Analyse de l'impact des restrictions d'accès à l'information scientifique sur la qualité des données d'entraînement des LLM

Robert Viseur[1]

1. Service TIC, FWEG, UMONS
17 place Warocqué, B-7000 Mons, Belgique
robert.viseur@umons.ac.be

RÉSUMÉ. Cet article examine l'impact, sur la qualité des données d'entraînement des grands modèles de langage (LLM), des mesures de blocage, par les éditeurs de revues scientifiques, des robots d'explorations exploités par les producteurs d'intelligences artificielles génératives (IAG) comme OpenAI. Des données de mauvaise qualité pour entraîner des LLM peuvent en effet conduire à la mésinformation scientifique. Une analyse empirique basée sur cinq hypothèses a été réalisée pour étudier les pratiques de blocage des robots d'IAG. Des scripts Python ont permis de collecter et d'analyser les fichiers robots.txt de différents sites, comparant les taux de blocage entre revues prédatrices et non prédatrices ainsi qu'entre revues de différents niveaux de classement. Il ressort que les robots d'IAG sont davantage bloqués que ceux des moteurs de recherche traditionnels, surtout par les éditeurs scientifiques de haut niveau. A contrario les revues prédatrices bloquent moins ces robots. Ces modalités de blocage entraînent donc une surreprésentation potentielle de contenus de moindre qualité dans les datasets d'entraînement des IAG. Cela crée un « biais de validation », augmentant le risque de mésinformation dans les réponses des IAG. L'étude révèle un problème peu exploré concernant l'accès inégal aux sources de qualité pour l'entraînement des IAG. Elle souligne l'impact potentiel des politiques de blocage sur la propagation de la mésinformation.

ABSTRACT. This article examines the impact, on the quality of data used to train large language models (LLMs), of measures taken by scientific publishers to block exploration robots used by generative artificial intelligence (GAI) producers such as OpenAI. Poor quality data used to train LLMs can lead to scientific misinformation. An empirical analysis based on five hypotheses was carried out to study the blocking practices of IAG robots. Python scripts were used to collect and analyse robots.txt files from different sites, comparing blocking rates between predatory and non-predatory journals and between journals at different ranking levels. Generally speaking, IAG robots are blocked more than those of traditional search engines, especially by high-level scientific publishers. On the other hand, predatory journals block these robots less. These blocking methods therefore lead to a potential over-representation of lower quality content in the IAG training datasets. This creates a 'validation bias', increasing the risk of misinformation in IAG responses. The study reveals a little-explored problem concerning unequal access to quality sources for training IAGs. It highlights the potential impact of blocking policies on the spread of misinformation.

Mots-clés : intelligence artificielle, biais, jeux de données, mésinformation, revue prédatrice.

KEYWORDS: artificial intelligence, bias, datasets, misinformation, predatory journals.

1. Introduction

L'année 2023 a été celle du développement commercial des intelligences artificielles génératives (IAG) avec l'essor des *chatbots* comme ChatGPT et plus largement celui des grands modèles de langage (LLM) comme GPT. GPT est « *un modèle linguistique autorégressif de troisième génération qui utilise l'apprentissage profond pour produire des textes semblables à ceux des humains* » (Floridi & Chiriatti, 2020). Les performances de ces modèles dépendent notamment de la disponibilité, en vue de leur entraînement, d'« *un ensemble de données non étiquetées composé de textes, tels que Wikipédia et de nombreux autres sites, principalement en anglais, mais aussi dans d'autres langues* » (Floridi & Chiriatti, 2020). Le Common Crawl, soit plus moins 300 TB avant filtrage[1], représenterait ainsi 60 % des données d'entraînement de GPT-3 (Brown et al., 2020). La qualité des données est également importante. Dodge et ses co-auteurs (2021) montrent ainsi l'appétit des producteurs d'IAG pour les données issues de la presse en ligne, des revues scientifiques et de l'encyclopédie collaborative Wikipédia. Malheureusement, Viseur et Delcoucq (2024) ont montré que les politiques de blocage des sites web de la presse en ligne, via le protocole d'exclusion des robots, étaient largement adoptées.

L'intelligence artificielle (IA) contribue au phénomène de désinformation (Bontridder & Poullet, 2021). La désinformation « *est une information fausse, inexacte ou trompeuse qui est diffusée dans l'intention de tromper le destinataire* », contrairement à la mésinformation, qui « *désigne une information fausse, inexacte ou trompeuse partagée sans intention de tromper* » (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* » (Maleki et al., 2024 ; 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* »).

Ce phénomène d'hallucination peut s'expliquer par différentes causes incluant la formulation des *prompts*, les limitations des modèles et les données d'entraînement. En particulier, les hallucinations peuvent découler du fait que les données sont « *biaisées, non actuelles, incomplètes ou inexactes* » (Hannigan et al., 2024). Le caractère non fiable des données pourra conduire le modèle à générer une information erronée basée sur des informations réellement présentes dans le *dataset*. L'absence de données relatives à une thématique pourra amener le modèle à broder pour composer une réponse cohérente, plausible, mais contenant des informations inexactes, inventées. Ces problèmes relèvent donc, soit de la fidélité (« *faithfulness* »), soit de l'exactitude des faits (« *factualness* ») (Ye et al., 2023).

Nous nous intéressons en particulier dans cette recherche aux risques de dégradation de la qualité des jeux de données d'entraînement des grands modèles de langage, et dès lors de mésinformation en matière d'information scientifique par les

[1] Voir https://commoncrawl.github.io/cc-crawl-statistics/.

IA génératives. 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 non validées voire frauduleuses ? L'indisponibilité d'informations scientifiques fiables et complètes pourrait en effet conduire non seulement à des biais mais aussi à des problèmes importants de qualité dans l'information scientifique générée par, d'une part, les agents conversationnels, d'autre part, les plateformes s'appuyant sur des modèles génératifs. Cela inclut par exemple les *newsbots* (Viseur, 2024).

Cet article est organisé en quatre 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. Cette dernière s'appuie sur la mesure du blocage des robots d'exploration des producteurs d'IAG par les éditeurs scientifiques. Les cinq hypothèses testées sont ensuite présentées. La troisième présente les résultats pour chaque hypothèse. La quatrième, précédant la conclusion, discute plus globalement les limitations des IAG dues aux jeux de données.

2. Revue de la littérature

Les producteurs d'IA génératives collectent de vastes ensembles de données à l'aide de robots d'exploration (*crawlers*) qui parcourent le Web (Viseur & Delcoucq, 2024). Cependant, ces robots sont susceptibles d'être bloqués de manière, soit passive, soit active (Dinzinger & Granitzer, 2024 ; Viseur & Delcoucq, 2024 ; Amin Azad et al., 2020 ; Sun et al., 2007). Le blocage passif s'appuie sur le protocole d'exclusion des robots[2]. Ce dernier permet de préciser les sections du site qui peuvent être parcourues et celles qui doivent être ignorées (Viseur & Delcoucq, 2024 ; Sun et al., 2007). Seuls les robots dits éthiques respectent ce principe d'*opt-out*. L'utilisateur d'un robot d'exploration peut ainsi choisir d'ignorer sciemment ces consignes et de collecter malgré tout les contenus publiés en ligne. A contrario, le blocage actif conduit à une détection du robot puis à son blocage[3]. La détection peut être réalisée simplement en s'appuyant sur des listes de *user agents* ou d'adresses IP. Ces dernières peuvent être fournies spontanément par les propriétaires des robots[4] ou alimentées par les gestionnaires de sites web. La détection est également possible par le calcul d'empreinte (« *browser fingerprinting* ») et l'analyse du comportement des terminaux accédant aux sites web (Amin Azad et al., 2020). Ce type d'approche plus sophistiquée est notamment retenu par le service commercial Cloudflare[5] (Amin Azad et al., 2020). Une fois repéré, le robot peut aussi être soumis à la résolution d'un *captcha* (Amin Azad et al., 2020). Enfin, une redirection peut également être utilisée par l'exploitation de la fréquente incapacité des robots (sauf s'ils s'appuient sur un « *headless browser* » comme feu PhantomJS ou Selenium par exemple) d'exécuter les codes Javascript. Les dispositifs de blocage actif sont dès lors nombreux, et aisément accessibles aux éditeurs.

---

[2] Voir https://robots-txt.com/ et https://datatracker.ietf.org/doc/rfc9309/.

[3] Pour une synthèse illustrée à destination des praticiens, voir par exemple https://www.willmaster.com/library/tutorials/ways-to-redirect-bots-and-browsers.php.

[4] Voir par exemple la page d'information fournie par OpenAI : https://platform.openai.com/docs/bots. Cette page inclut l'accès à des fichiers JSON documentant les IP utilisées par les différents robots.

[5] Voir https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/.

Cette collecte de données non négociée est considérée comme une forme de prédation par certains éditeurs de contenus. Elle conduit donc à des politiques de blocage par les propriétaires des sites présentant des contenus originaux (Viseur & Delcoucq, 2024 ; Dinzinger & Granitzer, 2024). Viseur et Delcoucq (2024) ont en particulier 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 sont fréquents, basés sur le protocole d'exclusion des robots, et que cela occasionne de nombreux biais, notamment linguistiques et culturels (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, ChatGPT-User (utilisé pour les actions dans ChatGPT ou les customs ChatGPT[6]), Google-Extended et CCbot (associé au Common Crawl). Fin 2024, OpenAI a rajouté OAI-SearchBot associé à son fonctionnement comme moteur de recherche. Des listes plus complètes existent[7] . 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 de la recherche 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 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.

L'automatisation de l'exploration de la littérature scientifique a précédé le développement des LLM. Deux jeux de données scientifiques antérieurs aux premières IAG commerciales ressortent ainsi de la littérature en NLP : S2ORC (Lo et al., 2020) et Microsoft Academic Graph (Wang et al., 2020). S2ORC est un vaste

[6] Voir https://platform.openai.com/docs/bots.
[7] Voir par exemple https://github.com/ai-robots-txt/ai.robots.txt.

corpus contenant à sa création 81,1 millions d'articles. Pour une minorité d'articles en *open access* (8,1 millions), le texte intégral est disponible ; pour la plupart, seules les métadonnées, le résumé et les références sont fournies. S2ORC est associé au moteur de recherche scientifique Semantic Scholar (Kinney et al., 2023), qui référence aujourd'hui plus de 200 millions de documents. Microsoft Academic Graph (MAG) prend la forme d'un graphe d'articles (incluant les résumés) régulièrement mis à jour. Ces deux jeux de données ne sont pas spécifiquement liés à l'univers des LLM. Cependant, à l'instar de certains outils basés sur l'IA générative, ils ont été conçus comme des outils de NLP (*Natural Langage Processing*) capables de « *supporter la recherche sur des documents académiques* » (Lo et al., 2020) au travers notamment d'« *agents logiciels* » capables d'explorer automatiquement la littérature scientifique disponible sur le Web (Wang et al., 2020). Microsoft Academic Graph (Wang et al., 2020 ; Sinha et al., 2015) a ultérieurement été fusionné avec AMiner (Tang et al., 2008) pour former Open Academic Graph (OAG). Sa réutilisation est possible à des fins de recherche uniquement.

La littérature existante donne quelques éclairages sur les *datasets* proposant de l'information scientifique pour l'entraînement des LLM (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 une information soumise à un processus strict de *peer reviewing*. 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 (Cabanac, 2024). 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*. 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 (Cardon, 2013), qui tient compte à la fois du nombre de citations reçues par une revue et du prestige des revues d'où proviennent les citations. 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.

En complément des *datasets* publics, les producteurs de LLM construisent aussi des *datasets* internes. Ceux-ci sont alimentés par, soit leurs propres robots, soit des accords signés avec les éditeurs (Gibney, 2024). Dès lors, les données d'entraînement sont composées, d'une part, de données publiques, utilisées après filtrage, collectées sur le Web par des tiers (p. ex. Common Crawl) ou par les producteurs eux-mêmes (p. ex. GPTbot), d'autre part, de données achetées auprès des éditeurs. Gibney (2024) mentionne ainsi Taylor & Francis[8] (accord avec Microsoft) et Wiley (partenaire inconnu), pour l'édition scientifique, ainsi que Financial Times (accord avec OpenAI), pour la presse spécialisée. Le phénomène concernerait cependant davantage d'éditeurs[9] tels que Elsevier, Springer Nature et De Gruyter Brill (Kwon, 2024). Le Generative AI Licensing Agreement Tracking tente ainsi d'inventorier les accords souscrits, souvent sans identification des bénéficiaires. Cette politique d'octroi de licences de gré à gré motive aussi les blocages mis en œuvre par les éditeurs (Kwon, 2024). De manière plus surprenante, des données volées, par exemple issues de Library Genesis (LibGen), seraient ponctuellement utilisées, par Meta (modèles LlaMa) et OpenAI (modèles GPT) notamment[10]. Le *dataset* Books3[11] serait ainsi incriminé. La composition précise des jeux de données d'entraînement demeure donc au final une inconnue, excepté pour des projets open-sources comme LUCIE[12], où la transparence est au cœur du projet.

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

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

[8] Voir https://www.ccn.com/news/technology/microsoft-taylor-francis-secret-ai-publishing-deal-outrages-academics/ et https://www.informa.com/globalassets/documents/investor-relations/2024/informa-plc---market-update.pdf pour plus d'informations.

[9] L'exploitation des contenus scientifiques pourrait être facilitée par la structuration des articles selon le format *Journal Article Tag Suite* (JATS) Voir Kleidermacher et Zou (2025) pour un exemple d'exploitation du format JATS par une IA générative.

[10] Voir https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ (incluant les liens vers les dossiers juridiques) et https://authorsguild.org/news/you-just-found-out-your-book-was-used-to-train-ai-now-what/ pour plus d'informations.

[11] Voir par exemple https://aicopyright.substack.com/p/the-books-used-to-train-llms pour une analyse du contenu de ce jeu de données.

[12] Voir https://lucie.chat/ et https://huggingface.co/collections/OpenLLM-France/lucie-llm-67099ba7b992dee2c32b1f92 pour plus d'informations.

qualité académique élémentaires tandis que le niveau 2 rassemble les meilleurs canaux de publications. Le niveau 0 contient des revues non scientifiques, dédiées par exemple à la vulgarisation, mais aussi des revues prédatrices, telles que « 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. En effet, 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.

Les bases de données obtenues en sortie sont ensuite utilisées pour collecter les fichiers *robots.txt* puis, après analyse de ces derniers, obtenir un Top 10 des robots les plus fréquemment cités, identifier les pratiques de blocage (pas de fichier *robots.txt*, liste blanche, liste noire…) et calculer le taux de blocage par robot ainsi que le biais global. Le biais global 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 ; Sun et al., 2007). Il est ainsi possible d'estimer la discrimination des robots d'IAG comparativement aux robots des moteurs de recherche. Le calcul du biais global utilise la procédure simplifiée définie par Viseur et Delcoucq (2024). Les résultats sont enregistrés dans un fichier journal. Ces fichiers peuvent ensuite être ingérés par ChatGPT pour la production de tableaux de synthèse spécifiques.

4. Résultats

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 dans ChatGPT ou dans les « customs » ChatGPT 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 internationaux bloquent davantage les robots d'IA génératives.

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 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, par 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 %)

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 %

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). 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.

Toutes nos hypothèses (H1, H2, H3, H4 et H5) sont donc corroborées par les valeurs calculées des taux de blocage et des biais globaux.

5. Discussion

Hannigan et ses co-auteurs (2024) rappellent que « *les chatbots génératifs ne s'intéressent pas à la connaissance intelligente mais à la prédiction* ». Les grands modèles de langage (LLM, *Large Language Model*) sont en effet entraînés sur de vastes ensembles de données, souvent collectées à partir du Web, pour prédire des contenus. Ils sont capables de générer « *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). Deux choses méritent d'être soulignées à ce stade. D'une part, les grands modèles de langage ne disposent pas de la capacité à dégager un consensus scientifique par la compréhension profonde d'un corpus de documents. D'autre part, la qualité des contenus prédits dépend fortement de la qualité des données fournies en entraînement.

Or, si les robots d'exploration sont bloqués par les revues scientifiques bien classées et les grandes plateformes comme Springer ou ScienceDirect, le contenu provenant de ces sources de haute qualité risque d'être sous-représenté dans les jeux de données obtenues par *scraping*. En revanche, les revues prédatrices, qui ne bloquent pas ces robots, risquent d'y voir leur contenu surreprésenté. Ce déséquilibre entraîne un risque de diminution de la qualité de l'information scientifique que les *chatbots* (ou d'autres applications génératives) peuvent fournir. Les revues prédatrices publient souvent des articles sans processus rigoureux de validation par les pairs. Les LLM entraînés sur ces données sont davantage susceptibles de produire des informations erronées. Cette tendance est amplifiée par la moindre accessibilité des données issues des revues les mieux classées. Dès lors, en relayant des informations issues de sources peu fiables, les *chatbots* peuvent involontairement contribuer à la propagation de la mésinformation scientifique

(« *botshit* »). Cela est particulièrement préoccupant dans des domaines sensibles comme la santé, l'environnement ou la technologie, où des informations erronées peuvent avoir des conséquences graves. En outre, cela peut influencer négativement la perception de certains sujets scientifiques.

La composition des jeux de données d'entraînement impacte l'existence de biais parmi les réponses des intelligences artificielles (Chu et al., 2024 ; Hannigan et al., 2024 ; Ferrara, 2023 ; Navigli & Conia, 2023). Navigli et Conia (2023) introduisent ainsi le « *biais de sélection de données* », qu'ils définissent comme « *le biais causé par le choix des textes qui composent un corpus d'entraînement* », en complément des biais sociaux (sexisme, âgisme, racisme...). Ce biais se produit lorsque « *les textes sont identifiés, ou lorsque les données sont filtrées et nettoyées* ». En lien avec ce biais de sélection de données, notre recherche nous permet d'enrichir 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. De son côté, Ferrara (2023) identifie sept 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. Les défauts des données d'entraînement ne sont pas irréversibles. Ils supposent cependant un travail fastidieux de rééquilibrage des *datasets*, c'est-à-dire de filtrage des données problématiques (Navigli & Conia, 2023). Cela peut d'ailleurs être vecteur de nouveaux biais (voir Dodge et al., 2021, par exemple, concernant l'introduction de biais démographiques lors du filtrage de contenus jugés grossiers).

Les politiques différenciées de blocage par les éditeurs scientifiques peuvent-elles 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 manuel, 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[13]. 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 remonte à 1991[14].

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

[13] Voir https://platform.openai.com/docs/models/gpt-4o.
[14] Voir https://en.wikipedia.org/wiki/ArXiv.

prédatrices). Dans les deux cas, le déséquilibre géographique est source de biais démographiques et de biais culturels, sans que l'impact soit facilement évaluable.

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., 2017). Il en résulte que les risques de mésinformation scientifique au sein des IAG varient probablement en fonction de la discipline.

Parmi les jeux de données à orientation scientifique, nous avons aussi vu que l'usage des *preprints* était répandu. Ce choix, notamment dictés par les facilités d'accès, est-il pénalisant du point de vue de la qualité des données collectées ? Dans le domaine de l'informatique, les pratiques de diffusion de recherches sous la forme de *preprints* sur arXiv a été étudiée par Lin et ses co-auteurs (2020). Leur recherche a nécessité un travail complexe de réconciliation des *preprints* et des versions publiées dans des actes de conférences ou des revues à comité de lecture. Leur recherche montre que près de 80 % des articles identifiés sont publiés dans un second temps. Les différences constatées concernent « *des révisions adéquates, des auteurs multiples, un résumé et une introduction détaillés, des références étendues et faisant autorité et un code source disponible* » (Lin et al., 2020). L'étude note cependant une tendance à la baisse du taux de publication des *preprints* (passé de 80 % à 75 % en quelques années). Ce haut taux de publication laisse cependant préjuger de la bonne qualité globale des recherches initiales. L'étude de Carneiro et ses co-auteurs (2020), appliquée à la littérature biomédicale (bioRxiv, PubMed), va dans le même sens. Les auteurs notent ainsi que la qualité des rapports est équivalente. Un léger avantage en faveur des versions publiées est cependant souligné. Même dans le contexte de la pandémie, propice à la mésinformation voire à la désinformation, les *preprints* sont apparus comme un allié précieux, par exemple pour la compréhension des mécanismes de transmission (Majumder & Mandl, 2020). Les *preprints* semblent donc ressortir comme un moyen efficace pour alimenter les jeux de données en informations scientifiques fiables et récentes.

Pour terminer, les problèmes de qualité des *datasets* peuvent également s'analyser, et être synthétisés, sous l'angle de la pertinence des représentations définie en management des systèmes d'information (Reix et al., 2011). Les données sont-elles, pour l'utilisateur, une représentation fiable de l'état de la connaissance à un instant donné ? Le premier problème est celui de l'accessibilité puisque certaines sources sont inaccessibles aussi bien pour les robots d'exploration collectant les données que pour les utilisateurs interagissant avec le *chatbot*. Surtout, ces *datasets* ne sont pas accessibles à des fins d'audit. Le second est celui de la fiabilité. Les données les plus facilement accessibles pour entraîner les intelligences artificielles génératives ne sont pas nécessairement les plus fiables comme le montrent les taux de blocage des éditeurs de presse ou des plateformes de l'édition scientifique. Cette question de la fiabilité est donc liée à celle de l'exactitude et de l'exhaustivité. La disponibilité accrue des contenus problématiques, couplée à la tendance naturelle aux hallucinations, conduit à des défauts d'exactitude. L'exhaustivité est impossible

dès lors qu'une partie de l'information échappe à l'utilisateur du fait de ces blocages. Le délai de constitution et de filtrage des jeux de données d'entraînement conduit à un défaut d'actualité. La mise à jour épisodique du modèle engendre par ailleurs un défaut de ponctualité. Dès lors, ces défauts de pertinence aboutissent à un ensemble d'erreurs et de biais dans les réponses.

Cette recherche souffre de cinq limitations. Premièrement, elle ne permet pas de nuancer 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 (p. ex. classement français FNEGE[15] 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 est relativement ancienne puisque sa mise à jour a été stoppée en 2017 (Richtig et al., 2018). Aussi serait-il intéressant de recalculer la mesure de biais global sur une liste de revues prédatrices, faisant consensus et surtout plus récente. Nous pensons par exemple aux listes d'éditeurs et de journaux publiées sur le site predatoryjournals.org. Quatrièmement, la méthode retenue analyse une partie seulement des données accessibles aux producteurs pour entraîner leurs LLM. En effet, à côté des jeux de données publics, les développeurs construisent des jeux de données internes, notamment alimentés par des données achetées auprès des éditeurs (Gibney, 2024). L'ampleur de ces acquisitions semble actuellement limitée (Kwon, 2024). Cependant, elle est difficile à estimer en pratique dès lors que les contractants communiquent peu ou prou sur les accords. Cet accès privilégié aux données limite donc le biais de sélection pour certains modèles entraînés par des *bigtechs* (Google, Microsoft, OpenAI...). Le problème demeure cependant entier pour des modèles entraînés par des acteurs plus modestes ne disposant pas de cet accès privilégié aux données. Cette situation illustre par ailleurs le manque de transparence sur les données utilisées, dès lors la difficulté d'évaluer les risques inhérents à l'utilisation d'un modèle. Cinquièmement, seul le blocage passif par protocole d'exclusion est pris en compte par nos mesures. Or, nous avons vu que les éditeurs disposaient d'une vaste panoplie de dispositifs de blocage. Cependant, cette approche nous semble fiable dès lors que l'*opt-out* représente un premier niveau de régulation, adapté aux robots éthiques, préalable à l'utilisation de méthodes plus sophistiquées.

6. Conclusion

L'étude met en lumière l'impact des restrictions d'accès appliquées par les éditeurs scientifiques aux robots d'exploration des intelligences artificielles

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

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 prédatrices dans les données d'entraînement des IAG, créant ainsi un biais de validation. S'il n'est pas corrigé, par exemple via l'acquisition de contenus sous licence, 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 modèles d'IAG. Notre recherche contribue ainsi à l'identification des biais dans les LLM ainsi qu'à leur mesure, leur compréhension et leur évitement (Chu et al., 2024 ; Ferrara, 2023 ; Navigli & Conia, 2023).

Cette recherche présente deux 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. 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 ?

7. Références

Amin Azad, B., Starov, O., Laperdrix, P., & Nikiforakis, N. (2020). Web runner 2049: Evaluating third-party anti-bot services. In Detection of Intrusions and Malware, and Vulnerability Assessment: 17th International Conference, DIMVA 2020, Lisbon, Portugal, June 24–26, 2020, Proceedings 17 (pp. 135-159). Springer International Publishing. https://doi.org/10.1007/978-3-030-52683-2_7.

Banks, M. (2016). What Sci-Hub is and why it matters. American Libraries, 47(6), 46-49. https://www.jstor.org/stable/26380679.

Beall, J. (2010). "Predatory" open-access scholarly publishers. The Charleston Advisor, 11(4), 10-17.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 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). Fake Science: Misconduct Galore and Proposed Counterattack. Doctoral. IPBS, France. 2024. https://ut3-toulouseinp.hal.science/hal-04129541/.

Cardon, D. (2013). Dans l'esprit du PageRank: une enquête sur l'algorithme de Google. Réseaux, (1), 63-95. https://shs.cairn.info/revue-reseaux-2013-1-page-63.

Carneiro, C. F., Queiroz, V. G., Moulin, T. C., Carvalho, C. A., Haas, C. B., Rayêe, D., ... & Amaral, O. B. (2020). Comparing quality of reporting between preprints and peer-

reviewed articles in the biomedical literature. Research Integrity and Peer Review, 5, 1-19. https://doi.org/10.1186/s41073-020-00101-3.

Chawla, D. S. (2017). Publishers take academic networking site to court. Science, vol. 358, issue 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, 26(1), 34-48. https://doi.org/10.1145/3682112.3682117.

Dinzinger, M., & Granitzer, M. (2024). A longitudinal study of content control mechanisms. In Companion Proceedings of the ACM on Web Conference 2024 (pp. 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 langage models. arXiv preprint arXiv:2304.03738. https://doi.org/10.5210/fm.v28i11.13346.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694. https://doi.org/10.1007/s11023-020-09548-1.

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.

Gershenson, S., Polikoff, M. S., & Wang, R. (2020). When paywall goes AWOL: The demand for open-access education research. Educational Researcher, 49(4), 254-261. https://doi.org/10.3102/0013189X20909834.

Gibney, E. (2024). Has your paper been used to train an AI model? Almost certainly. Nature, 632(8026), 715-716. https://doi.org/10.1038/d41586-024-02599-9.

Hannigan, T. R., McCarthy, I. P., & Spicer, A. (2024). Beware of botshit: How to manage the epistemic risks of generative chatbots. Business Horizons. https://doi.org/10.1016/j.bushor.2024.03.001.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., ... & Weld, D. S. (2023). The semantic scholar open data platform. arXiv preprint arXiv:2301.10140. https://doi.org/10.48550/arXiv.2301.10140.

Kleidermacher, H. C., & Zou, J. (2025). Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers. arXiv preprint arXiv:2502.17882. https://doi.org/10.48550/arXiv.2502.17882.

Kwon, D. (2024). Publishers are selling papers to train AIs-and making millions of dollars. Nature, 636(8043), 529-530. https://doi.org/10.1038/d41586-024-04018-5.

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. PloS one, 10(6), e0127502. https://doi.org/10.1371/journal.pone.0127502.

Lin, J., Yu, Y., Zhou, Y., Zhou, Z., & Shi, X. (2020). How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. Scientometrics, 124(1), 555-574. https://doi.org/10.1007/s11192-020-03430-8.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The semantic scholar open research corpus. arXiv preprint arXiv:1911.02782. https://doi.org/10.48550/arXiv.1911.02782.

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: a misnomer worth clarifying. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 133-138). IEEE. https://doi.org/10.1109/CAI59869.2024.00033.

Majumder, M. S., & Mandl, K. D. (2020). Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. The lancet global health, 8(5), e627-e630. https://doi.org/10.1016/S2214-109X%2820%2930113-3.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality, 15(2), 1-21. https://doi.org/10.1145/3597307.

Reix, R., Fallery, B., Kalika, M., & Rowe, F. (2011). Systèmes d'information et management des organisations. Vuibert. ISBN : 9782711743810.

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, 32(9), 1441-1449. https://doi.org/10.1111/jdv.15039.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), 243-246. ACM. https://doi.org/10.1145/2740908.2742839.

Sun, Y., Zhuang, Z., Councill, I. G., & Giles, C. L. (2007). Determining bias to search engines from robots. txt. In IEEE/WIC/ACM International Conference on Web Intelligence (WI'07) (pp. 149-155). IEEE. https://doi.org/10.1109/WI.2007.98.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008), 990-998. https://doi.org/10.1145/1401890.1402008.

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, 68(9), 2116-2127. https://doi.org/10.1002/asi.23687.

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.

Viseur, R. (2024). Analyse de l'impact des IA génératives sur la presse en ligne : anatomie d'un newsbot basé sur GPT. Actes des conférences AIM. Montpellier (France). https://aim.asso.fr/fr/publications/actes-conferences/id-1809-aim2024_557858.

Wang, K., Shen, Z., Huang, C., Wu, C. H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. Quantitative Science Studies, 1(1), 396-413. https://doi.org/10.1162/qss_a_00021.

Xia, J., Li, Y., & Situ, P. (2017). An overview of predatory journal publishing in Asia. Journal of East Asian Libraries, 2017(165), 4. https://scholarsarchive.byu.edu/jeal/vol2017/iss165/4.

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, 66(7), 1406-1417. https://doi.org/10.1002/asi.23265.

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.