrir de manière itérative de la littérature ou encore de faciliter la sélection des articles en interagissant avec leur contenu ou en produisant des résumés. Ces potentialités s'accompagnent cependant de limitations quant à la fiabilité des informations renvoyées dans les réponses par l'agent conversationnel. Leur diffusion doit donc s'accompagner d'une montée en compétences chez les enseignants-chercheurs puis d'une formation des étudiants et d'un suivi de l'utilisation de ces outils.

Cette recherche présente trois perspectives. Premièrement, la recherche actuelle ne particularise pas ses conclusions en fonction des disciplines scientifiques. Le taux de blocage des robots des producteurs d'IAG est-il homogène parmi l'ensemble de ces disciplines, ou bien certaines disciplines sont-elles davantage touchées que d'autres, et dès lors davantage exposées au risque de mésinformation scientifique ? Deuxièmement, le biais de validation a fait l'objet d'une estimation au niveau de la constitution des jeux de données brutes. L'analyse n'a pas été poussée jusqu'à des jeux de données filtrées, lesquels sont rarement publiés. La contamination par des contenus de faible qualité, voire prédateurs, se vérifie-t-elle dans les jeux de données réellement utilisés, et dans quelles proportions ?

6. Bibliographie

AZZOPARDI L. (2021), « Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval », *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, p. 27-37. https://doi.org/10.1145/3406522.3446023.

BEALL J. (2010), « "Predatory" Open-Access Scholarly Publishers », *The Charleston Advisor*, vol. 11, no 4, p. 10-17.

BANKS M. (2016), « What Sci-Hub Is and Why It Matters », *American Libraries*, vol. 47, no 6, p. 46-49. https://www.jstor.org/stable/26380679.

BONTRIDDER N., POULLET Y. (2021), « The Role of Artificial Intelligence in Disinformation », *Data & Policy*, vol. 3, e32. https://doi.org/10.1017/dap.2021.20.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., ... AMODEI D. (2020), « Language Models Are Few-Shot Learners », *arXiv preprint*, arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165.

CABANAC G. (2024), « Chain Retraction: How to Stop Bad Science Propagating Through the Literature », *Nature*, vol. 632, no 8027, p. 977-979. https://www.nature.com/articles/d41586-024-02747-1.

CHAWLA D. S. (2017), « Publishers Take Academic Networking Site to Court », *Science*, vol. 358, no 6360, p. 161. https://www.science.org/doi/pdf/10.1126/science.358.6360.161.

CHU Z., WANG Z., ZHANG W. (2024), « Fairness in Large Language Models: A Taxonomic Survey », *ACM SIGKDD Explorations Newsletter*, vol. 26, no 1, p. 34-48. https://doi.org/10.1145/3682112.3682117.

DE WYNTER A., WANG X., SOKOLOV A., GU Q., CHEN S. Q. (2023), « An Evaluation on Large Language Model Outputs: Discourse and Memorization », *Natural Language Processing Journal*, vol. 4. https://doi.org/10.1016/j.nlp.2023.100024.

DIGNUM, V. (2019). Responsible artificial intelligence: how to develop and use AI in a responsible way (Vol. 2156). Cham: Springer. https://doi.org/10.1007/978-3-030-30371-6.

DINZINGER M., GRANITZER M. (2024), « A Longitudinal Study of Content Control Mechanisms », *Companion Proceedings of the ACM on Web Conference 2024*, p. 1382-1387. https://doi.org/10.1145/3589335.3651893.

DODGE J., SAP M., MARASOVIĆ A., AGNEW W., ILHARCO G., GROENEVELD D., ... GARDNER M. (2021), « Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus », *arXiv preprint*, arXiv:2104.08758. https://doi.org/10.48550/arXiv.2104.08758.

FERRARA E. (2023), « Should ChatGPT Be Biased? Challenges and Risks of Bias in Large Language Models », arXiv preprint, arXiv:2304.03738. https://doi.org/10.5210/fm.v28i11.13346.

GAO L., BIDERMAN S., BLACK S., GOLDING L., HOPPE T., FOSTER C., ... LEAHY C. (2020), « The Pile: An 800GB Dataset of Diverse Text for Language Modeling », *arXiv preprint*, arXiv:2101.00027. https://doi.org/10.48550/arXiv.2101.00027.

FLORIDI L., CHIRIATTI M. (2020), « GPT-3 : Its Nature, Scope, Limits, and Consequences », *Minds and Machines*, vol. 30, p. 681-694. https://doi.org/10.1007/s11023-020-09548-1.

GERSHENSON S., POLIKOFF M. S., WANG R. (2020), « When Paywall Goes AWOL: The Demand for Open-Access Education Research », *Educational Researcher*, vol. 49, no 4, p. 254-261. https://doi.org/10.3102/0013189X20909834.

GOLDSTEIN P., STUETZLE C., BISCHOFF S. (2024), « Kneschke vs. LAION – Landmark Ruling on TDM Exceptions for AI Training Data – Part 1 », *Kluwer Copyright Blog*. https://copyrightblog.kluweriplaw.com/2024/11/13/kneschke-vs-laion-landmark-ruling-on-tdm-exceptions-for-ai-training-data-part-1/.

HAMET J., MICHEL S. (2018), « Les questionnements éthiques en systèmes d'information », *Revue française de gestion,* no 2, p. 99-129. https://doi.org/10.3166/RFG.2018.00221.

HU K. (2023), « ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note », *Reuters*, 2 février 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

HUTSON J. (2024), « Rethinking Plagiarism in the Era of Generative AI », *Journal of Intelligent Communication*, vol. 4, no 1, p. 20-31. https://doi.org/10.54963/jic.v4i1.220.

KACPERSKI C., BIELIG M., MAKHORTYKH M., SYDOROVA M., ULLOA R. (2023), « Examining Bias Perpetuation in Academic Search Engines: An Algorithm Audit of Google and Semantic Scholar », *arXiv preprint*, arXiv:2311.09969. https://doi.org/10.48550/arXiv.2311.09969.

LARIVIÈRE V., HAUSTEIN S., MONGEON P. (2015), « The Oligopoly of Academic Publishers in the Digital Era », PLOS ONE, vol. 10, no 6, e0127502. https://doi.org/10.1371/journal.pone.0127502.

MALEKI N., PADMANABHAN B., DUTTA K. (2024), « AI Hallucinations: A Misnomer Worth Clarifying », *2024 IEEE Conference on Artificial Intelligence (CAI)*, p. 133-138, IEEE. https://doi.org/10.1109/CAI59869.2024.00033.

NAVIGLI R., CONIA S., ROSS B. (2023), « Biases in Large Language Models: Origins, Inventory, and Discussion », *ACM Journal of Data and Information Quality*, vol. 15, no 2, p. 1-21. https://doi.org/10.1145/3597307.

RAHMAN M., TERANO H. J. R., RAHMAN N., SALAMZADEH A., RAHAMAN S. (2023), « ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples », *Journal of Education, Management and Development Studies*, vol. 3, no 1, p. 1-12. http://doi.org/10.52631/jemds.v3i1.175.

RAWTE V., PRIYA P., TONMOY S. M., ZAMAN S. M., SHETH A., DAS A. (2023), « Exploring the Relationship Between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness », *arXiv preprint*, arXiv:2309.11064. https://doi.org/10.48550/arXiv.2309.11064.

RICHTIG G., BERGER M., LANGE-ASSCHENFELDT B., ABERER W., RICHTIG E. (2018), « Problems and Challenges of Predatory Journals », *Journal of the European Academy of Dermatology and Venereology*, vol. 32, no 9, p. 1441-1449. https://doi.org/10.1111/jdv.15039.

SHARMA N., LIAO Q. V., XIAO Z. (2024), « Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking », *Proceedings of the CHI Conference on Human Factors in Computing Systems*, p. 1-17. https://doi.org/10.1145/3613904.3642459.

SHEN X., CHEN Z., BACKES M., ZHANG Y. (2023), « In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT », *arXiv preprint*, arXiv:2304.08979. https://doi.org/10.48550/arXiv.2304.08979.

SOUNG S., DUMOUCHEL G. (2019), « Les pratiques de recherche d'information des étudiants aux cycles supérieurs en éducation », *Revue internationale des technologies en pédagogie universitaire*, vol. 16, no 3, p. 73-92. https://doi.org/10.18162/ritpu-2019-v16n3-05.

SUN Y., ZHUANG Z., COUNCILL I. G., GILES C. L. (2007), « Determining Bias to Search Engines from Robots.txt », *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, p. 149-155, IEEE. https://doi.org/10.1109/WI.2007.98.

TEPLITSKIY M., LU G., DUEDE E. (2017), « Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science », *Journal of the Association for Information Science and Technology*, vol. 68, no 9, p. 2116-2127. https://doi.org/10.1002/asi.23687.

THELWALL M., KOUSHA K. (2017), « ResearchGate Articles: Age, Discipline, Audience Size, and Impact », *Journal of the Association for Information Science and Technology*, vol. 68, no 2, p. 468-479. https://doi.org/10.1002/asi.23675.

VISEUR R., DELCOUCQ L. (2024), « Exploration des pratiques de régulation des IA génératives par le protocole d'exclusion des robots », *INFORSID*, 28-31 mai 2024, Nancy (France). http://inforsid.fr/actes/2024/inforsid24-89-104.pdf.

WALSH I., RENAUD A., MEDINA M. J., BAUDET C., MOURMANT G. (2022), « ARTIREV: An Integrated Bibliometric Tool to Efficiently Conduct Quality Literature Reviews », *Systèmes d'information & management*, vol. 27, no 4, p. 5-50. https://revuesim.org/index.php/sim/article/view/1217.

XIA J., HARMON J. L., CONNOLLY K. G., DONNELLY R. M., ANDERSON M. R., HOWARD H. A. (2015), « Who Publishes in "Predatory" Journals? », *Journal of the Association for Information Science and Technology*, vol. 66, no 7, p. 1406-1417. https://doi.org/10.1002/asi.23265.

YAO S., YU D., ZHAO J., SHAFRAN I., GRIFFITHS T., CAO Y., NARASIMHAN K. (2024), « Tree of Thoughts: Deliberate Problem Solving with Large Language Models », *Advances in Neural Information Processing Systems*, vol. 36. https://doi.org/10.48550/arXiv.2305.10601.

YE H., LIU T., ZHANG A., HUA W., JIA W. (2023), « Cognitive Mirage: A Review of Hallucinations in Large Language Models », arXiv preprint, arXiv:2309.06794. https://doi.org/10.48550/arXiv.2309.06794.

ZHOU T., LI S. (2024), « Understanding User Switch of Information Seeking: From Search Engines to Generative AI », *Journal of Librarianship and Information Science*, 09610006241244800. https://doi.org/10.1177/09610006241244800.