

# Régulation des IA génératives par les gestionnaires de sites web : une étude exploratoire

*Robert Viseur, FWEG, Université de Mons, Belgique*

## Résumé :

L'année 2023 fut notamment celle de l'essor des IA génératives capables de générer des images (Stable Diffusion, Midjourney...) ou des textes (ChatGPT, Bard...) originaux. Ces nouveaux outils ont amené leur lot de polémiques. Parmi celles-ci, la question des droits d'auteur des contenus utilisés pour l'entraînement de ces modèles 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 ensembles 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 ainsi que les limitations actuelles du protocole d'exclusion des robots.

## Mots clés :

intelligence artificielle ; LLM ; ChatGPT ; robots d'exploration ; propriété intellectuelle

## Regulation of generative AI by web site managers: an exploratory study

## Abstract :

The year 2023 saw the rise of generative AI capable of generating original images (Stable Diffusion, Midjourney...) or texts (ChatGPT, Bard...). These new tools have brought 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 also highlight certain inconsistencies in the blocking measures and the current limitations of the robot exclusion protocol.

## Keywords :

artificial intelligence ; LLM ; ChatGPT ; search engine crawler ; intellectual property

# Régulation des IA génératives par les gestionnaires de sites web : une étude exploratoire

## 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 formes 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 Adversarial 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. 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 & Smyrniotis, 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 : une revue de la littérature, la présentation de la méthodologie et des données utilisées, les résultats et leur discussion.

## 2. Revue de la littérature

[ChatGPT](#), développé par OpenAI, est un agent conversationnel (*chatbot*) généraliste basé sur l'architecture GPT (*Generative Pre-trained Transformer*)<sup>1</sup>. Il se distingue par sa faculté à interpréter des directives formulées en langage naturel, appelées « *prompts* », et à engendrer

---

1 Cf. <https://help.openai.com/en/articles/6783457-what-is-chatgpt>.

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 Language 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* ».

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. 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](#)). D'autres agents conversationnels (p. ex. [Google Bard](#) renommé Gemini) et d'autres modèles de type LLM (p. ex. [Llama](#) et [Mistral](#)) font aujourd'hui concurrence aux produits d'OpenAI (Jiang et al., 2023 ; Singh et al., 2023).

Les IA génératives telles que ChatGPT se distinguent donc 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 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 d'une paternité 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 de savoir si l'utilisation par les producteurs d'IA générative de ces documents a priori protégés relève du « *fair use* » (dans les pays, tels que 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<sup>2</sup> des producteurs d'IA génératives tels qu'OpenAI (Lucchi, 2023).

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). Ce dernier peut prendre deux formes. La première 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). La seconde 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* »). 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). Sa syntaxe 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).

Dans le cas de ChatGPT, plusieurs robots interviennent dans le fonctionnement de l'outil<sup>3</sup>. 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<sup>4</sup>. Le Common Crawl est

---

2 Voir <https://originality.ai/blog/openai-chatgpt-lawsuit-list> pour une liste des actions en justice intentées contre OpenAI. 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).

3 Voir <https://platform.openai.com/docs/plugins/bot>.

4 Voir <https://commoncrawl.org/faq>.

notamment utilisé par les chercheurs en *machine learning* pour entraîner leurs modèles. Dans le cas de Google Bard, le robot d'exploration est nommé Google-Extended<sup>5</sup>.

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. 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 plateformes (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.

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.

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

---

5 Voir <https://blog.google/technology/ai/an-update-on-web-publisher-controls/>.

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 (« *log* ») 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* ».

#### 4. Analyse des résultats

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

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

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.

	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 %
	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 %

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

Les gestionnaires des sites du Top 100 Alexa 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).

	Robot	Citations	Allowed/Citations	Disallowed/Robotstxt
Site		27		

	Robot	Citations	Allowed/Citations	Disallowed/Robotstxt
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 %

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

L'inclusion des consignes spécifiques aux robots d'exploration des IA génératives est davantage développée chez les sites de journaux en ligne, 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).

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 & Smyrniotis, 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 connus 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<sup>6</sup>), 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

6 Cf. <https://openai.com/blog/axel-springer-partnership>.

sites de presse français). Le CCBot permet la constitution du CommonCrawl, 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 anthropic-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<sup>7</sup>. Parmi les motivations possibles à ces blocages citons la protection de contenus originaux protégés par droit d'auteur (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](#)).

## 5. 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. Six 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

---

<sup>7</sup> Voir <https://meta.stackoverflow.com/questions/421831/temporary-policy-generative-ai-e-g-chatgpt-is-banned>.

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 ouverts 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.

## 6. Références

- Binctin, N. (2022). Droit de la propriété intellectuelle, droit d'auteur, brevet, droits voisin, marque, dessins et modèles. LGDJ. ISBN : 978-2275108339.
- 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.
- 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.
- 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>.
- 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.
- 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., & Smyrniaios, 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>.
- 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>.
- Zirpoli, C. T. (2023). *Generative Artificial Intelligence and Copyright Law*. Congressional Research Service. <https://crsreports.congress.gov/product/pdf/LSB/LSB10922>.