

ÉTUDE DE RECHERCHE PPN WORLD N° 1 · ÉDITION 1

# 94 % des communiqués de presse sont ouverts aux robots d'IA

Un recensement préenregistré des règles visant les robots d'IA sur 1 503 sites de communiqués



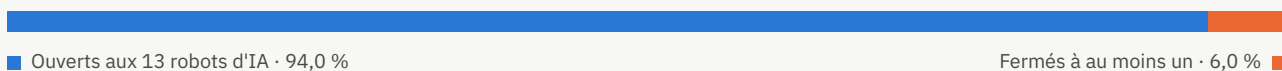
Par **Émile Bolduc** · PPN World  
Publiée le 8 octobre 2026

Nous avons lu le fichier robots.txt derrière chacun des 500 129 communiqués de presse. 94,0 % d'entre eux sont ouverts à tous les robots d'IA testés. Quand un éditeur refuse, il refuse d'entraîner l'IA — pas d'être cité par elle.

Édition 1 · Fichiers lus le 7 octobre 2026 · Publiée le 8 octobre 2026 · Préenregistrée  
<https://www.ppnworld.com/fr/research/press-releases-ai-crawlers> · Textes, figures et données : CC BY 4.0

## 94,0 %

des 500 129 communiqués de presse sont ouverts à tous les robots d'IA testés



500 129

COMMUNIQUÉS VÉRIFIÉS

1 503

SITES

13

ROBOTS D'IA TESTÉS



PRÉENREGISTRÉE

## RÉSUMÉ

Les assistants et les moteurs de réponse reprennent désormais les communiqués de presse, ce qui pose une question pratique à ceux qui les publient : ces communiqués sont-ils ouverts aux robots d'IA ? Avant de lire le moindre fichier, nous avons enregistré trois hypothèses. Nous avons ensuite lu le fichier robots.txt de chaque site derrière les 500 129 communiqués de l'archive de PPN World du 1er juillet 2026 au 6 octobre 2026 — 1 503 sites d'entreprises, de diffuseurs, de gouvernements, de villes, d'universités et de ligues — et appliqué chaque fichier, comme le prévoit la RFC 9309, à l'adresse de chaque communiqué, pour 13 robots d'IA ainsi que pour ceux de Google et de Bing. 94,0 % des communiqués se trouvent sur des sites qui laissent entrer chacun de ces robots d'IA, et 99,0 % sont ouverts aux trois robots qui alimentent les réponses de recherche par IA. Les refus sont rares et visent l'entraînement : 3,3 % des sites écartent le robot d'entraînement d'OpenAI, contre 0,6 % pour Googlebot (H1), et 28 sites refusent ce robot d'entraînement tout en admettant le robot de recherche d'OpenAI, et aucun ne fait l'inverse (H2). Organismes publics et éditeurs commerciaux ne diffèrent pas (3,8 % contre 4,2 % ; H3 non confirmée), et les copies archivées montrent le même portrait douze mois plus tôt. Fin 2023, près de la moitié des sites d'information les plus lus bloquaient déjà les robots d'OpenAI (Fletcher, 2024) ; les communiqués, eux, sont restés ouverts.

## SOMMAIRE

- 01 Principaux résultats    02 Pourquoi le mesurer    03 Comment nous avons mesuré
- 04 Presque tous les communiqués sont ouverts    05 L'entraînement refusé, la réponse admise    06 Qui refuse    07 Un an plus tôt
- 08 L'autre barrière : le pare-feu    09 Ce que cela signifie pour les communicants    10 Notes de méthode
- 11 Données, préenregistrement et citation

## 01 Principaux résultats

01

# 94,0 %

### 94 % des communiqués sont ouverts à tous les robots d'IA

73 des 1 443 sites qui ont répondu (5,1 %) refusent au moins un des 13 robots d'IA testés ; pondéré par le nombre de communiqués, 6,0 %. Les cinq plus grands sites qui ont répondu — 42,1 % de ces communiqués — n'en refusent aucun.

02

# 99,0 %

### La recherche par IA est encore plus ouverte

Seuls 34 sites (2,4 %) écartent ne serait-ce qu'un des robots qui alimentent les réponses de recherche par IA — ceux d'OpenAI, d'Anthropic ou de Perplexity. 99,0 % des communiqués sont ouverts aux trois.

03

# 3,3 % contre 0,6 %

### Les robots d'IA sont plus souvent refusés que Google (H1)

48 sites (3,3 %) refusent GPTBot ; 9 (0,6 %) refusent Googlebot. 39 refusent GPTBot et admettent Googlebot ; aucun ne fait l'inverse ( $p < 0,001$ ). Attendu, et désormais répliqué à pleine échelle.

04

# 28 contre 0

### Les éditeurs refusent l'entraînement, pas la réponse (H2)

28 sites refusent GPTBot, le robot d'entraînement d'OpenAI, tout en admettant OAI-SearchBot, qui alimente la recherche de ChatGPT. Aucun ne fait l'inverse ( $p < 0,001$ ). Tout site qui écarte un robot d'IA écarte un robot d'entraînement.

05

# 3,8 % contre 4,2 %

### Gouvernements et entreprises se comportent de la même façon (H3, non confirmée)

3,8 % des sites d'organismes publics et 4,2 % des sites commerciaux refusent GPTBot (écart de  $-0,3$  point, IC à 95 %  $-3,4$  à  $+2,0$  points). Les diffuseurs le refusent plus souvent que les salles de presse des entreprises (8,3 % contre 2,0 %), mais c'est rare des deux côtés.

06

# 4,3 % → 4,7 %

### Un an n'a rien changé

Comparés aux copies archivées des mêmes fichiers à l'automne 2025 (1 152 sites), les blocages visant l'IA passent de 4,3 % à 4,7 % des sites : 21 ont commencé, 17 ont cessé ( $p 0,63$ ).

07

# 75 sites

### Le pare-feu est l'autre barrière

75 sites (5,0 %) ont refusé notre client automatisé sur la page du communiqué elle-même, même lorsqu'il se présentait comme un navigateur — à peu près autant que ceux qui refusent un robot d'IA dans le robots.txt. Qu'un vrai robot d'IA franchisse ces pare-feu ne se voit pas de l'extérieur.

## 02 Pourquoi le mesurer

Un communiqué de presse est écrit pour être repris. En 2026, la reprise est de plus en plus le fait de l'IA : des assistants qui répondent aux questions, des moteurs qui résument, et les modèles derrière eux. Qu'un communiqué puisse les atteindre dépend d'abord d'un simple fichier texte, le robots.txt, par lequel un site indique à chaque robot où il peut aller. Le fichier distingue les robots par leur nom : un éditeur peut refuser celui qui collecte des données d'entraînement et admettre celui qui va chercher une page pour répondre à une question.

Les éditeurs de presse ont tranché tôt et ouvertement : fin 2023, 48 % des sites d'information les plus utilisés dans dix pays bloquaient les robots d'OpenAI (Fletcher, 2024), et les restrictions visant les robots d'IA se sont vite répandues parmi les sources les plus utilisées du web (Longpre et al., 2024). Les communiqués, eux, ont fait l'objet de débats plutôt que de mesures : on entend à la fois que les réponses de l'IA citent désormais les communiqués et que les éditeurs ferment la porte à l'IA. Cette étude le mesure, pour une population complète de sites de communiqués, avec une analyse enregistrée avant la lecture du moindre fichier.

## 03 Comment nous avons mesuré

La population comprend chaque communiqué de l'archive de PPN World daté du 1er juillet 2026 au 6 octobre 2026, issu des sources que l'archive diffuse : 500 129 communiqués avec un lien utilisable, sur 1 503 sites. Un site est ici une origine — schéma, hôte et port — puisque c'est ce que régit un fichier robots.txt. Chaque site prend le type d'éditeur des sources qui y renvoient : 781 relèvent de gouvernements et d'organismes publics (nationaux, régionaux et locaux) et 346 d'entreprises et de diffuseurs commerciaux.

Le 7 octobre 2026, nous avons demandé le robots.txt de chaque site en nous identifiant comme client de recherche, puis une seconde fois comme navigateur en cas de refus. Chaque fichier a été appliqué exactement comme le prévoit la RFC 9309 (Koster et al., 2022) — le groupe qui nomme le robot, sinon le groupe \* ; la règle la plus longue l'emporte — à l'adresse de chaque communiqué du site, et pas seulement à sa page d'accueil, car un site peut fermer un répertoire et laisser le reste ouvert. Un site refuse un robot quand plus de la moitié de ses communiqués lui sont fermés. Les 60 sites dont le fichier était inaccessible sont exclus des taux et comptés comme des refus dans un contrôle de robustesse.

Nous avons testé 13 robots d'IA, répartis d'avance en trois familles d'après la documentation de chaque exploitant (Anthropic, s.d. ; Apple, s.d. ; Common Crawl, s.d. ; Google, s.d. ; Meta, s.d. ; OpenAI, s.d. ; Perplexity, s.d.) : ceux qui collectent des données d'entraînement (GPTBot, ClaudeBot, CCBot, Meta-ExternalAgent, Bytespider, et les signaux d'entraînement Google-Extended et Applebot-Extended) ; ceux qui indexent des

pages pour les réponses de recherche par IA (OAI-SearchBot, Claude-SearchBot, PerplexityBot) ; et ceux qui ouvrent une page à la demande d'un utilisateur (ChatGPT-User, Claude-User, Perplexity-User). Googlebot et Bingbot servent de référence. Trois hypothèses ont été enregistrées, avec une correction de Holm : les robots d'IA sont plus souvent refusés que Googlebot (H1) ; le robot d'entraînement d'OpenAI plus souvent que son robot de recherche (H2) ; organismes publics et éditeurs commerciaux différent (H3).

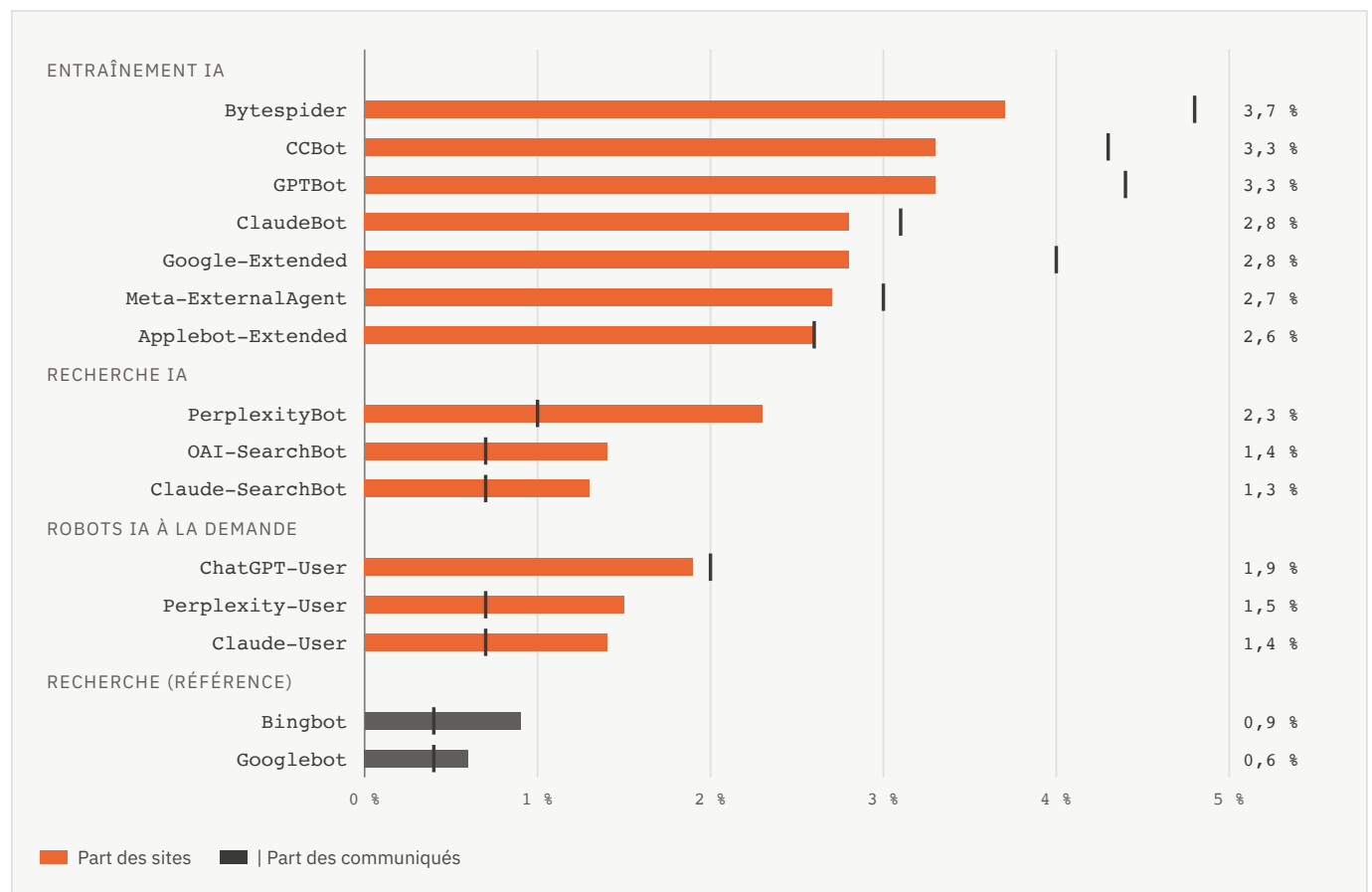
## 04 Presque tous les communiqués sont ouverts

Presque aucun fichier ne ferme la porte. 73 des 1443 sites qui ont répondu (5,1 %, IC à 95 % 4,0 % à 6,3 %) refusent au moins un robot d'IA, et 34 refusent un robot de recherche par IA. Pondérés par le nombre de communiqués de chaque site, 94,0 % des communiqués sont ouverts à tous les robots d'IA et 99,0 % aux trois robots de recherche par IA. Les cinq plus grands sites qui ont répondu — 42,1 % de ces communiqués — n'en refusent aucun.

FIGURE 1

### Les robots que refusent les sites de communiqués

Part des sites dont le robots.txt ferme la plupart de leurs adresses de communiqués à chaque robot (barres), et part des communiqués qui lui sont fermés (traits). Sites ayant répondu pour leur robots.txt.



Les robots les plus souvent refusés sont ceux qui collectent des données d'entraînement : Bytespider, GPTBot et CCBot arrivent en tête. Les moins refusés sont les robots de recherche par IA et ceux qui ouvrent une page pour un utilisateur. Googlebot est refusé par 9 sites, dont 8 parce qu'ils ferment leurs communiqués à tous les robots de la liste. H1 est confirmée : 39 sites refusent GPTBot tout en admettant Googlebot, aucun ne fait l'inverse (test exact de McNemar, p ajustée par Holm < 0,001).

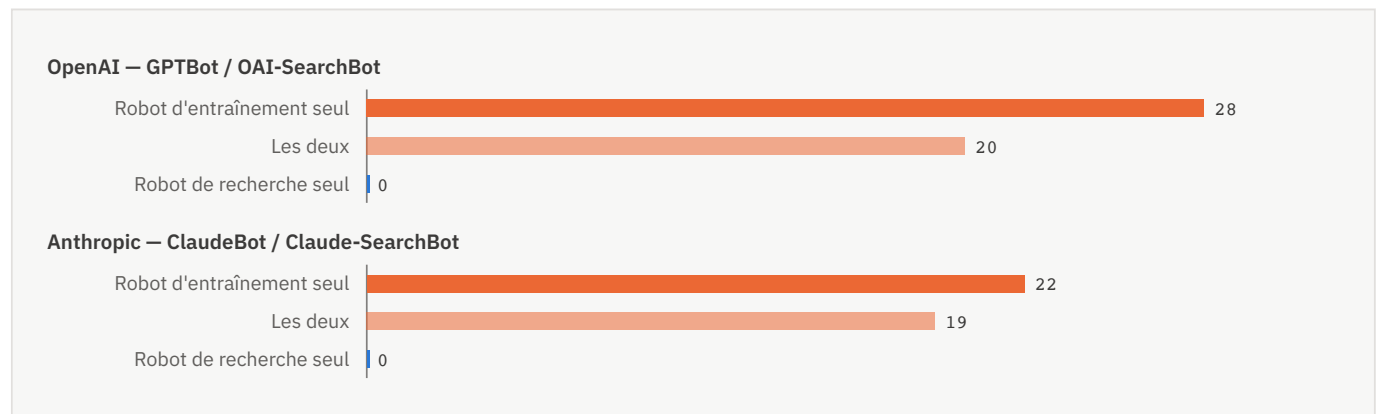
## 05 L'entraînement refusé, la réponse admise

H2 demandait si les éditeurs distinguent l'entraînement de la réponse. Ils le font, dans un seul sens. Parmi les sites qui ont répondu, 20 refusent les deux robots d'OpenAI ; 28 refusent GPTBot mais admettent OAI-SearchBot, le robot qui, selon OpenAI, fait remonter des sites dans la recherche de ChatGPT (OpenAI, s.d.) ; et aucun ne fait l'inverse (test exact de McNemar, p ajustée par Holm < 0,001 ; H2 confirmée). La paire d'Anthropic se partage de la même façon : 22 sites refusent ClaudeBot seul, aucun Claude-SearchBot seul.

FIGURE 2

### L'entraînement refusé, la réponse admise

Sites refusant le robot d'entraînement de chaque entreprise, son robot de recherche, ou les deux. Aucun site ne refuse le seul robot de recherche.



Une partie de l'écart tient à la manière d'écrire les fichiers. 35 des 48 sites qui refusent GPTBot le nomment dans leur fichier ; les robots de recherche, plus récents, sont rarement nommés et relèvent donc du groupe \*, qui les admet en général. Certains éditeurs font ce choix en toutes lettres — gov.ie, en Irlande, nomme GPTBot, ClaudeBot, Google-Extended et Meta-ExternalAgent et laisse les robots de recherche tranquilles. Dans un cas comme dans l'autre, un communiqué fermé à l'entraînement reste, le plus souvent, ouvert à la citation.

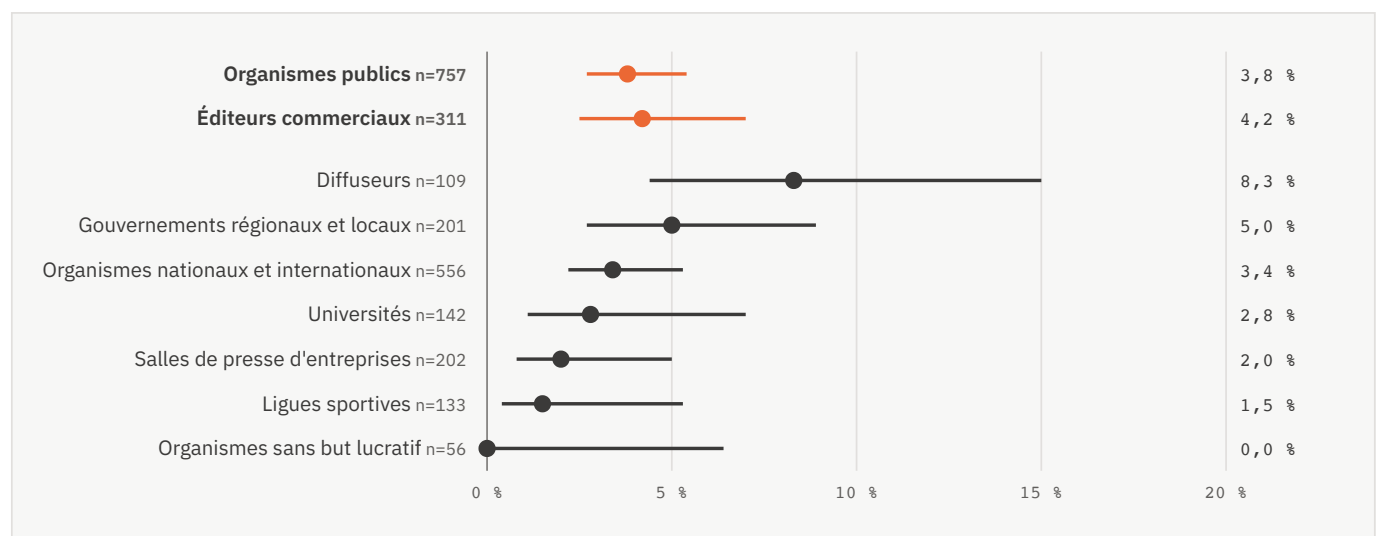
## 06 Qui refuse

H3 ne trouve aucune différence entre organismes publics et éditeurs commerciaux. 3,8 % des sites d'organismes publics (29 sur 757) et 4,2 % des sites commerciaux (13 sur 311) refusent GPTBot ; l'écart, -0,3 point (IC à 95 % -3,4 à +2,0 points), ne se distingue pas de zéro (p ajustée par Holm 0,86 ; H3 non confirmée). Dans le groupe commercial, les sites de diffusion refusent GPTBot plus souvent que les salles de presse des entreprises : 9 sur 109 (8,3 %) contre 4 sur 202 (2,0 %).

FIGURE 3

### Qui refuse GPTBot

Part des sites refusant GPTBot, par type d'éditeur, avec intervalles à 95 %. Les deux lignes regroupées forment la comparaison de H3. Sites ayant répondu pour leur robots.txt.



Les organismes publics relevant du domaine public, ceux qui refusent peuvent être nommés. Ceux qui portent le plus de communiqués sont le gouvernement de la ville de Moscou, la Generalitat valencienne, le conseil municipal de Barcelone, gov.ie en Irlande, l'autorité espagnole de la concurrence (CNMC), le bureau du gouverneur du Texas, le ministère autrichien de l'Intérieur et l'Organisation mondiale du commerce. Le service de nouvelles régional de la Toscane va plus loin et ferme ses pages de communiqués à tous les robots, Google compris.

Parmi les pays comptant au moins 20 sites, refuser un robot d'IA est plus fréquent dans une partie de l'Europe — Pays-Bas (16,7 %), Espagne (16,1 %), Allemagne (13,0 %) — qu'aux États-Unis (4,2 %), au Canada (4,4 %) ou au Japon (1,9 %). Les groupes sont petits et ces chiffres sont descriptifs.

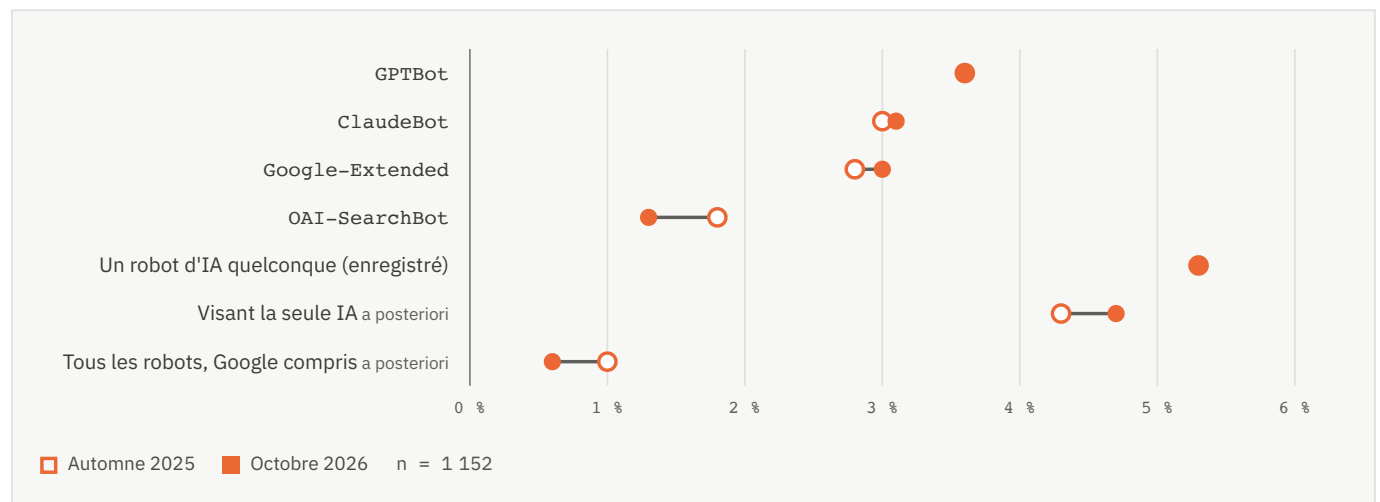
## 07 Un an plus tôt

Pour voir si la situation évolue, nous avons lu, dans la Wayback Machine de l'Internet Archive (Internet Archive, s.d.), la copie archivée du robots.txt de chaque site la plus proche du 7 octobre 2025, à deux mois près, et l'avons appliquée aux mêmes adresses de communiqués. 1 152 sites (80 % de ceux qui ont répondu) avaient une copie utilisable ; la copie médiane se situait à 6 jours de la date visée.

FIGURE 4

### Douze mois plus tard, la même chose

Part des sites refusant chaque robot dans la copie archivée de leur robots.txt la plus proche du 7 octobre 2025 (creux) et aujourd'hui (plein), sur les sites ayant une copie utilisable. Les lignes marquées a posteriori ont été séparées après avoir vu les données.



La part des sites refusant GPTBot était de 3,6 % à l'époque et de 3,6 % aujourd'hui. Selon la mesure enregistrée — un robot d'IA refusé, quel qu'il soit —, 61 sites refusaient alors et 61 aujourd'hui, 23 ayant commencé et 23 cessé. Ce va-et-vient mêle deux choses, que nous avons séparées après l'avoir constaté : les blocages généraux qui ferment un site à tous les robots, Google compris (11 sites alors, 7 aujourd'hui), et les blocages visant la seule IA, passés de 4,3 % à 4,7 % des sites (21 ont commencé, 17 ont cessé ; p 0,63). Les éditeurs modifient leurs fichiers dans les deux sens ; le total n'a pas bougé.

## 08 L'autre barrière : le pare-feu

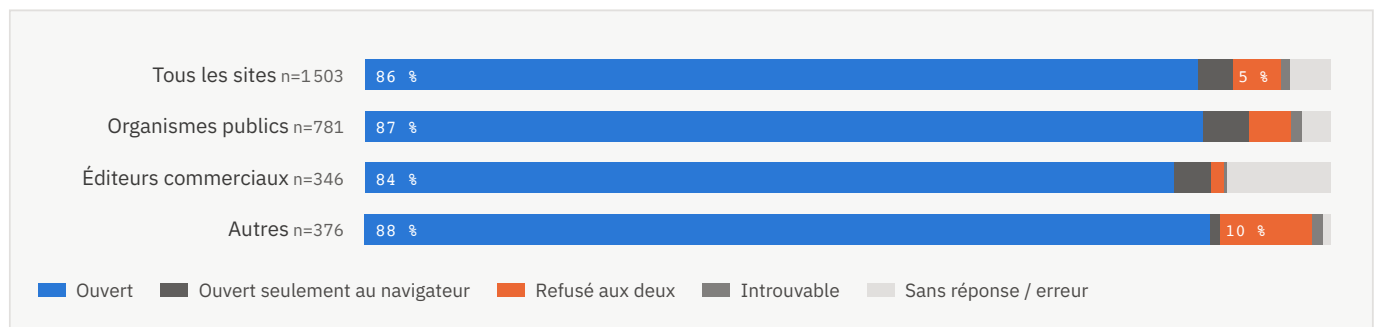
Le robots.txt est une demande, et elle est écrite. Un pare-feu devant un site s'applique, et il n'est écrit nulle part où un robot pourrait le lire. En ouvrant le communiqué le plus récent de chaque site, 75 sites (5,0 %) ont refusé notre client d'emblée, qu'il s'identifie honnêtement ou comme un navigateur, et 54 autres ont refusé l'identité honnête mais ouvert à celle du navigateur. 60 sites n'ont donné aucune réponse exploitable pour leur robots.txt (délais dépassés, erreurs de serveur, certificats invalides) ; un second client

HTTP, ajouté après la collecte, a montré que 30 d'entre eux — tous des sites d'entreprises, pour la plupart des pages de relations avec les investisseurs — laissent sans réponse un client de recherche identifié et opposent un 403 à une identité de navigateur.

FIGURE 5

### Ce qu'un client automatisé a trouvé sur la page du communiqué

Résultat de l'ouverture du communiqué le plus récent de chaque site : ouvert à un client identifié honnêtement ; ouvert seulement à une identité de navigateur ; refusé aux deux ; introuvable ; sans réponse ou erreur du serveur.



Un vrai robot d'IA n'est pas notre client. OpenAI et Google publient les adresses réseau qu'utilisent leurs robots, et un pare-feu peut les admettre tout en refusant les autres. Ce qu'on peut dire, c'est qu'à peu près autant de sites de communiqués placent un pare-feu anti-automates devant leurs communiqués qu'il y en a qui refusent un robot d'IA dans le robots.txt — et que le second est une ligne de texte écrite par l'éditeur, alors que le premier est souvent un réglage par défaut qu'il n'a jamais vu.

Les autres signaux sont plus rares encore. 29 des 1 351 pages de communiqués ouvertes demandent aux moteurs de recherche de ne pas les indexer ; 1 porte la balise officielle noai ; 10 fichiers robots.txt contiennent des Content Signals (Cloudflare, 2025), dont 5 refusent l'entraînement de l'IA. En revanche, 81 sites (5,4 %) publient un fichier llms.txt (Howard, 2024) — un guide rédigé pour les systèmes d'IA —, soit plus qu'il n'y en a qui refusent un robot d'IA.

## 09 Ce que cela signifie pour les communicants

### Les sites de communiqués bloquent-ils les robots d'IA ?

Rarement. Parmi les 1 443 sites qui ont répondu, derrière 500 129 communiqués, 5,1 % refusent au moins un robot d'IA dans leur robots.txt. 94,0 % des communiqués sont ouverts à tous les robots d'IA testés, et 99,0 % aux robots qui alimentent les réponses de recherche par IA.

### **Bloquer GPTBot écarte-t-il un communiqué de la recherche de ChatGPT ?**

Non. OpenAI documente des robots distincts (OpenAI, s.d.) : GPTBot collecte des données d'entraînement, OAI-SearchBot fait remonter des sites dans la recherche de ChatGPT, et chacun suit ses propres règles du robots.txt. 28 sites de communiqués refusent le premier et admettent le second ; aucun ne fait l'inverse. De même, Google-Extended régit l'entraînement et l'ancrage de Gemini, pas la recherche Google (Google, s.d.).

### **Si le robots.txt autorise les robots d'IA, peuvent-ils lire le communiqué ?**

Pas forcément. Un pare-feu peut refuser les clients automatisés quoi qu'en dise le robots.txt : 75 des 1 503 sites ont refusé notre client d'emblée, et 30 sites d'entreprises, pour la plupart de relations avec les investisseurs, n'ont donné aucune réponse à un client identifié. Testez la salle de presse avec un client qui n'est pas un navigateur, pas seulement le fichier robots.txt.

### **Quelle copie d'un communiqué un communicant doit-il vérifier ?**

Chaque site où vit le communiqué — la salle de presse, le diffuseur, toute reprise —, car chacun a son propre fichier et son propre pare-feu. Les cinq plus grands sites de notre population qui ont répondu (42,1 % de ces communiqués) ne refusent aucun robot d'IA ; la salle de presse de l'entreprise est la copie qu'elle contrôle.

---

## **10 Notes de méthode**

Statistiques. H1 et H2 reposent sur des tests exacts de McNemar (McNemar, 1947) sur le résultat apparié de chaque site ; H3 sur le test exact de Fisher, avec un intervalle hybride de Newcombe (Newcombe, 1998) pour l'écart de proportions. La correction de Holm (Holm, 1979) couvre les trois tests. Les intervalles sur les parts de sites sont des intervalles de Wilson à 95 % (Wilson, 1927). La population est un recensement des sites de l'archive : les parts de communiqués sont donc données sans intervalle.

Robustesse, entièrement publiée dans les données. Compter les fichiers inaccessibles comme des refus porte la part des sites refusant un robot d'IA à 8,8 % (GPTBot 7,2 %). Exclure les cinq plus grands sites porte la part pondérée par les communiqués à 10,3 %. Ne retenir que la réponse au client identifié donne 3,1 % pour GPTBot. Compter un site comme refusant dès qu'un seul de ses communiqués est fermé donne 3,7 % pour GPTBot et 1,2 % pour Googlebot. Chaque fichier a été relu au moins 6 heures après la première lecture : sur les 1 438 sites qui ont répondu les deux fois, aucun n'a changé de verdict pour GPTBot (5 n'ont pas répondu la seconde fois). Aucune conclusion ne change.

## Écarts par rapport à l'enregistrement

1. Paramètres de suivi. L'enregistrement appliquait chaque fichier à l'adresse complète du communiqué, paramètres compris. Certains fils ajoutent des balises de suivi de campagne (utm\_ et similaires) à chaque lien, et une règle générale contre les paramètres se lit alors comme le refus de communiqués en réalité ouverts — toute la production d'un diffuseur basculait ainsi. Ces balises (5 660 adresses de communiqués) sont retirées avant l'application des règles. En les conservant, comme prévu, GPTBot est refusé par 3,5 % des sites et un robot d'IA quelconque par 5,2 % ; 2 sites changent, aucune conclusion.
2. Copies archivées. L'enregistrement prévoyait la capture de statut 200 la plus proche du 7 octobre 2025. L'index de recherche de la Wayback Machine répondait en une quarantaine de secondes par requête ; chaque capture a donc été localisée par sa redirection vers la capture la plus proche, et conservée seulement si elle avait été archivée avec le statut 200. Un site dont la capture la plus proche était une erreur est compté comme non couvert.
3. Les signaux au niveau de la page ne sont lus que sur les pages qui se sont ouvertes. Une page de refus porte sa propre balise noindex, qui ne dit rien du communiqué derrière elle.
4. Ajoutés après avoir vu les données, et signalés comme tels : la séparation de l'évolution sur un an entre blocages généraux et blocages visant l'IA ; la même comparaison sans les sites dont le fichier actuel est refusé à tout client automatisé (48 alors, 54 aujourd'hui) ; et la vérification des sites inaccessibles par un second client.

## Limites

- Le robots.txt est une demande. Cette étude rapporte ce que demande chaque fichier, pas si chaque robot s'y conforme ; OpenAI et Perplexity indiquent que leurs robots déclenchés par un utilisateur peuvent ne pas suivre le robots.txt (OpenAI, s.d. ; Perplexity, s.d.).
- Notre client n'est pas un robot d'IA. Les pare-feu qui admettent les robots vérifiés par leur adresse sont invisibles de l'extérieur : les chiffres sur les pare-feu mesurent l'exposition aux clients automatisés en général, pas à l'IA.
- Les fichiers sont ceux servis un jour donné. Un fichier peut changer le lendemain matin.
- La population est l'archive de PPN World — environ 1 500 sites, à dominante anglophone et riche en organismes publics —, pas l'ensemble des éditeurs de communiqués du monde.
- Un communiqué peut vivre sur plusieurs sites. Nous testons la copie vers laquelle renvoie l'archive ; la salle de presse de l'émetteur ou une reprise peuvent répondre autrement.
- Quelques communiqués d'organismes publics renvoient à des plateformes tierces (Google Actualités, X, Instagram, Mailchimp), dont les fichiers sont ceux de la plateforme, pas de l'organisme.
- Les copies archivées couvrent quatre sites sur cinq et ont été capturées par le robot de l'Internet Archive, à qui l'on n'a peut-être pas servi la même chose qu'aux autres.

## Références

- Anthropic. (s.d.). Does Anthropic crawl data from the web, and how can site owners block the crawler? Claude Help Center. Consulté le 8 octobre 2026, à l'adresse <https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler>
- Apple. (s.d.). About Applebot. Apple Support. Consulté le 8 octobre 2026, à l'adresse <https://support.apple.com/en-us/119829>
- Cloudflare. (2025, September 24). Content Signals Policy. <https://contentsignals.org/>
- Common Crawl. (s.d.). CCBot. Consulté le 8 octobre 2026, à l'adresse <https://commoncrawl.org/ccbot>
- Fletcher, R. (2024, February 22). How many news websites block AI crawlers? Reuters Institute for the Study of Journalism. <https://doi.org/10.60625/risj-xm9g-ws87>
- Google. (s.d.). Google's common crawlers. Google for Developers. Consulté le 8 octobre 2026, à l'adresse <https://developers.google.com/crawling/docs/crawlers-fetchers/google-common-crawlers>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <https://www.jstor.org/stable/4615733>
- Howard, J. (2024, September 3). The /llms.txt file. llms-txt. <https://llmstxt.org/>
- Internet Archive. (s.d.). Wayback Machine. Consulté le 8 octobre 2026, à l'adresse <https://web.archive.org/>
- Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022, September). Robots Exclusion Protocol (RFC 9309). Internet Engineering Task Force. <https://doi.org/10.17487/RFC9309>
- Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A. M., Dinh, A., Chitongo, C., Yin, D., . . . Pentland, S. (2024). Consent in crisis: The rapid decline of the AI data commons. *Advances in Neural Information Processing Systems*, 37, 108042–108087. <https://doi.org/10.52202/079017-3431>
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. <https://doi.org/10.1007/BF02295996>
- Meta. (s.d.). Meta web crawlers. Meta for Developers. Consulté le 8 octobre 2026, à l'adresse <https://developers.facebook.com/documentation/sharing/webmasters/web-crawlers>
- Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, 17(8), 873–890. [https://doi.org/10.1002/\(SICI\)1097-0258\(19980430\)17:8%3C873::AID-SIM779%3E3.0.CO;2-I](https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8%3C873::AID-SIM779%3E3.0.CO;2-I)
- OpenAI. (s.d.). Overview of OpenAI crawlers. OpenAI Developers. Consulté le 8 octobre 2026, à l'adresse <https://developers.openai.com/api/docs/bots>
- Perplexity. (s.d.). Perplexity crawlers. Perplexity Docs. Consulté le 8 octobre 2026, à l'adresse <https://docs.perplexity.ai/docs/resources/perplexity-crawlers>
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.1080/01621459.1927.10502953>

## 11 Données, préenregistrement et citation

Chaque résultat par site derrière les figures est publié en CSV sous licence CC BY 4.0, avec la date de lecture de chaque fichier. Les organismes publics y figurent par adresse ; les entreprises et les diffuseurs par type et par rang, car PPN World ne nomme pas les sources commerciales avec lesquelles il n'a pas d'entente. Le préenregistrement est publié sans retouche.

### ↓ Télécharger les données (CSV)

Une ligne par site : type, pays, communiqués, statut du robots.txt, robots refusés aujourd'hui et douze mois plus tôt.

### Lire le préenregistrement

Enregistré le 7 octobre 2026, avant la lecture du moindre fichier robots.txt, et servi tel qu'il a été versé — il désigne donc cette étude par son numéro de travail, n° 03, et mentionne une étude pilote non publiée.

### Citer cette étude

Libre de citation, de reproduction graphique et de republication avec attribution. Citations suggérées :

#### APA

Bolduc, É. (2026). 94 % des communiqués de presse sont ouverts aux robots d'IA : un recensement préenregistré des règles visant les robots d'IA sur 1 503 sites de communiqués (PPN World Research Paper No. 1, Edition 1). PPN World. <https://www.ppnworld.com/fr/research/press-releases-ai-crawlers>

#### MLA

Bolduc, Émile. 94 % des communiqués de presse sont ouverts aux robots d'IA : un recensement préenregistré des règles visant les robots d'IA sur 1 503 sites de communiqués. PPN World, 8 Oct. 2026, [www.ppnworld.com/fr/research/press-releases-ai-crawlers](https://www.ppnworld.com/fr/research/press-releases-ai-crawlers).

#### CHICAGO

Bolduc, Émile. 2026. 94 % des communiqués de presse sont ouverts aux robots d'IA : un recensement préenregistré des règles visant les robots d'IA sur 1 503 sites de communiqués. PPN World Research Paper 1. PPN World. <https://www.ppnworld.com/fr/research/press-releases-ai-crawlers>.

### Licence

Textes, figures et données sont publiés sous CC BY 4.0

### Corrections

Aucune pour l'instant. Toute correction sera publiée ici avec sa date ; un chiffre cité n'est jamais modifié en silence.

## À PROPOS DE L'AUTEUR

**Émile Bolduc**

Centre de recherche PPN World

Émile Bolduc a écrit cette étude pour PPN World, qui suit les communiqués des entreprises, des diffuseurs et des organismes publics du monde entier. Une question de méthode ou de données, ou une extraction que nous n'avons pas publiée : écrivez au Centre de recherche.

[Profil LinkedIn](#)

[Écrire au Centre de recherche](#)