Exploration des pratiques de régulation des IA génératives par le protocole d'exclusion des robots

Robert Viseur[1], Landelin Delcoucq[2]

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

2. Service MIT, FPMs, UMONS
9 rue de Houdain, B-7000 Mons, Belgique
landelin.delcoucq@umons.ac.be

RÉSUMÉ. L'année 2023 fut notamment celle de l'essor des IA génératives capables de produire des images (Stable Diffusion, Midjourney…) ou des textes (ChatGPT, Bard…) originaux. Ces nouveaux outils ont amené leur lot de polémiques. Parmi celle-ci, la question des droits d'auteurs des contenus utilisés pour l'entraînement de ces modèle a rapidement touché les scènes médiatiques puis judiciaires. Dans cette recherche exploratoire, nous avons utilisé un robot d'exploration pour analyser les fichiers « robots.txt » de plusieurs ensemble de sites web incluant le Top 100 Alexa, des sites de presse en ligne et des sites d'éditeurs scientifiques. L'objectif était d'analyser le recours à cette norme technique, soit le protocole d'exclusion des robots, pour traiter cette question de la violation de la propriété intellectuelle. Nos résultats montrent une forte utilisation des mesures de blocage par les sites vivant de la publication de contenus. Ils mettent cependant en évidence certaines incohérences dans les mesures de blocage, des limitations dans le protocole d'exclusion des robots et des biais (pour lesquels une nouvelle mesure est proposée) que les politiques de blocage différenciées risquent d'introduire lors de l'entraînement des IA génératives.

ABSTRACT. The year 2023 saw the rise of generative AI capable of producing original images (Stable Diffusion, Midjourney...) or texts (ChatGPT, Bard...). These new tools have brought with them their share of controversy. One of these has been the issue of copyright in the content used to train these models, which has rapidly made its way into the media and legal arena. In this exploratory study, we used a crawler to analyse the "robots.txt" files of several sets of websites, including the Alexa Top 100, online press sites and scientific publishers' sites. The aim was to analyse the use of this technical standard, the robot exclusion protocol, to address the issue of intellectual property infringement. Our results show a high level of use of blocking measures by sites publishing content. However, they highlight certain inconsistencies in the blocking measures, limitations in the bot exclusion protocol and the biases (for which a new measure is proposed) that differentiated blocking policies are likely to cause when training generative AIs.

Mots-clés : IA, biais, LLM, ChatGPT, robots d'exploration, propriété intellectuelle.

KEYWORDS: AI, bias, LLM, ChatGPT, crawler, intellectual property.

1. Introduction

L'intelligence artificielle peut être vue comme « *un artefact informatique construit grâce à l'intervention humaine, qui pense ou agit comme les humains, ou comme nous nous attendons à ce que les humains pensent ou agissent* » (Dignum, 2019). Elle couvre différentes approches techniques incluant le *machine learning*. Ce dernier a connu des progrès sensibles ces dernières années avec l'essor des réseaux de neurones profonds, ou *deep learning* (Heudin, 2016). Dopé par les ressources de calcul disponibles (GPU, GPU as a Service…), le *deep learning* a notamment permis d'améliorer les dispositifs de reconnaissance de forme dans les images (p. ex. réseaux de neurones convolutifs). Plus récemment, les intelligences artificielles basées sur le *deep learning* ont démontré des aptitudes à la créativité, d'abord avec les *Generative Adversial Networks* (GAN), puis avec les *Latent Diffusion Models* (LDM), pour la création d'images, et les *Large Language Models* (LLM), pour la création de textes (Goodfellow et al., 2014 ; Floridi & Chiriatti, 2020 ; Roombach et al., 2022). Ces nouveaux outils (Stable Diffusion, Midjourney, OpenAI DALL-E, OpenAI ChatGPT, Google Bard…) se sont, en moins de 2 ans, rapidement diffusés auprès du grand public. À côté de celle sur l'authenticité de cette créativité (Chomsky, 2023), une polémique relative à la propriété intellectuelle a rapidement émergé, notamment dans le secteur de la presse (McKenzie & Arvanitis, 2023). Plusieurs actions en justice sont ainsi en cours contre OpenAI[1]. Parmi les acteurs attentifs à la défense de leur propriété intellectuelle face à ces nouveaux entrants citons en particulier le New York Times (Weatherbed, 2023). En pratique, ces IA sont entraînées sur des volumes massifs de données. Ces dernières sont souvent collectées par des robots d'exploration sur le Web sans concertation préalable et explicite avec les gestionnaires des sites web. Certains jeux de données, comme LAION pour les images et Common Crawl pour le texte, sont ainsi maintenus et publiés par des organisations sans but lucratif. Ce conflit quant à la réutilisation de contenus publiés en ligne n'est pas radicalement nouveau. Il rappelle ainsi les conflits récurrents entre la presse en ligne et Google autour, notamment, de son service Google Actualités (Rebillard & Smyrnaios, 2010 ; Ouakrat, 2020 ; Galloway, 2018). Ce différend s'est réglé par un mélange de négociations individuelles, de régulation et de normes techniques incluant le protocole d'exclusion des robots (Sun et al., 2007). Ce dernier permet aux gestionnaires de sites de limiter l'activité des robots éthiques sur leurs sites. Nous proposons dans cette recherche exploratoire d'étudier comment les gestionnaires de sites web recourent au protocole d'exclusion des robots et d'en discuter l'efficacité. La suite de notre papier est découpée en quatre sections : la revue de la littérature, la présentation de la méthodologie et des données utilisées, les résultats puis leur discussion (en particulier sur le plan des risques d'introduction de biais).

2. Revue de la littérature

Cette section présente les intelligences artificielles génératives puis les questions de propriété intellectuelle qu'elles soulèvent. Elle se termine par une description du protocole d'exclusion des robots suivie d'une présentation des robots utilisés par les principales IA génératives.

[1] Voir https://originality.ai/blog/openai-chatgpt-lawsuit-list.

2.1. Intelligences artificielles génératives

ChatGPT, développé par OpenAI, est un agent conversationnel (*chatbot*) généraliste basé sur l'architecture GPT (*Generative Pre-trained Transformer*)[2]. Il se distingue par sa faculté à interpréter des directives formulées en langage naturel, appelées « *prompts* », et à engendrer des réponses textuelles cohérentes en adéquation avec ces directives. 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). En tant que *Large Langage Model* (LLM), il est capable de « *prédire statistiquement des séquences de mots* ». Ce modèle linguistique est formé par entraînement « *sur 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, filtré, représente ainsi 60 % des données d'entraînement de GPT-3 (Brown et al., 2020). ChatGPT est actuellement disponible en deux versions : ChatGPT (gratuit), basé sur GPT-3.5, et ChatGPT Plus (20 dollars / mois), donnant accès à GPT-4 ainsi qu'à des fonctionnalités complémentaires incluant l'accès au Web (navigation) ou l'accès à des extensions tierces (Hackett, 2023 ; OpenAI).

ChatGPT s'est distingué par son rythme de croissance extrêmement rapide. En deux mois, il atteignait ainsi les cent millions d'utilisateurs actif (Hu, 2023). D'autres agents conversationnels en mode SaaS sont cependant venus concurrencer ChatGPT. Ils incluent notamment Bard, rebaptisé Gemini début 2024, chez Google, et Claude, chez Anthropic (Singh et al., 2023). La concurrence s'est également développée du côté des LLMs. La jeune pousse Hugging Face s'est illustrée avec sa communauté éponyme dédiée à la diffusion de jeux de données et de modèles open-sources[3]. Deux modèles ont particulièrement fait l'actualité. Le premier est le modèle proposé par la jeune pousse française Mistral (Jiang et al., 2023). Le second est le modèle Llama publié par META (Touvron et al., 2023). L'utilisation locale de modèles open-sources s'est par ailleurs trouvée simplifiée par la publication de plateformes technologiques telles qu'Ollama.

2.2. Questions de propriété intellectuelle

Les craintes suscitées en matière de droit d'auteur par les IA génératives portent, d'une part, sur les réponses aux *prompts*, d'autre part, sur les informations utilisées, lors de la phase d'entraînement, pour créer le modèle. Les IA génératives telles que ChatGPT se distinguent en effet par leur « *capacité à générer des textes dans n'importe quelle langue, dans n'importe quel format et sur n'importe quel sujet en quelques secondes* » (Lucchi, 2023). La question se pose donc de savoir si ces réponses sont elles-mêmes soumises au droit d'auteur. Dès lors que les États-Unis et l'Union européenne imposent la présence d'un humain en tant que créateur de l'œuvre, les productions des IA génératives (soit des machines) ne sont pas protégées par un droit d'auteur (Lucchi, 2023 ; Zirpoli, 2023). Deux cas particuliers doivent cependant être distingués. Le premier concerne la production d'une réponse trop proche des données d'entraînement. Dans ce cas, la similitude peut conduire à la reconnaissance d'une contrefaçon de l'œuvre originale. Le second concerne

[2] Voir https://help.openai.com/en/articles/6783457-what-is-chatgpt.
[3] Voir https://huggingface.co/datasets et https://huggingface.co/models.

l'existence d'une éventuelle paternité dans le cas où l'utilisateur fournit des données (« *input* ») au sein du *prompt* ou conçoit une séquence de *prompts* (« *instruction* ») particulièrement élaborée (Lucchi, 2023). La reconnaissance liée au contrôle par les instructions tend actuellement à être rejetée ; au contraire, cette reconnaissance serait avérée dès lors que l'humain apporte des modifications suffisamment créatives (Zirpoli, 2023).

La création d'un LLM passe par l'entraînement du modèle sur base d'importantes quantités de données textuelles. Ces données sont notamment collectées sur le Web. En tant qu'œuvres de l'esprit, ces documents bénéficient, par le seul acte de création, sans dépôt ni formalité, d'une protection par le droit d'auteur (Mattatia, 2017 ; Binctin, 2022). Certains de ces contenus sont par ailleurs couverts par des contrats autorisant certaines réutilisations, comme la famille des licences Creative Commons (Mattatia, 2017). Le droit d'auteur comporte également des exceptions. La question se pose ainsi de savoir si l'utilisation par les producteurs d'IA générative de ces documents a priori protégés relève du « *fair use* » (dans les pays, comme les États-Unis, où cette notion est reconnue) ou d'autres exceptions (p. ex. fouille de textes et de données) au droit d'auteur (Lucchi, 2023). Ces incertitudes juridiques, couplées à l'absence de partage de revenus, ont conduit des créateurs de contenus, d'une part, à revoir leurs conditions d'utilisation (Weatherbed, 2023), d'autre part, à attaquer en justice des producteurs d'IA génératives tels qu'OpenAI (Lucchi, 2023).

Les modèles LLM tels que GPT sont affectés par la problématique des biais. Dans ce contexte, un bias est défini comme « *la présence de déformations systématiques, d'erreurs d'attribution ou de distorsions factuelles qui ont pour effet de favoriser certains groupes ou certaines idées, de perpétuer des stéréotypes ou de rendre plus difficile l'accès à l'information* » (Ferrara, 2023). Plusieurs facteurs expliquent ce phénomène. Citons les jeux de données, les algorithmes d'apprentissage, l'annotation humaine des données et les décisions réglementaires (Ferrara, 2023). Les LLM mettent en œuvre une technique d'apprentissage auto-supervisé (Kalyan, 2023). L'auto-supervision implique que le modèle apprend à partir de données non étiquetées en générant ses propres étiquettes à partir des données elles-mêmes. La qualité du modèle généré dépend cependant de celle des données elles-mêmes (Ferrara, 2023). Il en résulte un important travail de nettoyage (Dodge et al., 2021). L'intervention humaine est utile pour assainir le jeu de données utilisé pour l'entraînement ou analyser les éventuelles défaillances lors de la génération de textes. C'est notamment ce qui explique le recours à des « *travailleurs du clic* » pour identifier les propos haineux ou violents dans ces données (Douet, 2023 ; Casilli, 2019 ; Tubaro, Casilli & Coville, 2020). En fonction des données utilisées, les biais vont prendre différentes formes. Elles incluent les biais démographiques, les biais culturels, les biais linguistiques, les biais temporels, les biais de confirmation ainsi que les biais idéologiques et politiques (Ferrara, 2023).

2.3. Protocole d'exclusion des robots

La collecte de données sur le Web passe par l'utilisation de logiciels spécialisés appelés « *robots* ». Sur base d'un ensemble d'hyperliens fournis par un utilisateur, le robot va en sauvegarder le contenu, découvrir d'autres hyperliens et généralement continuer son exploration de manière récursive. Il peut être utilisé pour la création de collections, soit à des fins personnelles (p. ex. sauvegarde locale d'un site web à

l'aide de wget ou d'HTTrack), soit à des fins commerciales (p. ex. moteur de recherche : Googlebot, Bingbot...), pour l'archivage (p. ex. Internet Archive), pour la recherche personnelle ou pour la création de statistiques (p. ex. Netcraft). L'activité des robots d'exploration sur les sites web peut être régulée à l'aide de mesures actives, telles que la détection puis le blocage, ou passives, telles que le protocole d'exclusion des robots (Yang & Liao, 2010).

Le protocole d'exclusion des robots peut prendre deux formes. La première est un format de balise HTML, soit la balise META robots, permettant notamment d'indiquer si une page peut être indexée (« *index* ») ou non (« *noindex* »). La seconde est un fichier structuré nommé « *robots.txt* » placé à la racine du site web. Il documente les fichiers ou les répertoires accessibles (« *allow* ») ou non (« *disallow* ») aux robots ainsi que certains souhaits plus spécifiques comme le délai entre deux requêtes (Sun, Zhuang & Giles, 2007 ; Sun, 2008). Il présente l'allure suivante :

```
User-Agent: *
Allow: /news
User-Agent: Googlebot
Disallow: /bin
Disallow: /log
Disallow: /src
User-Agent: wget
Disallow: /
```

Ces conventions peuvent être utilisées pour exclure certains robots, soit que leur activité soit jugée nuisible aux performances du site web (p. ex. surcharge du serveur sans apport d'audience), soit que le *webmaster* bloque les robots afin d'éviter un usage non souhaité de contenus protégés par droit d'auteur (Yang & Liao, 2010). Autoriser explicitement l'accès aux robots d'un moteur de recherche pourrait donc être vu comme une licence implicite d'accéder à ces contenus (Yang & Liao, 2010). La syntaxe de ce protocole reste cependant ambiguë quant aux droits conférés, et les tentatives de clarification telles que l'*Automated Access Content Protocol* (ACAP), n'ont pas connu le succès (Sire, 2015).

2.4. Robots d'exploration des IA génératives

Dans le cas de ChatGPT, plusieurs robots interviennent dans le fonctionnement de l'outil[4]. Le premier, nommé GPTBot, annoncé par OpenAI en août 2023 (David, 2023), est le robot d'exploration d'OpenAI (« *web crawling* »). Il intervient dans la collecte de données utilisées pour entraîner l'IA. Il peut être identifié par le gestionnaire du site web à partir de son *user-agent* ou de ses adresses IP. Le second, nommé ChatGPT-User, est principalement utilisée pour répondre à une requête d'un utilisateur imposant à ChatGPT d'utiliser sa fonction de navigation (« *user browsing* »). Actuellement, OpenAI gère ces robots comme un seul robot : le blocage d'un des deux entraîne donc le blocage des deux robots. À ces deux robots opérés par OpenAI doit être ajouté CCBot. Ce dernier est le robot d'exploration, basé sur le moteur de recherche open-source Nutch, utilisé par Common Crawl, une fondation à but non lucratif fournissant une copie du Web à destination des chercheurs[5]. Le Common Crawl est notamment utilisé par les chercheurs en *machine*

[4] Voir https://platform.openai.com/docs/plugins/bot.
[5] Voir https://commoncrawl.org/faq.

learning pour entraîner leurs modèles. Dans le cas de Google Bard, le robot d'exploration est nommé Google-Extended[6].

Ces problématiques ne sont pas totalement nouvelles. La régulation du comportement des robots d'exploration s'est en effet déjà posée à l'occasion d'un différend judiciaire qui a opposé en Belgique Copiepresse, un organisme représentant notamment la presse francophone belge, et Google, opérant le service Google Actualités (Yang & Liao, 2010). Le moteur de recherche avait notamment publié le contenu de certains articles au sein de son cache sans que cet usage n'ait été accordé (via la balise META robots « *archive* ») par les gestionnaires des sites web concernés, ce qui constitue un acte de contrefaçon au sens du droit d'auteur en Europe et a entraîné le paiement d'astreintes[7] par l'entreprise étasunienne. Ce type de contentieux a pris une dimension européenne avec la nouvelle directive européenne sur les droits voisins (Ouakrat, 2020). Les droits d'auteur ont ainsi été complétés par des droits voisins au profit des éditeurs et des agences de presse. Elle permet, contre rémunération, la reproduction et la diffusion totale ou partielle des contenus par les plates-formes (Ouakrat, 2020). Des accords ont ainsi été signés en 2023 entre Google et des organismes de gestion collective des droits, tels que DVP en France ou Corint Media en Allemagne, après une série d'accords individuels (Agence France Presse, Le Monde...).

3. Méthodologie et données

Afin de mieux comprendre comment les gestionnaires de sites web abordent l'arrivée sur le marché des outils d'IA générative, nous avons conçu un robot d'exploration capable de lire le contenu des fichiers « *robots.txt* » de plusieurs ensembles de sites web incluant des journaux français (27), des journaux belges francophones (10) et néerlandophones (8), des éditeurs de journaux scientifiques (16) et les sites appartenant à la dernière version publiée (2022) du Top 100 Alexa. Pour chaque ensemble nous avons analysé les directives concernant tous les robots (*) puis certains robots spécifiques incluant Bingbot, Googlebot, Googlebot-News, Wget, HTTrack, Google-Extended, CCBot, GPTbot et ChatGPT-User.

Cette sélection de robots se justifie comme suit. Bingbot et Googlebot sont les robots d'exploration liés aux moteurs de recherche Google Web Search et Bing. Leur activité ne pose généralement pas de problèmes aux yeux des gestionnaires de sites web à la recherche de visibilité dans les moteurs de recherche généralistes. Googlebot-News est le robot d'exploration spécifique à Google Actualités. L'utilisation des informations collectées par ce robot a donné lieu à des conflits judiciaires entre la presse et Google. Il permet donc de voir si les webmasters lui appliquent un traitement particulier. Wget et HTTrack sont des outils permettant l'aspiration de sites. Cette activité n'est généralement pas appréciée des gestionnaires de sites web dès lors que ces outils occasionnent une sollicitation des serveurs sans réelle contrepartie sur le plan du trafic ou des revenus. Les quatre derniers robots (soit Google-Extended, CCBot, GPTbot et ChatGPT-User) sont des robots utilisés par des producteurs de jeux de données (CCBot) ou de modèles de type LLM. C'est principalement la politique de régulation de ces robots qui nous intéresse dans le cadre de cette recherche.

[6] Voir https://blog.google/technology/ai/an-update-on-web-publisher-controls/.
[7] Voir https://www.copiepresse.be/judiciaire.php?classement=03.

Notre script d'analyse récupère les fichiers « *robots.txt* » des ensembles de sites web puis calcule si chaque robot est ou non cité. Ensuite, le script regarde si chaque robot fait l'objet d'un blocage complet. Le blocage complet est généralement notifié par l'instruction « *disallow* » suivie de « / ». Plusieurs exécutions ont été réalisées : 01-12-2023 et 21-01-2024. Chaque exécution sauvegarde localement les fichiers « *robots.txt* » exploités et alimente un fichier journal permettant de visualiser les résultats. Cela permet notamment de détecter d'éventuels dysfonctionnements liés à des erreurs de programmation ou à des spécificités de formatage de certains fichiers « *robots.txt* ». Ce script a par la suite été modifié pour ajouter une mesure de biais.

4. Analyse des résultats

Si l'on se concentre sur les 10 robots les plus fréquents (cf. Tableau 1), GPTBot ressort comme le robot d'exploration le plus cité. Les robots d'exploration des producteurs d'IA génératives font d'ailleurs l'objet d'une attention particulière des gestionnaires de sites puisque quatre d'entre eux se retrouvent dans le Top 10. Les autres robots fréquemment cités sont ceux des deux moteurs de recherche dominants, des régies publicitaires de Google, de Twitter et de l'Internet Archive. On constate également une augmentation des citations liées aux robots d'IA génératives entre les deux extractions, ce qui montre une attention croissante des gestionnaires de sites web à l'activité de ces robots.

Tableau 1. Statistiques de mentions de robots d'exploration (01-12-2023 & 21-01-2024).

Robots	**Mentions (#)**		**Robots**	**Mentions (#)**	
*	137	137			
gptbot	34	39	bingbot	21	23
twitterbot	32	32	chatgpt-user	21	23
googlebot	29	29	ccbot	19	23
mediapartners-google	23	24	google-extended	19	23
adsbot-google	22	23	ia_archiver	17	17

Les gestionnaires des sites du Top 100 Alexa (cf. Tableau 2) se satisfont globalement de consignes globales fournies à l'aide du joker (« * »). Lorsqu'un robot d'exploration est cité, il s'agit le plus généralement d'un robot de moteur de recherche. GPTbot est cependant bloqué par 10,2 % des sites web (soit 29,5 % en ajoutant les blocages par défaut via le joker).

Tableau 2. Traitement des robots par les sites du Top 100 Alexa (21-01-2024).

	Robot	**Citations**	**Allowed/Total**	**Disallowed/ Robotstxt**
Site		99		
Pas de « robots.txt »		11		
	*	85	80,0 %	19,3 %
	Bingbot	16	93,8 %	1,1 %
	Googlebot	21	90,5 %	2,3 %

	Robot	Citations	Allowed/Total	Disallowed/ Robotstxt
	Googlebot-News	1	100,0 %	0,0 %
	Wget	1	0,0 %	1,1 %
	HTTrack	1	0,0 %	1,1 %
	Google-Extended	3	33,3 %	2,3 %
	CCBot	5	0,0 %	5,7 %
	GPTBot	9	0,0 %	10,2 %
	ChatGPT-User	1	0,0 %	1,1 %

L'inclusion des consignes spécifiques aux robots d'exploration des IA génératives est davantage développée chez les sites de journaux en ligne (cf. Tableau 3), qu'il s'agisse de journaux belges ou de journaux français. Nous trouvons ainsi 44,4 % des journaux francophones belges, 75 % des journaux néerlandophones belges et 40,7 % des journaux français qui procèdent au blocage de GPTBot. Ce dernier s'accompagne généralement du blocage de ChatGPT-User (mais il se révèle moins systématique). Le constat est similaire chez les sites des éditeurs scientifiques (avec une certaine tolérance à l'égard de ChatGPT-User).

Tableau 3. Traitement des robots par les sites de presse française (21-01-2024).

	Robot	Citations	Allowed/Citations	Disallowed/ Robotstxt
Site		27		
Pas de « robots.txt »		0		
	Google-Extended	6	16,7 %	18,5 %
	CCBot	8	0,0 %	29,6 %
	GPTBot	13	0,0 %	44,4 %
	ChatGPT-User	12	0,0 %	40,7 %

Cette politique spécifique aux producteurs de contenus est cohérente avec leur modèle d'affaires. D'une part, les éditeurs vivent de leurs contenus. Les IA génératives permettent de venir concurrencer ces contenus, grâce à leurs modèles entraînés sur ces contenus, sans qu'aucun partage de revenus n'ait été négocié. La situation de dépendance est sensiblement différente de celle vécue, par exemple, entre Google et les éditeurs de presse (Ouakrat, 2020 ; Rebillard & Smyrnaios, 2010). En effet, il existe une dépendance mutuelle : Google a besoin des contenus des éditeurs pour alimenter ses bases de données, les éditeurs de presse ont besoin de Google pour que les utilisateurs retrouvent le contenu souhaité parmi les milliards de pages constituant le Web. Dans le cas des IA génératives, la relation de dépendance est à l'avantage des éditeurs dès lors que les producteurs d'IA génératives ont impérativement besoin de contenus produits par des humains pour maintenir la qualité de leur outil (Loukides, 2023). De même, un robot tel que ChatGPT-User permet de déléguer l'analyse des contenus à l'agent conversationnel. Il prive donc le site web d'une partie de son audience. Or cette dernière influence directement les revenus publicitaires des sites de presse (Schiff, 2006).

Cette politique de blocage contribue, en plus des actions en justice, à l'établissement d'un rapport de force favorable vis-à-vis des producteurs d'IA génératives. En effet, le blocage entraîne différentes conséquences dommageables à leur proposition de valeur. Premièrement, les blocages par protocole d'exclusion des

robots sont connues pour entraîner un biais dans l'information relayée par les moteurs de recherche (Sun, Zhuang, Councill & Giles, 2007) ; suffisamment suivies, ces politiques contribueront également aux biais dans les réponses fournies par les IA génératives. Deuxièmement, elles dégradent le service associé à la version payante de ChatGPT puisque les traitements automatisés depuis le client ChatGPT Plus (p. ex. synthèses de documents) deviennent impossibles une fois bloqué un des deux robots d'OpenAI. Des accords individuels commencent d'ailleurs à être contractés (p. ex. Axel Springer[8]), comme jadis entre la presse et les moteurs de recherche (Ouakrat, 2020).

Par contre, il est surprenant de constater que d'autres robots intervenant dans la production de jeux de données d'entraînement soient moins cités dans les fichiers « *robots.txt* ». C'est en particulier le cas de CCBot et de Google-Extended (ce constat est surtout valable pour les sites de presse française). Le CCBot permet la constitution du Common Crawl, notamment utilisé par de nombreux producteurs de LLM (voir par exemple les *datasets* des modèles hébergés sur Hugging Face), tandis que Google-Extended est utilisé par Google pour l'alimentation de Google Bard. Cela signifie que bloquer GPTBot mais pas CCBot n'empêchera pas l'alimentation du jeu de données d'entraînement utilisé par GPT. Plusieurs explications pourraient être proposées. Dans le cas de Google-Extended, la presse tend (difficilement) à normaliser ses relations avec Google, d'abord via des accords individuels, puis des accords sectoriels suite aux récentes évolutions légales dans l'Union européenne. De plus, Google utilise l'IA, d'une part, pour l'agent conversationnel Google Bard, d'autre part, pour son moteur de recherche (via *Search Generative Experience*), ce qui peut conduire les gestionnaires de sites à plus de prudence dans leur politique de blocage sélective. Concernant CCBot, le moindre blocage pourrait s'expliquer, d'une part, par une tolérance liée à l'usage du CCBot dans le monde de la recherche (incluant des projets sans lien direct avec les LLM commerciaux), d'autre part, par un manque de connaissance des jeux de données réellement utilisés par les producteurs d'IA génératives.

Ces résultats corroborent l'étude antérieure réalisée par Originality.ai (2023). Celle-ci relève cependant un taux de blocage sensiblement plus élevé pour le GPTBot (25.9% sur un Top 1000 mondial). Cette différence peut s'expliquer par la composition différente de l'échantillon de sites puisque les pratiques de blocage semblent varier fortement d'une catégorie de sites à une autre. Elle confirme par contre l'augmentation du blocage au fil des semaines ainsi qu'un blocage plus important de GPTBot comparativement à CCBot, à Google-Extended et à anthoropic-ai (le robot d'exploration associé au *chatbot* Claude). Ces blocages sélectifs sont susceptibles, comme pour les moteurs de recherche (Sun, 2008) d'induire des biais au niveau des contenus générés. Parmi les sites bloquant ChatGPT nous y trouvons notamment Amazon, Quora, NYTimes, Shutterstock, Wikihow et CNN. Nous pouvons y ajouter Stackoverflow[9]. Parmi les motivations possibles à ces blocages citons la protection de contenus originaux soumis au droit d'auteurs (journaux, stocks photos, QA), potentiellement liée à la défense de la qualité de partenariats antérieurs (p. ex. accord New York Times – Google ; Weatherbed, 2023), la lutte contre la contamination par des contenus hallucinés (QA) et la préservation d'avantages concurrentiels (p. ex. Amazon Bedrock).

[8] Voir https://openai.com/blog/axel-springer-partnership.

[9] Voir https://meta.stackoverflow.com/questions/421831/temporary-policy-generative-ai-e-g-chatgpt-is-banned.

5. Estimation des biais

L'approche utilisée dans l'article de Sun, Zhuang, Councill et Giles (2007) a été initialement conçue pour mesurer le biais des robots des moteurs de recherche. Cette méthode s'est avérée utile non seulement pour cet objectif initial mais aussi pour une analyse plus générale des fichiers « *robots.txt* » de différents sites web. Grâce à cette analyse, il est possible de quantifier le biais et de comparer la (dé)favorisation entre les différents moteurs de recherche mais aussi, dans notre cas, entre les différents robots d'exploration des IA génératives. La méthode se concentre sur deux aspects : le premier est le calcul du biais absolu d'un robot sur un site web spécifique, mesurant ainsi directement son comportement. Le second aspect évalue le biais relatif d'un robot par rapport à d'autres, en examinant les interactions sur un éventail de sites étudiés. Cette approche permet d'obtenir une vision à la fois directe et comparative du comportement des sites par rapport aux robots.

Deux types de biais sont calculés : le biais local et le biais global. Le biais local fournit une information absolue sur la (dé)favorisation d'un robot sur un site en particulier. Un robot est dit favorisé s'il peut accéder à un plus grand nombre de répertoires que le robot universel et inversement. Le robot universel étant définit comme tout robot qui ne correspond à aucun des noms *User-Agent* spécifiques dans le fichier « *robots.txt* ». L'algorithme permettant de calculer le biais d'un robot en particulier a donc besoin de trois informations : l'ensemble des répertoires du site, l'ensemble des répertoires accessibles par le robot universel et l'ensemble des répertoires accessibles par ce robot en particulier. L'ensemble des répertoires du site n'est pas une donnée accessible de façon évidente. Sun et ses co-auteurs approximent dès lors cet ensemble comme l'ensemble des répertoires présents dans un fichier « *robots.txt* ».

$$Biais(x) = \frac{N_{fav}(x) - N_{defav}(x)}{N} \quad (1)$$

La mesure du biais global, de façon absolue, donne une information utile. Cependant, elle reste difficile à interpréter et est limitée à la comparaison de robots sur un même site. La mesure du biais globale, basée sur cette dernière, permet de donner un indicateur relatif, normalisé entre -1 (biais défavorable) et +1 (biais favorable) pour un robot sur un ensemble de sites. Ce biais global se formule, pour un robot « *x* », comme indiqué dans l'équation (1), dans laquelle *N* représente le nombre de sites considérés, $N_{fav}(x)$, le nombre de sites pour lequel le biais est positif et $N_{defav}(x)$, le nombre de sites pour lequel le biais local est négatif. Le résultat représente donc la différence des proportions entre les sites qui favorisent le robot sélectionné et ceux qui le défavorisent. La valeur représente le pourcentage, en valeur absolue, de sites (de l'échantillon) qui favorisent (signe positif) ou défavorisent (signe négatif) le robot.

La mesure du biais global pose cependant un problème, en ce sens que la mise en œuvre du protocole d'exclusion des robots dans notre échantillon de site s'appuie très souvent sur une politique de liste blanche (blocage par défaut du robot universel puis autorisation sélective) ou de liste noire (autorisation par défaut du robot universel puis blocage sélectif). Or le blocage de la totalité d'un site passe par le blocage du répertoire racine (« / »), soit un seul répertoire. L'implémentation de la mesure du biais local par Sun et ses co-auteurs conduit donc à considérer comme favorisé un site dont la racine serait bloquée face au robot universel dont seuls

quelques répertoires seraient bloqués. Dans notre cas, lorsque les robots des IA génératives sont cités, nous avons par ailleurs vu que cela était pour les bloquer. Nous pouvons donc utiliser l'algorithme simplifié suivant pour calculer directement le biais global des robots d'exploration des IA génératives (cf. Algorithme 1).

Algorithme 1. Calcul du biais global.

```
Si robot pas cité :
  biais = 0
Si robot cité :
  Si robot universel bloqué :
    Si robot avec autorisation :
      biais = 1
  Sinon:
    Si robot cité avec blocage :
      biais = -1
    Sinon :
      biais = 0
```

La généralisation de cet algorithme peut être envisagée en considérant les combinaisons possibles de configuration des accès pour le robot universel et pour le robot spécifique considéré (cf. Tableau 4). En première approximation, nous avons considéré équivalents les accès partiels différents pour deux robots (cellules grisées). Ces valeurs (biais défavorable, neutralité, biais favorable) peuvent ensuite être combinées avec l'équation (1) afin d'obtenir une estimation du biais global pour chaque robot. Cette algorithme a été mis en œuvre avec la langage Python. Après un calcul manuel sur les sites de presse française, un calcul automatisé a été réalisé pour l'ensemble des sites (cf. Tableau 5). En pratique, les cas avec autorisation partielle représentent moins de 5 % des entrées analysées dans les fichiers « *robots.txt* ».

Tableau 4. Détermination des biais locaux en fonction des modalités d'accès.

Robot spécifique : / **Robot universel :**	**Rien**	**Disallow (total)**	**Disallow (partiel)**	**Allow + Disallow**	**Allow (partiel)**	**Allow (total)**
Rien	0	-1	-1	-1	-1	0
Disallow (total)	0	0	1	1	1	1
Disallow (partiel)	0	-1	0	0	0	1
Allow + Disallow	0	-1	0	0	0	1
Allow (partiel)	0	-1	0	0	0	1
Allow (total)	0	-1	-1	-1	-1	0

Le calcul du biais global (cf. Tableau 5) concernant les robots d'exploration des IA génératives permet de mettre en évidence le traitement défavorable aux produits d'OpenAI (GPTBot et ChatGPT-User) comparativement à ceux de Google (Google Bard). Ce biais est même amplifié par le fait qu'OpenAI bloque les deux robots dès lors que l'un des deux robots est bloqué. Une fois cette règle intégrée, le biais pour GPTBot et ChatGPT-User s'élève en réalité à -0,41 (OpenAI) pour les sites de presse française. À titre de comparaison, les biais globaux pour Googlebot et Bingbot pour l'ensemble des sites s'élèvent respectivement à 0,07 et 0,05.

Tableau 5. Calcul du biais global (exécution du 21-01-2024).

Robot	Biais global (presse française)	Biais global (tous les sites)
Google-Extended	-0,15	0
CCBot	-0,15	-0,14
GPTBot	-0,37	-0,23
ChatGPT-User	-0,37	-0,13

Cette différence de traitement est susceptible d'accroître les biais chez toutes ou partie de ces IA génératives. Si l'on s'appuie sur la typologie proposée par Ferrara (2023), plusieurs risques peuvent ainsi être identifiés. Premièrement, nous pouvons voir que certains contenus qualitatifs, tels que les articles de presse ou les articles scientifiques, étaient fréquemment bloqués. Si l'on admet que ces contenus, soumis à des règles déontologiques et à des processus de relecture, sont en moyenne moins sujets à des biais démographiques, culturels ou idéologiques, cela pourrait aboutir à des modèles davantage biaisés. Deuxièmement, au vu des différences de traitement entre agents, certaines IA pourraient se révéler davantage biaisées que d'autres. Nous constatons par ailleurs une corrélation inversée entre la popularité de l'outil associé au robot et le traitement favorable qui lui est réservé. Si Sun et al. (2007) constatent que les robots d'exploration des moteurs de recherche les plus populaires (en termes de parts de marché) sont les plus favorisés, l'inverse se produit avec les IA génératives, où l'outil le plus populaire (ChatGPT) fait l'objet du traitement le plus défavorable. Cela entraîne que l'outil le plus utilisé est aussi celui qui est le plus menacé par les biais liés aux déséquilibres dans les données d'entraînement. Troisièmement, les politiques de blocage ne sont pas nécessairement homogènes dans toutes les zones géographiques. Ces disparités pourraient conduire à des biais linguistiques, en particulier pour des langues déjà sous-représentées dans les jeux de données. Quatrièmement, le blocage plus important des robots d'OpenAI est partiellement compensé par l'accès plus large réservé à CCBot. OpenAI conserve dès lors un accès aux données via le Common Crawl. Par contre, cette différence handicape la mise à jour des données par OpenAI, ce qui peut introduire des biais temporels du fait des difficultés d'accès à certaines informations récentes.

Tableau 6. Biais global par orientation politique et par robot.

Robot	Extrême gauche	Gauche	Centre gauche	Centre	Centre droit	Droite	Extrême droite	Complo-tiste
CCBot	0,000	-0,313	-0,320	-0,212	-0,439	-0,214	0,000	0,000
GPTBot	-0,222	-0,375	-0,420	-0,308	-0,512	-0,500	0,000	-0.091

Pour développer ces éléments de discussion, nous avons généré avec ChatGPT (sous GPT-4) une base de données de sites de presse, pour chaque pays de l'Union européenne et le Royaume-Uni. Pour chaque site, nous disposons du nom, de l'URL, du pays, de la langue (codes ISO) et de l'orientation politique (suivant les catégories suivantes : extrême gauche, gauche, centre gauche, centre, centre droit, droite, extrême-droite). Notons que, par défaut, ChatGPT n'inclut jamais de site d'extrême-gauche ou d'extrême-droite dans ses réponses. Un *prompt* spécifique a donc été utilisé pour générer ces deux listes. ChatGPT refusant de créer une liste de journaux

d'extrême-droite puis générant finalement, après justification de l'usage de ces données (*sic*), une telle liste en changeant l'orientation (droite conservatrice), nous avons finalement créé une liste de site de journaux étiquetés à l'extrême-droite en recourant à l'agent conversationnel Le Chat proposé par Mistral. Les catégorisations ont été contrôlées par un échantillonage aléatoire. Au final, un ensemble de 190 fichiers robots (sur 206 sites listés) ont pu être explorés. Pour chaque site, et pour 6 robots (ccbot, gptbot, chatpt-user, google-extended, googlebot, bingbot), les biais locaux ont ensuite été calculés à l'aide du script Python (-1, 0 ou +1 ; cf. Tableau 4). Ces données calculées ont ensuite été traitées dans LibreOffice.org Calc pour estimer le biais global par pays, par langue et par orientation politique (cf. Tableau 6). Les valeurs de biais global permettent notamment de constater un moindre blocage du côté des journaux aux positionnements extrêmes (cf. Tableau 6). Les blocages varient également fortement par pays (biais global fortement défavorable, soit inférieur ou égal à -0,5, pour le Common Crawl concernant l'Allemagne, la Belgique, le Danemark, la Finlande, le Luxembourg et les Pays-Bas) et par langue (biais global fortement défavorable pour le néerlandais, le danois et le finnois). Cette exploration relative à des sources de données qualitatives (presse) permet donc de mettre en évidence les biais culturels, linguistiques et politiques induits par ces pratiques de blocage sélectif. Notons que le biais global est généralement davantage marqué pour GPT (gptbot) et nettement moins pour Bard/Gemini (google-extended). Un test supplémentaire sur 22 fichiers robots de sites complotistes proposés par Mistral a également révélé un biais global très faible (cf. Tableau 6).

La stimulation des biais par le recours différencié au protocole d'exclusion des robots pourrait par ailleurs conduire à un effet boule de neige. En effet, les LLM (*Large Language Model*) tels que GPT sont de plus en plus souvent utilisés comme générateurs de données pour l'entraînement, soit de SLM (*Small Language Model*) (Eldan & Li, 2023), soit de systèmes de *machine learning* plus classiques capables de réaliser, par exemple, des tâches de classification (Yu et al., 2023 ; Ye et al., 2022). Eldan et Li (2023) réussissent ainsi à développer un SLM (TinyStories) capable de rivaliser, sur le plan des capacités de génération de texte, avec un LLM (GPT-2) grâce à un effort sur la conception du jeu de données généré. La méthode s'appuie, d'une part, sur des *prompts* permettant de préserver les éléments essentiels du langage naturel dans le jeu de données généré, d'autre part, sur un *framework* basé sur GPT (GPT-4) permettant l'évaluation des textes générés par le SLM. Ye et ses co-auteurs (2022) montrent la possibilité de générer un jeu de données utilisable pour réaliser des tâches d'analyse de sentiments. Yu et ses co-auteurs (2023) font de même pour des tâches de classification. Ils montrent par ailleurs l'importance de la structure du *prompt* pour éviter de limiter la diversité des données générées et propager « *les biais systématiques hérités des LLM* ».

6. Conclusion

Cette recherche exploratoire a permis d'identifier les enjeux associés aux IA génératives de type *text-to-text* en matière de propriété intellectuelle puis d'analyser les stratégies de régulation mises en place par les gestionnaires de sites web. Leur conséquence a par ailleurs fait l'objet d'une discussion. Sept constats ont ainsi pu être dégagés.

Premièrement, et comparativement à des études antérieures telles que Sun (2008), l'utilisation du protocole d'exclusion des robots semble aujourd'hui

pratiquement généralisée. Deuxièmement, les gestionnaires de sites web se sont montrés réactifs face à l'émergence des IA génératives en intégrant des directives spécifiques aux robots d'exploration des IA génératives dans leurs fichiers d'exclusion de robots. Troisièmement, si l'on prend le Top 100 Alexa comme un étalon des pratiques courantes en matière d'exclusion des robots, l'activité des robots d'exploration des IA génératives apparaît peu ou prou problématique aux yeux des gestionnaires des sites web. Quatrièmement, les sites d'éditeurs de contenus (édition scientifique et, plus encore, presse) bloquent fréquemment les robots d'exploration des IA génératives, en particulier ceux clairement identifiés comme œuvrant à la création de vastes corpus d'entraînement. Cinquièmement, les stratégies de blocage de ces éditeurs se focalisent parfois trop sur OpenAI et négligent les outils de collecte associés à d'autres producteurs d'IA générative ou, plus grave, de jeux de données ouvertes utilisés par ces producteurs, ce qui réduit leur efficacité dans la constitution d'un rapport de force favorable vis-à-vis des producteurs d'IA génératives. Sixièmement, la réaction des gestionnaires de sites face aux IA génératives remet en lumière les limitations du protocole d'exclusion des robots et, d'une part, l'intérêt d'une désambiguïsation des consignes d'accès aux contenus en ligne (Yang & Liao, 2010 ; Sire, 2015), d'autre part, l'utilité de régulations spécifiques. Septièmement, les différences de politiques lors de l'exclusion de certains robots d'IA générative entraîne un risque accru de biais lors de l'entraînement de ces dernières.

Cette recherche ouvre la perspective d'une étude plus ambitieuse des biais potentiellement introduits par les pratiques de blocage des robots d'exploration des producteurs de LLM. Celle-ci passe par l'utilisation d'un annuaire de sites web, plus volumineux et diversifié, permettant d'analyser de manière systématique les biais induits, par exemple sur les plans culturels (p. ex. blocage différencié par pays ou par région) ou linguistiques (p. ex. blocage différencié par langue).

7. Références

Binctin, N. (2022). Droit de la propriété intellectuelle, droit d'auteur, brevet, droits voisin, marque, dessins et modèles. LGDJ. ISBN : 978-2275108339.

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.

Casilli, A. (2019). En attendant les robots. Enquête sur le travail du clic. Seuil. ISBN : 978-2-0214-0188-2.

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The False Promise of ChatGPT. The New York Times, 8. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

David, E. (2023). Now you can block OpenAI's web crawler. The Verge, 7 août 2023. https://www.theverge.com/2023/8/7/23823046/openai-data-scrape-block-ai.

Dignum, V. (2019). Responsible artificial intelligence: how to develop and use AI in a responsible way (Vol. 2156). Cham: Springer. ISBN : 978-3-030-30373-0.

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.

Douet, M. (2023). Au Kenya, des « entraîneurs » de ChatGPT s'élèvent contre leurs conditions de travail. Le Monde, 19 octobre 2023. https://www.lemonde.fr/afrique/article/2023/10/19/au-kenya-des-entraineurs-de-chatgpt-s-elevent-contre-leurs-conditions-de-travail_6195464_3212.html.

Eldan, R., & Li, Y. (2023). TinyStories: How Small Can Language Models Be and Still Speak Coherent English?. arXiv preprint arXiv:2305.07759. https://doi.org/10.48550/arXiv.2305.07759.

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. First Monday, Vol.28, N°11. 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.

Galloway, S. (2018). The four - Le règne des quatre : la face cachée d'Amazon, Apple, Facebook et Google. Quanto. ISBN : 978-2889152469.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27. https://doi.org/10.48550/arXiv.1406.2661.

Hackett, C. (2023). OpenAI's ChatGPT plus: an electronic resources librarian's review. Journal of Electronic Resources Librarianship, 35(4), 299-304. https://doi.org/10.1080/1941126X.2023.2271373.

Heudin, J.-C. (2016). Comprendre le deep learning: Une introduction aux réseaux de neurones. Science-e-Book. ISBN : 979-1091245449.

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/.

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.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. arXiv preprint. https://doi.org/10.48550/arXiv.2310.06825.

Kalyan, K. S. (2023). A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal, 100048. https://doi.org/10.1016/j.nlp.2023.100048.

Layton, D. (2023). ChatGPT — Show me the Data Sources. Medium. https://medium.com/@dlaytonj2/chatgpt-show-me-the-data-sources-11e9433d57e8.

Loukides, M. (2023). Model Collapse: An Experiment. O'Reilly, 24 octobre 2023. https://www.oreilly.com/radar/model-collapse-an-experiment/.

Lucchi, N. (2023). ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. European Journal of Risk Regulation, 1-23. https://doi.org/10.1017/err.2023.59.

Mattatia, F. (2017). Droit d'auteur & propriété intellectuelle dans le numérique. Eyrolles. ISBN : 978-2-416-00813-9.

McKenzie, S. & Arvanitis, L. (2023). Les "newsbots" montent au front : des sites d'actualité générés par l'IA se multiplient en ligne. NewsGuard, 01 mai 2023. https://www.newsguardtech.com/fr/special-reports/bots-ia-generative-sites/.

Originality.ai (2023). Websites That Have Blocked OpenAI's GPTBot CCBot Anthropic Google Extended - 1000 Website Study. https://originality.ai/ai-bot-blocking.

Ouakrat, A. (2020). Négocier la dépendance? Google, la presse et le droit voisin. Sur le journalisme, 9(1), 44-57. https://doi.org/10.25200/SLJ.v9.n1.2020.417.

Rebillard, F., & Smyrnaios, N. (2010). Les infomédiaires, au coeur de la filière de l'information en ligne: les cas de Google, Wikio et Paperblog. Réseaux, (2), 163-194. https://doi.org/10.3917/res.160.0163.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695). https://doi.org/10.48550/arXiv.2112.10752.

Schiff, F. (2003). Business models of news Web sites: A survey of empirical trends and expert opinion. First Monday. https://doi.org/10.5210/fm.v8i6.1061.

Singh, S. K., Kumar, S., & Mehra, P. S. (2023). Chat GPT & Google Bard AI: A Review. In 2023 International Conference on IoT, Communication and Automation Technology (ICICAT) (pp. 1-6). IEEE. https://doi.org/10.1109/ICICAT57735.2023.10263706.

Sire, G. (2015). Inclusion exclue: le code est un contrat léonin: Enquête sur la valeur technique et juridique du protocole robots. txt. Réseaux, (1), 187-214. https://doi.org/10.3917/res.189.0187.

Sun, Y. (2008). A comprehensive study of the regulation and behavior of web crawlers. Doctoral dissertation, The Pennsylvania State University. https://www.proquest.com/openview/22e8942f3aa45b2c043dba62f33ef3a1/1.

Sun, Y., Zhuang, Z., & Giles, C. L. (2007). A large-scale study of robots. txt. In Proceedings of the 16th international conference on World Wide Web (pp. 1123-1124). https://doi.org/10.1145/1242572.1242726.

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.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971.

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. Big Data & Society, 7(1), 2053951720919776. https://doi.org/10.1177/2053951720919776.

Weatherbed, J. (2023). The New York Times prohibits using its content to train AI models. The Verge, 14 avril 2023. https://www.theverge.com/2023/8/14/23831109/the-new-york-times-ai-web-scraping-rules-terms-of-service.

Yang, C., & Liao, H. J. (2010). Using the Robots. txt and Robots Meta tags to implement online copyright and a related amendment. Library hi tech, 28(1), 94-106. http://dx.doi.org/10.1108/07378831011026715.

Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., ... & Kong, L. (2022). Zerogen: Efficient zero-shot learning via dataset generation. arXiv preprint arXiv:2202.07922. https://doi.org/10.48550/arXiv.2202.07922.

Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A., Krishna, R., ... & Zhang, C. (2023). Large language model as attributed training data generator: A tale of diversity and bias. arXiv preprint arXiv:2306.15895. https://doi.org/10.48550/arXiv.2306.15895.

Zirpoli, C. T. (2023). Generative Artificial Intelligence and Copyright Law. Congressional Research Service. https://crsreports.congress.gov/product/pdf/LSB/LSB10922.