Est-ce que les données constituent une barrière à l'entrée sur le marché des IA génératives ? Un stage au barreau dans la matière du droit des sociétés Une épreuve orale de simulation de présentation en conseil d'administration

Auteur : Akhmedov, Roman
Promoteur(s) : Auer, Dirk
Faculté : Faculté de Droit, de Science Politique et de Criminologie
Diplôme : Master en droit, à finalité spécialisée en droit économique et social
Année académique : 2025-2026
URI/URL : http://hdl.handle.net/2268.2/26412

LIÈGE université
Droit, Science Politique & Criminologie
Département de Droit

Est-ce que les données constituent une barrière à l'entrée sur le marché des IA génératives ?

Roman AKHMEDOV

Travail de fin d'études

Master en droit à finalité spécialisée en droit économique et social

Année académique 2025-2026

Recherche menée sous la direction de :

Monsieur Dirk AUER

Professeur ordinaire

RESUME

L'essor fulgurant des systèmes d'intelligence artificielle générative soulève d'importantes questions en droit européen de la concurrence. Ces modèles reposent sur trois intrants essentiels : les données, la puissance de calcul et les talents humains. Les autorités de concurrence ont très tôt identifié la donnée comme un intrant stratégique susceptible de générer des barrières à l'entrée, en raison notamment du contrôle de vastes corpus par quelques grandes plateformes numériques, des pratiques d'accès exclusif ou discriminatoire et des effets de réseau liés aux boucles de rétroaction des données. Ce mémoire se propose de déterminer si, du point de vue du droit de la concurrence, la donnée constitue réellement une barrière à l'entrée absolue et déterminante sur ce marché.

REMERCIEMENTS

Tout d’abord, je souhaite remercier mon promoteur, Dirk AUER, qui m’a accompagné pendant la rédaction de ce TFE.

Ensuite, je souhaite aussi remercier ma mère, Наталья (Natalya), et Sylejman.

Je souhaite aussi remercier ceux avec qui j’ai étudié Laurane, Yassine, François, Simon, Bastien, Sasha, Matteo, Vernier et Aloïs. C’était sympathique. Merci à Laurane et Sokol d’avoir relu ce travail.

Merci à Alexandre, *Bukayo*, *Lionel*, Quentin, Sokol, Temuujin et Tom.

Merci à Quentin de m’avoir conseiller d’introduire ce recours.

Merci à Mathéa de l’avoir relu.

Enfin, je souhaite adresser mes remerciements les plus sincères et les plus profonds au V. Sans son soutien financier décisif, ce parcours académique n'aurait tout simplement pas été possible. Cette aide a changé le cours de ma vie, et je ne l'oublierai jamais.

Table des matières

Introduction

Avant l'émergence des modèles d'IA comme ChatGPT ou Claude, les systèmes d'intelligence artificielle étaient conçus pour accomplir des tâches spécifiques. Si l'essor du World Wide Web et les progrès de la puissance de calcul, dès le début des années 2000, ont permis la constitution de jeux de données massifs, les architectures algorithmiques de l'époque ne permettaient pas encore d'exploiter pleinement ce volume.

Ce verrou technologique a sauté avec l'avènement de l'apprentissage profond dans les années 2010. Cette évolution a favorisé le développement d'algorithmes capables d'absorber et de traiter efficacement ces masses de données. Concernant ces anciens systèmes d'IA, une étude a démontré[1] que plus il y avait de données, plus les systèmes d'IA étaient performants. Ainsi, le rôle critique de la donnée comme intrant essentiel de l'IA était déjà établi bien avant l'explosion des modèles génératifs grand public[2].

Avec le lancement de ChatGPT au mois de novembre 2022, le grand public découvre une nouvelle catégorie de systèmes d'IA générative capable de générer différents types de contenus (du texte, des images, des vidéos ou de la musique) à partir d'instructions données en langage naturel. Ces systèmes d'IA générative s'appuient en pratique sur des modèles d'IA à usage général[3].

Ces modèles reposent sur trois intrants essentiels : les données, la puissance de calcul et les talents. Leur développement nécessite la combinaison de ces trois éléments. Les autorités de la concurrence à travers le monde ont d'ailleurs identifié leur caractère stratégique et soulignent qu'ils font l'objet d'une compétition particulièrement intense entre les entreprises spécialisées en IA générative.

Ce travail de fin d'études vise précisément à déterminer si la donnée constitue réellement un intrant aussi critique que le soutiennent les autorités de la concurrence. Les principales préoccupations d'ordre concurrentiel portent essentiellement sur les pratiques de refus

[1] M.BANKO et E. BRILL, *Scaling to Very Very Large Corpora for Natural Language Disambiguation*, t. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 2001, p. 26–33

[2] S. Russell et P. Norvig, *Intelligence artificielle : une approche moderne*, 4e éd., Pearson, 2021, p.24

[3] Article 3 du Règlement (UE) 2022/1925 du Parlement européen et du Conseil du 14 septembre 2022 relatif aux marchés contestables et équitables dans le secteur numérique et modifiant les directives (UE) 2019/1937 et (UE) 2020/1828 (règlement sur les marchés numériques), *J.O.U.E.*, L 265, 12 octobre 2022.

d'accès ou d'accès discriminatoire aux données de la part d'entreprises disposant d'un avantage concurrentiel significatif en la matière.

Dans un premier temps, une présentation succincte de l'intelligence artificielle générative sera proposée, accompagnée d'une synthèse des positions adoptées par les différentes autorités de la concurrence. Dans un deuxième temps, une analyse approfondie de l'accès aux données sur le marché de l'IA générative sera menée. Enfin, les réponses apportées par l'Union européenne en matière d'accès aux données seront examinées.

Chapitre 1 L'IA générative et les préoccupations des autorités de la concurrence

Ce premier chapitre présente, d'une part, le fonctionnement des systèmes d'IA générative (section 1) et, d'autre part, la manière dont les autorités de la concurrence appréhendent les enjeux concurrentiels soulevés par ces technologies (section 2).

Section 1 L'IA générative

§1 Fonctionnement

L'IA générative repose sur des réseaux de neurones artificiels, c'est-à-dire des modèles composés de couches de "neurones" interconnectés [4]. Les données d'entrée sont progressivement transformées au fil de ces couches, ce qui permet au modèle d'apprendre des schémas de plus en plus complexes et des représentations de plus en plus abstraites. Son fonctionnement est fondamentalement probabiliste : à partir d'un contexte donné en langage naturel, le modèle prédit la suite la plus probable, mot par mot[5]. Ainsi, après la séquence « Messi a gagné huit », un modèle correctement entraîné attribuera une probabilité élevée au terme « ballons d'Or ».

[4] · X, « Artificial Neurons: Structure, Operation, and Practical Applications », disponible sur ogi.digital, consulté le 02/02/2026

[5] FRANCE NUM, « L'intelligence artificielle (IA), à quoi ça sert ? », disponible sur www.francenum.gouv.fr, consulté le 10/02/2026

§2 Entrainement

L'entraînement des modèles d'IA générative a fait l'objet de nombreuses analyses doctrinales et institutionnelles, auxquelles il n'y a pas lieu de revenir en détail ici. Il convient néanmoins d'en rappeler les grandes lignes, telles que synthétisées par l'avis n° 24-A-05 de l'Autorité de la concurrence de juin 2024 qui précise notamment ce qui suit : « La modélisation de l'IA générative implique deux étapes principales. La phase d'entraînement du modèle vise à lui apprendre des capacités générales permettant au modèle de produire le contenu (texte, images ou autres) qui constitue la réponse la plus probable à une question posée. Cette étape peut être complétée par une spécialisation sur des tâches spécifiques, appelée réglage fin. Enfin, la production de contenu à partir de ce modèle, aussi appelée inférence, nécessite la mise à disposition du modèle aux utilisateurs finaux. Chaque étape nécessite une puissance de calcul et un apport en données distincts»[6].

Parmi les principaux modèles d'IA générative à usage général aujourd'hui figurent notamment GPT-5 développé par OpenAI, Claude 4.7 par Anthropic, Gemini 3.1 par Google DeepMind, les différentes versions de LLaMA proposées par Meta ou encore son dernier modèle Muse Spark, ainsi que les modèles de la famille Mistral développés par Mistral AI. Ces modèles se distinguent non seulement par leurs performances, mais aussi par leur mode de diffusion. Certains sont dits « fermés » ou propriétaires : leur code source, leurs poids et leurs données d'entraînement ne sont pas rendus publics, ce qui est le cas de GPT-5 ou de Gemini 3.1. D'autres sont diffusés en open source, ce qui signifie que leur poids et, dans certains cas, leur code sont librement accessibles et réutilisables par des tiers : c'est notamment le cas de certains modèles LLaMA de Meta et de plusieurs modèles Mistral. Cette distinction est loin d'être anodine du point de vue concurrentiel : elle conditionne en effet la capacité des acteurs tiers à s'appuyer sur ces modèles pour développer leurs propres applications ou pour les affiner sur des données spécialisées, sans avoir à repartir de zéro ni disposer de ressources comparables à celles des grandes plateformes.

Le secteur connaît actuellement une véritable course au modèle d'IA le plus performant : de nouveaux modèles, aux capacités sans cesse accrues, sont lancés presque chaque mois,

[6] Avis n° 24-A-05 de l'Autorité de la concurrence *relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative*, 28 juin 2024, point 15

chaque acteur cherchant à conquérir ou à conserver le leadership sur le marché de plus en plus concurrentiel.

Section 2 Les positions des autorités de concurrence

§1 Une vigilance anticipative face à l'IA générative

L'essor rapide de l'IA générative suscite en effet des préoccupations croissantes, en particulier quant aux risques de concentration du marché, de renforcement du pouvoir de marché et d'effets discriminatoires. Par conséquent, les différentes autorités de la concurrence à travers le monde ont légitimement fait part de leurs inquiétudes.

L'Autorité française de la concurrence a publié en 2024[7] un avis, dont nous avons tiré un extrait ci-avant, et dans lequel elle recense plusieurs barrières potentielles à l'entrée sur le marché du développement des modèles de fondation. Elle y souligne notamment le recours nécessaire à des processeurs graphiques ou à d'autres puces spécialisées, le cloud, l'accès aux données ainsi que la disponibilité des talents comme autant de facteurs susceptibles de freiner l'entrée ou l'expansion de nouveaux acteurs sur le marché de l'IA générative. De nombreuses autorités de concurrence ont convergé vers une analyse similaire. Ainsi, la CMA[8], l'AdC[9] et la FTC[10] ont toutes identifié, à des degrés divers, les risques liés au contrôle de ces inputs critiques. De plus, l'OCDE[11] et la commission européenne[12] ont dernièrement publié des travaux qui convergent aussi vers ces barrières. Enfin, une déclaration conjointe de la FTC, de la CMA et de la Commission européenne synthétise cette approche commune, en mettant en avant la nécessité de surveiller les effets de la concentration sur les mêmes intrants critiques afin de préserver la contestabilité des marchés de l'IA générative[13].

---

[7] Avis n° 24-A-05 de l'Autorité de la concurrence, ibidem.

[8] COMPETITION AND MARKETS AUTHORITY (CMA), *CMA AI strategic update*, 29 avril 2024, disponible sur https://www.gov.uk/ , consulté le 17/11/2025

[9] AUTORIDADE DA CONCORRÊNCIA (ADC), *Competition and Generative AI – Zooming in on Data*, Short Papers Series n°1, 27 septembre 2024, disponible sur https://www.concorrencia.pt/

[10] BUREAU OF COMPETITION ET OFFICE OF TECHNOLOGY (FTC), « Generative AI Raises Competition Concerns », *Tech at FTC*, 29 juin 2023, disponible sur www.ftc.gov, consulté le 11/04/2026

[11] R. MAY, *Intelligence artificielle, données et concurrence*, OCDE, DAF/COMP(2024)2, mai 2024, disponible sur www.oecd.org, consulté le 26/11/2025

[12] K. KOWALSKI, C. VOLPIN et Z. ZOMBORI, « Competition in Generative AI and Virtual Worlds », *Competition Policy Brief*, n°3, 2024, disponible sur competition-policy.ec.europa.eu.

[13] M. VESTAGER et al., « Joint Statement on Competition in Generative AI Foundation Models and AI Products », 23 Juillet 2024, disponible sur competition-policy.ec.europa.eu.

§2 Les risques liés au contrôle des données

Les autorités identifient en particulier plusieurs risques récurrents liés aux données. D'une part, elles soulignent que les entreprises disposant d'un accès massif à des données critiques peuvent restreindre ou refuser l'accès à ces ressources, ce qui est susceptible d'ériger des barrières à l'entrée. D'autre part, elles attirent l'attention sur le risque de monopolisation progressive de certains corpus via des accords d'accès privilégiés, voire exclusifs, qui réservent à quelques acteurs l'usage de données stratégiques pour l'entraînement de leurs modèles. Elles mentionnent également la concentration des données d'utilisateurs entre les mains de quelques grandes plateformes, ce qui pourrait renforcer leur avantage concurrentiel dans la fourniture de services d'IA générative.

Les autorités de la concurrence, marquées par leur intervention tardive sur les marchés numériques, tentent aujourd'hui d'adopter une approche anticipative face à l'IA générative, notamment par des régulations *ex ante* comme le montre l'adoption du DMA. Cette posture les conduit à identifier très tôt certains intrants, notamment la donnée, comme des sources potentielles de pouvoir de marché. Toutefois, dans un secteur encore émergent, cette anticipation pourrait conduire à une surestimation du rôle de la donnée comme barrière à l'entrée.

Chapitre 2 Données et concurrence dans l'IA générative

Ce deuxième chapitre est consacré à la donnée et à son rôle dans l'IA générative. Il examinera dans un second temps les conditions d'accès aux données dans l'écosystème de l'IA générative, en analysant la diversité des sources et des modes d'obtention, ainsi que les coûts et contraintes liés à leur exploitation, afin d'apprécier dans quelle mesure elles peuvent constituer une barrière à l'entrée.

Section 1 La donnée : définition et enjeux

Pour déterminer l'objet de cette étude, il convient d'abord de préciser ce qu'est une « donnée ». Avant l'adoption du règlement sur la gouvernance des données (Data Governance Act ou DGA), il n'existait pas de définition de la donnée au niveau de l'Union européenne.

Dans un premier temps, le législateur européen s'est concentré sur la notion de donnée personnelle avec l'adoption du RGPD en 2016. C'est grâce à l'arrivée de la Commission von der Leyen que la donnée sera finalement définie[14]. Cela a conduit à l'adoption du DGA, qui est entré en vigueur le 23 juin 2022. L'article 2, paragraphe 1, du Règlement définit la donnée comme « toute représentation numérique d'actes, de faits ou d'informations et toute compilation de ces actes, faits ou informations, notamment sous la forme d'enregistrements sonores, visuels ou audiovisuels ». La même définition sera ultérieurement reprise par d'autres instruments européens comme le Data Act[15].

Les données se sont désormais imposées comme un actif stratégique majeur pour les entreprises en leur permettent notamment de mieux anticiper les comportements, d'orienter les décisions, d'améliorer leurs produits et services ou d'en développer de nouveaux, conférant ainsi un avantage concurrentiel à celles qui en disposent et savent les exploiter. Dans certains cas, cet avantage a conduit des entreprises à mettre en œuvre des pratiques susceptibles de restreindre la concurrence, ce qui a amené la Commission européenne à intervenir, à plusieurs reprises, afin d'encadrer l'accès aux données et leur utilisation.

En effet, les grandes plateformes numériques ont accumulé, depuis plus de 20 ans, un nombre conséquent de données. Alphabet possède, par exemple, toutes les données issues de son index de recherche Google Search, de l'utilisation de Google Chrome, Google Ads ou Google Maps, ainsi que de YouTube ou Google Books. Meta en possède également via ses célèbres réseaux sociaux Facebook, Instagram et WhatsApp. Microsoft est le propriétaire de l'index de recherche qui alimente le moteur de recherche Bing et de GitHub, la plateforme de référence du partage de code entre développeurs[16].

Section 2 L'accès aux données

Si ces éléments suggèrent l'existence d'avantages liés aux données, leur capacité à constituer de véritables barrières à l'entrée demeure incertaine et mérite d'être nuancée.

[14] L. Pailler et C. Brunerie, *La cohérence du droit des données,* Paris, Lefebvre Dalloz, 2026, p. 11

[15] Article 2 du Règlement (UE) 2023/2854 du Parlement européen et du Conseil du 13 décembre 2023 concernant des règles harmonisées portant sur l'équité de l'accès aux données et de l'utilisation des données et modifiant le règlement (UE) 2017/2394 et la directive (UE) 2020/1828, 22 décembre 2023

[16] Avis n° 24-A-05 de l'Autorité de la concurrence, ibidem, point 254.

§1 Non-rivalité et multiplicité des sources

La donnée présente une caractéristique économique fondamentale qui la distingue des ressources physiques plus traditionnelles comme le pétrole : en effet contrairement à cette ressource, la donnée est un bien non-rival[17]. Le fait pour une entreprise de détenir ou d'exploiter un certain jeu de données n'empêche donc pas, en principe, d'autres entreprises d'accéder aux mêmes informations ou à des informations équivalentes : une même base de données peut être copiée et utilisée simultanément par plusieurs acteurs sans être « consommée ». À l'inverse, un baril de pétrole ne peut être utilisé qu'une fois ; si une entreprise le vend, elle s'en trouve immédiatement dépossédée.

Par ailleurs, le caractère non rival de la donnée implique également que la cession ou le partage de données ne prive pas nécessairement le vendeur de la possibilité de continuer à les utiliser. Le partage des données profitera à chacun. On a déjà pu lire à cet égard qu'« un bien numérique en général, et la donnée en particulier, permet de générer des externalités positives : le partage du bien dégage naturellement de la valeur ajoutée »[18]. Les biens numériques se prêtent en effet à des usages répétés et multipliés. Leur diffusion et leur mise en commun peuvent créer de la valeur additionnelle plutôt que de la détruire[19].

Cette non-rivalité de principe facilite le partage de la donnée entre les différents acteurs du marché et atténue les possibilités d'appropriation exclusive. Elle concourt ainsi, en théorie, à réduire les barrières à l'entrée dans le secteur de l'IA générative.

L'évolution récente du marché vient confirmer empiriquement cette analyse. Effectivement, OpenAI est le leader sur le segment des IA génératives, ce qui lui a permis de capter un volume considérable de données d'utilisateurs. Toutefois, depuis 2025, plusieurs concurrents gagnent en importance : Google avec la famille Gemini, dont l'adoption progresse rapidement, et Anthropic avec les modèles Claude connaissent une croissance notable en particulier sur le segment des entreprises. Cette évolution suggère que, si les effets de réseau et les données de retour utilisateur peuvent initialement avantager le premier entrant, ils n'empêchent pas

[17] A. STAPP, « Why Data Is Not the New Oil », disponible sur https://truthonthemarket.com/, consulté le 05/02/2025

[18] N. GLADY, « La Data n'est pas le nouveau pétrole », disponible sur https://www.hbrfrance.fr/ consulté le 29/10/2025

[19] S. BROOKS et G. ROESNER, « *Data use externalities* », disponible sur www.gov.uk

l'émergence de modèles concurrents efficaces capables de s'appuyer sur leurs propres écosystèmes d'utilisateurs et leurs canaux de distribution sur un marché qui demeure hautement concurrentiel.

§2 Diversité des modes d'accès aux données

Il existe une multiplicité de sources de données. Même lorsque certains acteurs disposent de bases très importantes, d'autres entreprises peuvent accéder à des données substituables ou reconstituer des corpus similaires par d'autres moyens.

a) Donnée publique et web scraping

i. Donnée publique

Une première voie d'approvisionnement consiste à exploiter les données publiques, en accès libre et gratuit, et sous un format numérique facilement réutilisable [20]. Dans l'Union européenne, cette catégorie recouvre en particulier les données ouvertes du secteur public, publiées via des portails d'open data nationaux ou européens[21].

Les développeurs de modèles d'IA peuvent également recourir à des dépôts de données publics gérés par des acteurs privés ou communautaires. On peut citer, par exemple, la plateforme Kaggle[22], qui héberge des milliers de jeux de données publics dans des domaines variés (texte, images, vidéos, etc.) ou encore Hugging Face qui offre un point d'accès unifié à un grand nombre de jeux de données communautaires, déjà structurés pour l'entraînement des IA génératives[23].

Enfin, certains ensembles massifs de données ouvertes ont été conçus spécifiquement pour les besoins de l'IA. Le projet Common Crawl[24] fournit ainsi des états mensuels du Web mondial (HTML, texte, métadonnées, graphes de liens), largement utilisés comme bases de données

[20] M. Hoang, *Les données publiques : qu'en est-il de leur ouverture et de leur réutilisation ?,* Cahiers français, 2021, p.36-45

[21] Commission européenne, « Open Data Portals », disponible sur digital-strategy.ec.europa.eu, consulté le 7 avril 2026

[22] X, « Kaggle Datasets », disponible sur www.kaggle.com, consulté le 7 avril 2026.

[23] X, « Hugging Face Datasets », disponible sur www.huggingface.co , consulté le 7 avril 2026.

[24] X, « Common Crawl Overview », disponible sur www.commoncrawl.org , consulté le 7 avril 2026.

textuelles pour l'entraînement de grands modèles de langage. De même, l'initiative LAION publie de grands jeux de données extraits du Web, formatés pour faciliter leur utilisation dans l'entraînement de modèles de vision et de multimodalité[25]. Ces ressources illustrent comment l'ouverture de données, combinée à des initiatives spécialisées, alimente directement le développement de systèmes d'IA à très grande échelle.

ii. Web scraping

Le « web scraping » permet d'extraire directement du contenu public depuis les sites Web. Cette technique consiste à explorer des sites sur Internet à l'aide d'outils automatisés pour y extraire des contenus tels que des textes, des images, des sons ou des vidéos. Cependant, cette technique ne s'avère pas toujours légal lorsqu'elle extrait des informations personnelles ou protégé par le droit d'auteur. Ainsi, de nombreuses interrogations juridiques sont soulevées concernant le droit d'auteur, la conformité à l'IA Act ou encore au regard du RGPD. Cette technique a notamment poussé le New York Times à introduire une action en justice contre OpenAI pour violation du droit d'auteur[26].

Cependant, cette technique s'avère de moins en moins efficace. Les premiers modèles de type GPT ont pu être entraînés sur de larges corpus issus du Web dans un contexte où relativement peu de sites se préoccupaient de limiter l'accès par des robots et où la réutilisation à des fins d'entraînement n'était pas encore prise en compte. Désormais, un nombre croissant de sites, en particulier les grands médias, mettent en place des mesures pour restreindre ou encadrer le web scraping[27]. Par exemple, Reddit, dont les flux de discussions constituent une ressource précieuse pour l'entraînement des modèles de langage, a ainsi fortement restreint l'accès gratuit à son API.[28] De même, la plateforme X a revu sa politique d'accès et interdit à des tiers

[25] X, « LAION Projects », disponible sur laion.ai, consulté le 7 avril 2026.
[26] T. MERGEL, « NYT v. OpenAI: The Times's About-Face », *Harvard Law Review Blog*, disponible sur https://harvardlawreview.org/ , consulté le 8 avril 2026.
[27] A. BOCHAROV et al., « Declaring your AIndependence: Block AI bots, scrapers and crawlers with a single click », disponible sur blog.cloudflare.com, consulté le 7 avril 2026.
[28] B. CHERREY, « Reddit va faire payer l'accès à son API à certains utilisateurs : mauvaise nouvelle pour les entreprises d'IA », disponible sur www.zdnet.fr, consulté le 5 avril 2026.

de collecter à grande échelle des données de son application ou de son API[29]. Pour autant, ces fermetures ne concernent qu'une petite fraction du Web.

L'obtention de données via les portails publics et le web scraping ne constituent pas deux mondes étanches. En pratique, les portails d'open data ont toutefois vocation à offrir des contenus déjà structurés, documentés afin de pouvoir les réutiliser[30], ce qui tend à réduire les coûts de préparation et d'analyse pour l'entraînement des modèles. À l'inverse, une partie des données accessibles par scraping ne bénéficie pas forcément du même degré de structuration ni des mêmes garanties juridiques explicites[31], de sorte que leur réutilisation à grande échelle s'accompagne souvent de coûts supplémentaires de nettoyage et de risques accrus en matière de propriété intellectuelle ou de conformité contractuelle.

b) Accès contractuel aux données

Les entreprises peuvent également acquérir des données par la voie contractuelle, soit en ayant recours à des data brokers, acteurs spécialisés dans la collecte, l'agrégation et la revente de vastes ensembles de données, soit en négociant directement des accords de licence avec les ayants droit. Ainsi, OpenAI a multiplié les partenariats avec des groupes de presse d'envergure tels que News Corp[32], Axel Springer[33], ainsi que le Washington Post[34]. Sur le marché francophone, OpenAI a également signé un accord notable avec le journal « Le Monde ». Les autres géants du secteur suivent une trajectoire similaire et seront demain suivis par d'autres. En effet, Alphabet a conclu un accord avec la plateforme Reddit[35] pour accéder à ses flux de données en temps réel. De son côté, Meta a noué des liens avec CNN et Fox News[36]. Le secteur n'est pas limité aux acteurs américains, puisque l'entreprise française

[29] X, « X interdit l'utilisation de son contenu ou de son API pour affiner ou entraîner des grands modèles d'IA », disponible sur opentermsarchive.org, consulté le 8 avril 2026.

[30] Commission Européenne, « *Open Data Goldbook for Data Managers and Data Holders* », disponible sur data.europa.eu, 2020, consulté le 7 avril 2026.

[31] OECD (2025), « Intellectual property issues in artificial intelligence trained on scraped data », *OECD Artificial Intelligence Papers*, No. 33, OECD Publishing, Paris, https://doi.org/10.1787/d5241a23-en.

[32] Newscorp, « News Corp and OpenAI Sign Landmark Multi-Year Global Partnership », disponible sur newscorp.com , consulté le 8 avril 2026.

[33] Reuters, « Global news publisher Axel Springer partners with OpenAI in landmark deal », disponible sur www.reuters.com, consulté le 15 avril 2026.

[34] Open AI, « The Washington Post partners with OpenAI », disponible sur openai.com, consulté le 15 avril 2026

[35] Avis n° 24-A-05 de l'Autorité de la concurrence, ibidem, p.49.

[36] New York Post, « Meta reaches AI deals with CNN, Fox News, other media outlets », disponible sur nypost.com,

Mistral AI s'est rapprochée de l'Agence France-Presse[37]. Amazon a également entamé des collaborations, notamment avec le New York Times. Meta, Google, Apple et Microsoft concluent moins de contrats que leurs concurrents parce qu'ils possèdent déjà et de toute évidence une grande base de données. À ce jour, aucun de ces accords n'est exclusif, de sorte qu'aucun acteur ne bénéficie, sur cette seule base, d'un accès unique à ces corpus de données. Ils contribuent néanmoins à structurer un marché émergent de la licence de contenus pour l'IA, où les grandes entreprises technologiques sont les mieux placées pour conclure des accords avec les principaux ayants droit, ce qui pourrait, à terme, accroître les coûts d'entrée pour des concurrents de plus petite taille.

c) Donnée synthétique

Une autre manière de répondre aux besoins en données consiste à recourir à des données synthétiques, dès lors qu'elles reproduisent de manière suffisamment réaliste les caractéristiques de données réelles[38]. Il s'agit de données générées artificiellement, souvent par des modèles d'IA eux-mêmes, qui peuvent ensuite être utilisées pour entraîner ou affiner d'autres modèles, voire le modèle d'origine. Cette approche offre un substitut aux données réelles dans des contextes où celles-ci sont coûteuses, sensibles ou difficiles à obtenir. Le site d'« IBM » indique que les avantages que procurent les données synthétiques sont nombreux. Elles permettent une personnalisation des données, une confidentialité et l'obtention de données plus riches[39]. Pour autant, un entraînement basé uniquement sur des données synthétiques ne sera pas concluant. Il faut effectivement combiner des données réelles et des données synthétiques pour maximiser l'efficacité du système d'IA[40].

Il est important de préciser que le recours massif aux données synthétiques s'accompagne de limites importantes. Lorsque ces données sont produites à partir de modèles imparfaits, elles risquent de reproduire et d'amplifier les biais, erreurs ou lacunes présents dans le modèle

[37] Le monde, « Meta signe de nouveaux accords avec des médias pour intégrer leurs contenus à son assistant IA », disponible sur https://www.lemonde.fr/, consulté le 13/02/2026

[38] R. Alaily. « The New AI Economy : Understanding the Technology, Competition, and Impact for Societal Good » disponible sur concurrence.com , consulté le 13/02/2026

[39] IBM, « Qu'est-ce que les données synthétiques ? », disponible sur www.ibm.com, consulté le 15 avril 2026.

[40] Sha Sajadieh et al., « The AI Index 2026 Annual Report », AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, Avril 2026

générateur. La littérature récente a notamment mis en évidence le risque de « model collapse »[41], un phénomène dégénératif par lequel des modèles entraînés de manière récurrente sur leurs propres productions perdent progressivement en diversité et en qualité. Des travaux plus récents[42] nuancent toutefois ce constat, en montrant que ce risque peut être significativement atténué lorsque les données synthétiques ne font qu'accompagner les données réelles.

En définitive, les données synthétiques contribuent à abaisser certaines barrières à l'entrée liées à l'accès aux données réelles. Elles offrent aux nouveaux entrants la possibilité d'augmenter la taille et la diversité apparentes de leurs jeux de données sans devoir négocier systématiquement des licences coûteuses ou collecter eux-mêmes des volumes massifs d'informations sensibles, ce qui réduit les contraintes financières et juridiques pesant sur l'entraînement de modèles d'IA générative. Ces techniques permettent aux plus petites structures de se faire une place sur un marché concurrentiel déjà chargé.

d) Modèle en open source

L'essor des modèles en open source constitue un facteur supplémentaire qui pourrait tempérer l'idée selon laquelle la donnée représenterait une barrière à l'entrée insurmontable. Plusieurs entreprises rendent en effet librement accessibles certains de leurs modèles d'IA[43]. C'est le cas de Mistral AI, qui met à disposition différents modèles spécialisés, ou encore de Google. Ces modèles présentent un avantage concret pour les nouveaux entrants : ayant déjà été préentraînés sur de larges corpus de données, ils peuvent être repris et spécialisés sur des jeux de données plus ciblés[44], sans que l'entreprise concernée ait à supporter le coût d'un entraînement complet. Certains modèles, comme Meta's LLaMA or Hugging Face's BLOOM,

[41] I. Shumailov et al., « AI models collapse when trained on recursively generated data », disponible sur https://www.nature.com/, consulté le 14/03/2026.

[42] Sha Sajadieh et al., « The AI Index 2026 Annual Report », AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, Avril 2026.

[43]CNIL, « Les pratiques open source en intelligence artificielle » , disponible sur https://cnil.fr/fr ; M. Zúñiga et D. Auer, « Don't Believe the Hype (on Competition and AI) », disponible sur https://truthonthemarket.com/, consulté le 26/03/2026.

[44] T. Oeyen et Y.Yargici, « Uncharted territories : Generative AI, merger control and the Microsoft-Open AI saga », disponible sur concurrence.fr, consulté le 18/03/2026.

arrivent à surpasser des modèles fermés dans certaines tâches[45]. La mise à disposition de modèles en open source peut conduire à l'innovation, ce qui fera baisser, in fine, les barrières à l'entrée du marché. Cette dynamique est activement encouragée par les pouvoirs publics : les États-Unis y voient un levier d'innovation dans leur AI Action Plan[46], et le président Macron a exprimé un soutien comparable à l'Open source comme vecteur de souveraineté technologique européenne[47].

Cela étant, l'open source ne suffit pas à lui seul à garantir une concurrence effective. L'Autorité française de la concurrence souligne que la mise à disposition de modèles ouverts n'élimine pas les asymétries liées aux données[48]. Les modèles open source ne fournissent, par ailleurs, pas nécessairement l'accès aux jeux de données d'entraînement, ni aux données de retour utilisateur, de sorte que la capacité à constituer et exploiter des corpus de haute qualité demeure largement concentrée entre les mains des grands acteurs[49].

e) Qualité des données et rendements décroissants

L'efficacité des modèles ne dépend pas uniquement du volume de données. Au-delà d'un certain seuil, l'ajout de données supplémentaires produit des rendements décroissants, voire négatifs.

Dès lors, la qualité de la donnée devient primordiale. Une mauvaise qualité des données utilisées lors de l'entraînement conduit à des modèles moins précis et plus sujets aux hallucinations, c'est-à-dire à la production de réponses qui seront factuellement incorrectes[50]. Le principe consacré est alors le « garbage in, garbage out »[51]. Il est en effet primordial pour que la sortie de l'IA soit efficace, que celle-ci soit entrainée sur des données de qualité. Il est, par conséquent, essentiel que les données d'entraînement fassent l'objet d'un travail

[45] M. ALMADA et al., « Competition in and through artificial intelligence », *Research Handbook on Competition and Technology,* Cheltenham, Edward Elgar Publishing, 2025, p. 111.
[46] OFFICE OF SCIENCE AND TECHNOLOGY POLICY, *Winning the Race AMERICA'S AI ACTION PLAN*, 2025
[47] A. Piquard, « Intelligence artificielle : la bataille des modèles en open source » disponible sur lemonde.fr, consulté le 29/03/2026.
[48] Avis n° 24-A-05 de l'Autorité de la concurrence, ibidem, p.56
[49] R. Longo et M. Rocha, « Generative AI : The new digital frontier for competition », disponible sur concurrence.com, consulté le 30/03/2026
[50] X, « Qu'est-ce que la qualité des données pour l'IA ? », disponible sur www.ibm.com, s.d., consulté le 27 avril 2026.
[51] M. ALMADA et al., *Ibidem*, p.109

de « curation ». Par curation des données, on entend l'ensemble des opérations visant à préparer les données pour l'entraînement : sélection des sources, filtrage des contenus non pertinents ou de mauvaise qualité, nettoyage et, le cas échéant, une annotation manuelle humaine, etc.. Ce travail de curation est coûteux mais crucial car il conditionne directement la pertinence, la fiabilité et la sécurité des modèles d'IA générative[52].

Le modèle « Koala », développé en 2023 par des chercheurs de l'université de Berkeley, en est un exemple. Affiné sur une base de données soigneusement curée, ses résultats ont montré qu'il pouvait rivaliser avec les modèles ChatGPT dans plus de la moitié des cas. Cela a conduit ses auteurs à conclure que la clé pour construire de bons modèles réside davantage dans la présence de données de haute qualité et diversifiées que dans le simple volume[53] de sorte qu'il serait plus intéressant de viser la qualité plutôt que la quantité.

Cet exemple a une implication concurrentielle directe et nuance considérablement la thèse selon laquelle la quantité de données constitue une barrière insurmontable à l'entrée. Si la qualité des données et l'efficacité de l'algorithme peuvent compenser le volume, alors la barrière à l'entrée se déplace partiellement. Elle ne tient plus seulement à la possession de grandes quantités de données, mais à la capacité à les sélectionner, les préparer et les exploiter intelligemment, ce qui, certes, reste coûteux, mais potentiellement accessible à des acteurs disposant d'une expertise technique suffisante, même sans les ressources des grandes plateformes établies.

f) Les coûts et contraintes liés au cycle de vie des données

Même lorsque l'accès aux données n'est pas, en tant que tel, restreint, leur exploitation soulève des contraintes techniques, économiques et juridiques significatives. Autrement dit, la question concurrentielle ne se limite pas à l'obtention des données, mais s'étend à leur gestion tout au long de leur cycle de vie.

En premier lieu, le stockage et la gestion des données à grande échelle impliquent des coûts importants, qui pourraient constituer une barrière à l'entrée sur le marché[54]. Les entreprises

[52] X, « Qu'est-ce que la curation de données ? », disponible sur www.ibm.com, consulté le 27 avril 2026.
[53] Y. GENG et al., « Koala: A Dialogue Model for Academic Research », disponible sur bair.berkeley.edu, consulté le 25/02/2026
[54] M. ALMADA, *ibidem*, p.109

doivent disposer d'infrastructures capables de conserver, de traiter et de rendre accessibles des volumes massifs de données, ce qui suppose des capacités conséquentes en matière de stockage et de puissance de calcul. La mise en place de telles infrastructures représente un investissement financier significatif. À cet égard, certains grands acteurs, présents à plusieurs niveaux de la chaîne de valeur de l'IA, bénéficient d'importantes économies d'échelle, ce qui réduit leurs coûts unitaires et leur confère un avantage concurrentiel.

En second lieu, la gestion des données suppose la mise en place d'une gouvernance structurée, visant à organiser l'accès aux données, à en assurer la qualité, la traçabilité et la sécurité, ainsi qu'à documenter leur origine et leur transformation. Ces exigences revêtent une importance toute particulière dans le contexte de l'intelligence artificielle, dans la mesure où la qualité, la représentativité et la fiabilité des données conditionnent directement les performances des modèles. Ce besoin de gouvernance est d'ailleurs consacré et renforcé par le cadre réglementaire de l'Union européenne. Le Règlement sur l'intelligence artificielle impose, en particulier pour les systèmes à haut risque, des exigences spécifiques en matière de qualité et de gouvernance des données ainsi que la mise en place de procédures formalisées de contrôle, de documentation et de gestion du cycle de vie des jeux de données, afin de pouvoir démontrer leur conformité.

Ces différents éléments tendent à montrer que la gestion des données constitue une source de coûts fixes non négligeable, susceptible d'avantager les acteurs déjà établis. Toutefois, ces contraintes ne doivent pas être interprétées comme des barrières à l'entrée insurmontables. En effet, le recours à des services de cloud permet aux nouveaux entrants d'accéder à des infrastructures performantes sans supporter l'intégralité des coûts initiaux[55].

Ainsi, si le cycle de vie des données introduit des frictions et des coûts qui peuvent influencer la structure concurrentielle du marché, il ne suffit pas, à lui seul, à verrouiller l'accès au marché de l'IA générative. Il confirme plutôt l'idée que l'avantage concurrentiel ne réside pas uniquement dans la détention de données, mais dans la capacité à les exploiter efficacement, une capacité qui, bien que coûteuse, demeure accessible à une pluralité d'acteurs.

[55] A. Manganelli, *Foundation models and generative AI applications: what competitive concerns ?*, European Competition Journal, p3.

e) Les boucles de rétroaction des données

La littérature distingue deux types de boucles de rétroaction des données : les boucles de « *within-user learning »*, où les données d'un utilisateur améliorent principalement le service pour ce même utilisateur en générant des coûts de changement, et les boucles d' « *across-user learning »*, où les données de chaque utilisateur contribuent à améliorer le service pour l'ensemble des utilisateurs. Ces dernières étant susceptibles de générer des effets de réseau plus significatifs et, partant, des dynamiques de type *winner-takes-all* davantage préoccupantes du point de vue du droit de la concurrence[56].

L'exploitation des données des utilisateurs est ainsi souvent présentée comme un puissant mécanisme de rétroaction : plus un service est utilisé, plus il génère de données susceptibles d'affiner les modèles sous-jacents, ce qui améliore la qualité perçue et renforce, en retour, son attractivité. Cette dynamique est d'autant plus préoccupante qu'elle s'appuie sur une structure de coûts typique des marchés numériques : des coûts de développement élevés, mais un coût marginal quasi nul pour chaque utilisateur supplémentaire. Ainsi, plus le nombre d'utilisateurs augmente, plus le coût moyen diminue, renforçant l'avantage des acteurs déjà en place.[57]. Certaines autorités de concurrence, parmi lesquelles l'Autorité française de la concurrence et la Commission européenne, soulignent précisément le risque que ces boucles de rétroaction renforcent la position des acteurs déjà bien dotés en données, en particulier lorsque les données d'usage alimentent automatiquement et de manière continue l'amélioration des modèles, rendant la contestabilité du marché théoriquement possible mais pratiquement difficile. Cette analyse doit cependant être nuancée au regard des spécificités techniques et économiques du marché des modèles d'IA générative. Hagiu et Wright démontrent en effet que, contrairement à ce que suggère une approche purement théorique, les données des utilisateurs ne sont ni systématiquement exploitées, en raison de contraintes juridiques, contractuelles ou de confidentialité, notamment liées au RGPD, ni nécessairement uniques, et ne génèrent pas toujours des signaux suffisamment riches pour conférer un avantage concurrentiel durable. Dans le contexte spécifique des grands modèles de langage, les données de rétroaction recueillies auprès des utilisateurs demeurent faibles, consistant

[56] A. HAGIU et J. WRIGHT, « Artificial intelligence and competition policy », *International Journal of Industrial Organization*, 2025, p.7.

[57] M. ALMADA, J. MARANHAO, et G. SARTOR, ibidem, p.116

essentiellement en des évaluations binaires « j'aime/je n'aime pas ». De plus, les entreprises déployant ces modèles via des interfaces ou des API ne rétrocèdent généralement pas les données d'usage au fournisseur du modèle, que ce soit pour des raisons stratégiques ou de conformité réglementaire. L'évolution récente du marché des modèles d'IA générative semble conforter cette analyse : malgré son avance initiale avec ChatGPT, OpenAI n'a pas empêché l'émergence de concurrents comme Google ou Anthropic, qui ont réussi à capter des parts de marché significatives en s'appuyant sur leurs propres écosystèmes, canaux de distribution et choix technologiques distincts. Dans ce secteur, la boucle de rétroaction liée aux données semble donc offrir surtout un avantage supplémentaire limité, dont l'importance diminue rapidement une fois qu'un certain volume de données est atteint, plutôt que de constituer une véritable barrière à l'entrée au sens du droit européen de la concurrence.

L'ensemble de ces éléments suggère que, si les données peuvent conférer des avantages concurrentiels réels, elles ne constituent pas, en l'état, une barrière à l'entrée absolument déterminante et durable sur le marché de l'IA générative. La diversité des sources, les différents modes d'accès et l'importance croissante de la qualité et de la curation atténuent la portée des asymétries de volume pur, sans toutefois les faire entièrement disparaître. Ces constats expliquent que le législateur européen ait fait des données un axe central de sa stratégie numérique, en cherchant à encadrer leur circulation et à corriger certaines asymétries d'accès : c'est l'objet du chapitre suivant.

Chapitre 3 La stratégie européenne pour les données

L'Union européenne affiche l'ambition de « créer un marché unique des données qui garantira la compétitivité mondiale et la souveraineté de l'Europe en matière de données. Cela conduira à la création d'espaces européens communs des données. Ils veilleront à ce que davantage de données soient disponibles pour une utilisation dans l'économie et la société, tout en gardant les entreprises et les individus qui génèrent les données sous contrôle »[58]. Reste à déterminer dans quelle mesure ces textes permettent effectivement de faciliter l'accès aux données, en particulier à celles qui sont pertinentes pour l'entraînement et l'utilisation des modèles d'IA

[58] COMMISSION EUROPÉENNE, « Une stratégie européenne pour les données », disponible sur digital-strategy.ec.europa.eu, consulté le 16/04/2026

générative. Ce chapitre examine d'abord les instruments horizontaux que sont le Data Governance Act (DGA) et le Data Act (section 1), qui visent à organiser le partage et la gouvernance des données dans l'économie européenne. Il se penche ensuite sur le Digital Market Act (DMA), dont certaines obligations de partage de données imposées aux contrôleurs d'accès peuvent contribuer, au moins partiellement, à limiter les avantages concurrentiels des grandes plateformes (section2), avant d'aborder l'exemple sectoriel de l'Espace européen des données de santé (section3).

Section 1 DGA et Data Act

Ces deux instruments appréhendent la donnée comme un bien porteur de valeur économique[59]. Ils visent, en ce sens, à encourager le partage de données dans des conditions plus sûres et plus prévisibles, afin de lever certains freins contractuels ou organisationnels à leur circulation. Parallèlement, ils poursuivent un objectif d'amélioration de la gouvernance des données, en imposant des exigences plus strictes en matière de qualité, de traçabilité, de documentation et de répartition des droits, de manière à encadrer l'accès aux données et leur usage tout au long de leur cycle de vie.

Le DGA crée un cadre horizontal destiné à faciliter le partage de données en instaurant notamment la catégorie des « services d'intermédiation de données », définis comme des services visant à « établir des relations commerciales de partage de données entre un nombre indéterminé de personnes concernées ou de détenteurs de données, d'une part, et des utilisateurs de données, d'autre part » [60] . Il entend ainsi promouvoir l'émergence d'intermédiaires de confiance chargés d'organiser la mise à disposition de données dans des conditions transparentes et non discriminatoires. Il facilite aussi le partage de donnée sur une base volontaire.

Le Data Act vise principalement les données générées par l'usage de produits et services connectés (Internet des objets ou IoT). Il poursuit un double objectif : renforcer le contrôle

[59] C. FISCHER, « Gouvernance par le consentement, big data et nouveau droit européen des données : dans quelle mesure le Data Act et le Data Governance Act favorisent-ils le big data ? », communication présentée au colloque *Les plateformes numériques et l'intelligence artificielle* (Conférence du Jeune Barreau de Bruxelles), Bruxelles, 22 mars 2024, point 51.

[60] Article 2, 11° du Règlement (UE) 2022/868 du Parlement européen et du Conseil du 30 mai 2022 portant sur la gouvernance européenne des données et modifiant le règlement (UE) 2018/1724, *J.O.U.E.*, L 152, 3 juin 2022.

des utilisateurs sur les données qu'ils contribuent à générer et faciliter le partage de ces données entre différents acteurs, qu'il s'agisse de relations B2B ou B2C. Ainsi, il impose notamment que les utilisateurs aient la possibilité de recevoir les données ou d'en autoriser la transmission à des tiers, sur la base de droits d'accès clairement définis et de conditions contractuelles encadrées, sans pour autant reposer systématiquement sur le consentement au sens du célèbre RGPD.

Dans le contexte de l'IA générative, cela signifie que l'utilisation de données issues de l'Internet des objets à grande échelle impliquerait, en pratique, d'obtenir le consentement d' un nombre très important d'utilisateurs ou de partenaires contractuels afin de partager ces données dans le respect de ce cadre juridique, ce qui limite l'accès aux données.

Cependant, ni le DGA ni le Data Act ne sont conçus pour organiser un accès généralisé aux corpus utilisés pour l'entraînement des modèles d'IA générative : le premier se concentre sur des mécanismes d'intermédiation et de partage volontaire, tandis que le second porte sur les données issues de l'IoT. Leur contribution à la réduction des asymétries d'accès aux données pertinentes pour les modèles de fondation reste donc indirecte et limitée, ce qui laisse subsister, en pratique, une partie des barrières identifiées par les autorités de concurrence.

Section 2 DMA

Le DMA a pour objet d'instaurer, dans l'ensemble des États membres, un cadre harmonisé destiné à prévenir certaines pratiques abusives des grandes plateformes numériques qualifiées de « contrôleurs d'accès ». Ces derniers sont soumis à un ensemble de contraintes comportementales destinées à garantir la contestabilité et l'équité des marchés numériques concurrentiels.

Pour être désignée comme contrôleur d'accès, une entreprise doit avoir « un poids important sur le marché intérieur », fournir « un service de plateforme essentiel qui constitue un point d'accès majeur permettant aux entreprises utilisatrices d'atteindre leurs utilisateurs finaux » parmi lesquels figurent expressément les moteurs de recherche en ligne et jouir « d'une position solide et durable, dans ses activités, ou jouira, selon toute probabilité, d'une telle

position dans un avenir proche »[61]. C'est à ce titre qu'Alphabet, maison mère de Google, a été désignée contrôleur d'accès par la Commission dès le 6 septembre 2023. Parmi ses obligations, l'article 6(11) du DMA impose aux contrôleurs d'accès de partager, à des conditions équitables, raisonnables et non discriminatoires, les données générées par l'usage de leurs services de plateforme essentiels.

Moteur de recherche Google

Google concentre aujourd'hui près de 90 % du marché mondial des moteurs de recherche, une position dominante construite et consolidée sur plus d'une décennie. Cependant, l'apparition des IA génératives a le potentiel de redistribuer les cartes[62]. Les Etats-Unis et l'Europe ont imposé à Google de partager certaines données issues de son moteur de recherche.

C'est d'abord aux États-Unis que la question du partage des données de recherche comme instrument de remédiation concurrentielle a été tranchée. Dans son opinion rendue le 2 septembre 2025 dans l'affaire United States et al. v. Google LLC[63], le tribunal fédéral du District de Columbia a ordonné à Google de partager avec les « concurrents qualifiés » certaines données d'index de recherche et certaines données utilisateurs (requêtes, clics, impressions). Le tribunal fonde cette obligation sur le constat que l'avantage de Google en matière d'échelle (Google recevant neuf fois plus de requêtes quotidiennes que l'ensemble de ses rivaux réunis, et dix-neuf fois plus sur mobile) constitue un « fruit » direct de ses accords de distribution exclusifs jugés contraires au Sherman Act, et non le seul produit de sa supériorité technique. Plus encore, anticipant les mutations du marché, le tribunal étend explicitement le périmètre de ces obligations aux acteurs de l'intelligence artificielle générative, citant OpenAI, Anthropic et Perplexity, en modifiant la définition de « concurrent qualifié » pour y inclure tout fournisseur de produits système d'IA générative susceptible de concurrencer Google dans la fonction d'accès à l'information. Le tribunal constate qu'Eddy Cue, vice-président senior

[61] Article 3 du Règlement (UE) 2022/1925 du Parlement européen et du Conseil du 14 septembre 2022 relatif aux marchés contestables et équitables dans le secteur numérique et modifiant les directives (UE) 2019/1937 et (UE) 2020/1828 (règlement sur les marchés numériques), *J.O.U.E.*, L 265, 12 octobre 2022.

[62] C. CABRAL et al., « The challenge of AI encroachment on online search and advertising », *Bruegel Working Paper*, n° 19/2024, disponible sur www.bruegel.org, consulté le 14/04/2026

[63] U.S. Dist. Ct. D.D.C., *United States et al. v. Google LLC*, 2 septembre 2025.

d'Apple, a témoigné que le volume de requêtes Google Search dans le navigateur Safari avait diminué pour la première fois en vingt-deux ans, vraisemblablement sous l'effet de l'essor des systèmes d'IA générative. La décision américaine, fondée sur une approche contentieuse *ex post*, reconnaît que la position dominante dans le domaine de la recherche en ligne peut désormais entraîner une domination dans le secteur de l'intelligence artificielle. Elle souligne ainsi que pour y remédier efficacement, il est nécessaire d'élargir le champ des obligations au-delà du marché strictement défini lors du jugement en responsabilité.

Cette approche trouve un prolongement, quelques mois plus tard, dans la démarche engagée par la Commission européenne sur le fondement du DMA, dans une logique *ex ante*.

C'est sur ce fondement que la Commission a notifié à Google, le 16 avril 2026, des mesures préliminaires précisant que celui-ci devrait permettre à des moteurs de recherche tiers d'accéder aux données de classement, de requêtes, de clics et de consultations. Consciente que l'essor récent des IA génératives a le potentiel de reconfigurer les usages de la recherche en ligne, ces systèmes étant désormais capables d'interroger le Web ou des moteurs de recherche via des API, la Commission observe les premiers signes d'une érosion de la part de marché de Google et a expressément inclus les services d'intelligence artificielle dotés de fonctionnalités de recherche parmi les bénéficiaires de ces obligations de partage. Elle poursuit ainsi la même logique que le tribunal américain : prévenir que la position dominante historique de Google dans la recherche en ligne ne se transfère et ne se consolide dans l'ère des grands modèles de langage. Une décision finale contraignante est attendue au plus tard le 27 juillet 2026.

En réduisant l'asymétrie de données dont bénéficie Google, ce partage vise précisément à accroître la concurrence sur l'accès à ces données critiques, en facilitant le développement de services de recherche concurrents, y compris ceux qui reposent sur des modèles d'IA générative.

Section 3 Espace européen des données de santé

Dans le domaine de la santé, l'Union européenne a engagé une démarche spécifique pour faciliter le partage et la réutilisation des données de santé tout en maintenant un niveau élevé

de protection des patients. Elle a ainsi adopté le Règlement (UE) 2025/327[64] relatif à l'Espace européen des données de santé (EHDS), qui vise à établir un cadre commun pour l'utilisation et l'échange de données de santé dans l'ensemble de l'UE. Le règlement poursuit un double objectif : d'une part, améliorer l'accès transfrontalier aux données de santé pour les soins (usage primaire) ; d'autre part, permettre la réutilisation sûre et fiable de ces données à des fins de recherche, d'innovation, d'élaboration des politiques publiques et de réglementation (usage secondaire)[65]. Dans cette perspective, le législateur européen entend stimuler l'économie des données de santé et favoriser l'émergence de nouvelles solutions en matière de soins et d'innovations fondées sur l'IA ; le considérant 61 mentionne explicitement l'entraînement, le test et l'évaluation d'algorithmes comme exemple d'usage secondaire autorisé.

Le règlement prévoit la création d'organismes d'accès aux données de santé. Ils doivent être créés par chaque Etat membre. Ils auront pour mission de faciliter l'accès aux données de santé par la délivrance d'un permis de données si certaines conditions sont remplies[66].

La difficulté principale des données de santé est que ce sont, majoritairement et par essence, des données à caractère personnel, souvent sensibles au regard du RGPD, ce qui rend leur partage et leur utilisation particulièrement encadrés. L'EHDS ne remet pas en cause ce régime de protection, mais organise un partage simplifié. En effet, les données peuvent être mises à disposition de tiers, pour des finalités précisément définies et sous des conditions strictes.

Sur le plan temporel, le règlement EHDS est entré en vigueur le 26 mars 2025, mais ses dispositions ne s'appliqueront que progressivement. Les premières obligations opérationnelles liées à l'usage secondaire, y compris celles qui concernent l'entraînement d'algorithmes d'IA ne prendront effet qu'à partir de 2029[67] pour certaines catégories de données, dans le cadre d'un calendrier de mise en œuvre échelonné jusqu'en 2035. Cette montée en charge graduelle signifie que, si l'EHDS est appelé à devenir un instrument d'accès

[64] Règlement (UE) 2025/327 du Parlement européen et du Conseil du 11 février 2025 relatif à l'espace européen des données de santé, *J.O.U.E.*, L, 21 février 2025.

[65] COMMISSION EUROPEENE, « Règlement relatif à l'espace européen des données de santé », disponible sur europa.eu, consulté le 19/04/2026

[66] T. Minssen et al., « Governing AI in the European Union: emerging infrastructures and regulatory ecosystems in health », *Research Handbook on Health, AI and the Law,* Cheltenham, Edward Elgar Publishing Ltd; 2024, p.328.

[67] X, « L'Espace Européen des Données de Santé », disponible sur hda.belgium.be, consulté le 30/04/2026

régulé aux données de santé pour l'IA, son impact concret sur les barrières d'accès à court terme reste encore limité et dépendra largement de la manière dont les États membres mettront en place les structures d'accès et les environnements techniques prévus par le règlement.

L'EHDS s'inscrit clairement dans la stratégie européenne visant à structurer un véritable marché unique des données, en particulier dans un secteur sensible comme la santé. En organisant un cadre commun de partage pour les acteurs publics et privés, il facilite l'accès aux données de santé, y compris l'entraînement d'algorithmes d'IA.

Conclusion

Ce travail avait pour ambition de déterminer si les données constituent une barrière à l'entrée sur le marché des intelligences artificielles génératives. L'analyse menée à travers trois axes complémentaires, la compréhension du marché et des préoccupations des autorités de concurrence ; l'examen de l'accessibilité liés aux données ; et l'évaluation de la réponse réglementaire européenne conduit à une réponse nuancée.

Les autorités de concurrence ont identifié très tôt les données comme un intrant stratégique susceptible de conférer un avantage concurrentiel décisif aux acteurs qui en disposent en grande quantité. Cette vigilance anticipée se fonde sur des risques réels : la concentration de vastes corpus entre les mains de quelques grandes plateformes numériques, les effets de réseau générés par les boucles de rétroaction de données, et les pratiques d'accès exclusif ou discriminatoire à des données stratégiques. Ces préoccupations sont légitimes et ne doivent pas être minimisées.

Toutefois, l'analyse approfondie des caractéristiques économiques et techniques des données révèle que leur rôle de barrière à l'entrée est loin d'être absolu. La non-rivalité fondamentale des données implique que leur détention par un acteur n'en prive pas structurellement les autres. La multiplicité des sources d'approvisionnement offre aux nouveaux entrants des voies d'accès alternatives qui réduisent sensiblement les asymétries initiales. Par ailleurs, la recherche récente démontre que, passé un certain seuil, l'ajout de données supplémentaires génère des rendements décroissants. C'est la qualité des données, davantage que leur volume brut, qui détermine in fine la performance des modèles.

Face à ces enjeux, l'Union européenne déploie un arsenal réglementaire ambitieux destiné à faciliter l'accès aux données et à préserver la contestabilité des marchés numériques. Le DGA et le Data Act posent les fondements d'un partage plus fluide des données entre acteurs, tout en renforçant la gouvernance et la traçabilité. Le DMA impose aux contrôleurs d'accès des obligations de partage à des conditions équitables et non discriminatoires : l'injonction adressée à Google de partager ses données de recherche, adoptée conjointement par le tribunal fédéral américain et la Commission européenne, illustre concrètement la manière

dont le droit de la concurrence peut intervenir pour atténuer une asymétrie de données jugée préjudiciable. Enfin, l'espace européen des données de santé (EHDS) ouvre, à un horizon plus lointain, des possibilités d'accès régulé à un corpus de données particulièrement stratégiques pour le développement de l'IA dans le secteur médical.

Cet arsenal réglementaire reste néanmoins partiel. Ni le DGA ni le Data Act ne sont conçus pour résoudre spécifiquement les asymétries d'accès aux corpus d'entraînement pertinents pour les modèles de fondation. L'EHDS n'entrera en phase opérationnelle qu'à partir de 2029 pour certaines catégories de données. Et si le DMA constitue l'instrument le plus immédiatement efficace pour contraindre les contrôleurs d'accès à partager leurs données, son champ d'application reste limité aux grandes plateformes désignées.

En définitive, la question posée en ouverture de ce travail appelle une réponse mesurée et nuancée: les données constituent une barrière à l'entrée réelle, mais relative et non déterminante. Elles confèrent des avantages concurrentiels tangibles aux acteurs qui en disposent en grande quantité et qui savent les exploiter efficacement. Mais elles ne suffisent pas à elles seules à verrouiller durablement le marché, dès lors que d'autres intrants comme la puissance de calcul, les talents et, surtout, la capacité d'innovation algorithmique jouent un rôle tout aussi structurant dans la compétitivité des acteurs de l'IA générative. Plutôt que de traiter la donnée comme une barrière à l'entrée en soi, il convient donc d'adopter une approche empirique et au cas par cas, attentive aux conditions concrètes d'accès aux données stratégiques, à la nature des pratiques qui pourraient les monopoliser, et à l'efficacité réelle des instruments réglementaires mis en place pour en garantir un accès équitable. C'est à cette condition que le droit de la concurrence pourra pleinement accompagner, sans entraver, la dynamique d'innovation qui caractérise l'ère des intelligences artificielles génératives.

Bibliographie

I. Doctrine (Livres et Articles)

ALMADA M., MARANHÃO J., et SARTOR G., « Competition in and through artificial intelligence », *Research Handbook on Competition and Technology*, Cheltenham, Edward Elgar Publishing, 2025 ;

BANKO M. and BRILL E., *Scaling to Very Very Large Corpora for Natural Language Disambiguation*, t. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, 2001 ;

FISCHER, C., « Gouvernance par le consentement, big data et nouveau droit européen des données : dans quelle mesure le Data Act et le Data Governance Act favorisent-ils le big data ? », communication présentée au colloque *Les plateformes numériques et l'intelligence artificielle (Conférence du Jeune Barreau de Bruxelles)*, Bruxelles, 22 mars 2024 ;

HAGIU A. et WRIGHT J., « Artificial intelligence and competition policy », *International Journal of Industrial Organization*, 2025 ;

HOANG, M., « Les données publiques : qu'en est-il de leur ouverture et de leur réutilisation ? », *Cahiers français*, 2021 ;

MANGANELLI, A., « Foundation models and generative AI applications: what competitive concerns? », *European Competition Journal* ;

MINSSEN, T. et al., « Governing AI in the European Union: emerging infrastructures and regulatory ecosystems in health », *Research Handbook on Health, AI and the Law*, Cheltenham, Edward Elgar Publishing Ltd, 2024 ;

PAILLER, L., & BRUNERIE, C., *La cohérence du droit des données*, Paris, Lefebvre Dalloz, 2026 ;

RUSSELL, S. et NORVIG, P., *Intelligence artificielle : une approche moderne*, 4e éd., Pearson, 2021 ;

II. Textes Législatifs et Jurisprudence

Règlement (UE) 2023/2854 du Parlement européen et du Conseil du 13 décembre 2023 concernant des règles harmonisées portant sur l'équité de l'accès aux données et de l'utilisation des données (Data Act) ;

Règlement (UE) 2022/868 du Parlement européen et du Conseil du 30 mai 2022 portant sur la gouvernance européenne des données (Data Governance Act) ;

Règlement (UE) 2022/1925 du Parlement européen et du Conseil du 14 septembre 2022 relatif aux marchés contestables et équitables dans le secteur numérique (Digital Markets Act) ;

Règlement (UE) 2025/327 du Parlement européen et du Conseil du 11 février 2025 relatif à l'espace européen des données de santé ;

U.S. Dist. Ct. D.D.C., United States et al. v. Google LLC, 2 septembre 2025 ;

III. Rapports et Avis des Autorités

AUTORIDADE DA CONCORRÊNCIA (ADC), « Competition and Generative AI – Zooming in on Data », Short Papers Series n°1, 27 septembre 2024 ;

AUTORITÉ DE LA CONCURRENCE, Avis n° 24-A-05 relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative, 28 juin 2024 ;

COMMISSION EUROPÉENNE, « Une stratégie européenne pour les données », digital-strategy.ec.europa.eu, 2026 ;

COMPETITION AND MARKETS AUTHORITY (CMA), « CMA AI strategic update », 29 avril 2024 ;

KOWALSKI, K., VOLPIN, C. et ZOMBORI, Z., « Competition in Generative AI and Virtual Worlds », *Competition Policy Brief*, n°3, 2024 ;

MAY, R., « Intelligence artificielle, données et concurrence », OCDE, DAF/COMP(2024)2, mai 2024 ;

OECD (2025), « Intellectual property issues in artificial intelligence trained on scraped data », *OECD Artificial Intelligence Papers*, No. 33, OECD Publishing ;

OFFICE OF SCIENCE AND TECHNOLOGY POLICY, « Winning the Race: AMERICA'S AI ACTION PLAN », 2025 ;

IV. Documentation internet (Articles en ligne, Blogs et Presse)

ALAILY R., « The New AI Economy : Understanding the Technology, Competition, and Impact for Societal Good », disponible sur https://www.concurrences.com/en/review/numeros/no-2-2024/on-topic/artificial-intelligence-and-antitrust, consulté le 13/02/2026 ;

BOCHAROV A., VARGAS S., MARTINETTI A., TATORIS R. et AZEVEDO C., « Declaring your AIndependence: Block AI bots, scrapers and crawlers with a single click », disponible sur https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/ , consulté le 7/04/2026 ;

BROOKS S. et ROESNER, G., « Data use externalities », disponible sur https://www.gov.uk/government/publications/value-of-data-research-reports-2022, consulté le 20/02/2026 ;

MARTENS B., « The challenge of AI encroachment on online search and advertising », *Bruegel Working Paper*, n° 19/2024, disponible sur https://www.bruegel.org/working-paper/challenge-ai-encroachment-online-search-and-advertising, consulté le 14/04/2026 ;

CHERREY B., « Reddit va faire payer l'accès à son API à certains utilisateurs », disponible sur https://www.zdnet.fr/actualites/reddit-va-faire-payer-l-acces-a-son-api-a-certains-utilisateurs-mauvaise-nouvelle-pour-les-entreprises-d-ia-39957350.htm, consulté le 05/04/2026 ;

CNIL, « Les pratiques open source en intelligence artificielle », cnil.fr , consulté le 26/03/2026 ;

FRANCE NUM, « L'intelligence artificielle (IA), à quoi ça sert ? », disponible sur https://www.francenum.gouv.fr/guides-et-conseils/intelligence-artificielle/generation-de-contenus-texte-image-son-video/une, consulté le 10/02/2026 ;

GENG X., GUDIBANDE A., LIU H., WALLACE E., ABBEEL P., LEVINE S. et SONG D., « Koala: A Dialogue Model for Academic Research », disponible sur https://bair.berkeley.edu/blog/2023/04/03/koala/, consulté le 25/02/2026 ;

GLADY N., « La Data n'est pas le nouveau pétrole », disponible sur https://www.hbrfrance.fr/innovation/la-data-nest-pas-le-nouveau-petrole-6232, consulté le 19/10/2025 ;

IBM, « Qu'est-ce que les données synthétiques ? », disponible sur https://www.ibm.com/fr-fr/think/topics/synthetic-data, consulté le 15/02/2026 ;

LE MONDE, « Meta signe de nouveaux accords avec des médias pour intégrer leurs contenus à son assistant IA », disponible sur https://www.lemonde.fr/pixels/article/2026/03/14/meta-signe-de-nouveaux-accords-avec-des-medias-pour-integrer-leurs-contenus-a-son-assistant-ia_6671107_4408996.html, consulté le 13/02/2026 ;

LONGO R. et ROCHA M., « Generative AI : The new digital frontier for competition », *https://www.concurrences.com/en/review/numeros/no-2-2024/on-topic/artificial-intelligence-and-antitrust*, consulté le 30/03/2026 ;

MERGEL T., « NYT v. OpenAI: The Times's About-Face », disponible sur https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/, consulté le 08/04/2026 ;

NEW YORK POST, « Meta reaches AI deals with CNN, Fox News, other media outlets », disponible sur https://nypost.com/2025/12/05/business/meta-reaches-ai-deals-with-cnn-fox-news-other-media-outlets/ , consulté le 13/02/2026 ;

NEWSCORP, « News Corp and OpenAI Sign Landmark Multi-Year Global Partnership », disponible sur https://newscorp.com/2024/05/22/news-corp-and-openai-sign-landmark-multi-year-global-partnership/, consulté le 08/04/2026 ;

OEYEN T. et YARGICI Y., « Uncharted territories : Generative AI, merger control and the Microsoft-Open AI saga », disponible sur https://www.concurrences.com/en/review/numeros/no-2-2024/on-topic/artificial-intelligence-and-antitrust , consulté le 18/03/2026 ;

OPEN AI, « The Washington Post partners with OpenAI », disponible sur https://openai.com/global-affairs/the-washington-post-partners-with-openai/, consulté le 15/04/2026 ;

PIQUARD A., « Intelligence artificielle : la bataille des modèles en open source », disponible sur https://www.lemonde.fr/economie/article/2025/03/13/intelligence-artificielle-la-bataille-des-modeles-en-open-source_6580360_3234.html, consulté le 29/03/2026 ;

REUTERS, « Global news publisher Axel Springer partners with OpenAI in landmark deal », disponible sur https://www.reuters.com/business/media-telecom/global-news-publisher-axel-springer-partners-with-openai-landmark-deal-2023-12-13/, consulté le 15/04/2026 ;

SAJADIEH S., FATTORINI L., PERRAULT R., GIL Y., PARLI V., SANTARLASCI L., PAVA J., MASLEJ N., ALTMAN R., BRYNJOLFSSON E., BRODLEY C., CLARK J., DIGNUM V., KUMAR V., LANDAY J., LYONS T., MANYIKA J., NIEBLES J. C., SHOHAM Y., TABASSI E., WALD R., WALSH T. et WELD D, « The AI Index 2026 Annual Report », Institute for Human-Centered AI, Stanford University, disponible sur https://hai.stanford.edu/ai-index/2026-ai-index-report, consulté le 10/04/2026 ;

SHUMAILOV I., SHUMAYLOV Z., ZHAO Y., PAPERNOT N., ANDERSON R. et GAL Y., « AI models collapse when trained on recursively generated data », disponible sur https://www.nature.com/articles/s41586-025-08905-3, consulté le 14/03/2026 ;

VESTAGER M., CARDELL S., KANTER J. et KHAN L. M, « Joint Statement on Competition in Generative AI Foundation Models and AI Products », disponible sur https://competition-policy.ec.europa.eu/about/news/joint-statement-competition-generative-ai-foundation-models-and-ai-products-2024-07-23_en, consulté le 12/02/2026 ;

X, « Common Crawl Overview », disponible sur https://commoncrawl.org/, consulté le 07/04/2026 ;

X, « Hugging Face Datasets », disponible sur https://huggingface.co/, consulté le 07/04/2026 ;

X, « Kaggle Datasets », disponible sur *www.kaggle.com*, consulté le 07/04/2026 ;

X, « LAION Projects », disponible sur https://laion.ai/, consulté le 07/04/2026 ;

X, « Qu'est-ce que la curation de données ? », disponible sur https://www.ibm.com/fr-fr/think/topics/data-curation, consulté le 27/04/2026 ;

X, « X interdit l'utilisation de son contenu ou de son API pour entraîner des grands modèles d'IA », disponible sur https://opentermsarchive.org/fr/memos/x-interdit-lutilisation-de-son-contenu-ou-de-son-api-pour-affiner-ou-entrainer-des-grands-modeles-dia/, consulté le 05/04/2026 ;

ZÚÑIGA M. et AUER D., « Don’t Believe the Hype (on Competition and AI) », disponible sur https://truthonthemarket.com/2024/08/13/dont-believe-the-hype-on-competition-and-ai/, consulté le 26/03/2026 ;