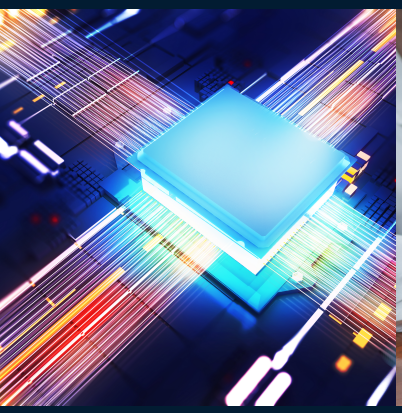
RÉPUBLIQUE FRANÇAISE

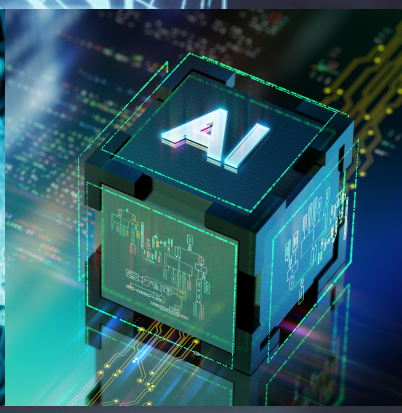
Autorité
de la concurrence

AVIS 24-A-05

du 28 juin 2024

relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative



RÉPUBLIQUE FRANÇAISE

Autorité de la concurrence

Avis n° 24-A-05 du 28 juin 2024 relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative

L'Autorité de la concurrence (section IA),

Vu la décision n° 24-SOA-01 du 7 février 2024, enregistrée sous le numéro 24/0007 A relative à une saisine d'office pour avis portant sur le secteur de l'intelligence artificielle générative ;

Vu le livre IV du code de commerce ;

Vu le document de consultation publique relative à la saisine d'office pour avis portant sur le secteur de l'intelligence artificielle générative publié par l'Autorité de la concurrence le 8 février 2024 ;

Vu les contributions reçues jusqu'au 22 mars 2024 ;

Vu les autres pièces du dossier ;

Les représentants des sociétés Google, Microsoft, Mistral AI, Orange, France Digitale et de la Direction générale des entreprises entendus sur le fondement des dispositions du deuxième alinéa de l'article L. 463-7 du code de commerce ;

Les rapporteurs, le chef du service de l'économie numérique et le commissaire du Gouvernement entendus lors de la séance du 29 mai 2024 ;

Adopte l'avis suivant :

Résumé[1]

Depuis le lancement public de l'agent conversationnel ChatGPT, créé par l'entreprise OpenAI, en novembre 2022, l'intelligence artificielle (ci-après « IA ») générative a pris une place centrale dans le débat public et économique. Les questions qu'elle soulève vont de l'éthique au respect de la propriété intellectuelle ou encore à son impact sur le marché du travail et la productivité. Elle offre de nombreuses possibilités aux entreprises en termes, par exemple, de création de contenu, de conception graphique, de collaboration entre salariés ou de service aux clients.

Les bénéfices de l'IA générative ne se matérialiseront que si l'ensemble des ménages et des entreprises ont accès à une diversité de modèles adaptés à leurs cas d'usage. Il est dès lors essentiel que le fonctionnement concurrentiel du secteur soit favorable à l'innovation et permette la présence d'une multiplicité d'acteurs.

L'IA générative

L'IA est définie par le Parlement européen comme tout outil utilisé par une machine afin de « *reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité* ». **On parle d'IA générative pour désigner des modèles d'IA capables de créer de nouveaux contenus qui peuvent par exemple être du texte, des images, du son ou des vidéos**.

Deux phases essentielles de la modélisation de l'IA générative sont à différencier :

- **l'*entraînement* désigne le processus d'apprentissage initial d'un modèle (souvent appelé « modèle de fondation », qui inclut les grands modèles de langage, en anglais *Large Language Model* ou LLM)**, à l'occasion duquel ses paramètres, appelés « poids », seront déterminés. Cet entraînement nécessite à la fois une grande puissance de calcul et un volume important de données, généralement publiques. Il peut, éventuellement, être complété par une phase de spécialisation, aussi appelée réglage fin (en anglais, « *fine tuning* »), pendant laquelle le modèle va être adapté à une tâche particulière, comme par exemple répondre à des questions des utilisateurs finaux, ou à un ensemble de données spécialisées (par exemple, juridiques ou de santé). La spécialisation se fait généralement à partir d'un ensemble plus petit de données propriétaires, et peut faire appel à l'expertise humaine ;
- **l'*inférence* correspond à l'utilisation du modèle, une fois entrainé, pour créer du contenu**. Le modèle peut être mis à disposition des utilisateurs via une application, comme ChatGPT de la société OpenAI ou Le Chat de Mistral AI, ou une interface de programmation pour les développeurs. La puissance de calcul nécessaire dépend du nombre d'utilisateurs. À l'inverse de nombreux services numériques, le coût marginal de l'IA générative n'est pas négligeable, compte tenu du coût de ce calcul. L'inférence peut également être complétée par l'apport de nouvelles données (qui n'ont pas été utilisées pour l'entraînement), pour ancrer le modèle (en anglais « *grounding* »), par exemple dans l'actualité la plus récente.

[1] Ce résumé a un caractère strictement informatif. Seuls font foi les motifs de l'avis numérotés ci-après.

Les acteurs de la chaîne de valeur

La chaîne de valeur de l'IA générative va, en amont, de la conception, l'entraînement et l'inférence des modèles jusqu'à leur usage, en aval, par des utilisateurs finaux. Les acteurs actifs dans cette chaîne de valeur sont :

- les **grands acteurs du numérique** : Alphabet et Microsoft sont intégrés verticalement et de manière conglomérale sur toute la chaîne de valeur, tandis qu'Amazon, Apple, Meta et Nvidia sont seulement présents sur certaines couches spécifiques ;
- les **développeurs de modèles** : il peut s'agir de *start-ups* ou de laboratoires de recherche spécialisés en IA, comme Anthropic, HuggingFace Mistral AI, OpenAI. Ces développeurs ont souvent noué des partenariats avec un ou plusieurs géants du numérique, comme OpenAI avec Microsoft et Anthropic avec Amazon et Google. Ils peuvent adopter une approche plus ou moins ouverte quant à l'information disponible sur leurs modèles et la possibilité de les réemployer et de les adapter.

À l'amont, plusieurs types d'acteurs sont présents :

- les **fournisseurs de composants informatiques** développent des processeurs graphiques (en anglais, *Graphics Processing Unit* ou GPU) et des accélérateurs d'IA, qui sont des composants indispensables à l'entraînement de modèles d'IA générative. Outre Nvidia, principal acteur du secteur et première capitalisation boursière mondiale à la date de publication de cet avis, et des grands acteurs du numérique qui conçoivent leurs propres accélérateurs d'IA, ce secteur comprend également des acteurs historiques comme Advanced Micro Devices (AMD) et Intel ;
- les **fournisseurs de services d'informatique en nuage (*cloud*)** jouent un rôle important dans le développement de nouvelles technologies d'IA car ils fournissent les capacités de stockage, de traitement des données et de calcul nécessaires, notamment, aux développeurs de modèle de langage. Ils comprennent à la fois des géants du numérique, appelés « *hyperscalers* », comme Amazon Web Services (AWS), Google Cloud Platform (GCP) et Microsoft Azure, des fournisseurs de *cloud* comme 3DS Outscale, IBM et OVHcloud ainsi que des fournisseurs spécialisés dans l'IA comme CoreWeave. Le secteur du *cloud* a été décrit par l'Autorité dans son avis n° 23-A-08 du 29 juin 2023. Les ressources de calcul nécessaires peuvent également être fournies par les **supercalculateurs publics** (comme le supercalculateur Jean Zay en France), historiquement dédiés au calcul de haute performance et qui se sont diversifiés pour accueillir des projets de recherche en IA.

À l'aval, de nombreux acteurs commercialisent de nouveaux services fondés sur l'IA générative à destination du grand public (comme ChatGPT), des entreprises et des acteurs publics et/ou intègrent l'IA générative dans leurs services existants (comme Zoom).

Une priorité croissante des autorités publiques

Le secteur de l'IA générative fait l'objet d'une attention accrue à travers le monde.

En France, le Gouvernement a lancé en 2018 une **stratégie nationale pour l'IA** visant à doter la France de capacités de recherche compétitives et à diffuser les technologies d'IA au sein de l'économie. En mars 2024, la Commission de l'IA mise en place par le Premier ministre a présenté 25 recommandations, préconisant notamment de faire de la France un pôle majeur de la puissance de calcul, de faciliter l'accès aux données et la mise en place d'une gouvernance mondiale de l'IA.

En Europe, plusieurs textes législatifs encadrant le développement du secteur de l'IA ont été adoptés ces deux dernières années. En particulier, le règlement européen sur l'IA (en anglais, **AI Act**), en cours de publication au *Journal officiel* de l'Union européenne, sera applicable pour l'essentiel à partir de 2026. Il impose notamment des obligations de transparence aux fournisseurs de grands modèles d'IA générative et la mise en place de procédures permettant de garantir le respect de la législation européenne sur les droits d'auteur lorsqu'ils entrainent leurs modèles. Ces obligations ne s'appliquent pas aux modèles publiés sous licence libre et ouverte et dont les paramètres sont rendus publics (sauf si ceux-ci présentent un risque systémique). Bien que publiés avant l'essor de l'IA générative, le règlement européen sur les marchés numériques (en anglais, Digital Markets Act ou DMA) et le règlement européen sur les données (en anglais, Data Act) auront également un impact sur le secteur.

Dans le monde, une série d'initiatives sur l'IA ont été adoptées, comme la déclaration commune de Bletchley au Royaume-Uni en novembre 2023 lors du sommet consacré à la sécurité dans l'IA. Le **prochain sommet mondial aura lieu en France les 10 et 11 février 2025**. D'autres initiatives ont été prises par le G7, les États-Unis, le Royaume-Uni ou la Chine, par exemple.

Les barrières à l'entrée sur ce secteur sont élevées

L'accès à une puissance de calcul suffisante pour effectuer un grand nombre d'opérations en parallèle et dotée d'une forte précision pour déterminer plusieurs milliards de paramètres, est essentiel pour développer un modèle de fondation. Les processeurs graphiques développés par Nvidia (associés au logiciel CUDA) ou les accélérateurs d'IA développés par les grands acteurs du numérique (comme les unités de traitement tensoriel – en anglais, *Tensor Processing Unit* ou TPU – de Google) sont indispensables pour l'entraînement, le réglage fin et l'inférence des modèles d'IA générative. Ils sont également très coûteux. Ce secteur est caractérisé depuis 2023 par des difficultés d'approvisionnement dues à une explosion de la demande.

Au-delà des géants du numérique et de quelques entreprises disposant de centres de données suffisamment grands en interne (comme Meta ou Samsung), **le *cloud* apparait comme un passage obligé pour accéder à la puissance de calcul** nécessaire à l'entraînement de modèle. En effet, il permet aux développeurs d'accéder à des services d'infrastructure et de plateforme spécifiques à l'IA, qui correspondent à leurs besoins, tout en évitant des investissements initiaux massifs dans l'infrastructure informatique. **Le *cloud* est également un vecteur de diffusion des modèles en aval sur des places de marché.**

En outre, l'entraînement des grands modèles d'IA générative nécessite de **grands volumes de données**. Celles-ci sont principalement issues de sources publiquement accessibles, comme les pages internet, ou de jeux de données comme les archives Web de Common Crawl, organisation dont l'objectif est de fournir gratuitement des données issues d'Internet depuis 2008. Le nettoyage et le traitement de ces données constituent un facteur différenciant, les acteurs devant les filtrer pour ne garder que le contenu de qualité.

Les acteurs du secteur interrogés dans le cadre de cet avis ont exprimé des **inquiétudes sur l'accès aux données**. D'une part, les modèles sont de plus en plus grands et leur entraînement nécessite toujours davantage de données, ce qui fait craindre que les données publiquement accessibles ne suffisent pas à l'avenir, et que les données propriétaires détenues par un petit nombre d'acteurs prennent plus de poids. D'autre part, l'accès à certaines données publiquement accessibles soulève des **incertitudes juridiques**, comme l'illustrent les actions en justice intentées par plusieurs ayants droit, telle la plainte déposée par le New York Times contre OpenAI et Microsoft.

Enfin, l'entraînement de grands modèles nécessite également des **compétences techniques très poussées** en apprentissage automatique (« *machine learning* »), ainsi qu'une expérience empirique qui ne peut être acquise qu'en travaillant avec ces modèles.

Le besoin en puissance de calcul, en données et en talents entraine **un besoin de financement important** des acteurs du secteur de l'IA générative. Les investissements dans ce secteur ont ainsi été multipliés par 6 entre 2022 et 2023 pour s'établir à plus de 20 milliards d'euros.

Des évolutions techniques ou organisationnelles et certaines politiques publiques peuvent permettre de limiter ces barrières à l'entrée

Premièrement, il est possible d'accéder à la puissance de calcul des **supercalculateurs publics**. En effet, en contrepartie d'une contribution à la science ouverte (publication des travaux dans une revue académique par exemple), l'accès à ces ressources est gratuit, ce qui contribue à réduire les barrières à l'entrée pour certains acteurs, notamment du monde de la recherche. Une équipe de chercheurs de l'école CentraleSupélec a ainsi pu entraîner un modèle appelé CroissantLLM sur le supercalculateur français Jean Zay. L'initiative conjointe EuroHPC œuvre au développement de ces supercalculateurs sur tout le territoire européen, et prévoit **l'installation d'un nouveau supercalculateur en France** en 2025.

Deuxièmement, **certaines** innovations technologiques permettent d'ores et déjà de réduire le besoin en données et en puissance de calcul :

- **des innovations dans l'architecture des modèles** d'IA générative permettent d'améliorer l'entraînement ou le réglage fin afin que celui-ci soit plus performant et moins coûteux. C'est par exemple le cas des combinaisons d'experts (« *Mixture of Experts* » ou MoE) ou de l'adaptation de faible rang (« *Low Rank Adaptation* » ou LoRA) ;
- le secteur s'oriente vers des **modèles de taille réduite**, plus facilement utilisables pour l'inférence, qui peuvent par exemple être utilisés sur des smartphones ;
- **l'utilisation de données synthétiques** (elles-mêmes créées par IA) peut remplacer en partie les données réelles et réduire les contraintes liées à l'utilisation de données personnelles. Néanmoins, elle s'accompagne de certains risques, comme la propagation de biais ou l'augmentation du taux d'erreur.

Enfin, de nombreux développeurs choisissent **une approche ouverte** (dite *open source*) afin de contribuer à l'amélioration de la connaissance globale sur cette technologie, permettant ainsi à d'autres acteurs de réutiliser les modèles ou de les spécialiser. Cependant, l'*open source* recouvre des situations très variables, allant de celles où seuls les poids du modèle sont rendus publics (*open-weights*), la plus répandue, aux modèles totalement ouverts où l'ensemble du code, de l'architecture, des données d'apprentissage, des poids et du processus d'apprentissage sont mis à disposition. Si la publication des poids du modèle peut avoir un impact bénéfique sur la concurrence pour le réglage fin et l'inférence, elle ne réduit cependant pas (ou peu) les barrières pour un acteur souhaitant entraîner un modèle de fondation. L'ouverture d'autres éléments du modèle serait nécessaire pour sa reproduction, comme par exemple le code et les données permettant l'entraînement ou les données utilisées.

Certaines entreprises peuvent bénéficier d'avantages liés à leurs activités sur d'autres marchés numériques

Les grandes entreprises du numérique bénéficient d'un accès privilégié aux intrants nécessaires pour l'entraînement et le développement des modèles de fondation. Ces avantages ne sont pas aisément réplicables par les développeurs de modèles de fondation concurrents qui n'ont pas accès à ces intrants dans les mêmes conditions.

Elles ont en effet **un accès facilité à la puissance de calcul** en tant que partenaires et concurrentes des fournisseurs de puces pour l'IA. D'une part, elles ont la capacité d'acheter en grande quantité et négocier des accords préférentiels avec des fournisseurs de processeurs graphiques comme Nvidia. D'autre part, la plupart de ces grandes entreprises développent également en interne des accélérateurs d'IA spécifiquement adaptés à leurs écosystèmes, comme les TPU de Google ou Trainium d'AWS. Des alternatives au logiciel Cuda de Nvidia commencent également à être développées par ces grandes entreprises.

Ces entreprises bénéficient également **d'un accès privilégié à un large volume de données** (ainsi, YouTube offre à Alphabet une source majeure de données d'entraînement pour les modèles d'IA producteurs de vidéos). Elles peuvent également accéder à des **données associées à l'utilisation de leurs services**. Elles peuvent aussi utiliser leur puissance financière pour conclure des accords avec des propriétaires de données tiers. Google s'est ainsi engagée à payer 60 millions de dollars (environ 55 millions d'euros) par an pour accéder aux données de Reddit, un site communautaire américain de discussions et d'actualités sociales.

Par ailleurs, **un grand nombre de talents** sont attirés par les salaires attractifs ainsi que les perspectives de travail au sein des grandes entreprises du numérique, compte tenu de leur réputation en matière d'innovation, leur positionnement global et la profondeur de leur catalogue de services.

Au-delà d'un accès inégalé aux intrants nécessaires pour l'entraînement de modèles d'IA générative, les grandes entreprises du numérique bénéficient **d'avantages liés à leur intégration verticale et conglomérale qui leur garantit l'accès aux utilisateurs, entreprises et consommateurs.** Le secteur est en effet caractérisé par d'importants coûts fixes occasionnés par l'entraînement initial d'un modèle de fondation, ce qui donne lieu à **des économies d'échelle,** les acteurs cherchant à amortir ces coûts sur le plus grand nombre d'utilisateurs possibles. Les produits d'IA générative se caractérisent également par **des économies de gamme** car, une fois développé, un modèle de fondation peut servir à une grande variété d'applications. Le secteur de l'IA générative peut également conduire à **la présence d'effets de réseaux de nature cumulative**, puisque les données de retour des utilisateurs permettent d'affiner les modèles et d'améliorer leur performance ou de proposer de nouveaux services.

L'Autorité constate également que ces entreprises commencent à intégrer les outils d'IA générative **dans leurs écosystèmes de produits et de services**. Ainsi, Microsoft déploie ses propres modèles et ceux de son partenaire OpenAI dans la fonction « Copilot » afin d'améliorer la fonctionnalité de recherche de Microsoft Bing et avec « Copilot for Microsoft 365 », un assistant IA conçu pour fonctionner avec l'offre Microsoft 365. Par ailleurs, **les places de marché** de ces grandes entreprises permettent d'accéder à des modèles d'IA générative, propriétaires ou de tiers, conçus pour fonctionner dans leur écosystème.

Il existe des risques concurrentiels à l'amont de la chaîne de valeur

S'il semble prématuré de tirer des conclusions définitives sur la définition des marchés pertinents et sur le pouvoir de marché de certains acteurs, il convient néanmoins de rester vigilant car l'accès de ces grandes entreprises du numérique à des intrants clés et les avantages liés à leur intégration verticale et conglomérale créent les conditions d'une forte concentration à leur profit et renforcent leur puissance sur des marchés distincts mais liés, ou connexes, tels que les logiciels de productivité de bureau, les moteurs de recherche ou la publicité en ligne. **Dans certains cas, l'analyse concurrentielle pourra donc utilement reposer sur la constitution ou le renforcement d'écosystèmes plutôt que sur une analyse marché par marché.**

Le recours aux outils traditionnels du droit de la concurrence, **tels que le droit des ententes et surtout l'abus de position dominante**, conserve toute sa pertinence. D'autres outils juridiques pourraient également être mobilisés comme **l'abus de dépendance économique** dans des situations où la position dominante n'est pas caractérisée, ou, s'agissant des pratiques contractuelles, le **droit des pratiques restrictives de concurrence,** dont la mise en œuvre relève, principalement, de la compétence de la Direction générale de la concurrence, de la consommation et de la répression des fraudes (ci-après « DGCCRF ») et des juridictions commerciales.

L'Autorité identifie plusieurs risques d'abus

- **Un risque d'abus au niveau des composants informatiques**

France Digitale, une association représentant un grand nombre de *start-ups* et investisseurs français du numérique, fait état de risques potentiels tels que des **fixations de prix, des restrictions d'approvisionnement, des conditions contractuelles déloyales ou des comportements discriminatoires.** Par ailleurs, des préoccupations relatives à la dépendance du secteur envers le logiciel de programmation de puces **CUDA** de Nvidia, seul environnement parfaitement compatible avec les GPU devenus incontournables pour le calcul accéléré, ont été exprimées. Les récentes annonces d'investissements de Nvidia dans des fournisseurs de services *cloud* spécialisés dans l'IA, tels que Coreweave, suscitent également des inquiétudes.

Ce secteur, qui a fait l'objet d'une opération de visite et saisie inopinée en septembre 2023, est attentivement scruté par les services d'instruction de l'Autorité.

- **Des risques de verrouillage par les grands fournisseurs de services *cloud***

L'Autorité constate que plusieurs pratiques de verrouillage financier et technique, qu'elle a déjà identifiées dans son avis n° 23-A-08 sur le *cloud*, semblent perdurer et même s'intensifier afin d'attirer le plus grand nombre de *start-ups* actives dans le secteur de l'IA générative.

Tout d'abord, des **offres de crédits *cloud* particulièrement élevées** sont proposées notamment à destination des entreprises innovantes du secteur. Des pratiques de **verrouillage technique** ont également été identifiées.

Outre qu'elles pourraient être qualifiées en droit de la concurrence, notamment d'abus de position dominante, certaines de ces pratiques sont également encadrées par la loi n° 2024-449 du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique ou par le règlement européen sur les données (« *Data Act* »).

❖ **Des préoccupations de concurrence concernant l'accès aux données**

Les entreprises innovantes du secteur pourraient être confrontées à des pratiques de **refus d'accès ou d'accès discriminatoire** de la part d'entreprises disposant d'un accès significatif aux données, comme, par exemple, un index web.

Par ailleurs, des accords par lesquels de grandes entreprises du numérique se réserveraient l'exclusivité de l'accès aux données des créateurs de contenu ou leur verseraient des rémunérations importantes difficilement réplicables par leurs concurrents pourraient constituer des pratiques anticoncurrentielles d'entente ou d'abus.

L'accès aux données des utilisateurs constitue également un enjeu majeur. En effet, plusieurs acteurs rapportent que les grandes entreprises du secteur continuent d'utiliser diverses stratégies pour limiter l'accès des tiers aux données de leurs utilisateurs, en faisant un usage abusif de règles juridiques, comme la protection des données personnelles, ou encore de préoccupations de sécurité.

Il convient enfin de relever que les éditeurs expriment de grandes préoccupations liées à l'exploitation de leurs contenus par les fournisseurs de modèles de fondation, **sans l'autorisation des détenteurs de droits**. Dans sa décision n° 24-D-03 dans le dossier des « droits voisins », l'Autorité a ainsi établi que Google avait utilisé, aux fins d'entraînement de son modèle de fondation Gemini (agent conversationnel basé sur le modèle de fondation du même nom et, anciennement « Bard »), des contenus des éditeurs et agences de presse, sans les avertir et sans leur offrir la possibilité effective d'exercer leur droit de retrait. Si cette question soulève des questions de respect des droits de propriété intellectuelle qui vont au-delà du champ d'étude du présent avis, le droit de la concurrence pourrait, sur le principe, appréhender ces questions sur le fondement d'une atteinte à la loyauté des transactions, par exemple, et donc de l'abus d'exploitation.

❖ **Une vigilance particulière s'impose sur les risques liés à l'accès à une main-d'œuvre qualifiée**

En droit de la concurrence, les pratiques mises en œuvre sur les marchés du travail font l'objet d'une vigilance particulière des autorités de contrôle. Au-delà des accords de fixation des salaires entre entreprises, les accords de non-débauchage (« *no-poach* ») peuvent également constituer des pratiques anticoncurrentielles prohibées.

Un sujet de préoccupation additionnel consiste dans **le recrutement, par les géants du numérique, d'équipes entières** (comme par exemple le recrutement par Microsoft d'une grande partie des 70 employés de la *start-up* Inflection) **ou d'employés stratégiques de développeurs de modèles** (comme le bref recrutement par Microsoft de M. X…, le fondateur d'OpenAI avant qu'il ne soit finalement réintégré dans la société). Si cette pratique peut être examinée sous l'angle du contrôle des concentrations, elle peut également s'analyser en une tentative d'exclusion de concurrents.

S'il ressort de l'instruction du présent avis que de telles restrictions ne semblent pas, pour l'instant, soulever de préoccupations particulières des parties prenantes, l'Autorité estime qu'une vigilance s'impose sur ces questions.

❖ **Les modèles en accès libre peuvent entraîner des risques concurrentiels**

Si les modèles en accès libre peuvent permettre d'abaisser les barrières à l'entrée, ils peuvent également susciter des préoccupations de concurrence. En effet, dans certains cas, les conditions d'accès et de réutilisation des modèles ou de certains de leurs composants peuvent conduire au verrouillage des utilisateurs.

❖ Les risques liés à la présence d'entreprises sur plusieurs marchés distincts

L'intégration verticale de certains acteurs du numérique, et leurs écosystèmes de services, sont susceptibles de donner lieu à plusieurs pratiques abusives.

À l'amont, les développeurs de modèles pourraient se voir opposer **un refus ou des limites d'accès à des puces ou des données nécessaires pour entraîner des modèles de fondation concurrents.** Cette pratique pourrait avoir pour effet d'entraîner des retards ou de conduire à la mise en place de modèles moins ambitieux, nuisant ainsi au maintien d'une concurrence effective sur le marché.

Plusieurs acteurs s'inquiètent également **des accords d'exclusivité** entre fournisseurs de services *cloud* et développeurs de modèles de fondation. Selon eux, ces accords viseraient en effet à s'assurer que les développeurs dépendent exclusivement de ces fournisseurs pour l'accès aux services *cloud* nécessaires et pour la distribution aux clients et seraient ainsi susceptibles d'avoir **un impact sur l'innovation** et la concurrence entre fournisseurs, surtout lorsqu'un modèle particulier occupe une position significative sur le marché.

D'autres risques découlent de l'utilisation en aval des modèles d'IA générative, par **des pratiques de ventes liées.** Les entreprises détenant des positions prééminentes ou dominantes sur des marchés connexes à l'IA pourraient lier la vente de ces produits ou services à celle de leurs propres solutions d'IA. **L'intégration d'outils d'IA générative sur certains supports, comme les *smartphones,* suscite notamment des inquiétudes.** Ce type de pratiques pourrait consolider durablement le secteur de l'IA générative autour d'entreprises numériques déjà dominantes.

Les concurrents en aval pourraient également être lésés par des pratiques **d'autopréférence** de la part d'acteurs verticalement intégrés, affectant la capacité des développeurs de modèles non intégrés verticalement à les concurrencer.

L'ensemble de ces comportements pourraient permettre à certaines entreprises d'utiliser le pouvoir de marché qu'elles détiennent sur des marchés distincts mais liés au détriment d'acteurs alternatifs, restreignant les choix offerts aux utilisateurs et l'incitation à développer des solutions alternatives.

Les prises de participation minoritaires et partenariats des géants du numérique peuvent également soulever des préoccupations de concurrence

Dans un secteur comme l'IA, où les investissements sont très élevés compte tenu du coût d'accès aux intrants, seuls quelques grands acteurs disposent des capacités financières pour conclure des accords avec de jeunes entreprises innovantes ou prendre des participations en leur sein. Les investissements et les partenariats entre acteurs du secteur ne sont pas condamnables en soi. Ils peuvent en effet permettre aux *start-ups* de bénéficier de ressources financières et technologiques des grandes entreprises et, ainsi, favoriser l'innovation. Pour l'acquéreur, ces investissements permettent de se diversifier ou d'avoir accès à des technologies innovantes de nature à améliorer la qualité de ses services. Par exemple, Microsoft a noué un partenariat exclusif avec la société OpenAI sous la forme d'un investissement pluriannuel.

Mais ils présentent néanmoins des risques concurrentiels non négligeables qui requièrent une vigilance particulière de la part des autorités de concurrence. Ils peuvent en effet **affaiblir l'intensité concurrentielle** entre les deux entités, entraîner **des effets verticaux**, un **renforcement de la transparence** sur le marché ou un **verrouillage** de certains acteurs.

Les prises de participation minoritaires des grands acteurs peuvent être appréhendées par les autorités de concurrence sur plusieurs fondements. D'une part, ces opérations peuvent être soumises à autorisation préalable dans le cadre du contrôle des concentrations si elles confèrent aux investisseurs un contrôle de fait et dépassent les seuils communautaires et nationaux de notification. Elles peuvent également être examinées, sous certaines conditions, si elles sont en dessous de ces seuils ou en marge d'une opération de concentration. D'autre part, elles peuvent être appréhendées *ex post* sous l'angle du droit des pratiques anticoncurrentielles, sur le fondement du droit des ententes ou de l'abus de position dominante (y compris collective). L'Autorité constate toutefois **un manque de transparence** sur ces accords qui ne permettent pas toujours de déterminer si ceux-ci peuvent nuire à la concurrence et donc aux consommateurs. Ces inquiétudes sont partagées par les autorités de concurrence dans le monde, comme le montrent notamment les investigations en cours concernant Alphabet, Amazon, Anthropic, Microsoft et OpenAI.

Les risques de collusion entre entreprises du secteur

Si la quasi-totalité des parties prenantes interrogées lors de la consultation publique de l'Autorité n'ont pas fait état d'inquiétudes spécifiques sur cette question, l'utilisation de l'IA générative pourrait néanmoins favoriser des pratiques concertées déjà connues et qui ont fait l'objet en novembre 2019 d'une étude commune de l'Autorité et du Bundeskartellamt allemand, comme l'utilisation parallèle d'algorithmes individuels distincts ou le recours à des algorithmes d'apprentissage automatique. Là encore, la vigilance s'impose.

Perspectives

L'Autorité constate que le secteur de l'IA générative est loin d'avoir atteint son potentiel. Moins de deux ans après le lancement de ChatGPT, de nombreux acteurs établis ont investi dans ce domaine et une multitude de jeunes entreprises sont apparues pour accélérer la recherche et diffuser cette technologie auprès des entreprises et des consommateurs.

La course à l'innovation et au développement de nouveaux modèles d'IA générative devrait se poursuivre sur au moins deux axes : la taille des modèles et leur optimisation à taille constante. Il convient de noter que la taille des modèles influe également sur **l'impact environnemental de l'IA** générative.

L'Autorité observe également une tendance à la **« plateformisation** » dans le secteur de l'IA générative. Les places de marché apparaissent comme des points de passage obligatoires pour les développeurs de modèles qui souhaitent atteindre les consommateurs ou les entreprises utilisatrices d'IA.

Un des principaux enjeux pour le bon développement de la concurrence dans le secteur de l'IA générative réside dans la diffusion de ressources ouvertes. Si le secteur bénéficiait de critères plus précis pour qualifier l'ouverture d'un modèle, les acteurs qui le souhaitent pourraient faire valoir cette qualité comme un avantage concurrentiel.

Recommandations

La dynamique concurrentielle du secteur pourrait être renforcée par les recommandations suivantes, qui, pour la plupart, ne nécessitent pas d'initiative législative au niveau français ou européen.

L’Autorité appelle à la pleine utilisation du cadre réglementaire applicable au secteur.

La Commission devrait porter une attention particulière au développement des services permettant l’accès aux modèles d’IA générative dans le *cloud* (en anglais, *Model as a Service* ou MaaS) et évaluer la possibilité de désigner les entreprises fournissant de tels services en tant que contrôleurs d'accès pour ce qui concerne ces services dans le cadre du règlement européen sur les marchés numériques. Un certain nombre de comportements problématiques identifiés *supra* seraient ainsi prohibés *ex ante*.

L’Autorité encourage par ailleurs la DGCCRF à accorder une attention particulière à l’utilisation des avoirs d’informatique en nuage dans le domaine de l’IA, notamment dans le cadre de la mise en œuvre de la loi du 21 mai 2024 visant à sécuriser et réguler l’espace numérique.

Enfin, le futur Bureau européen de l'IA et l'autorité nationale compétente en France, qui sera désignée en application de l’article 70 du règlement sur l’IA, devront s’assurer d'une part que la mise en œuvre du règlement ne freine pas l’émergence ou l’expansion d’opérateurs de taille plus modeste et d'autre part que les plus grands acteurs du secteur ne détournent pas le texte à leur avantage.

L'Autorité appelle aussi au concours des autorités compétentes et à l’utilisation de tous les outils disponibles. Elle s’engage ainsi à rester vigilante dans le secteur de l'IA générative, au côté de la DGCCRF, afin de mobiliser, si nécessaire, l'ensemble de leurs outils respectifs pour agir de manière rapide et efficace.

S’agissant de l’accès à la puissance de calcul, l'Autorité est favorable, comme de nombreux acteurs publics, **au développement des supercalculateurs publics**, qui constituent une alternative aux fournisseurs de *cloud* et permettent à des acteurs académiques notamment d'accéder à la puissance de calcul, ce qui est bénéfique pour l'innovation. L’Autorité est également favorable à leur ouverture, dans certaines conditions, à des opérateurs privés, contre rémunération.

Concernant les données, les autorités publiques, notamment dans le cadre de la mission confiée par la ministre de la Culture au Conseil supérieur de la propriété littéraire et artistique, pourraient inciter les ayants droit à tenir compte de **la valeur économique des données** selon les cas d’usage (en introduisant par exemple des prix différenciés), et à proposer des offres groupées réduisant les coûts de transactions, ceci afin de garantir les capacités d'innovation des développeurs de modèles.

Enfin, l'Autorité appelle à une **transparence accrue** sur les prises de participation minoritaires dans les entreprises innovantes, sur la base de l'article 14 du DMA, qui permet de demander aux entreprises désignées des informations sur leurs opérations d’acquisition.

SOMMAIRE

Introduction

1. Depuis le lancement public de l'agent conversationnel ChatGPT, créé par l'entreprise OpenAI, en novembre 2022, l'intelligence artificielle (ci-après « IA ») générative a pris **une place centrale dans le débat public et économique**. Les questions qu'elle soulève vont de l'éthique au respect de la propriété intellectuelle ou encore à son impact sur le marché du travail et la productivité. Elle offre de nombreuses possibilités aux entreprises en termes, par exemple, de création de contenu, de conception graphique, de collaboration entre salariés ou de support clients. Pour la Commission de l'intelligence artificielle mise en place en France par le Premier ministre (ci-après « Commission de l'IA »), « *l'IA générative constitue un tournant majeur de* [l'] *histoire de l'innovation* ». En effet, « [c]*es caractéristiques de l'IA générative* [Réalisme, simplicité, rapidité, aptitudes] *permettent l'automatisation d'un certain nombre de tâches qui étaient difficilement automatisables auparavant. Par exemple, elles facilitent la personnalisation des offres commerciales, simplifient l'analyse de données financières, accélèrent la recherche scientifique, etc. Ces mêmes caractéristiques laissent penser que l'IA pourrait prendre la suite des ordinateurs personnels, des réseaux sociaux et des smartphones comme « la » plateforme numérique dominante, la couche technologique sur laquelle tous les autres nouveaux services sont construits* »[2].

2. Selon une étude de la direction générale du Trésor[3], il est encore **trop tôt** pour identifier un effet macroéconomique de l'IA sur la croissance. De nombreuses études essaient d'estimer l'impact de l'IA sur la productivité du travail. Selon certaines, bien que l'impact de l'IA générative soit incertain et conditionné aux avancées technologiques, cette innovation seule pourrait augmenter la productivité du travail aux États-Unis de presque 1,5 point de pourcentage par an sur une période de 10 ans après une adoption généralisée[4]. D'autres auteurs s'attendent à un impact plus modeste, inférieur à un point de productivité cumulé sur dix ans[5]. Pour des tâches spécifiques, de premières études suggèrent des effets positifs de l'IA (notamment générative) sur la productivité individuelle de certains travailleurs. Au sein de la profession des conseillers clientèle par exemple, une étude constate un gain de productivité moyen de 14 % pour les conseillers ayant accès à des agents conversationnels, et plus important encore pour les travailleurs les moins expérimentés[6]. Cependant, les emplois les plus qualifiés pourraient également être touchés par l'IA générative, et les emplois de services pourraient être plus affectés que les emplois industriels, ce qui distingue l'IA de précédentes vagues d'innovation qui ont d'abord touché les emplois industriels peu qualifiés.

---

[2] Commission de l'intelligence artificielle, IA : notre ambition pour la France, mars 2024.

[3] Direction générale du Trésor, Trésor-Eco : les enjeux économiques de l'intelligence artificielle, avril 2024.

[4] Goldman Sachs (2023), "The potentially large effects of artificial intelligence on economic growth", Global Economics Analyst.

[5] Daron Acemoglu, The simple macroeconomics of artificial economics, MIT, avril 2024.

[6] Erik Brynjolfsson & Danielle Li & Lindsey R. Raymond, 2023. "Generative AI at Work," NBER Working Papers 31161, National Bureau of Economic Research, Inc.

3. Dans ce contexte, l'Autorité de la concurrence (ci-après « l'Autorité ») a décidé, le 8 février 2024, de **s'autosaisir pour avis** sur le fonctionnement concurrentiel du secteur de l'IA générative[7], sur le fondement de l'article L. 462-4 du code de commerce.

4. Cet avis vise à fournir aux acteurs du secteur une analyse concurrentielle du fonctionnement de ce marché en plein développement. Il se concentre plus particulièrement sur les stratégies mises en place par les grands acteurs du numérique visant à consolider leur pouvoir de marché **à l'amont de la chaîne de valeur de l'IA générative**, **c'est-à-dire dans la conception, l'entraînement et la spécialisation des grands modèles de langage**, ou à tirer parti de ce pouvoir de marché pour se développer dans ce secteur en plein essor. Ainsi, l'Autorité s'intéresse en particulier aux pratiques mises en œuvre par les acteurs déjà présents sur l'infrastructure d'informatique en nuage (*cloud*) et aux problématiques liées à l'accès à ces infrastructures, à la puissance de calcul, aux données et à une main-d'œuvre qualifiée. Elle examine également **les prises de participation et les partenariats des grands acteurs du numérique,** notamment dans des entreprises innovantes spécialisées dans l'IA générative. Elle n'aborde par conséquent qu'à titre incident les pratiques des acteurs à l'aval de la chaîne de valeur, c'est-à-dire au contact du consommateur final, et pas du tout les conséquences de l'IA pour le fonctionnement concurrentiel de l'ensemble de l'économie – question d'importance majeure et qui méritera des analyses ultérieures.

5. L'objet d'un tel avis n'est pas de qualifier des comportements sur un marché au regard des articles 101 et 102 du Traité sur le fonctionnement de l'Union européenne (ci-après « TFUE ») et des articles L. 420-1 et L. 420-2 du code du commerce. Il vise plutôt à améliorer la compréhension du secteur, proposer des éléments d'analyse, esquisser les risques potentiels d'un point de vue concurrentiel, et le cas échéant faire, des recommandations permettant d'en améliorer le fonctionnement.

6. L'Autorité a lancé une consultation publique, ouverte du 8 février au 22 mars 2024, visant à approfondir sa compréhension du secteur.

7. Le document de consultation publique invitait les acteurs du secteur à se prononcer sur les ressources nécessaires au développement des modèles de fondation, le paysage concurrentiel et les pratiques susceptibles d'être mises en place par les acteurs du secteur, les participations minoritaires ainsi que les perspectives du marché. Cette consultation a permis **à plus d'une quarantaine d'acteurs et une dizaine d'associations d'acteurs**, d'une variété de taille et de secteurs d'activité, d'exprimer leur position et leurs éventuelles préoccupations concurrentielles.

8. L'Autorité a conduit en parallèle de nombreux entretiens sur le fondement de l'article L. 450-3 du code de commerce. Elle s'est ainsi entretenue avec des acteurs privés français et internationaux (fournisseurs, clients, associations ou autres) et des acteurs institutionnels (services ministériels, autorités de régulation sectorielle, autorités de concurrence étrangères, etc.). L'Autorité a en particulier échangé avec les autorités ayant mené des travaux approfondis sur les enjeux concurrentiels soulevés par le secteur de l'IA générative, notamment :

[7] Communiqué de presse de l'Autorité, Intelligence artificielle générative : l'Autorité s'autosaisit pour avis et lance une consultation publique jusqu'au vendredi 22 mars, 8 février 2024.

- l'Autorité de la concurrence portugaise (*Autoridade da Concorrência*), qui a publié une étude sur le secteur de l'IA générative le 5 novembre 2023[8] ;
- la Commission européenne (ci-après « Commission »), qui a lancé un appel à contributions sur l'IA générative le 9 janvier 2024[9] ;
- l'Autorité de concurrence et de consommation américaine (*Federal Trade Commission*, ci-après « FTC »), qui a lancé des enquêtes sur les investissements et partenariats dans le secteur de l'IA générative le 25 janvier 2024[10] ;
- l'Autorité de la concurrence et des marchés du Royaume-Uni (*Competition and Markets Authority*, ci-après « CMA »), qui a publié un premier rapport sur les modèles de fondation le 18 septembre 2023, puis une mise à jour le 11 avril 2024[11].

[8] Autoridade da Concorrência, AdC warns of competition risks in the Generative Artificial Intelligence sector, 5 novembre 2023.

[9] Communiqué de presse de la Commission européenne, La Commission lance des appels à contributions sur la concurrence dans les mondes virtuels et l'IA générative, 9 janvier 2024.

[10] FTC, FTC Launches Inquiry into Generative AI Investments and Partnerships, 25 janvier 2024.

[11] CMA, AI Foundation Models : Initial report, 18 septembre 2023 et AI Foundation Models : Update Paper, 11 avril 2024.

I. Le secteur de l'IA générative

A. DEFINITIONS

9. L'IA est définie par le Parlement européen comme tout outil utilisé par une machine afin de « *reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité* »[12]. Cette définition inclut de nombreuses tâches automatisables, comme par exemple la classification, la recommandation de contenus (fréquente sur les médias sociaux), la prédiction ou encore la génération de données.

10. Le futur règlement européen sur l'IA définit un « *modèle d'IA à usage général* » comme « *un modèle d'IA, y compris lorsque ce modèle d'IA est entraîné à l'aide d'un grand nombre de données utilisant l'auto-supervision à grande échelle, qui présente une généralité significative et est capable d'exécuter de manière compétente un large éventail de tâches distinctes, indépendamment de la manière dont le modèle est mis sur le marché, et qui peut être intégré dans une variété de systèmes ou d'applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant leur mise sur le marché* »[13].

11. Le règlement considère les grands modèles d'IA générative comme un exemple typique de modèle d'IA à usage général « *étant donné qu'ils permettent la production flexible de contenus, tels que du texte, de l'audio, des images ou de la vidéo, qui peuvent aisément s'adapter à un large éventail de tâches distinctes* »[14]. Les modèles d'IA à usage général (comprenant les modèles d'IA générative) sont fréquemment appelés **modèles de fondation** (voir glossaire).

12. Au niveau français, la Commission de l'IA considère que « [l]*'IA est qualifiée de générative, car elle permet de générer de nouveaux contenus sous la forme de texte, d'image, de son, de vidéo ou de code* »[15].

13. Les modèles d'IA générative se différencient selon le type de données acceptées en entrée et les données produites en sortie. Ces modèles ont vocation à produire du contenu – texte, images ou vidéos par exemple – généralement à partir d'une requête. Dans le cadre de la génération de texte, on parle fréquemment de **grands modèles de langage** (ci-après « LLM », voir glossaire), mais l'IA générative ne se limite pas au contenu textuel.

14. Les modèles d'IA générative peuvent également être multimodaux, capables de combiner différents types de données en entrée et/ou en sortie. Par exemple, les modèles de génération d'images prennent souvent en entrée du texte, et produisent des images fondées sur ce dernier. D'autres modèles peuvent accepter en entrée une combinaison de texte et d'image.

---

[12] Parlement européen, Intelligence artificielle : définition et utilisation, 7 septembre 2020.

[13] Règlement sur l'intelligence artificielle, article 3(63), 16 avril 2024 (définition inchangée dans la version du 13 juin 2024).

[14] Règlement sur l'IA précité, considérant 99, 16 avril 2024.

[15] Rapport de la Commission IA précité.

B. DEVELOPPEMENT D'UN MODELE D'IA GENERATIVE

15. 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 (en anglais « *fine tuning* », voir glossaire). 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.

1. PHASE D'ENTRAINEMENT D'UN MODELE D'IA GENERATIVE

16. Un modèle d'IA générative est d'abord entrainé pour apprendre des capacités générales, puis il peut être spécialisé sur une tâche spécifique.

a) L'entraînement initial du modèle

17. Selon la Commission nationale de l'informatique et des libertés (CNIL), « *l'entraînement est le processus de l'apprentissage automatique pendant lequel le système d'intelligence artificielle construit un modèle à partir de données* »[16]. Cet entraînement initial permet de déterminer les paramètres, aussi appelés « poids » (en anglais « *weights* », voir glossaire), du modèle.

18. D'après un acteur du secteur « [l]*'entraînement du modèle est basé sur l'évaluation répétitive des prédictions actuelles par rapport aux valeurs cible. Les paramètres sont alors ajustés en mesurant ces résultats, le but étant de construire de façon progressive un modèle de plus en plus performant. Plus la complexité des tâches à réaliser est élevée, plus le modèle nécessite de paramètres* ».

19. Les modèles d'IA générative étant très complexes, ils peuvent avoir plusieurs centaines de millions à plusieurs centaines de milliards de paramètres. Ils mettent en œuvre des techniques telles que l'apprentissage profond (voir glossaire) et les réseaux de neurones (voir glossaire).

20. La majorité des modèles de fondation actuels pour la génération de texte sont développés en utilisant un algorithme d'apprentissage profond intitulé *Transformers,* introduit en 2017 par une équipe de chercheurs de Google[17]. Cet algorithme a amélioré les techniques existantes de l'époque en y ajoutant un mécanisme d'autoattention, qui permet une meilleure prise en compte du caractère séquentiel de certains types de données, notamment le langage naturel.

21. Les modèles de génération d'images peuvent quant à eux utiliser d'autres architectures, telles que les modèles de diffusion ou les réseaux antagonistes génératifs (en anglais « *Generative Adversarial Networks* » ou « GAN ») ou des modèles de diffusion (en anglais « *Diffusion Models* »). Néanmoins, avec l'évolution rapide de ces technologies, de nouvelles méthodes d'entraînement et architectures pourraient apparaître et remplacer des architectures de modèles existantes.

---

[16] Glossaire de la CNIL, définition de l'entraînement (ou apprentissage).

[17] Vaswani et al., Attention is all you need, juin 2017.

22. Quels que soient le modèle et l'architecture retenus, la phase d'entraînement initial requiert **une puissance de calcul très importante** et un **vaste ensemble de données générales**, souvent issues de sources publiques.

Une puissance de calcul très importante

23. Le calcul comprend « *l'ensemble des services offrant une capacité à traiter un grand nombre d'informations en même temps*. »[18]. Les besoins en puissance de calcul de l'IA générative sont importants et nécessitent l'utilisation de matériel informatique spécifique, permettant d'effectuer un grand nombre d'opérations de grande précision de manière simultanée. Les processeurs graphiques (ci-après « GPU », voir glossaire) sont fréquemment utilisés pour cette tâche.

24. À l'origine construits pour faire des calculs permettant l'affichage d'images, les processeurs graphiques ont évolué ces dernières années pour optimiser les tâches de calcul en IA. Ces processeurs sont particulièrement adaptés aux tâches liées à l'IA, étant capables d'effectuer plusieurs milliers d'opérations mathématiques en parallèle, ce qui les rend plus performants que les processeurs centraux (« CPU »).

25. D'autres processeurs, appelés accélérateurs d'IA, permettent également d'effectuer ces tâches de calcul liées à l'IA. Ce sont des circuits intégrés spécifiques à une application (en anglais « *Application Specific Integrated Circuit* » ou « ASIC ») conçus et optimisés pour les charges de travail en IA. Le terme puce ou « *AI chips* » en anglais regroupe communément les GPU et les accélérateurs d'IA.

26. Trois voies principales permettent aux entreprises d'accéder à la puissance de calcul : l'informatique en nuage (en anglais « *cloud »,* la voie la plus utilisée, voir *infra*), le développement d'une infrastructure sur site ou l'utilisation de ressources de calcul partagées (comme un supercalculateur public). Les besoins en puissance de calcul dépendent du nombre de paramètres du modèle mais également de la quantité de données utilisées pour l'entraînement du modèle.

27. L'entraînement des modèles nécessite l'utilisation d'un matériel informatique conséquent permettant d'effectuer de nombreux calculs simultanément. Un acteur indique que l'ordre de grandeur de la puissance de calcul est autour de « *1000/2000 GPUs pendant quelques semaines pour les modèles à l'état de l'art (autour de 70 milliards de paramètres)* ».

Un vaste ensemble de données générales

28. L'entraînement d'un modèle d'IA générative nécessite également de grandes quantités de données, dont il faut garantir la qualité et la diversité afin d'éviter l'apparition de biais, tout biais dans les données pouvant se retrouver dans le modèle. Ainsi, un acteur indique que « [p]*lus que pour d'autres types d'IA, les données sont l'élément le plus critique pour former et développer les modèles GPAI* [modèle d'IA à usage général]*, car des quantités massives de données sont nécessaires pour entraîner un algorithme* ». Les données fréquemment utilisées dans l'entraînement de modèles de fondation peuvent être différenciées selon leur type (texte, image, vidéos, etc.), ou leur source (publique, propriétaire ou tierce, voir *infra*).

29. Ces caractéristiques sont valables quel que soit le type de modèle entrainé : modèle de langage, de génération d'images ou de vidéos. Les différences dans l'entraînement de ces modèles portent sur les données fournies pour l'entraînement, ainsi que sur l'architecture du

[18] Voir l'avis n° 23-A-08 de l'Autorité, paragraphe 28, page 27.

modèle. Ainsi, un modèle texte-image permettant de produire des images à partir de requêtes textuelles nécessitera des données d'entraînement lui permettant d'apprendre à effectuer cette tâche, par exemple sous la forme d'images annotées.

Évaluation des modèles

30. Une fois l'entraînement du modèle de fondation terminé, celui-ci peut être évalué pour être ensuite comparé aux autres modèles existants. Différents tests (« *benchmarks* » ou critères de référence) sont utilisés par les acteurs du marché. Ces tests permettent d'évaluer la capacité d'un modèle sur différentes tâches, comme les connaissances générales en histoire, en droit ou en informatique, la résolution de problèmes mathématiques ou le raisonnement scientifique.

31. Au-delà des performances brutes des modèles d'IA générative sur ces tests de référence, les modèles sont aussi mis en concurrence sur de nombreuses autres caractéristiques, telles que la taille du modèle, représentée par le nombre de poids dont il dispose, ses capacités multimodales permettant d'accepter en entrée ou de produire en sortie différents types de données (texte, image, vidéos, etc.) ou encore la fenêtre contextuelle (en anglais « *context window* »), qui indique la quantité maximale de contenus qu'un modèle peut prendre en compte dans la requête initiale pour la production d'une réponse.

b) La spécialisation des modèles ou réglage fin

32. Une fois l'entraînement initial effectué, le modèle peut éventuellement être spécialisé pour réaliser certaines tâches spécifiques. Ce réglage fin, qui consiste en un ajustement des paramètres obtenus à la fin de l'entraînement initial, a pour objectif d'améliorer les compétences du modèle pour un cas d'usage spécifique sans affecter les capacités générales du modèle. Selon la CNIL, il s'agit une « *technique consistant à spécialiser un modèle d'IA pré-entraîné à l'accomplissement d'une tâche spécifique. Cela consiste généralement à entraîner le modèle dans son ensemble, ou seulement certaines couches d'un réseau de neurones, pour un faible nombre d'itérations sur un ensemble de données spécifiques correspondant à la tâche visée* ».[19]

33. Cette phase de réglage fin peut être effectuée par l'opérateur qui a réalisé l'entraînement initial du modèle, ou par tout autre opérateur ayant accès aux poids du modèle déterminés lors de l'entraînement initial.

34. Le réglage fin peut prendre plusieurs formes, telles que :

- **la spécialisation sur une activité sectorielle spécifique**. Dans ce cas, de nouvelles données spécifiques sont fournies au modèle dans le cadre d'une extension de son entraînement, pour le rendre plus performant pour des besoins sectoriels. Ainsi, un modèle de génération de texte généraliste peut être spécialisé sur un corpus de textes juridiques pour répondre à des questions nécessitant une connaissance approfondie du droit ;
- **l'apprentissage par renforcement et rétroaction humain** (ci-après « RLHF », voir glossaire). Il s'agit d'une « *approche d'apprentissage par renforcement* [apprentissage issu de l'expérience] *qui utilise les commentaires et les évaluations d'utilisateurs*

[19] Glossaire de la CNIL, définition de l'ajustement (*fine tuning*).

humains pour guider l'apprentissage d'un modèle d'intelligence artificielle »[20]. Le RLHF vise donc à apprendre au modèle à produire des contenus qui correspondent au mieux à ce qui est attendu par des humains, en s'appuyant sur des évaluateurs humains qui notent les contenus générés par un modèle d'IA générative. Comme indiqué par le Pôle d'Expertise et de Régulation du Numérique (ci-après « PEReN ») dans son éclairage sur l'IA générative[21], grâce au RLHF, « *les LLM conversationnels sont optimisés afin de satisfaire le plus possible les utilisateurs humains* », ce qui améliore le rendu des réponses.

35. Cette phase de spécialisation fait généralement appel à un volume de données plus faible et nécessite ainsi une puissance de calcul inférieure à celle de la phase d'entraînement. Un acteur précise que « *cette phase est moyennement consommatrice de moyens de calcul. Avec l'apparition de grands modèles généralistes on peut penser que cette phase va devenir la plus importante dans la diffusion de l'IA dans les entreprises et la société* ».

2. L'INFERENCE OU LA PRODUCTION DE CONTENU

36. Un modèle d'IA générative entrainé, et éventuellement spécialisé, a ensuite vocation à être utilisé pour générer des contenus. Cette dernière étape de production de contenu est aussi appelée l'inférence (voir glossaire) et nécessite de mettre le modèle à disposition des utilisateurs finaux.

37. Il existe de nombreux modes de déploiement des modèles d'IA générative auprès des utilisateurs finaux, dépendant généralement du degré d'ouverture souhaité par le développeur. Les modèles propriétaires peuvent être réservés à un usage interne, ou rendus accessibles via des applications spécifiques (Internet ou mobiles telles que ChatGPT d'OpenAI ou Le Chat de Mistral AI), des fenêtres de dialogue dans des applications (par exemple bureautiques ou collaboratives), des assistants vocaux ou des interfaces de programmation pour les développeurs (API, voir glossaire). Les utilisateurs n'ont en général pas accès au modèle lui-même et ne peuvent donc pas le réutiliser ou le modifier.

38. Bien que les acteurs utilisent fréquemment l'appellation *open source*, les modèles « ouverts » sont le plus souvent mis à disposition par le biais de la publication de leurs poids (approche *« open-weights »,* voir glossaire), permettant de les réutiliser et/ou de les modifier, parfois sous certaines conditions de licences. Cette mise à disposition peut également s'accompagner de l'ensemble des ressources ou partie des ressources qui ont servi à l'entraînement du modèle (code, données, etc.), se rapprochant ainsi d'une ouverture complète. Les défis soulevés par ce manque de définition consensuelle de l'*open source* pour l'IA seront discutés aux paragraphes 179 et suivants.

39. La puissance de calcul nécessaire à l'inférence dépend du nombre d'utilisateurs et de la taille (nombre de poids) du modèle : pour un petit modèle proposé à un faible nombre d'utilisateurs, quelques processeurs graphiques peuvent suffire. Cependant, si la taille du modèle et le nombre d'utilisateurs augmentent, les coûts d'utilisation peuvent être démultipliés. Un acteur indique ainsi que « *contrairement à d'autres innovations numériques majeures, la fourniture de LLM de produits ou services d'IA générative implique un coût marginal significatif, principalement en raison du coût du calcul* ».

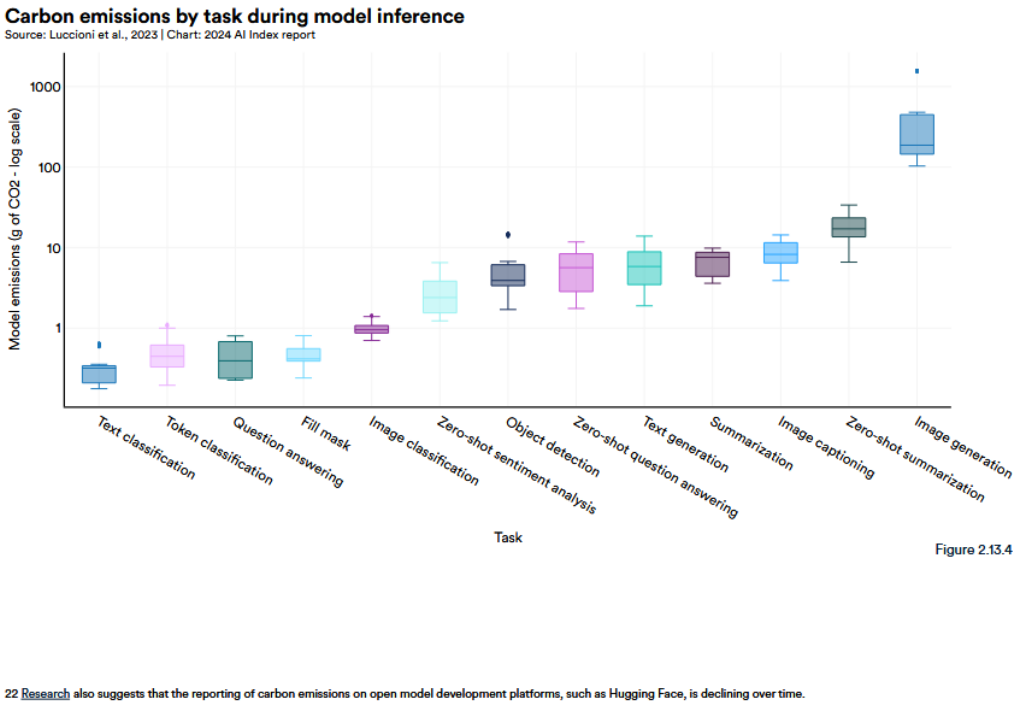
---

[20] Glossaire de la CNIL, définition de l'apprentissage par renforcement et rétroaction humaine.

[21] PEReN, Éclairage sur…n°6 – ChatGPT ou la percée des modèles d'IA conversationnels, 6 avril 2023.

40. Plus la puissance de calcul est élevée plus la consommation énergétique augmente. Dans la phase d'inférence, l'empreinte carbone varie selon l'application utilisée, et est par exemple beaucoup plus élevée pour la génération d'image que de texte.

Figure n° 1 : émissions de carbone par tâches durant la phase d'inférence



22 Research also suggests that the reporting of carbon emissions on open model development platforms, such as Hugging Face, is declining over time.

Source : Université de Stanford, Artificial intelligence index report 2024, page 156

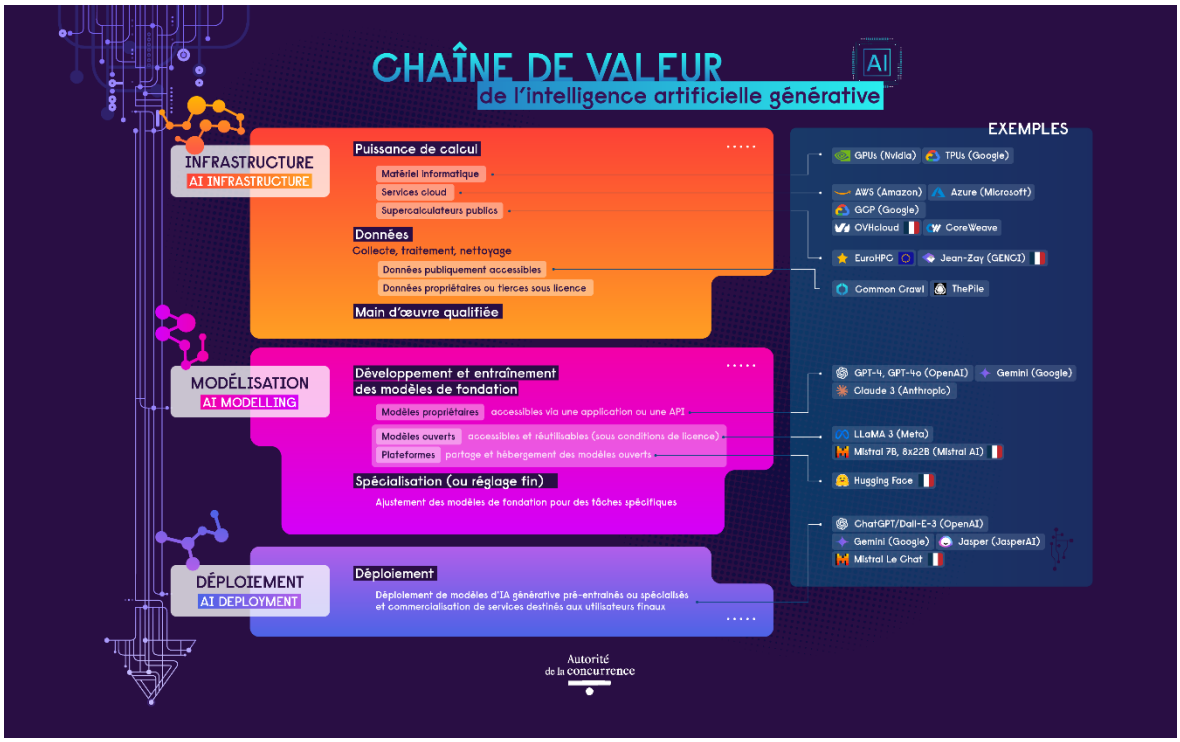
41. Dans sa forme la plus simple, l'inférence ne nécessite pas de données additionnelles, hormis celles fournies dans la requête, par exemple la question posée par un utilisateur à un agent conversationnel. Cependant, il existe des techniques permettant d'utiliser de nouvelles données lors de l'inférence telles que **la génération augmentée de récupération** (RAG, voir glossaire) qui vise à améliorer le résultat produit par un modèle de langage en incluant dans la requête des données externes fiables permettant d'ajouter du contexte ou de donner des éléments de réponse à la question posée par l'utilisateur. Des bases de connaissances spécifiques, comme des données internes d'entreprise, peuvent ainsi être utilisées à cette étape, en fournissant au modèle les données les plus pertinentes en plus de la question posée. Cette technique est notamment utilisée par les applications de type agent conversationnel, afin d'apporter au robot une connaissance des données nouvelles qui n'ont pas été utilisées lors de l'entraînement initial, comme les données d'actualité. Pour son application Gemini, Google décrit ainsi une étape **d'ancrage du modèle** (en anglais « *grounding* »), qui « *consiste en l'envoi d'une requête par Bard* [maintenant Gemini] *à Google Search afin d'obtenir des informations utiles pour répondre à la question posée par l'utilisateur* »[22].

[22] Voir la décision n° 24-D-03 de l'Autorité, paragraphe 166, page 37.

C. LES ACTEURS DE LA CHAÎNE DE VALEUR DE L'IA GENERATIVE

42. Le paysage concurrentiel du secteur de l'IA générative évoluant très rapidement, la présentation des acteurs effectuée dans cette partie est valable à la date de publication de l'avis. La figure n° 2 présente la chaîne de valeur de l'IA générative.

Figure n° 2 : principaux acteurs de la chaîne de valeur de l'IA générative



Source : Autorité de la Concurrence, inspiré de l'article ChatGPT, Bard & Co. : an introduction to AI for competition and regulatory lawyers de Thomas Höppner et Luke Streatfeild, 23 février 2023

43. Les grands acteurs du numérique semblent avoir adopté différentes stratégies en matière d'IA générative. Alphabet et Microsoft sont présents sur l'ensemble de la chaîne de valeur, en partie grâce à des partenariats noués avec des développeurs de modèles de fondation. Amazon, Apple, Meta et Nvidia sont présents à certains niveaux de la chaîne de valeur. Au-delà de ces grands acteurs, d'autres acteurs sont présents à l'amont et à l'aval.

1. LES GRANDS ACTEURS DU NUMERIQUE PRESENTS DANS LE SECTEUR

a) Alphabet

44. Google est une société américaine créée en 1998, dont les fondateurs ont inventé le moteur de recherche éponyme, qui est le plus utilisé en France et dans le monde. Les activités de Google ont été regroupées en 2015 au sein du conglomérat Alphabet, qui regroupe notamment la fourniture de services de recherche en ligne et de systèmes d'exploitation, la publicité en ligne et la fourniture de services *cloud*.

45. Alphabet fournit des services *cloud* via Google Cloud Platform (ci-après « GCP »). De plus, Alphabet a développé ses propres processeurs tensoriels (ci-après « **TPU** », voir glossaire), fabriqués par Broadcom et qui sont, selon Google, des « *accélérateurs d'IA spécialement*

conçus et optimisés pour l'entraînement et l'inférence de modèles d'IA volumineux »[23]. Ces processeurs sont utilisés pour les besoins internes d'Alphabet dans ses centres de données depuis 2015 (bien qu'ils n'aient été officiellement annoncés qu'en 2016) et sont proposés et commercialisés depuis 2018 aux clients de GCP sur des machines virtuelles. Alors que la première génération des TPU permettait uniquement l'inférence, la cinquième version introduite en décembre 2023[24], a permis l'entraînement du modèle de fondation Gemini, et une sixième version plus performante a été annoncée en mai 2024[25]. GCP propose également une plateforme appelée **Model Garden** [26] proposant l'accès à plus de 130 modèles de fondation, incluant les modèles d'IA générative propriétaires de Google, ainsi que d'autres modèles tiers.

46. En termes de données, Alphabet possède notamment le plus grand index de recherche pour son moteur de recherche Google, ainsi que la plus grande base de vidéos au monde avec YouTube, qui hébergerait plus de 10 milliards de vidéos publiques[27].

47. Alphabet est un acteur de l'IA depuis de nombreuses années, notamment depuis le rachat en 2014 de DeepMind, un laboratoire de recherche en IA fondé en 2010. Devenu Google DeepMind, il est notamment connu pour le lancement d'AlphaGo (première IA à avoir battu le champion du monde de Go) et AlphaFold (prédiction de la structure des protéines). Ce laboratoire a fusionné en 2023 avec Google Brain, le laboratoire qui a créé le cadriciel (en anglais « *framework* », voir glossaire) de développement TensorFlow, très utilisé en IA. Les différents laboratoires de recherche d'Alphabet ont permis le développement de nombreux modèles de fondation, tels que Bert (2018), Imagen (mai 2022), PaLM 2 (mai 2023), Gemini (décembre 2023) ou Gemma (février 2024).

48. Alphabet a lancé en mars 2023 dans le monde, et en juillet 2023 en Europe, son agent conversationnel Bard (renommé en février 2024 **Gemini**), accessible à tous les utilisateurs. Elle a annoncé, lors de sa présentation « Google I/O 2024 », l'ajout des fonctionnalités d'IA générative dans d'autres services tels que son moteur de recherche (via l'AI Overview), ses outils de bureautique Workspace et l'intégration de Gemini dans son système d'exploitation mobile Android[28] ou d'une version Nano de Gemini dans les téléphones Pixel 8 Pro.

b) Amazon

49. Fondée en 1994, Amazon est une société américaine dont l'activité principale est le commerce en ligne via sa place de marché de vente de produits amazon.com. Cette activité s'est ensuite diversifiée avec la fourniture de services informatiques *cloud*, via sa filiale Amazon Web Services (ci-après « AWS ») ou d'objets connectés comme Alexa. AWS est un des acteurs majeurs de la fourniture de services *cloud* en France et dans le monde[29].

---

[23] Google, Accélérez le développement de l'IA avec les TPU Google Cloud.

[24] Enabling next-generation AI workloads: Announcing TPU v5p and AI Hypercomputer, 7 décembre 2023.

[25] Google, Introducing Trillium, sixth-generation TPUs, 15 mai 2024.

[26] Google Cloud, Model Garden sur Vertex AI.

[27] McGrady, R., Zheng, K., Curran, R., Baumgartner, J., & Zuckerman, E. (2023). Dialing for Videos: A Random Sample of YouTube. *Journal of Quantitative Description: Digital Media 3*, *20 décembre 2023*.

[28] Frandroid, La Google I/O 2024 résumée en 15 annonces : Gemini 1.5 Pro, Project Astra, AI Overview, Gmail, Andoid 15, Veo, etc., 14 mai 2024.

[29] Voir l'avis n° 23-A-08 de l'Autorité, paragraphes 91 à 94.

50. AWS met à disposition de ses clients de services *cloud* des puces spécialisées pour l'IA appelées **Trainium** (pour l'entraînement) et **Inferentia** (pour l'inférence) depuis 2018. Ces puces ont été développées pour améliorer les performances des modèles d'IA sur AWS tout en réduisant le coût et la consommation énergétique[30].

51. Amazon propose également plusieurs services *cloud* spécifiques à l'IA générative. Par exemple, **Amazon SageMaker** fournit aux clients des outils pour créer, entraîner et déployer leurs propres modèles de fondation. Amazon se positionne également comme développeur de modèles, via sa gamme de modèles Titan, accessible notamment via sa plateforme **Amazon Bedrock**, permettant aux développeurs d'accéder à de multiples modèles de fondation développés par Amazon et par des tiers. Amazon Data Exchange est un service *cloud* d'AWS regroupant de nombreux jeux de données tierces qu'Amazon met à disposition de ses utilisateurs *cloud*. Ce service couvre de nombreux types de données, notamment financières ou sectorielles.

52. Amazon intègre des systèmes d'IA générative dans plusieurs de ses produits, comme l'assistant vocal Alexa. D'autres produits sont développés autour de l'IA générative, comme Rufus, un assistant d'achat alimenté par l'IA pour sa plateforme d'e-commerce, ou Amazon Q, un agent conversationnel, notamment capable d'écrire du code, de le tester et de solliciter d'autres services *cloud* comme AWS S3 pour les développeurs sur AWS.

c) Apple

53. Apple est une société américaine fondée en 1976 spécialisée dans la conception, la fabrication et la commercialisation de produits électroniques (iPhone, iPad et Mac) et de logiciels.

54. L'activité d'Apple dans le secteur de l'IA générative est moins développée que celle des autres grands acteurs du numérique. Son président, Tim Cook, a cependant fait état de la volonté d'Apple de se positionner sur ce marché[31]. Apple a ainsi présenté au premier semestre 2024 ses premiers modèles développés en interne, avec le grand modèle propriétaire **MM1** et une gamme de petits modèles ouverts, OpenELM.

55. De par ses activités historiques dans les produits électroniques, Apple dispose d'une voie privilégiée de commercialisation de ses produits d'IA générative. Lors de la conférence des développeurs qui s'est tenue le 10 juin 2024, Apple a annoncé le lancement de fonctionnalités basées sur de l'IA générative appelées « Apple Intelligence » dans les dernières versions de ses produits (iPhone, iPad et Mac), ainsi qu'un partenariat avec **OpenAI**. Ainsi, dans Apple Intelligence, les requêtes les plus simples pourront être traitées directement sur l'appareil par un petit modèle de fondation d'Apple, tandis que les requêtes complexes seront traitées dans des serveurs *cloud* Apple (équipés de puces Apple) avec l'aide de ses modèles les plus performants et pour certaines tâches des modèles d'OpenAI[32]. Au vu des besoins en puissance de calcul de l'IA générative, seuls les iPhone équipés des puces A17 Pro Bionic peuvent supporter Apple Intelligence.

---

[30] Rapport Annuel 2023 d'Amazon.

[31] Reuters, Apple to disclose AI plans later this year, CEO Tim Cook says, 28 février 2024.

[32] Apple, Introducing Apple's On-Device and Server Foundation Models, 10 juin 2024.

d) Meta

56. Créée en 2004, Meta (anciennement Facebook Inc. jusqu'en octobre 2021) est une société américaine spécialisée dans les services et les produits liés à l'Internet. Meta exploite plusieurs réseaux sociaux comme Facebook, Instagram ou WhatsApp.

57. À l'inverse des autres grandes entreprises américaines verticalement intégrées (Amazon, Google et Microsoft), Meta n'a pas d'activité de fourniture de services *cloud*. En revanche, Meta dispose de larges infrastructures informatiques et **de ses propres centres de données** pour l'exploitation de ses plateformes.

58. Meta développe et conçoit ses propres accélérateurs pour l'IA (« *Meta Training and Inference Accelerator* ») destinés à améliorer l'efficacité des charges de travail, tels que la recommandation de contenus dans le fil d'actualité, dans la publicité, ou l'IA générative. Une première version de ces accélérateurs a été annoncée par Meta en 2023. Compte tenu de son activité sur les réseaux sociaux, Meta dispose également de vastes bases de données notamment d'images et de vidéos.

59. Au sein du groupe Meta, le Facebook AI Research (FAIR) est un groupement de laboratoires de recherche dédié à la recherche fondamentale en IA et ayant pour objectif de faire avancer la science ouverte. Ces laboratoires ont contribué au développement des modèles de langage de Meta, dont la première version intitulée **Llama** a été publiée en février 2023, suivie de Llama 2 en juillet 2023 et Llama 3 en avril 2024. Ces modèles ont tous été annoncés et publiés en *open-weights*, avec une licence permettant leur réutilisation commerciale, sauf pour les services ayant plus de 700 millions d'utilisateurs.

60. Meta a également lancé en avril 2024 **Meta AI**, un agent conversationnel basé sur la gamme de modèles Llama. Meta prévoie également d'intégrer des outils d'IA générative à ses principales plateformes, et a annoncé en mai 2024 le lancement d'un « bac à sable » pour tester l'intégration de l'IA générative aux outils publicitaires de Meta[33].

e) Microsoft

61. Microsoft Corporation (ci-après « Microsoft ») est une société américaine qui offre une large gamme de produits technologiques. Microsoft est un fournisseur historique de systèmes d'exploitation pour ordinateurs (Windows) et de logiciels de bureautique (la suite Microsoft 365, anciennement Office). Il possède également le moteur de recherche Bing.

62. Microsoft Azure est l'un des principaux fournisseurs de services *cloud* mondiaux. Il a annoncé en 2023 la sortie d'accélérateurs d'IA appelés **Maia** qui seront disponibles courant 2024. Ces puces ont été particulièrement optimisées pour être utilisées dans les infrastructures *cloud* de Microsoft Azure. Azure dispose également d'une plateforme de mise à disposition de modèles, **Azure AI Model Catalog**.

63. Microsoft a participé à l'entraînement de grands modèles de langage, comme Megatron-Turing-NLG (530 milliards de paramètres) en coopération avec Nvidia en 2021. À la suite de son partenariat et de ses investissements réalisés dans la société OpenAI (voir *infra*), Microsoft semble avoir effectué une transition vers le développement de petits modèles de langage (« *Small Langage Models* » ou SLM). En complément des modèles développés par OpenAI, Microsoft propose ainsi l'utilisation commerciale d'outils fondés sur ses modèles propriétaires de génération d'images (gamme Florence, dont Florence 2 publiée en

[33] Meta, Lancement de l'IA Sandbox pour les annonceurs et expansion de Meta Advantage Suite, 12 mai 2024.

novembre 2023) et de génération de texte (Orca et Phi, en ce compris Phi-3 annoncé et publié en *open-weights* en avril 2024).

64. Microsoft intègre des outils d'IA générative dans plusieurs de ses produits et services historiques. Depuis 2023, le moteur de recherche Bing propose un assistant Bing AI. Github, une plateforme de développement et de partage de code rachetée par Microsoft en 2018, commercialise depuis fin 2021 Github Copilot, un assistant de développement permettant la génération de code. Les logiciels de la suite Microsoft 365 intègrent également des assistants **Copilot** basés sur l'IA générative. Ces différents outils s'appuient majoritairement sur les modèles de fondation d'OpenAI et également les modèles propriétaires de Microsoft.

f) Nvidia

65. Nvidia est une société américaine spécialisée dans le développement et la conception de GPU et de circuits intégrés pour l'informatique. Elle a été fondée en 1993 et développe depuis 2018 une gamme de processeurs graphiques spécialisés dans le calcul pour les centres de données. Nvidia poursuit une forte dynamique depuis l'émergence de l'IA et est devenue, mi-juin 2024, la première capitalisation boursière mondiale à la date de publication de cet avis, devant Microsoft, Apple, Amazon, et Alphabet, avec une valorisation **de plus de 3 400 milliards de dollars** (soit environ 3 140 milliards d'euros)[34].

66. Nvidia ne produit pas elle-même les processeurs graphiques, mais fait notamment appel à la société taiwanaise TSMC pour cela. Les processeurs graphiques Nvidia les plus performants à l'heure actuelle sont les « **A100 » et les « H100 »**, et une nouvelle génération dénommée **Blackwell** a été annoncée en mars 2024. Nvidia est également connue pour son logiciel **Cuda** (voir glossaire), permettant la programmation sur ses propres GPU.

67. Nvidia a étendu sa présence sur la partie amont de la chaîne de valeur en nouant des partenariats avec de nombreux fournisseurs de services *cloud*, incluant par exemple AWS, GCP et Microsoft Azure. Nvidia a contribué au développement de modèles de fondation, en collaboration avec Microsoft (Megatron 530B), et propose une plateforme **Nvidia AI Foundation** permettant aux développeurs d'accéder à ses modèles propriétaires et à d'autres modèles tiers. Nvidia a également annoncé le 14 juin 2024 la publication d'une gamme de modèles appelée Nemotron-4 facilitant la génération de données synthétiques[35].

2. LES DEVELOPPEURS DE MODELES D'IA GENERATIVE

68. Les développeurs de modèles, qui conçoivent et entrainent les modèles d'IA générative, sont au cœur de la chaîne de valeur. Outre les grandes entreprises présentées *supra*, les développeurs sont principalement des laboratoires de recherche en IA ou des entreprises innovantes natives de l'IA.

69. Les principaux acteurs sont les suivants :

- **Anthropic** est un laboratoire de recherche en IA fondé en 2021 par d'anciens membres d'OpenAI, avec l'objectif de proposer des outils d'IA plus sécurisés et responsables.

[34] L'Opinion, L'entreprise américaine d'IA Nvidia devient la première capitalisation mondiale en Bourse, 18 juin 2024.

[35] Nvidia, NVIDIA Releases Open Synthetic Data Generation Pipeline for Training Large Language Models, 14 juin 2024.

Anthropic a développé une gamme de grands modèles de langage commercialisée sous le nom de Claude, dont la troisième version a été annoncée en mars 2024 ;

- **HuggingFace** est une entreprise franco-américaine fournissant une plateforme d'hébergement et de collaboration, qui met à disposition des développeurs la vaste majorité des modèles publiés en *open source* ou en *open-weights*, ainsi que les réutilisations qui en sont faites. Au 5 juin 2024, sa plateforme recense ainsi plus de 700 000 modèles et 158 000 jeux de données[36]. HuggingFace a notamment mené l'initiative BigScience (en collaboration avec d'autres acteurs comme le CNRS) pour développer un modèle appelé **Bloom** en 2021. Ce modèle, entrainé sur un supercalculateur public, est depuis accessible à tous en *open source* ;
- **Mistral AI** est une *start-up* française fondée en avril 2023 et valorisée à quasiment 6 milliards d'euros suite à sa troisième levée de fonds en juin 2024[37]. Elle est spécialisée dans le développement de modèles d'IA générative, comme Mistral 7B, Mistral 8x7B et Mixtral 8x22B, modèles dont les poids ont été publiés (*open-weights),* ou Mistral Large. Elle a également annoncé le lancement de Mistral Large en février 2024, un modèle propriétaire alimentant en partie son agent conversationnel appelé Le Chat[38] ;
- **OpenAI** est le premier acteur à avoir proposé son modèle d'IA générative au grand public via un agent conversationnel dénommé ChatGPT. Elle est à l'origine de l'explosion de l'intérêt pour cette technologie. Fondée en 2015 en tant qu'association à but non lucratif, elle a créé en 2019 une branche dite « à but lucratif plafonné »[39]. Les modèles d'IA générative alimentant ChatGPT sont GPT3.5, GPT3.5 Turbo et GPT-4[40]. OpenAI développe également des modèles de génération d'images avec DALL-E, ou de vidéos avec Sora, annoncé en février 2024.

70. De nombreux autres acteurs développent leurs propres modèles de fondation, comme **Aleph Alpha** (Allemagne), **Cohere** (États-Unis), **LightOn** (France), **Stability AI** (États-Unis) ou **xAI** (États-Unis).

---

[36] Site d'HuggingFace (modèles et données) consulté le 5 juin 2024.

[37] Le Monde Informatique, En levant 600 M€, Mistral AI frôle les 6 Md€ de valorisation, 11 juin 2024.

[38] Mistral AI, Le Chat, 26 février 2024

[39] Dans son post de blog daté du 11 mars 2019, Open AI indique à ce sujet : « [n]*ous voulons augmenter notre capacité à lever des capitaux tout en continuant à servir notre mission, et aucune structure juridique préexistante que nous connaissons ne permet d'atteindre le bon équilibre. Notre solution consiste à créer OpenAI LP en tant qu'hybride entre une société à but lucratif et une société à but non lucratif, que nous appelons une société à "but lucratif plafonné". L'idée fondamentale d'OpenAI LP est que les investisseurs et les employés peuvent obtenir un rendement plafonné si nous réussissons notre mission, ce qui nous permet de lever des capitaux d'investissement et d'attirer des employés avec des capitaux propres semblables à ceux d'une startup. Mais tout retour au-delà de ce montant - et si nous réussissons, nous nous attendons à générer des ordres de grandeur de valeur supérieure à ce que nous devons aux personnes qui investissent ou travaillent à OpenAI LP - est la propriété de l'entité OpenAI Nonprofit d'origine* » (traduction libre).

[40] Generative Pretrained Transformers (réseau de neurones pré-entrainés d'architecture Transformers).

3. LES PARTENARIATS ENTRE GRANDS ACTEURS DU SECTEUR ET DEVELOPPEURS DE MODELES

71. Le secteur de l'IA générative est caractérisé par de nombreux accords entre ses différents acteurs. Ces accords peuvent revêtir plusieurs formes comme des accords de fourniture de ressources de calcul et/ou des accords de licences avec les fournisseurs de modèles de fondation.

72. Un certain nombre d'accords incluent également des **prises de participation minoritaires** et la conclusion d'accords commerciaux, exclusifs ou non exclusifs, pour le développement ou la commercialisation de modèles de fondation au sein de leurs plateformes.

73. Les principaux accords, classés par montants d'investissements, sont les suivants :

- un partenariat exclusif entre Microsoft et la société **OpenAI** prenant la forme d'un « *investissement pluriannuel* [de Microsoft] *de plusieurs milliards de dollars pour accélérer les découvertes en IA* » au mois de janvier 2023[41]. Microsoft fournit à OpenAI des capacités de calcul intensif afin d'accélérer les recherches d'OpenAI et déploie les modèles d'OpenAI dans ses produits, notamment via le *cloud* Azure. Microsoft est devenu par ailleurs le fournisseur de services *cloud* exclusif d'OpenAI ;

- les investissements d'Amazon (4 milliards de dollars, soit environ 3,7 milliards d'euros, pour « *une participation minoritaire* »[42] non exclusive) et Google (2 milliards de dollars, soit environ 1,8 milliard d'euros pour une participation estimée à 10 %, selon certains articles[43]) dans **Anthropic** en 2023[44]. Ces partenariats permettent notamment à Anthropic d'accéder aux puces d'IA et aux clients d'AWS et Google ;

- l'investissement de 1,3 milliard de dollars (soit environ 1,2 milliard d'euros) dans **Inflection**, une entreprise innovante américaine visant à offrir des services d'IA personnalisés, réalisé en juin 2023 par des entreprises comme Microsoft et Nvidia[45]. En mars 2024, deux des cofondateurs d'Inflection, et une partie du personnel, ont quitté la société pour prendre la tête de la division Microsoft AI nouvellement créée qui sera en charge de Copilot, Bing et Edge[46] ;

- les investissements de plusieurs sociétés importantes telles qu'Amazon, Google ou Nvidia pour un montant de 235 millions de dollars (soit environ 220 millions d'euros) au mois d'août 2023 dans la société **HuggingFace** afin d'accélérer la formation, l'ajustement et le déploiement de grands modèles utilisés pour créer des applications d'IA

---

[41] Article de blog de Microsoft, Microsoft and OpenAI extend partnership, 23 janvier 2023. Cet investissement a été initié en 2019 lorsque Microsoft a investi 1 milliard de dollars dans OpenAI.

[42] Communiqué de presse d'Amazon, Amazon and Anthropic Announce Strategic Collaboration to Advance Generative AI, 25 septembre 2023.

[43] Le Monde informatique, Après Amazon, Google va investir jusqu'à 2 milliards de dollars dans Anthropic, 30 octobre 2023.

[44] TechCrunch, AI's proxy war heats up as Google reportedly backs Anthropic with $2B, 27 octobre 2023.

[45] TechCrunch, Inflection lands $1.3B investment to build more 'personal' AI, 29 juin 2023.

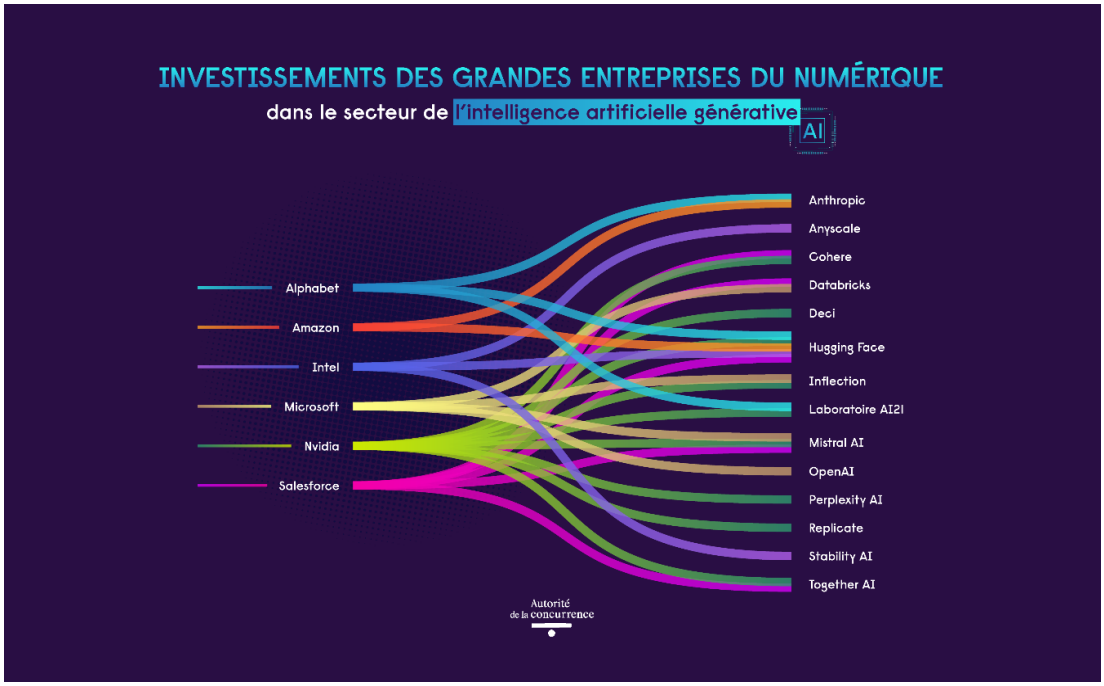
[46] Communiqué d'Inflection, the new Inflection: an important change to how we'll work, 19 mars 2024.

générative. Le 25 janvier 2024, HuggingFace a également annoncé un « *partenariat stratégique* » avec Google Cloud[47] ;

– l'investissement de 15 millions d'euros de Microsoft dans **Mistral AI** en 2024 par le biais d'une obligation convertible en actions et d'un partenariat[48] permettant à Mistral AI d'accéder à l'infrastructure de supercalculateurs Azure AI et de proposer ses modèles premium dans le catalogue de modèles d'Azure AI Studio et d'Azure Machine Learning en tant que Model-as-a-Service (ci-après « MaaS »). Mistral Large, le dernier modèle de génération de texte de MistralAI, sera également accessible aux clients de Microsoft. Par ailleurs, des entreprises comme Nvidia ou Salesforce ont participé à sa dernière levée de fonds du mois de juin 2024.

74. Le schéma suivant illustre les nombreux investissements croisés entre les grandes entreprises du numérique et les entreprises innovantes du secteur, notamment de la part de Nvidia, qui a investi dans de nombreuses entreprises du secteur de l'IA générative, et notamment des développeurs de modèles (Mistral AI, Inflection, Deci, HuggingFace, AI21 Labs, etc.).

Figure n° 3 : investissements des grandes entreprises du numérique au sein des entreprises innovantes du secteur (au mois de mai 2024)



Source : Autorité de la concurrence inspirée par l'article de S&P Global, Untangling the web of strategic tech investments in generative AI, 22 février 2024[49]

---

[47] Communiqué de Hugging Face, Hugging Face and Google partner for open AI collaboration Ce partenariat permettra notamment permettra aux clients de Google Cloud de pouvoir « *entraîner et déployer facilement des modèles Hugging Face au sein de Google Kubernetes Engine (GKE) et de Vertex AI. Les clients bénéficieront des capacités matérielles uniques disponibles dans Google Cloud, comme les instances TPU, les VM A3, alimentées par les GPU NVIDIA H100 Tensor Core, et les VM C3, alimentées par les CPU Intel Sapphire Rapid* » (traduction libre).

[48] Microsoft, Microsoft et Mistral AI annoncent un nouveau partenariat pour accélérer l'innovation en intelligence artificielle et dévoilent Mistral Large, disponible dès à présent sur Azure, 26 février 2024.

[49] Le schéma, qui n'a pas vocation à être exhaustif, repose sur la liste des principaux acteurs actifs dans l'IA générative en France fournie par un grand acteur du secteur. Par ailleurs, la liste des investissements provient

4. LES ACTEURS PRESENTS A L'AMONT DE LA CHAINE DE VALEUR

75. À l'amont de la chaîne de valeur de l'IA générative, d'autres acteurs sont présents, tels que les fournisseurs de composants informatiques, les fournisseurs de services *cloud* et les supercalculateurs publics.

a) Fournisseurs de composants informatiques

76. Les composants informatiques tels que les processeurs graphiques sont des pièces indispensables à l'entraînement de modèles d'IA générative. Au-delà de Nvidia et des grands acteurs du numérique (présentés ci-dessus), ce secteur regroupe quelques acteurs historiques et des nouveaux entrants.

77. Les principaux concurrents historiques de Nvidia sont Advanced Micro Devices (ci-après « **AMD** ») **et Intel**.

78. AMD est une société américaine fondée en 1969. Depuis le rachat d'ATI Technologies en 2006, AMD est active dans la production de microprocesseurs pour les cartes graphiques. Les processeurs graphiques appelés Instinct MI300X sont la dernière version de GPU d'AMD[50], avant la sortie des MI325X prévue en fin d'année 2024.

79. Intel est une entreprise américaine fondée en 1968, spécialisée dans le développement et la production de cartes mères et de microprocesseurs (appelés « x86 ») pour les ordinateurs, mais elle est également active dans le secteur des processeurs graphiques et des accélérateurs IA, en ce compris l'accélérateur d'IA appelé Gaudi 3 dévoilé en avril 2024[51].

80. Plusieurs autres entreprises, telles que Cerebras (États-Unis), Graphcore (Royaume-Uni), SambaNova (États-Unis) ou Groq (États-Unis), s'installent progressivement sur ce secteur en proposant des puces spécialisées pour l'IA.

b) Fournisseurs de services *cloud*

81. Au-delà des trois grands fournisseurs de services *cloud* (Microsoft Azure, AWS et GCP) présentés ci-dessus, de nombreuses autres entreprises comme 3DSOutscale, Alibaba Cloud, IBM, Oracle Cloud, OVHcloud, Scaleway fournissent des services *cloud* susceptibles d'être utilisés à la fois à l'amont pour l'entraînement et la spécialisation des modèles, mais également à l'aval pour l'inférence des modèles d'IA générative. Ces acteurs du *cloud* ont fait l'objet d'une présentation détaillée aux paragraphes 103 à 117 de l'avis n° 23-A-08, auxquels le présent avis renvoie.

82. D'autres acteurs spécialisés font également leur apparition afin de répondre aux besoins spécifiques de ressources de calcul du secteur. C'est notamment le cas du fournisseur

---

des recherches effectuées par l'Autorité et peut omettre des investissements si ceux-ci n'ont pas fait l'objet de communication publique. Enfin, il convient de préciser que certaines entreprises innovantes actives en France ne semblent pas avoir fait l'objet d'investissements par des grands acteurs du numérique (mais peuvent bénéficier de fonds de la part d'investisseurs institutionnels par exemple) et ne figurent donc pas dans le schéma. C'est le cas par exemple des développeurs de modèles comme Eleven Labs, LightOn, ou Naver.

[50] Usine Digitale, IA générative : Les profits d'AMD tirés vers le haut par son accélérateur Instinct MI300X, 31 janvier 2024.

[51] Usine Digitale, Intel dévoile Gaudi 3, sa dernière arme pour se lancer dans la bataille de l'IA générative, 10 avril 2024.

américain **CoreWeave**, spécialisé dans la fourniture de services de calcul de haute performance et, selon lui, « *partenaire préféré* »[52] de l'entreprise Nvidia. Sa dernière levée de fonds en mai 2024 a atteint plus d'un milliard de dollars (environ 930 millions d'euros), faisant monter sa valorisation à 19 milliards de dollars (soit environ 17,5 milliards d'euros)[53].

83. D'autres acteurs tels que Denvr Dataworks (Canada), Lambda Labs (États-Unis) ou encore TensorWave (États-Unis) sont positionnés sur ce segment de la fourniture de services *cloud* spécialisés en IA, souvent en partenariat avec des acteurs des composants informatiques, tels que Nvidia, AMD ou Intel.

c) Supercalculateurs publics

84. Un supercalculateur est défini par le Commissariat à l'énergie atomique (CEA) comme « *un très grand ordinateur, réunissant plusieurs dizaines de milliers de processeurs, et capable de réaliser un très grand nombre d'opérations de calcul ou de traitement de données simultanées* »[54]. Les supercalculateurs sont historiquement utilisés pour la recherche fondamentale, et des tâches telles que les prévisions météorologiques ou climatiques, ou encore des simulations en science des matériaux, en chimie ou dans le domaine médical. Leur but est de fournir des ressources de calcul aux chercheurs gratuitement.

85. Il existe un grand nombre de supercalculateurs publics dans le monde. Le classement TOP500[55], établi par une équipe de chercheurs du laboratoire national Lawrence-Berkeley et des universités du Tennessee et de Manheim, classe les supercalculateurs selon plusieurs critères de performance, parmi lesquels la puissance de calcul et la performance environnementale. Selon le classement établi en juin 2024, parmi les dix supercalculateurs les plus puissants, cinq sont situés aux États-Unis (dont un appartenant à Microsoft Azure et un autre à Nvidia), quatre en Europe (en Finlande, Suisse, Italie et Espagne) et un au Japon.

86. Au sein de l'Union européenne, **EuroHPC** (pour « European High Performance Computing »), une initiative conjointe entre acteurs publics et privés, a permis l'installation de huit supercalculateurs dans toute l'Europe (au-delà des trois supercalculateurs précités en Finlande, Italie et Espagne, cinq autres sont situés au Luxembourg et au Portugal, en République Tchèque, Bulgarie et Slovénie). Un nouveau supercalculateur dit exaflopique, c'est-à-dire ayant pouvant réaliser un milliard de milliards d'opérations de calcul par seconde est en cours de construction en Allemagne.

87. Un autre supercalculateur exaflopique dénommé **Jules-Verne** a été annoncé par EuroHPC en France en 2025. Ce supercalculateur sera le plus puissant sur le territoire français et viendra compléter l'offre déjà existante, avec les trois supercalculateurs **Jean Zay** (à l'Institut de développement et des ressources en information scientifique (IDRIS) du CNRS à Orsay, 190ème du TOP500), **Adastra** (implanté au Centre informatique national de l'enseignement supérieur (CINES) à Montpellier, 20ème du TOP500) et **Joliot-Curie** (au Très grand centre de calcul du Commissariat à l'énergie atomique (TGCC-CEA) à Bruyères-le-Châtel, 132ème du TOP500). La gestion de ces supercalculateurs est assurée par

---

[52] CoreWeave, CoreWeave Becomes NVIDIA's First Elite Cloud Services Provider for Compute.

[53] TechCrunch, CoreWeave's $1.1B raise shows the market for alternative clouds is booming, 5 mai 2024.

[54] CEA, L'essentiel sur les supercalculateurs, 7 mars 2022.

[55] Classement TOP500 de juin 2024.

le Grand Équipement National de Calcul Intensif (GENCI[56]) et les organismes de recherche partenaires (CNRS, CEA, etc.) qui hébergent ces centres de données.

5. LES ACTEURS PRINCIPALEMENT PRESENTS A L'AVAL DE LA CHAINE DE VALEUR

88. En aval, des acteurs proposent les produits et services fondés sur de l'IA générative aux utilisateurs finaux ou les intègrent aux flux de travail des entreprises.

a) Grands acteurs de la technologie intégrant les outils d'IA générative

89. Outre les grands acteurs du numérique déjà mentionnés ci-dessus, d'autres acteurs du secteur de la technologie ont commencé à intégrer ces nouveaux outils dans leurs produits et services existants. Par exemple :

- Adobe permet l'utilisation de fonctionnalités d'IA générative dans son outil Photoshop, utilisant son propre modèle propriétaire Firefly et d'autres modèles comme DALL-E ;
- Samsung a lancé en janvier 2024 sa gamme de smartphones Galaxy S24 qui inclut de nombreux outils d'IA générative (traduction en temps réel, retouche de photos, recherche instantanée, etc.) ;
- Zoom Workplace a intégré des outils d'IA générative (Zoom AI Companion) pour proposer des comptes rendus automatisés des appels et des réunions passés par sa plateforme.

b) Acteurs proposant des produits et services à destination des utilisateurs, des entreprises et des acteurs publics

90. Les outils d'IA générative peuvent être à destination des entreprises, des développeurs ou du grand public.

91. La majorité des développeurs de modèles (voir *ci-dessus*) proposent également une interface Internet permettant de tester et d'utiliser leur produit d'IA générative, le plus souvent sous la forme d'un agent conversationnel (pour les modèles de texte ou d'image) accessible gratuitement ou en payant pour accéder aux fonctionnalités plus avancées. Parmi les applications d'IA générative à destination du grand public, il est possible de citer **ChatGPT d'OpenAI, Gemini de Google et Le Chat de Mistral AI** pour la génération de texte, ou bien **MidJourney et StableDiffusion** pour la génération d'images.

92. Au niveau mondial, de nombreuses entreprises proposent des applications, outils ou plateformes fondées sur l'IA générative, ce qui fait dire à un acteur que « *le marché du déploiement est beaucoup plus concurrentiel puisque de très nombreuses entreprises recourent aujourd'hui à des modèles de fondation pour concevoir leurs propres systèmes spécialisés, soit par réentraînement (fine tuning), soit simplement en optimisant le prompting* [instruction] *ou en utilisant la RAG* ».

---

[56] GENCI est une société civile créée en 2007. Elle est détenue à 49 % par l'État via le ministère de l'enseignement supérieur et de la recherche et par le CEA (20 %), le CNRS (20 %), France Universités (10 %) et INRIA (1 %).

93. Plusieurs centaines d'entre elles proposent des applications à destination des entreprises, dans de nombreux types d'activités (tels que les ventes, le marketing, les ressources humaines, la finance, le juridique) et dans tous les secteurs, comme la banque, l'assurance, la santé, les transports, l'agriculture ou l'industrie[57]. Ces applications couvrent différentes modalités, allant de la génération de textes, d'images, de vidéos, à la génération de code informatique. Uniquement au niveau français, plus de 130 *start-ups* proposant des outils dans diverses catégories, comme le design, la productivité, la relation clients, la vente, la santé, la cybersécurité ou la gestion de la connaissance[58].

94. Plusieurs types d'acteurs proposent également des prestations de facilitateurs pour le compte de leurs clients. Tel est le cas d'entreprises de service du numérique (ci-après « ESN ») comme Accenture, Atos ou Capgemini, en compagnie de *start-ups,* comme par exemple Dust ou AleIA au niveau français.

D. UNE PRIORITE CROISSANTE DES AUTORITES PUBLIQUES

95. Bien que le secteur de l'IA générative se soit développé relativement récemment, les pouvoirs publics se sont rapidement mobilisés en France, en Europe et dans le reste du monde.

1. LA STRATEGIE FRANÇAISE POUR L'IA

96. À la suite du rapport « *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne* »[59], le Gouvernement français a lancé en 2018 une stratégie nationale pour l'IA visant à doter la France de capacités de recherche compétitives et à diffuser les technologies d'IA au sein de l'économie. Le lancement du plan « France 2030 », en octobre 2021, visait en outre à développer la compétitivité industrielle et les technologies d'avenir.

97. La première phase de la stratégie nationale (2018-2022) a consisté à renforcer les capacités de recherche de la France en favorisant la création et le développement d'un réseau d'instituts interdisciplinaires d'IA, le soutien à des chaires d'excellence en IA, le financement de programmes doctoraux et l'investissement dans les capacités de calcul de la recherche publique (supercalculateur Jean Zay).

98. La deuxième phase, lancée en 2022, a pour objectif de diffuser l'IA dans l'économie avec trois leviers principaux : la formation et la recherche, le soutien à une offre à l'état de l'art et le rapprochement entre l'offre et la demande en IA. Dans ce contexte, le Président de la République a notamment annoncé neuf lauréats de l'appel à manifestation d'intérêt « IA-clusters » et de nouveaux dispositifs de soutien à l'investissement[60].

---

[57] FirstMark, Machine Learning, AI and Data landscape, avril 2024.

[58] Wavestone, Radar 2023 des startups françaises de l'IA générative, janvier 2024.

[59] Rapport de C. Villani, Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne, remis au premier ministre le 28 mars 2018.

[60] DGE, Annonce de 9 nouveaux lauréats pour l'appel à manifestation d'intérêt « IA-clusters », 22 mai 2024.

99. Un premier appel à projets « *Communs numériques*[61] *pour l'intelligence artificielle générative* » lancé en 2023 avait pour ambition, d'une part, d'accélérer la création et la mise en accessibilité de communs numériques sur l'ensemble de la chaîne de valeur de l'IA générative à condition que le caractère incitatif de l'aide publique soit justifié et avéré et, d'autre part, de développer des produits ou services innovants.

100. Un deuxième appel à projets « *Accélérer l'usage de l'IA générative dans l'économie* », ouvert jusqu'au 2 juillet 2024, se concentre sur la partie aval de la chaîne de valeur en encourageant le développement de solutions d'IA générative intégrées, avec un niveau de fonctionnalité avancé et un horizon d'adoption à court terme. Le Gouvernement souhaite ainsi accompagner 500 petites et moyennes entreprises et établissements de tailles intermédiaires dans l'adoption et l'usage des solutions d'IA d'ici 2025.

101. Par ailleurs, le 19 septembre 2023, le Gouvernement a installé la **Commission de l'IA** précitée, sous la présidence de M. Y… et Mme Z…, réunissant des acteurs de différents secteurs (culturel, économique, technologique, de recherche) afin de « *contribuer à faire de la France un pays à la pointe de la révolution de l'IA* ». Publié le 14 mars 2024, le rapport a émis 25 recommandations dont le coût total est estimé à 27 milliards d'euros.

102. Il y est notamment préconisé :

- le lancement d'un plan de sensibilisation et de formation pour répondre aux besoins actuels et à venir ;
- l'investissement massif dans les entreprises du numérique et la transformation des entreprises pour soutenir l'écosystème français de l'IA et en faire l'un des premiers mondiaux ;
- la mise en place d'un pôle majeur de la puissance de calcul en France ;
- un accès facilité à des données de qualité (données personnelles ou données protégées par un droit de propriété littéraire ou artistique) ;
- l'amélioration des conditions de la recherche publique de l'IA de haut niveau en France ;
- la mise en place d'une gouvernance mondiale de l'IA, en ce compris la nécessité d'assurer un suivi de l'évolution des concentrations de marché et la mise en place rapide de la réglementation nécessaire pour éviter les abus de position dominante.

103. En avril 2024, le Premier ministre a par ailleurs annoncé la création d'un service d'IA baptisé « Albert », permettant notamment de simplifier les démarches administratives des citoyens et d'automatiser certaines tâches comme la retranscription de dépôts de plaintes[62].

---

[61] D'après l'appel à projets « Communs numériques pour l'intelligence artificielle générative », il est notamment entendu ici par « commun numérique » une ressource produite ou entretenue collectivement par une communauté d'acteurs, et gouvernée par des règles qui lui assurent son caractère collectif et partagé. Les communs numériques peuvent par exemple concerner des bases de données d'apprentissage et de test valorisant le patrimoine national de données.

[62] Les Échos, Qu'est-ce qu'Albert, l'intelligence artificielle française déployée par le gouvernement ?, 24 avril 2024.

2. AU NIVEAU EUROPEEN

a) Le règlement européen sur l'intelligence artificielle (« AI Act »)

104. Le 21 avril 2021, la Commission européenne a publié une proposition de règlement concernant l'IA[63]. Après l'approbation du Parlement européen, puis du Conseil de l'Union européenne le 21 mai 2024, le règlement devrait être publié prochainement.

105. D'après les dernières versions disponibles publiquement[64], ce règlement s'appliquera aussi bien aux acteurs du secteur public que du secteur privé, à l'intérieur comme à l'extérieur de l'UE, dès lors que le « *système d'IA* »[65] est mis sur le marché dans l'Union ou que son utilisation a une incidence sur des personnes situées dans l'UE. Il établit des obligations pour les systèmes d'IA en fonction de leurs risques potentiels. Les systèmes dits à risque inacceptable sont interdits s'ils menacent les droits des citoyens (comme l'exploitation de la vulnérabilité des personnes) et les systèmes « *à haut risque* » sont soumis à des obligations strictes[66].

106. Des obligations spécifiques sont par ailleurs imposées aux fournisseurs de « *modèles d'IA à usage général* », notamment les grands modèles d'IA générative (article 53). Ainsi, ils devront notamment rédiger et tenir à jour une documentation technique du modèle, y compris son processus d'entraînement et d'essai ainsi que les résultats de son évaluation afin de la fournir sur demande au Bureau de l'IA (*AI Office*, nouveau centre d'expertise en matière d'IA, chargé de mettre en œuvre les pouvoirs conférés par le règlement à la Commission, de promouvoir l'écosystème européen de l'IA et de collaborer avec les autorités compétentes des États membres dans le cadre de la gouvernance prévue par le règlement) et aux autorités nationales compétentes. Ils doivent mettre à la disposition des fournisseurs de systèmes d'IA qui ont l'intention d'intégrer le modèle d'IA à usage général dans leurs systèmes d'IA des informations et de la documentation. Ils doivent également mettre en place une politique visant à se conformer à la législation européenne sur les droits d'auteur et publier des résumés détaillés des contenus utilisés pour leur entraînement. Ces obligations ne s'appliquent pas aux modèles publiés dans le cadre d'une licence libre et ouverte et dont les paramètres sont rendus publics, sauf si ceux-ci présentent un risque systémique[67].

---

[63] Proposition de règlement de la Commission européenne établissant des règles harmonisées concernant l'intelligence artificielle du 21 avril 2021.

[64] Version du texte datée du 13 juin 2024 disponible sous le lien suivant : https://data.consilium.europa.eu/doc/document/PE-24-2024-REV-1/en/pdf.

[65] Le système d'IA est défini à l'article 3 comme suit : « *un système automatisé conçu pour fonctionner à différents niveaux d'autonomie, qui peut faire preuve d'une capacité d'adaptation après son déploiement et qui, pour des objectifs explicites ou implicites, déduit, à partir des données d'entrée qu'il reçoit, la manière de générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels* ».

[66] Cette classification dépend de la fonction exécutée par le système d'IA, ainsi que du but spécifique dans lequel le système est utilisé et des modalités de cette utilisation. L'annexe III du règlement fait par exemple référence aux systèmes utilisés en lien avec le domaine de la « *répression, dans la mesure où leur utilisation est autorisée par la législation nationale ou de l'Union applicable* » comme le profilage par exemple.

[67] Selon l'article 3(65) du projet de règlement, le risque systémique est défini comme « *un risque spécifique aux modèles d'IA à usage général ayant un impact significatif notamment sur le marché de l'Union en raison de leur portée, ou en raison d'effets négatifs réels ou raisonnablement prévisibles sur la santé et la sécurité*

107. Plusieurs organes de gouvernance sont créés au niveau de l'Union comme un bureau de l'IA, qui développe l'expertise et les capacités de l'Union dans le domaine de l'IA, et un Comité européen de l'IA, composé d'un représentant par État membre, à des fins notamment de coordination. Chaque État membre établit ou désigne en tant qu'autorités nationales compétentes au moins une autorité notifiante[68] et au moins une autorité de surveillance du marché.

108. Le règlement entrera en vigueur 20 jours après sa publication au *Journal officiel* de l'Union européenne et sera pleinement applicable 24 mois après son entrée en vigueur, soit a priori dans le courant de l'année 2026, à l'exception de certaines dispositions comme les règles de classification pour les systèmes à haut risque (qui seront applicables 36 mois après l'entrée en vigueur).

b) Autres règlements européens susceptibles d'avoir un impact sur l'IA

Le règlement sur les marchés numériques

109. Le règlement sur les marchés numériques (le *Digital Markets Act,* ci-après « **DMA** »)[69] a été adopté le 14 septembre 2022 pour encadrer les pratiques des géants du numérique. Il prévoit certaines obligations qui pourraient, sous réserve de l'appréciation de la Commission, s'appliquer dans le secteur de l'IA dès lors que des services essentiels de plateforme désignés par la Commission européenne, comme des moteurs de recherche, des réseaux sociaux ou des assistants vocaux, intègrent des services d'IA[70]. Il s'agit notamment des obligations qui suivent[71] :

- l'interdiction de combiner ou d'utiliser de manière croisée les données à caractère personnel provenant d'un service de plateforme essentiel avec les données provenant de tout autre service, sauf consentement des utilisateurs (article 5(2)) ;
- l'interdiction d'utiliser les données non publiques, y compris celles produites par les entreprises utilisatrices (article 6(2)) ;
- l'obligation pour les contrôleurs d'accès d'assurer la portabilité effective et gratuite des données fournies par l'utilisateur final ou produites par l'activité de l'utilisateur final dans le cadre de l'utilisation du service de plateforme essentiel concerné (article 6(9)) ;

---

publiques, les droits fondamentaux ou la société dans son ensemble, susceptibles d'être propagés à grande échelle sur la chaîne de valeur » (traduction libre). Selon les conditions posées par l'article 51 du projet de règlement, un modèle d'IA à usage général a un risque systémique s'il remplit deux conditions : le fait d'avoir des capacités à fort impact selon une méthodologie technique (« *est présumé avoir des capacités à fort impact* (…) *lorsque la quantité cumulée de calcul utilisée pour son entraînement mesurée en opérations en virgule flottante est supérieure à 10^25* ») ou une décision de la Commission.

[68] D'après l'article 28 du projet de règlement, chaque État membre désigne ou établit au moins une autorité notifiante chargée de mettre en place et d'appliquer les procédures nécessaires à l'évaluation, à la désignation et à la notification des organismes d'évaluation de la conformité, ainsi qu'à leur contrôle.

[69] Règlement (UE) 2022/1925 du Parlement européen et du Conseil du 14 septembre 2022 relatif aux marchés contestables et équitables dans le secteur numérique (publié le 12 octobre 2022).

[70] Les obligations du DMA ne peuvent s'appliquer qu'aux services de plateformes essentiels des contrôleurs d'accès qui sont visés à l'article 2 du DMA, parmi lesquels ne figurent pas les MaaS (models-as-a-service), voir développements *infra* et la proposition n° 1 de l'Autorité.

[71] L'obligation d'information de la Commission de tout projet de concentration (article 14) est examinée *supra* aux paragraphes 394 et suivants.

- l'obligation de fournir aux utilisateurs professionnels un accès gratuit, continu et en temps réel aux données fournies ou produites dans le cadre de l'utilisation du service de plateforme essentiel concerné, y compris les données personnelles (article 6(10)) ;
- l'obligation de partager les données concernant les classements, requêtes, clics et vues aux moteurs de recherche tiers (article 6(11)).

Le règlement européen sur les données

110. Le règlement européen portant sur l'équité de l'accès aux données et de l'utilisation des données (en anglais, « *Data Act* »)[72] a pour objectif de supprimer les obstacles à l'accès aux données, tant pour les organismes du secteur privé que pour les organismes du secteur public, tout en préservant les incitations à investir dans la production de données, en assurant un contrôle équilibré des données pour leurs créateurs.

111. Le règlement établit des règles en faveur de l'ouverture accrue des données issues des produits connectés afin de permettre le développement d'une économie de la donnée véritablement compétitive et équitable, qui profite à la fois aux utilisateurs et aux entreprises européennes. Il vise également à prévenir l'exploitation abusive de déséquilibres contractuels en ce qui concerne l'accès aux données et leur utilisation, comme des clauses permettant à une partie qui a imposé unilatéralement une clause d'avoir accès aux données de l'autre partie contractante et de les utiliser d'une manière qui porte atteinte notamment à ses droits de propriété intellectuelle (article 13 (b)). Il donne également aux autorités publiques le droit d'accéder aux données détenues par les entreprises dans des situations exceptionnelles et relevant de l'intérêt public. Il vise enfin à lever les principales barrières au recours à des services de *cloud* concurrents en instaurant par exemple la suppression progressive des frais de changement de fournisseur (article 29) et des mesures visant à faciliter le changement de fournisseur du point de vue technique (article 30).

112. En France, la loi n° 2024-449 du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique (ci-après « loi SREN ») a pour objectif d'anticiper certaines dispositions du règlement relatives au secteur du *cloud* (voir supra paragraphe 252). Elle a été adoptée le 21 mai 2024.

3. LES REGLES MISES EN PLACE DANS LE RESTE DU MONDE

113. Au-delà de la France et de l'Europe, des initiatives sont également mises en œuvre dans le reste du monde.

114. Plusieurs organisations internationales ont adopté des principes communs. Le 30 octobre 2023, les dirigeants du G7 ont ainsi adopté un accord portant sur des principes directeurs internationaux en matière d'IA et un code de conduite volontaire dans le cadre du processus d'Hiroshima[73]. La présidence italienne du G7, dont l'IA fait partie des priorités, a

[72] Règlement (UE) 2023/2854 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 (règlement sur les données).

[73] G7, Code de conduite international pour les systèmes d'IA avancés dans le cadre du processus Hiroshima, 30 octobre 2023. Le code de conduite contient 11 recommandations non contraignantes en faveur d'une « IA sûre, sécurisée et digne de confiance », y compris les modèles de fondation les plus avancés et les systèmes d'IA générative.

adopté une déclaration ministérielle le 15 mars 2024 visant à faire progresser les travaux en faveur d'une IA sûre et digne de confiance[74]. D'autres initiatives internationales ont vu le jour comme la convention-cadre sur l'IA adoptée récemment par le Conseil de l'Europe[75].

115. En novembre 2023, le premier sommet pour la sécurité dans l'IA (« *AI Safety Summit* »), organisé par le Royaume-Uni, a adopté « *la déclaration Bletchley* »[76] signée par 28 pays et l'Union européenne afin de favoriser une compréhension commune des risques technologiques posés par l'IA et développer des coopérations internationales sur la sécurité de ces systèmes. Après la Corée, **la prochaine édition de ce sommet aura lieu en France les 10 et 11 février 2025**[77].

116. Au niveau national, une série de mesures volontaristes sont mises en œuvre.

117. Après avoir privilégié une approche fondée sur des engagements volontaires des acteurs de l'IA[78], un décret présidentiel (« *executive order* ») portant sur la régulation de l'IA a été publié par l'administration Biden le 30 octobre 2023[79]. Le 27 avril 2024, le gouvernement américain a mis en place un conseil fédéral de l'IA destiné à fournir des recommandations pour garantir l'adoption sécurisée de l'IA aux États-Unis.[80] Ce conseil est composé des dirigeants des plus grandes entreprises du secteur (comme OpenAI, Microsoft, Google, ou Nvidia), de représentants du Gouvernement et de chercheurs.

118. Au mois de février 2024, le Royaume-Uni a précisé son approche sur la régulation de l'IA en confiant la régulation des systèmes d'IA aux régulateurs sectoriels qui s'appuieront notamment sur les cinq principes non statutaires édictés par le Gouvernement et destinés à guider leur action : la sûreté des systèmes, la transparence des systèmes, la comptabilité des systèmes avec les lois existantes, la responsabilité des systèmes et leur contestabilité. Il s'est également engagé à développer à plus long terme des règles contraignantes pour « *le petit nombre d'entreprises développant des systèmes d'IA générale* », sans pour autant à ce stade communiquer sur des mesures en ce sens[81].

119. En août 2023, l'administration chinoise en charge du cyberespace a publié des « mesures provisoires » destinées à imposer des obligations notamment aux fournisseurs d'IA

---

[74] Déclaration ministérielle du G7 des 14 et 15 mars 2024.

[75] Le 17 mai 2024, le Conseil de l'Europe a adopté la convention-cadre sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit. Cette convention, signée par 46 pays incluant les États membres de l'UE, ainsi que d'autres pays comme les États-Unis, le Canada et le Japon, vise à établir des règles de respect des droits fondamentaux, face aux risques de discrimination ou d'atteintes à la vie privée lors de l'utilisation de technologies d'IA.

[76] The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 novembre 2023.

[77] Le Monde, Emmanuel Macron veut faire de la France « un des pays champions de l'IA », 22 mai 2024.

[78] Contexte, À Washington, le gratin de l'IA promet de s'autoréguler, 24 juillet 2023.

[79] The White House, FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, 30 octobre 2023. Il vise notamment à (i) mettre en place une série de standards afin de garantir la création d'outils d'IA sûrs et sécurisés avant leur diffusion publique (ii) demander au Congrès d'adopter une loi sur la protection des données personnelles, (iii) s'attaquer aux questions de discrimination algorithmique, (iv) promouvoir l'innovation et la concurrence, sous l'autorité notamment de la Federal Trade Commission et (v) renforcer la coopération sur l'IA au niveau international.

[80] Le Monde, Intelligence artificielle : création d'un conseil fédéral pour aider le gouvernement américain, 27 avril 2024.

[81] Direction générale du Trésor, Brèves numériques Royaume-Uni, 14 décembre 2023 au 7 février 2024.

générative. Ceux-ci ont l'obligation de procéder à des évaluations de sécurité et de soumettre aux autorités des rapports sur leurs outils, notamment si ceux-ci sont susceptibles d'influencer l'opinion publique.[82] En complément, plusieurs mesures sur l'intelligence artificielle sont en cours d'élaboration, notamment pour favoriser le développement industriel.[83] Lors de la visite du Président de la République Populaire de Chine, il a été annoncé que la France participera à la Conférence mondiale pour l'IA et à la rencontre de Haut niveau sur la Gouvernance mondiale de l'IA que la Chine organisera en 2024[84].

120. Concernant les autorités de concurrence, de nombreux travaux ont été menés ou sont en cours sur les enjeux concurrentiels soulevés par le secteur de l'IA générative. Au-delà des autorités précitées (voir *infra* paragraphe 8), les autorités canadienne[85], indienne[86] et hongroise[87] ont également annoncé le lancement d'études sur le secteur.

[82] La Tribune, IA générative : la Chine instaure de nouvelles réglementations, 18 août 2023.

[83] Forbes, China's New Draft AI Law Prioritizes Industry Development, 22 mars 2024.

[84] Déclaration conjointe entre la République française et la République populaire de Chine sur l'intelligence artificielle et la gouvernance des enjeux globaux, 6 mai 2024.

[85] Bureau de la Concurrence du Canada, Le Bureau de la concurrence sollicite des commentaires sur l'intelligence artificielle et la concurrence, 20 mars 2024.

[86] The Telegraph, Competition Commission of India to undertake market study on AI, 6 mars 2024.

[87] Hungarian Competition Authority launches study on impact of AI on competition and consumers, 17 janvier 2024.

II. Analyse concurrentielle

121. Le secteur de l'IA générative est marqué par d'importantes barrières à l'entrée (A) susceptibles de favoriser les grands acteurs du numérique bénéficiant par ailleurs d'avantages liés à leurs activités sur d'autres marchés numériques (B). Malgré le développement récent du secteur, des risques concurrentiels peuvent apparaître, notamment à l'amont de la chaîne de valeur (C).

A. UN SECTEUR MARQUE PAR DES BARRIERES A L'ENTREE ELEVEES

122. Comme indiqué *supra*, l'IA générative nécessite trois intrants clés, dont l'accès ou la détention constituent autant de barrières à l'entrée (1). Ceux-ci nécessitent par ailleurs de lourds investissements (2). Certaines innovations pourraient toutefois limiter ces barrières à l'entrée (3).

1. LES INTRANTS NECESSAIRES AU DEVELOPPEMENT DES MODELES DE FONDATION PEUVENT CONSTITUER DES BARRIERES A L'ENTREE

123. L'entraînement et l'inférence de modèles d'IA générative nécessitent des intrants essentiels, comme la puissance de calcul, les données et les talents.

a) La nécessité d'avoir recours à des processeurs graphiques ou d'autres processeurs spécialisés pour l'IA

124. Comme indiqué *supra*, l'entraînement de modèles de fondation de l'IA générative implique des besoins en matériel informatique spécifiques. Les puces utilisées doivent être capables d'effectuer un grand nombre d'opérations en parallèle et nécessitent une forte précision pour déterminer précisément plusieurs milliards de paramètres. Les puces les plus utilisées pour réaliser ces tâches sont les GPU produites notamment par Nvidia, même si plusieurs grands acteurs du numérique développent leurs propres accélérateurs d'IA.

125. Il ressort de la consultation publique menée par l'Autorité que ces GPU et autres accélérateurs d'IA ne peuvent pas facilement être remplacés par des CPU, compte tenu de leur performance. Comme l'indiquent plusieurs acteurs du secteur, les GPU fournissent une puissance et une capacité de calcul supérieure à celles des CPU. Le secteur de l'IA se caractérise par une course à l'entraînement de modèles de plus en plus performants, où seules des GPU peuvent permettre un entraînement suffisamment rapide, compatible avec la vitesse d'évolution du marché. Un autre acteur du secteur considère que « *le temps d'apprentissage nécessaire pour le développement de services d'IA générative serait multiplié par 1 000 sans accès aux GPUs* ».

126. Compte tenu de l'explosion de la demande en calcul spécifique à l'IA depuis deux ans, les utilisateurs de GPU connaissent des difficultés d'approvisionnement. Plusieurs acteurs confirment que la création d'une infrastructure disposant d'une puissance de calcul suffisante est particulièrement coûteuse et difficile, compte tenu des pénuries générées par une forte demande et d'une offre limitée de semi-conducteurs.

127. Par ailleurs, au-delà du matériel informatique (le « *hardware* »), la création de modèles de fondation nécessite également une couche logicielle (le « *software* ») permettant d'exécuter le code informatique directement sur la carte graphique, afin de pouvoir distribuer le calcul au mieux sur la carte graphique et donc optimiser ses performances.

128. L'environnement logiciel propriétaire CUDA développé par Nvidia, exclusif à ses propres puces, est le plus utilisé par les acteurs du secteur, ce que confirme un acteur : « *CUDA, un cadriciel bas-niveau de calcul matriciel utilisé quasiment systématiquement par les bibliothèques d'apprentissage profond de plus haut niveau citées précédemment, est prédominant dans la pratique d'entraînement des LLMs ; il a été développé par NVIDIA et est exclusivement associé aux GPUs de cette marque. Son concurrent sur le segment des GPUs, AMD, a quant à lui développé un cadriciel analogue (ROCm) mais son utilisation semble très marginale. De manière plus large, il existe une intrication fondamentale entre le matériel* (…) *et les logiciels utilisables sur étagère pour effectuer de l'IA(G)* [IA générative] *: ils forment des écosystèmes dont la maturité et la taille des communautés sont de nature à entraîner des concentrations et de l'inertie* ».

b) Le *cloud*, un passage obligé pour accéder à la puissance de calcul

Une infrastructure sur site très coûteuse

129. À l'heure actuelle, la plupart des entreprises du secteur ne gèrent pas leurs propres infrastructures sur site en raison des coûts élevés qui y sont associés. Un acteur précise ainsi que « [l]*e coût d'acquisition de ce type de matériel n'est pas à la portée de toutes les entreprises. Le prix des GPU H100 de Nvidia, qui constituent actuellement la référence dans ce domaine, s'élève à 30 à 40 000 euros l'unité. Si un GPU peut suffire pour l'inférence, l'entraînement d'un modèle de fondation suppose de disposer de plusieurs milliers de tels processeurs (les plus grands modèles utilisant plusieurs dizaines de milliers de GPU)* ».

130. Ainsi, seuls quelques acteurs, comme Meta ou Samsung, peuvent atteindre la puissance de calcul nécessaire aux différentes étapes du développement de modèles d'IA générative avec une infrastructure sur site. En plus des coûts d'investissements initiaux, ces acteurs doivent prendre en charge l'exploitation, la maintenance et l'évolution des serveurs. Un acteur précise « [qu'a]*u-delà du coût d'acquisition des GPU, il existe des coûts d'usage (électricité, refroidissement) et de mise en œuvre (notamment local à aménager) importants. Les investissements nécessaires pour entrer et se développer sur ce marché sont donc considérables et leur amortissement par une entreprise qui se bornerait à concevoir un modèle de fondation pour ses propres besoins serait très incertain* ».

131. Ces contraintes limitent la possibilité pour un nouvel acteur de l'IA générative de développer son infrastructure interne pour entraîner ses propres modèles. Ainsi, un acteur spécifie « [qu']*en pratique, seul un petit nombre d'entreprises disposent des infrastructures suffisantes pour atteindre la puissance de calcul nécessaire au développement de l'IA* ». Un autre précise « [qu']*une entreprise spécialisée dans le développement de LLM n'a pas vocation à développer sa propre infrastructure* ».

132. Un acteur du marché indique également que « *la recherche dans les GPU étant extrêmement dynamique, ce matériel peut être frappé d'*« *obsolescence » à une échéance rapprochée. Il peut être judicieux dans ces conditions de recourir à de la puissance de calcul "as a service"* ». Cette obsolescence matérielle rapide est caractéristique du marché de l'IA, notamment générative, et renforce les contraintes liées à l'usage d'une infrastructure sur site.

133. En revanche, la puissance de calcul nécessaire pour le réglage fin est bien inférieure à celle requise pour l'entraînement. Ceci est lié au volume de données indispensable bien moindre pour cette phase.

134. De même pour la phase d'inférence, si les besoins en calcul sont dépendants de la sollicitation du modèle (par exemple liée au nombre d'utilisateurs), ils ne nécessitent pas des processeurs graphiques aussi performants (et donc aussi onéreux) que pour l'entraînement.

135. Ceci pourrait permettre à certaines grandes entreprises ayant déjà des infrastructures de calcul sur site dans le cadre de leur activité de les faire évoluer pour permettre le réglage fin ou l'inférence de modèles d'IA générative en interne.

Le cloud est la solution privilégiée pour l'entraînement ou la spécialisation des modèles, et permet également de faciliter le déploiement à l'aval

136. Le *cloud* donne aux entreprises accès à une infrastructure informatique à la demande, évolutive et adaptable en fonction des besoins. En complément de l'infrastructure, il facilite également l'accès des utilisateurs à de nombreux services managés (PaaS et SaaS).

137. D'après un acteur « *du point de vue des développeurs, il n'existe pas de différence technique fondamentale entre les ressources de calcul fournies par les fournisseurs de cloud et par une infrastructure sur site. Les développeurs de MF* [modèles de fondation] *sont souvent des « natifs du numérique » qui préfèrent utiliser une infrastructure cloud modulable et rentable plutôt que d'engager des coûts importants en investissant dans une infrastructure sur site* ».

138. L'Autorité a identifié, dans son avis n° 23-A-08 précité[88], les avantages et inconvénients du recours à une infrastructure *cloud.* Ceux-ci sont les mêmes pour l'entraînement de modèles d'IA générative et sont confirmés par les acteurs. Ainsi, le *cloud* offre l'avantage d'éviter les coûts d'investissements initiaux et de maintenance, et permet une tarification à l'usage dépendant uniquement des besoins de l'entreprise. Il offre également une flexibilité et un accès rapide aux technologies les plus avancées.

139. S'agissant du secteur de l'IA, les fournisseurs de service *cloud* (ci-après FSC) jouent un double rôle, à la fois en amont pour l'entraînement ou le réglage fin des modèles, mais également à l'aval où ils constituent une plateforme privilégiée pour le déploiement de modèles de fondation à disposition des entreprises. Ainsi, un acteur indique que « [p]*our une entreprise plus classique qui développerait une spécialisation sur des modèles pré-entrainés pour ses activités, il est très intéressant de recourir aux services "sur étagère" (modèles à dispo) des FSC qui permettent de partir de modèles pré-entrainés (sorte de marketplace de modèles IA) et de bénéficier de manière ponctuelle de la puissance de calcul nécessaire au moment de la spécialisation dans des conditions tarifaires raisonnables.*»

140. Les services de type MaaS tels que Model Garden sur Google Cloud, Amazon Bedrock sur AWS ou Azure AI sur Microsoft Azure semblent constituer des **points de contact majeurs entre les développeurs de modèles et les entreprises utilisatrices**, qui sont le plus souvent déjà clientes du fournisseur de services *cloud*. Ces services simplifient l'accès aux modèles d'IA pour les entreprises, souvent via une interface de programmation et leur permettent de déployer plus facilement des applications tirant parti de l'IA générative.

141. Les développeurs de modèles ont ainsi tout intérêt à ce que leurs modèles soient présents chez le plus grand nombre de fournisseurs de services *cloud* pour maximiser le nombre de

[88] Voir l'avis n° 23-A-08 de l'Autorité, paragraphes 18 à 22, pages 23 et 24.

leurs clients potentiels. Le rôle de point de passage obligé des fournisseurs de services *cloud*, à la fois pour l'entraînement, mais également pour l'inférence est, partant, renforcé.

142. Ainsi, l'IA générative est présente sur toutes les couches du *cloud*. Les services IaaS (calcul, stockage, etc.) et PaaS (bases de données vectorielles, outils d'IA, etc.) permettent l'entraînement et la spécialisation des modèles, tandis que de plus en plus de services SaaS incluent des outils d'IA générative. Les développeurs peuvent également proposer leurs modèles ou services d'IA générative via les places de marché du *cloud*.

c) L'entraînement des modèles nécessite un vaste ensemble de données

143. En l'état actuel de la technologie de l'IA générative, fondée sur les grands modèles de langage, les données sont indispensables pour l'entraînement et le réglage fin des modèles, ainsi que pour l'inférence, lorsque des techniques comme le RAG ou le « *grounding* » sont utilisées. Grâce aux données qui lui sont présentées, le modèle peut apprendre à créer du contenu.

Des données en grand nombre et de qualité suffisante sont nécessaires à l'entraînement de modèles d'IA générative

144. Le volume de données est crucial pour l'entraînement de modèles d'IA générative. En 2020, un article de recherche publié par des ingénieurs de la société OpenAI estimait que la performance des grands modèles de langage augmentait avec le nombre de données fournies à l'entraînement du modèle, lançant une course à la taille des modèles[89]. Les acteurs du secteur confirment l'importance d'une « *grande quantité de données hétérogènes comme du texte, des images, de l'audio, des vidéos en fonction du type de contenu que le modèle est destiné à générer* ».

145. Les développeurs de modèles communiquent peu sur les données utilisées pour l'entraînement. À titre d'exemple, Meta indique que plus de 15 000 milliards de *tokens* (voir glossaire) ont été utilisés pour l'entraînement de son modèle Llama 3, soit 7 fois plus que son prédécesseur Llama 2, publié moins d'un an auparavant pour le même nombre de paramètres[90].

146. Les contributions à la consultation publique insistent également sur la qualité des données, notamment pour les futurs modèles. Ainsi, un acteur indique qu'il « *s'attend à ce que le succès des futurs modèles en accès libre et propriétaires dépende davantage de la qualité ou de la pertinence des données par rapport à la tâche à accomplir ou de la plus grande performance des algorithmes que de l'utilisation de plus grands volumes de données* ».

147. La qualité des données résulte principalement de leur nettoyage et des traitements qui leur sont apportés, notamment pour exclure les données de mauvaise qualité. Ces traitements constituent une étape nécessaire avant l'entraînement et également un facteur de différenciation entre acteurs. Ainsi, un acteur indique « [qu']*il est rare que les concepteurs de modèles de fondation dévoilent de manière extensive la composition détaillée de leurs bases de données puisqu'il s'agit d'un avantage concurrentiel majeur* » et un autre confirme que cette étape initiale de nettoyage et de notation des données peut prendre plusieurs mois.

---

[89] Kaplan et al, Scaling Laws for Neural Langage Models, janvier 2020.

[90] Communiqué de Meta, Introducing Meta Llama 3 : The most capable openly available to date, 18 avril 2024.

148. Les acteurs du marché font la différence entre les données utilisées pour l'entraînement et celles utilisées pour le réglage fin. Ainsi, l'un d'entre eux indique « [qu']*en ce qui concerne le pré-entraînement, les données utilisées sont des données en accès libre, générales et nombreuses,* (…). *Ces données visent à entraîner le modèle de langue à l'acquisition des connaissances générales.* (…) *en ce qui concerne la spécialisation (réglage fin ou fine-tuning), les données généralement utilisées sont soit des données en accès libre, récentes, n'ayant pas été utilisées durant le pré-entraînement, soit des données internes à l'entreprise (données propriétaires, plus précises, spécialisées, techniques mais souvent de quantité moindre)* ». Lors de l'inférence, des techniques telles que le RAG peuvent nécessiter l'utilisation de données pertinentes selon le cas d'usage, qui peuvent être des données internes d'entreprises ou des données d'actualité. L'utilisation d'autres types de données, comme les données synthétiques, sera développée *infra*.

Les modèles d'IA générative sont principalement entrainés sur des données publiques

149. Il ressort des contributions des acteurs à la consultation publique que la majorité des modèles d'IA générative sont principalement entrainés sur des bases de données publiquement accessibles. Ces données, souvent décrites par les acteurs comme « *publiques* », contiennent les jeux de données publics ou le contenu accessible sur Internet, même si certaines de ces données peuvent être protégées, par exemple au titre du droit d'auteur. Un acteur indique à cet égard que « *les données provenant de ces sources publiquement disponibles constituent typiquement la majorité des données utilisées pour l'entraînement des MF. Ils sont souvent complétés par une plus petite quantité de données propriétaires et de données de tiers* ».

150. De nombreux jeux de données sont accessibles librement sur Internet, et sont fréquemment utilisés par les développeurs pour l'entraînement des modèles de fondation. Parmi les plus connus et utilisés, on peut citer :

- Common Crawl, organisation à but non lucratif fondée en 2007 aux États-Unis. Elle a pour mission de fournir gratuitement des archives d'Internet. Depuis 2008, de nombreux *crawl* (voir glossaire) d'Internet ont été réalisés, et le dernier, daté de mai 2024 regroupe quasiment 3 milliards de pages Internet, soit un peu moins de 400 téraoctets de données ;
- C4 (*Colossal Cleaned Crawl Corpus*), version filtrée du Common Crawl, uniquement en anglais (mC4 est la version multilingue) et publiée par Google ;
- LAION-5B (*Large Scale Artificial Intelligence Open Network*), un jeu de données publié en 2022, contenant quasiment 6 milliards d'images annotées avec des descriptions en anglais et dans d'autres langues. LAION est une organisation à but non lucratif qui fournit des données et des modèles pour encourager la recherche en IA ;
- The Pile, jeu de données textuelles ouvert composé de 22 sous-jeux de données, tels que des livres, des publications scientifiques, des blogs, etc., publié par EleutherAI, un groupe de recherche en IA à but non lucratif fondé en 2020.

151. Outre ces jeux de données historiques, de nombreux autres jeux de données sont fréquemment publiés, ce qui peut contribuer à réduire certaines barrières à l'accès aux données pour l'entraînement de modèles d'IA générative. Par exemple :

- Common Corpus, une collection de données multilingue ouverte contenant plus de 500 milliards de mots [91];
- FineWeb, un jeu de données de 15 000 milliards de *tokens* issus du Common Crawl publié par HuggingFace en mai 2024 ;
- YouTube Commons, un jeu de données publié en avril 2024 contenant la retranscription de plus de deux millions de vidéos YouTube.

152. En complément de ces jeux de données individuels, plusieurs plateformes permettent d'accéder à de nombreux jeux de données, comme **Github** ou **Hugging Face**. Cette dernière propose ainsi plus de 158 000 jeux de données de tout type sur sa plateforme, incluant certains des jeux de données présentés ci-dessus.

153. Au niveau européen, plusieurs initiatives d'espaces de **données sectorielles** (en anglais « *data spaces* ») ont vu le jour, telles que CatenaX dans le secteur de l'automobile, AgDataHub dans l'agriculture et l'agroalimentaire ou EonaX pour le transport aérien. Ces espaces de données peuvent par exemple permettre la spécialisation d'un modèle d'IA générative sur le secteur en question plutôt que d'avoir recours à des données **propriétaires ou tierces.**

154. Pour le réglage fin, si les réponses des acteurs confirment l'utilisation principale de données propriétaires non publiques, il existe également des exemples de jeux de données de réglage fin publics comme **OpenOrca**, un jeu de données publié en 2023 contenant des instructions pour permettre aux modèles de répondre à des questions d'utilisateurs.

L'accès à la donnée publique fait face à des incertitudes

155. Le manque de données publiquement accessibles et de qualité suffisante pourrait contraindre les acteurs à utiliser un volume plus important de données propriétaires pour l'entraînement de modèles d'IA générative.

156. Premièrement, les développeurs de modèles s'inquiètent des problèmes juridiques engendrés par l'utilisation de bases de données telles que le Common Crawl. En effet, plusieurs fournisseurs de contenu font désormais valoir leurs droits, ce qui a pour conséquence de réduire les facilités d'accès et/ou d'empêcher la collecte de données par des robots d'indexation. Certains éditeurs de presse comme le New York Times ont lancé **des actions en justice** contre Microsoft et OpenAI sur le fondement du droit d'auteur[92], tandis qu'en France, l'ADAGP (la société des auteurs dans les arts graphiques et plastiques) propose un guide pratique à destination des éditeurs pour procéder à l'*opt-out* de l'utilisation des contenus à des fins de recherche en IA[93]. Un acteur du secteur relève à cet effet « [qu']*on ne peut plus refaire le scraping "complet" d'internet comme ça a été fait par OpenAI pour entraîner GPT-3. Il est impossible de le faire pour des raisons de droits d'auteurs notamment, avec tous les procès autour d'OpenAI et Google à ce sujet* ».

[91] Le Common Corpus et Youtube Commons sont publiés par Pleias, une *start-up* française soutenue par le ministère de la culture et la Direction interministérielle du numérique (DINUM).

[92] Le Monde, Le « New York Times » poursuit en justice Microsoft et OpenAI, créateur de ChatGPT, pour violation de droits d'auteur, 27 décembre 2023.

[93] ADAGP, L'ADAGP prend des mesures pour protéger ses membres face à la menace des intelligences artificielles génératives, 23 février 2024.

157. Deuxièmement, certaines plateformes communautaires comme Reddit et Twitter ont augmenté le prix de leurs interfaces de programmation, souvent utilisées par les développeurs pour collecter des données, au cours de l'année 2023, dans l'objectif de mieux valoriser leurs données propriétaires, utilisées notamment pour l'entraînement de modèles d'IA générative[94].

158. Face à ces risques, certains développeurs de modèles nouent des **partenariats** avec des éditeurs et ayants droit. Par exemple, Google a signé des accords avec Reddit[95] et StackExchange[96]. OpenAI a noué des accords avec de nombreux fournisseurs de contenus et éditeurs de presse dans plusieurs pays, comme Associated Press aux États-Unis[97] et Le Monde en France[98]. Le tableau ci-dessous recense tous les accords relatifs à l'utilisation des données annoncés par OpenAI à ce jour. Des observateurs indiquent qu'OpenAI a approché plusieurs groupes de presse avec des offres allant de 1 à 5 millions de dollars par an[99], mais l'accord avec News Corp serait d'un un montant bien supérieur, proche des 250 millions de dollars (soit plus de 230 millions d'euros) sur 5 ans (soit 50 millions par an)[100].

Tableau : accords passés entre OpenAI et des fournisseurs de contenu (à jour au 20 juin 2024)

Fournisseur de contenu	Pays	Date accord
The Atlantic	États-Unis	29/05/2024
Vox Media Inc.	États-Unis	29/05/2024
News Corp	Royaume-Uni	22/05/2024
Reddit	États-Unis	16/05/2024
Dotdach Meredith	États-Unis	07/05/2024
Financial Times	Royaume-Uni	29/04/2024
Le Monde	France	13/03/2024
Prisa Media	Espagne	13/03/2024
Axel Springer	Allemagne	13/12/2023
Associated Press (AP)	États-Unis	13/07/2023

Source : Autorité de la concurrence et annonces d'OpenAI.

[94] Forbes, Death By API: Reddit Joins Twitter In Pricing Out Apps, 1er juin 2023.

[95] Le Figaro, IA : Reddit noue un accord de licence inédit avec Google pour 60 millions de dollars, 22 février 2024.

[96] Blog de StackOverflow, Stack Overflow and Google *Cloud* Announce Strategic Partnership to Bring Generative AI to Millions of Developers, 29 février 2024.

[97] Associated Press, AP, Open AI agree to share select news content and technology in new collaboration, 13 juillet 2023.

[98] Le Monde, Intelligence artificielle : un accord de partenariat entre « Le Monde » et OpenAI, 13 mars 2024.

[99] The Verge, OpenAI's news publisher deals reportedly top out at $5 million a year, 4 janvier 2024.

[100] Les Echos, « Méga accord » entre OpenAI et News Corp, 23 mai 2024.

d) Des compétences techniques rares et très recherchées

159. Les compétences techniques représentent une autre ressource nécessaire à la conception des modèles de fondation. Les développeurs de modèle doivent avoir des connaissances approfondies en science des données, en apprentissage automatique (« *machine learning* ») et en apprentissage profond, des compétences en traitement automatique du langage naturel (NLP, voir glossaire) ou en vision par ordinateur (« *computer vision* »), des connaissances en ingénierie et en développement opérationnel (« *DevOps* »), pour être capables à la fois de développer le code et de mettre en place l'architecture pour l'exécuter de manière performante. En plus de leur formation théorique, les ingénieurs doivent avoir travaillé avec des réseaux de neurones et plus spécifiquement de type *transformers*.

160. En effet, l'entraînement d'un modèle de fondation repose sur de nombreux hyperparamètres (voir glossaire). Leur maitrise et leur optimisation ne s'acquièrent que par l'expertise empirique, chacun ayant une incidence croisée sur la performance du modèle.

161. Ainsi, un acteur évoque la « *capacité à rester à la pointe de la recherche fondamentale (nouvelles architectures de modèles, nouvelles innovations de rupture dans les paradigmes d'apprentissage) et de la recherche plus appliquée (optimisations à la marge)* » comme compétence indispensable pour exceller dans ce secteur à la pointe de l'innovation. Un autre précise cependant « [qu']*un nombre croissant de scientifiques des données et d'ingénieurs ont été attirés vers ce secteur dynamique* » en vue du grand intérêt pour le secteur.

162. En revanche, il convient de préciser qu'il n'est pas nécessaire de disposer de très grandes équipes pour développer des modèles de langage. Ainsi, de nombreuses *start-ups* se sont construites autour d'un petit nombre de talents et ont développé des modèles par la suite, à l'instar de Mistral AI qui a annoncé son premier modèle en septembre, quelques mois après avoir été fondée, alors qu'elle ne comptait que 22 salariés en décembre 2023[101].

2. LA LOURDEUR DES INVESTISSEMENTS REQUIERT LA CONCLUSION D'ACCORDS ENTRE GRANDS ACTEURS ET DEVELOPPEURS DE MODELES DE FONDATION

163. Comme développé *supra*, le développement de modèles de fondation nécessite des investissements importants pour disposer d'une puissance de calcul suffisante, de très grands ensembles de données et l'accès à un savoir-faire hautement spécialisé.

164. L'ampleur des investissements requis créée des barrières à l'entrée significatives. Ainsi, selon des estimations, l'entraînement du modèle de fondation GPT-3 d'OpenAI aurait coûté à lui seul plus de 4 millions de dollars (environ 3,7 millions d'euros) et le modèle GPT-4 qui a suivi plus de 78 millions de dollars (environ 72 millions d'euros)[102].

165. Tel est d'autant plus le cas que ces investissements importants doivent être renouvelés continuellement. Selon plusieurs acteurs du secteur, les développeurs de modèles de fondation doivent investir en permanence pour améliorer leurs modèles et lancer des versions améliorées sur le marché.

166. C'est la raison pour laquelle **les investissements dans le secteur ont été multipliés par près de six entre 2022 et 2023**. Les entreprises du secteur ont ainsi levé plus de 22 milliards de dollars en 2023 (soit environ **20 milliards d'euros**), contre environ

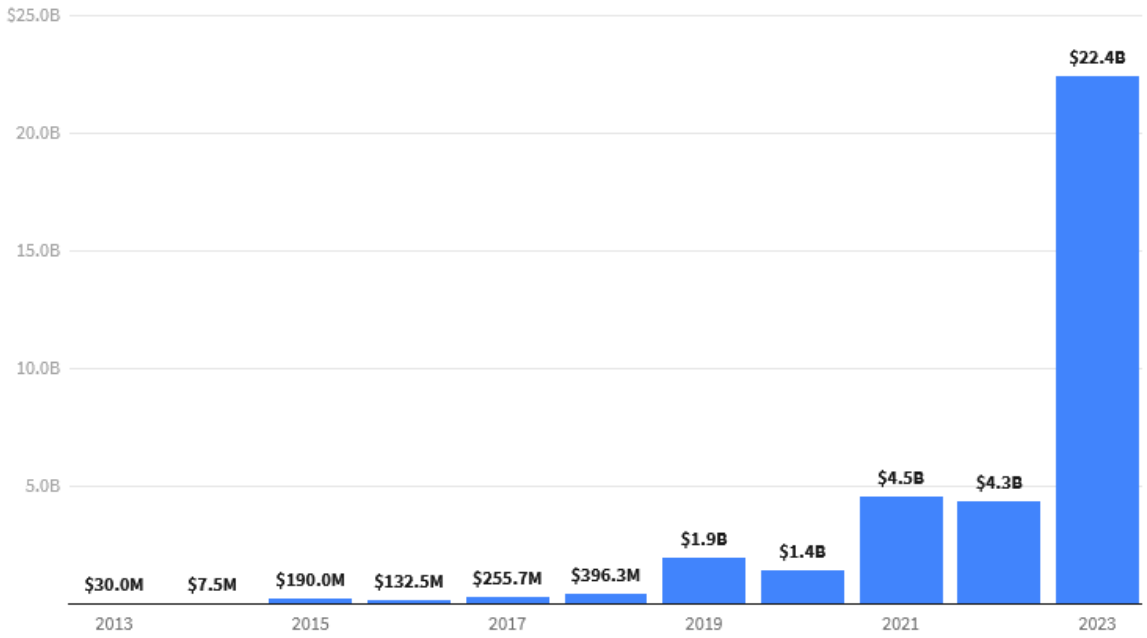
---

[101] Le Monde Informatique, Mistral lève 385 M€ et devient une licorne française, 11 décembre 2023.

[102] Stanford University, Artificial Intelligence Index Report 2024, page 64.

4 milliards de dollars en 2022 (soit environ 3,7 milliards d'euros). Plus de 70 % de ces investissements vont à des développeurs de modèles de fondation. Cependant, la Cour des Comptes européenne pointe la faiblesse des investissements privés au niveau européen (par rapport aux autres acteurs, États-Unis et Chine), et déplore le manque de gouvernance et de coordination des investissements publics dans l'IA.[103]

Figure n° 4 : les investissements dans le secteur de l'IA générative



Source : Dealroom, Generative AI, janvier 2024.

167. Ces investissements sont d'autant plus remarquables que, dans le même temps, les investissements ont globalement baissé de 38 % par rapport à 2023, une baisse constatée à tous les niveaux de financement. Outre l'IA, le secteur des semi-conducteurs et des batteries ont également vu une hausse des investissements en 2023[104].

168. En lien avec ces investissements, le secteur connaît **de nombreux accords de partenariat entre les grandes entreprises technologiques et les développeurs de modèles d'IA générative** (voir *supra*, paragraphes 71 et suivants). Comme l'indique Amazon pour annoncer sa collaboration avec la société Hugging Face « *la construction, l'entraînement et le déploiement de grands modèles de langage et de vision est un processus coûteux et chronophage qui nécessite une expertise approfondie en apprentissage machine (ML). Les modèles étant très complexes et pouvant contenir des centaines de milliards de paramètres, l'IA générative est largement hors de portée pour de nombreux développeurs* »[105] (traduction libre).

---

[103] Cour des comptes européenne, Rapport spécial 08/2024: L'UE face au défi de l'intelligence artificielle – Pas de progrès possibles sans une gouvernance renforcée et sans investissements plus importants et mieux ciblés, 29 mai 2024.

[104] Crunchbase, Global Startup Funding In 2023 Clocks In At Lowest Level In 5 Years, 4 janvier 2024.

[105] Amazon, AWS and Hugging Face collaborate to make generative AI more accessible and cost efficient, 21 février 2023.

3. LES EVOLUTIONS SUSCEPTIBLES DE LIMITER LES BARRIERES A L'ENTREE

a) Les supercalculateurs publics, une alternative pour l'entraînement des modèles

169. Traditionnellement centrés sur le calcul à haute performance (« *High Performance Calculation* », ci-après HPC), les supercalculateurs publics ont entamé une transition ces dernières années pour accueillir davantage de projets de recherche en IA. Cette transition a nécessité l'évolution du matériel, pour accueillir des processeurs graphiques en complément des CPU, historiquement utilisés pour le HPC. Par exemple, le supercalculateur Jean Zay, inauguré en 2020 en France, a connu depuis plusieurs évolutions de matériel et extensions pour ajouter des processeurs graphiques et permettre plus de projets spécialisés en IA. Une nouvelle extension prévue pour juin 2024 va ainsi lui conférer une puissance de calcul de plus de 125 péta FLOPS, (soit 125 millions de milliards d'opérations de calcul par seconde), grâce à l'ajout de 1 456 GPU Nvidia H100[106].

170. Cette puissance de calcul reste toutefois nettement inférieure à celles des supercalculateurs ayant permis l'entraînement des plus grands modèles, tels que GPT-4 ou Llama 3, qui comptent plusieurs dizaines de milliers de GPU[107]. La puissance de calcul d'un supercalculateur public doit de plus être partagée entre un grand nombre de projets d'IA. Pour les utilisateurs, elle offre néanmoins l'avantage de proposer un support technique, permettant par exemple l'optimisation des codes informatiques.

171. En contrepartie d'une contribution à la science ouverte (publication des travaux dans une revue académique par exemple), l'accès à ces ressources est gratuit, ce qui contribue à réduire les barrières à l'entrée pour les des acteurs, notamment du monde de la recherche. Par exemple, une équipe de chercheurs de l'école CentraleSupélec a utilisé les ressources du supercalculateur Jean Zay pour entraîner un modèle de langage bilingue français-anglais dénommé « **CroissantLLM** »[108]. Au niveau européen, un appel à candidatures a été lancé en mars 2024 pour permettre à des acteurs publics ou privés d'accéder à la puissance de calcul des supercalculateurs EuroHPC (MareNostrum, Leonardo et autres)[109].

172. L'usage de ressources publiques comme les supercalculateurs permet l'entraînement et le réglage fin de modèles. En revanche, les supercalculateurs ne constituent pas une solution pour l'inférence, en raison notamment des contraintes liées à l'accessibilité à la puissance de calcul, souvent accordée pour un créneau temporel limité (semaines ou mois). Ainsi, un acteur indique que « *beaucoup de startups françaises qui utilisent* [un supercalculateur] *pour l'apprentissage se tournent vers des acteurs privés (dont AWS qui en profite beaucoup) pour proposer des services d'inférence et commerciaux, faute d'offre souveraine intégrée* ».

---

[106] Communiqué de l'IDRIS, Extension de Jean Zay : 1456 GPU H100 pour franchir la barre des 125 PFlop/s, 28 mars 2024.

[107] Voir par exemple le communiqué de publication de Llama 3 : « *Nous avons effectué l'entraînement sur deux clusters de 24 000 GPUs chacun* ».

[108] Usine Digitale, CroissantLLM : Des chercheurs de CentraleSupélec lancent un modèle d'IA open source et bilingue, 4 mars 2024.

[109] EuroHPC, EuroHPC JU Access Call for AI and Data-Intensive Applications, 5 mars 2024.

b) Des innovations technologiques réduisant le besoin en puissance de calcul et en données

173. Dans un marché aussi dynamique que celui de l'IA générative, des **innovations technologiques** apparaissent continuellement afin de permettre le développement de modèles plus sobres, utilisant moins de données et donc une puissance de calcul plus limitée. Ceci permet de diminuer les barrières à l'entrée, en réduisant les coûts d'entraînement des modèles, la dépendance à des volumes de données particulièrement vastes et les coûts d'inférence lors de l'utilisation.

174. En ce qui concerne l'architecture, plusieurs modèles à l'état de l'art comme Mixtral 8x22B sont fondés sur des **combinaisons d'experts** (« MoE », voir glossaire). Cette architecture est divisée en plusieurs sous-ensembles de réseaux de neurones peu denses (en anglais « *sparse* ») appelés experts et spécialisés sur une tâche spécifique, et utilise un routeur pour déterminer quel expert doit être utilisé pour répondre à une requête.[110] Ce type d'architecture spécifique permet **d'améliorer l'efficacité de l'entraînement** : pour un budget de calcul donné, il sera plus efficace qu'une architecture classique. Il permet également de réduire significativement les coûts de l'inférence. En effet, au lieu d'utiliser tous les poids du modèle, seule une partie des experts et donc des poids (par exemple 2 experts sur 8) est utilisée pour répondre à une requête. Ainsi, la société Mistral AI met en avant l'efficacité de son modèle Mixtral 8x22B « *qui n'utilise que 39 milliards de paramètres actifs sur 141 milliards, ce qui offre une rentabilité inégalée pour sa taille* »[111].

175. De même que pour l'architecture des modèles, des innovations peuvent apparaître pour **améliorer l'efficacité des techniques de réglage fin** et réduire leurs coûts. Par exemple, plusieurs chercheurs de Microsoft ont présenté en 2021 une technique appelée LoRA (« *Low Rank Adaptation* »)[112], qui réduit les besoins en puissance de calcul pour le réglage.

176. Il est également possible de réduire les besoins en puissance de calcul en développant des modèles plus petits. Ceux-ci peuvent fournir des réponses à des requêtes plus spécialisées, qui nécessitent moins de puissance de calcul.

177. Il convient néanmoins de relativiser, à court terme, la possibilité que ces modèles plus économes puissent remplacer les grands modèles actuels. En premier lieu, si ces modèles sont moins coûteux, ils sont également à l'heure actuelle moins performants que les plus grands modèles, ce qui conduit certains acteurs du secteur à les réserver à des cas d'usages spécifiques (comme la recherche documentaire) et à s'interroger sur leur rentabilité économique. En second lieu, le rapport *AI Index* de Stanford observe que le recours à la puissance de calcul a augmenté de manière exponentielle pour les principaux modèles d'IA, surtout depuis les cinq dernières années. Il compare à cet effet le modèle « Transformer » de Google datant de 2017, qui a eu recours à 7 400 peta FLOPS, au modèle « Gemini Ultra » de Google, sorti à la fin de l'année 2023, qui a nécessité 50 milliards de peta FLOPS[113].

178. De nombreux acteurs évoquent également la possibilité d'utiliser des **données synthétiques** (voir glossaire), par exemple engendrées par un autre modèle de fondation, pour l'entraînement de modèles d'IA générative. Ainsi, de nouveaux types de réseaux de neurones,

---

[110] HuggingFace, Mixture of Experts explained, 11 décembre 2023.

[111] Mistral, Cheaper, Better, Faster, Stronger, 17 avril 2024.

[112] Hu, Shen et al., LoRA: Low-Rank Adaptation of Large Language Models, juin 2021.

[113] Université de Stanford, Artificial intelligence Index Report 2024, page 51.

comme les GAN ou les autoencodeurs variationnels (« *variational autoencoders* », VAE) permettent la génération de contenus ressemblant à la distribution du contenu fourni en entrée. Les données synthétiques peuvent être moins coûteuses à acquérir. Par exemple, l'université de Stanford a développé une version spécialisée du modèle Llama, dénommée « Alpaca-7B », en mars 2023, à l'aide de données synthétiques issues de ChatGPT, pour un coût de génération des données inférieur à 600 dollars (555 euros).[114] En outre, l'utilisation de données synthétiques permet présente de réduire les contraintes liées aux données personnelles. Néanmoins, elle s'accompagne de certains risques, comme la propagation de biais ou l'augmentation du taux d'erreurs[115].

c) Les modèles ouverts (*open source*) contribuent à réduire les barrières à l'entrée

179. Dans le secteur informatique, l'Autorité définit un logiciel *open source* comme « *un logiciel dans lequel le code source est à la disposition du grand public. Le développement de ces « logiciels libres » implique un effort de collaboration où les programmeurs améliorent ensemble le code source et partagent les changements au sein d'une communauté* »[116]. Ces conditions impliquent également la possibilité d'apporter des modifications au logiciel existant.

180. Dans le cadre de l'IA, l'*Open Source Initiative* (« OSI »), une organisation à but non lucratif ayant vocation à défendre les principes de *l'open source*, a lancé une réflexion autour d'une définition spécifique de ces principes appliqués à l'IA[117.] Cette réflexion porte sur les critères de transparence et de mise à disposition permettant à un modèle d'IA de se conformer aux critères de l'*open source*.

181. En pratique, dans le secteur de l'IA générative, l'*open source* fait référence à des situations très variables allant des modèles où seuls les poids sont rendus publics (*open-weights*) jusqu'aux modèles totalement ouverts où l'ensemble du code, de l'architecture, des données d'apprentissage, les poids et le processus d'apprentissage sont mis à disposition. L'absence de définition objective de l'*open source* dans le cadre de l'IA générative crée un risque de confusion pour les utilisateurs, voire de communication trompeuse de la part des développeurs (parfois appelé « *open-washing* »)[118].

---

[114] Stanford University, Alpaca: A Strong, Replicable Instruction-Following Model, 13 mars 2023.

[115] United Nations University, The Use of Synthetic Data to Train AI Models : Opportunities and Risks for Sustainable Development, 4 septembre 2023.

[116] Voir l'avis de l'Autorité n° 14-A-18, paragraphe 19 page 9.

[117] Open Source Initiative, The Open Source AI Definition – draft v. 0.0.8, consulté le 18 juin 2024.

[118] Liesenfeld, A., & Dingemanse, M. Rethinking open source generative AI: open-washing and the EU AI Act In the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). ACM.

Figure n° 5 : modes de publication de modèles d'IA générative

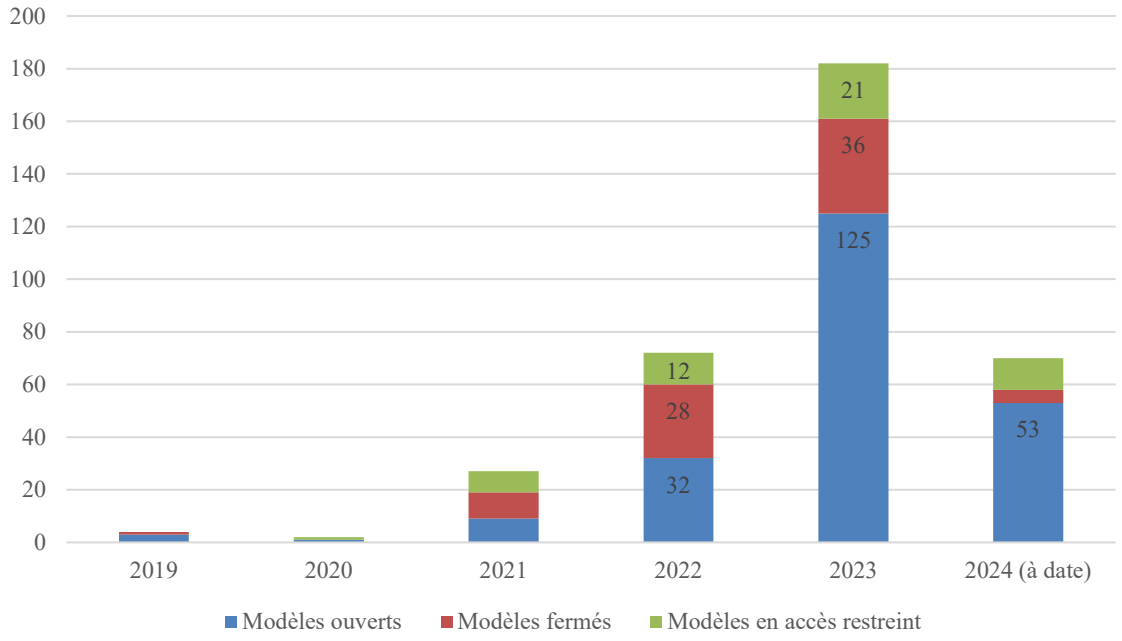
Degré d'ouverture	Modèle Fermé	Accès hébergé	Accès via API	Réglage fin via API	Poids Disponibles	Poids, données et code, avec restrictions	Poids, données et code, sans restrictions
Modèle (Développeur)	*Flamingo (Google)*	*Pi (Inflection)*	*GPT-4 (OpenAI)*	*GPT-3.5 (OpenAI)*	*Llama 3 (Meta)*	*Bloom (BigScience)*	*GPT-Neo-X (EleutherAI)*

Source : Traduction libre du graphique publié par le Stanford Institute en décembre 2023, lui-même issu du papier d'Irene Solaiman, The Gradient of Generative AI Release: Methods and Considerations, 2023

182. L'Autorité observe que parmi les développeurs de modèle, les acteurs non commerciaux publient généralement tous leurs travaux en *open source*, tandis que les acteurs commerciaux publient certains de leurs modèles en *open-weights* mais gardent leurs modèles les plus performants en modes propriétaire. À titre d'exemple, Mistral AI a publié en *open-weights* plusieurs de ses modèles, mais pas Mistral Large, son modèle le plus performant. Meta a, pour sa part, publié sa gamme de modèles Llama en *open-weights*, mais avec des conditions de licence restreintes aux utilisations commerciales dans des applications ayant plus de 700 millions d'utilisateurs.

183. Les grands acteurs verticalement intégrés comme Google ou Microsoft tendent également à mettre à disposition certains de leurs petits modèles de langage en *open-weights*, avec Gamma (Google) et Phi (Microsoft). La figure n° 6 décrit la forte augmentation du nombre de modèles d'IA générative publiés par les acteurs, ainsi que leur propension à se tourner de plus en plus largement vers la diffusion ouverte de leurs modèles, principalement via une stratégie *open-weights*.

Figure n° 6 : types de modèles de fondation d'IA générative publiés



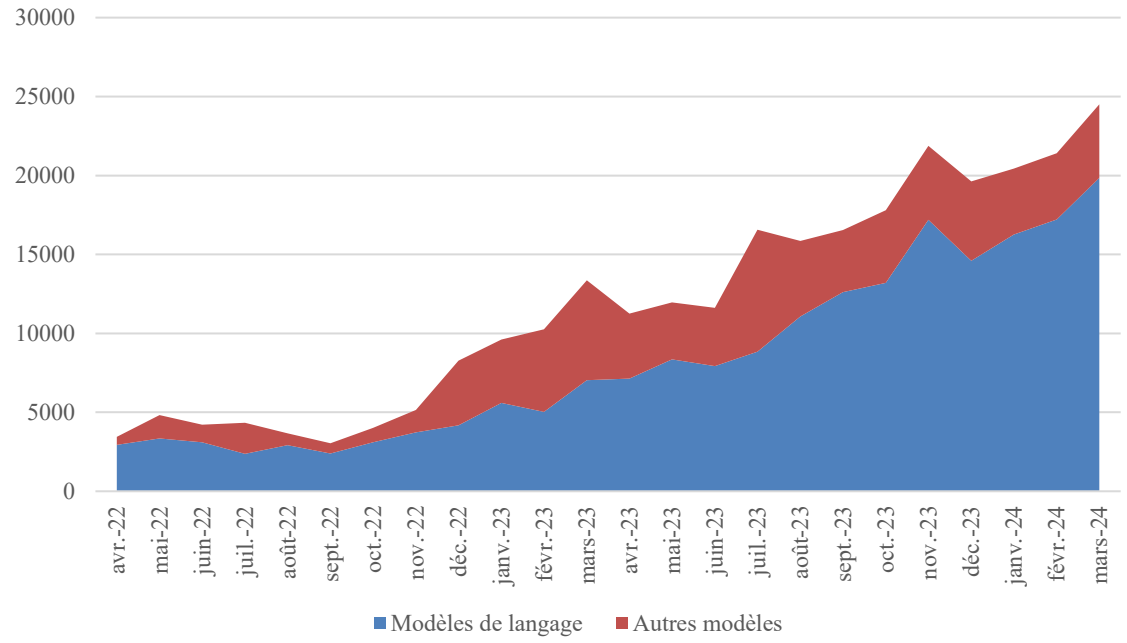
Source : Ecosystem Graphs: The Social Footprint of Foundation Models, Stanford CRFM, données consultées sur GitHub le 7 juin 2024

184. L'existence de nombreux modèles ouverts permet ainsi à un plus grand nombre d'acteurs d'entrer sur la partie aval de l'IA générative. C'est notamment le cas de nombreux acteurs (entreprises, administrations ou chercheurs par exemple) **qui n'ont pas les moyens de**

développer leurs propres modèles de fondation d'IA générative, et pour lesquels il est plus aisé de procéder à la spécialisation de modèles existants.

185. Ainsi, le réglage fin de modèles ouverts de fondation contribue au développement et à l'approfondissement de la recherche tout en réduisant, à l'aval, les barrières à l'entrée. Ainsi, alors que les modèles de fondation se comptent en centaines seulement, **le nombre de nouveaux modèles spécialisés s'élèvent à plusieurs dizaines de milliers chaque mois** sur la plateforme de collaboration HuggingFace. Au 5 juin, le nombre total de modèles (tous types confondus) a dépassé les 700 000 sur cette même plateforme. [119]

Figure n° 7 : nombre de nouveaux modèles d'IA publiés chaque mois sur HuggingFace



Données obtenues sur le site OECD.AI (2024) issue des données d'HuggingFace selon la méthodologie de comptage établie par l'OCDE

186. Toutefois, si la diffusion de modèles et de technologies en *open source* peut permettre de réduire les barrières à l'entrée, en permettant à un plus grand nombre d'acteurs d'accéder à ces technologies, elle ne supprime pas pour autant les barrières pour un acteur souhaitant développer son propre modèle de fondation ou pour obtenir la puissance de calcul suffisante au réglage fin d'un modèle d'IA générative. En effet, comme le relève un acteur, si la stratégie appliquée par la majorité des acteurs de l'IA générative, consistant à la mise à disposition des poids d'un modèle, « *permet une large réutilisation, elle ne permet en revanche pas de réduire la barrière à l'entrée de nouveaux acteurs souhaitant entraîner des modèles de fondation. En effet, la connaissance de ces poids n'apporte qu'une utilité marginale à des fins d'entraînement de modèles de fondation, dont l'objectif est précisément de créer de nouveaux modèles et leurs propres poids* ». La transparence sur d'autres éléments du modèle serait nécessaire pour sa reproduction par d'autres, comme par exemple le code et les données permettant l'entraînement, les données utilisées, etc.

[119] Site d'HuggingFace consulté le 5 juin 2024.

187. En outre, la publication de modèles d'IA générative très performants peut poser des problèmes de sécurité. En effet, à l'inverse des modèles accessibles via des applications ou des interfaces de programmation, qui sont généralement accompagnés de filtres de sécurité, les modèles publiés de manière ouverte peuvent être réutilisés par des acteurs malintentionnés pour produire du contenu problématique (tels que la pédopornographie ou la désinformation par exemple).

B. CERTAINES ENTREPRISES PEUVENT BENEFICIER D'AVANTAGES LIES A LEURS ACTIVITES SUR D'AUTRES MARCHES NUMERIQUES

188. La position de certains acteurs sur d'autres marchés en lien avec l'IA générative pourrait engendrer différents avantages concurrentiels. L'ensemble de ces avantages ne sont pas aisément réplicables par les développeurs de modèles de fondation.

1. UN ACCES PRIVILEGIE AUX INTRANTS NECESSAIRES POUR L'ENTRAINEMENT ET LE DEVELOPPEMENT DES MODELES DE FONDATION

a) Un accès facilité à la puissance de calcul

189. Comme examiné précédemment, la création d'une infrastructure de calcul suffisante est particulièrement coûteuse et difficile pour un nouvel entrant. Or, ainsi que l'a montré l'avis n° 23-A-08 de l'Autorité[120], **AWS, GCP et Microsoft Azure sont les trois principaux fournisseurs de services *cloud* en France et parmi les premiers au monde**. Ils disposent donc déjà des capacités financières, techniques et du savoir-faire requis pour créer et gérer une telle infrastructure.

190. Par ailleurs, **leur capacité à acheter en grande quantité et à négocier des accords préférentiels avec des fournisseurs comme Nvidia** peut leur permettre d'accéder à des ressources même en période de forte demande et de pénurie. Meta développe ainsi son infrastructure d'IA pour intégrer 350 000 GPU Nvidia H100 d'ici la fin de l'année 2024[121], soit un achat estimé à environ 9 milliards de dollars (soit environ 8,3 milliards d'euros)[122]. **Leur accès à une large base d'utilisateurs** permet également à ces entreprises de mieux optimiser leur infrastructure de calcul.

191. Cependant, même pour ces entreprises, la pénurie de GPU peut poser des défis. Dans un marché où la demande dépasse l'offre, les grandes entreprises peuvent aussi rencontrer des difficultés à obtenir suffisamment de GPU pour répondre à leurs besoins et à ceux de leurs clients, ce qui peut entrainer des retards dans le déploiement de nouveaux produits ou services.

192. C'est la raison pour laquelle **plusieurs géants du numérique se sont lancés dans le développement de leurs propres accélérateurs d'IA spécifiquement adaptés à leurs écosystèmes**, comme les TPU de Google, AWS Trainium ou Microsoft Maia. La fourniture

---

[120] Les services *cloud* des *hyperscalers* ont fait l'objet d'une présentation détaillée dans l'avis n° 23-A-08 de l'Autorité précité.

[121] Meta, Building Meta's GenAI Infrastructure, 12 mars 2024.

[122] Le Monde Informatique, Pour l'IA, Meta va acquérir 350 000 accélérateurs Nvidia H100, 19 janvier 2024.

de ces puces en interne présente l'avantage de ne pas être dépendante des cycles de développement de produits des fournisseurs de GPU externes et de réagir plus rapidement aux évolutions du marché. Les *hyperscalers* peuvent également concevoir des puces répondant spécifiquement à leurs besoins et à ceux de leurs clients, ce qui peut inclure des optimisations pour certaines tâches d'IA ou des fonctionnalités spécifiques n'existant pas dans les GPU externes. Un acteur du secteur confirme ainsi que « *l'optimisation de ses propres puces, pour ses propres centres de données, pour ses propres modèles, peut constituer un avantage concurrentiel significatif* ». Par ailleurs, bien que le développement d'accélérateurs d'IA en interne nécessite un investissement initial important compte tenu de la nécessité d'acquérir ces capacités de production en externe ou de l'acquisition de compétences, il peut être plus rentable à long terme de développer ses propres puces plutôt que de s'approvisionner auprès d'un seul acteur, surtout si les fournisseurs de services *cloud* peuvent produire ces puces à grande échelle. Ces atouts peuvent ainsi progressivement positionner favorablement ces puces par rapport à la concurrence[123] même si, à ce jour, ces grandes entreprises ne semblent pas vouloir les commercialiser à des tiers en dehors de leur *cloud*.

193. Au-delà du développement de puces, plusieurs grands acteurs travaillent à l'émergence d'alternatives au logiciel Cuda de Nvidia. Ainsi, les TPU de Google sont conçus pour être utilisées avec le logiciel TensorFlow de Google. OpenAI a développé Triton, dont « *l'objectif est de fournir un environnement open source permettant d'écrire du code rapide avec une productivité supérieure à celle de CUDA* »[124]. La Fondation UXL, regroupant des fournisseurs tels que Google et Intel, prévoit également de créer une suite de logiciels et d'outils en accès libre capables d'alimenter plusieurs types d'accélérateurs.[125]

194. Ces entreprises se trouvent donc dans une position ambigüe dans la mesure où elles sont à la fois **partenaires et concurrentes** vis-à-vis des fournisseurs de micro processeurs pour l'IA générative. Comme l'indique France Digitale, « [p]*armi les plus grands clients de Nvidia, on trouve les entreprises de cloud intégrées verticalement : AWS, Alibaba, Google et Microsoft. Ces acteurs occupent une position singulière, combinant les rôles de partenaires et de concurrents de Nvidia. Ces entreprises investissent dans la conception de leurs propres puces spécifiques pour réduire leur dépendance à Nvidia, et par ailleurs, nouent des partenariats stratégiques et annoncent des investissements importants dans l'achat de GPUs Nvidia* »[126].

b) Un accès privilégié aux données

195. Plusieurs acteurs incontournables du numérique bénéficient d'avantages importants concernant la collecte des données nécessaires au stade de l'entraînement et du réglage fin des modèles.

---

[123] Dans la présentation de ses résultats trimestriels, Amazon confirme : « *Nous avons le plus grand choix d'instances de calcul NVIDIA, mais la demande pour notre silicium personnalisé, Trainium et Inferentia, est assez élevée en raison de ses avantages en termes de prix et de performances par rapport aux autres solutions disponibles* » (page 5) (traduction libre).

[124] Traduction libre du projet Github de Triton.

[125] Le Monde informatique, La Fondation UXL peaufine une alternative au Cuda de Nvidia, 26 mars 2024.

[126] France Digitale, « Des puces aux applications, l'Europe peut-elle être une puissance de l'IA générative ? », avril 2024, page 13.

196. Tout d'abord, ils bénéficient du volume de données nécessaire au stade de l'entraînement des modèles de fondation.

197. À titre d'exemple :

- Alphabet, en plus d'être l'un des seuls acteurs entièrement intégrés verticalement dans ce secteur, en ayant accès à la fois à une infrastructure développée en interne (TPU) et à un LLM développé par ses soins (Gemini), a accès à un patrimoine considérable de données grâce notamment aux 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. Par exemple, YouTube hébergerait 10 milliards de vidéos, offrant à Alphabet une source majeure de données d'entraînement pour les modèles d'IA (vidéos ou de langage avec la retranscription textuelle des vidéos) ;
- Meta bénéficie d'un large ensemble de données grâce à ses plateformes Facebook, Instagram et WhatsApp. Mark Zuckerberg a récemment déclaré que Facebook et Instagram ont « *des centaines de milliards d'images partagées publiquement et des dizaines de milliards de vidéos publiques, ce qui, selon nos estimations, est supérieur à l'ensemble de données Common Crawl* (…) »[127] (traduction libre) ;
- 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.

198. **Leur position leur confère également un accès privilégié à une multitude de métadonnées et de données associées à l'utilisation de leurs services**. Cette situation favorise leur accès à des données indirectes auxquelles des acteurs de taille plus modeste ne peuvent avoir accès. En effet, l'utilisation du modèle en phase d'inférence permet à l'entreprise qui collecte les données, notamment sur la satisfaction de l'utilisateur, d'optimiser ce modèle, et les futurs modèles. Plus un modèle est utilisé, plus le développeur disposera de retours permettant d'en améliorer la performance. Compte tenu de cette boucle de rétroaction positive, une position concurrentielle forte est susceptible de se consolider rapidement, voire de muer en position dominante.

199. En plus du grand volume de données auquel ces acteurs ont accès, ceux-ci concluent également de nombreux accords avec des propriétaires de données tiers (voir *supra* paragraphe 158). Alphabet s'est ainsi engagée à payer 60 millions de dollars (soit environ 55 millions d'euros) par an pour accéder aux données de Reddit, un site communautaire américain de discussions et d'actualités sociales[128].

200. **L'accès à ces données ne s'effectue pas dans les mêmes conditions pour les développeurs de modèles et les grandes entreprises du numérique.** Tout d'abord, parce qu'ils n'ont pas la possibilité de conclure des accords avec des fournisseurs de contenus aux mêmes conditions financières que les grands acteurs du numérique. Ensuite, parce qu'ils ne

---

[127] Meta Platforms, Inc. (META), Fourth Quarter 2023 Results Conference Call, February 1st, 2024 (page 3).

[128] Le Figaro, IA : Reddit noue un accord de licence inédit avec Google pour 60 millions de dollars, 22 février 2024. Dans sa déclaration déposée devant la Security and Exchange Commission aux États-Unis, Reddit a indiqué : « *Nous sommes également aux premiers stades de la monétisation de notre nouvelle opportunité en matière de licences de données en permettant à des tiers d'accéder, de rechercher et d'analyser les données sur notre plateforme. En janvier 2024, nous avons conclu certains accords de licence de données d'une valeur contractuelle globale de 203 millions de dollars et d'une durée allant de deux à trois ans* (…) *Les données Reddit augmentent et se régénèrent constamment à mesure que les utilisateurs viennent interagir avec leurs communautés et entre eux. Nous pensons que les données croissantes de notre plateforme seront un élément clé dans la formation des principaux grands modèles linguistiques (« LLM ») et serviront de canal de monétisation supplémentaire pour Reddit* » (traduction libre).

peuvent pas aisément accéder aux données des grands acteurs compte tenu des conditions posées par eux.

201. Premièrement, si les grands acteurs du numérique ont accès à un grand volume de données au sein de leurs écosystèmes, cela ne signifie pas que ces données sont nécessairement libres de droits. Selon un acteur, « *il convient de bien différencier la notion de donnée qualifiée de « propriétaire » ou « first party » par les conglomérats technologiques – correspondant en substance à toute donnée passant par leur intermédiaire qui resterait au sein de leur écosystème avec des opérateurs avec lesquels ils ont des accords, les données « tierces parties » qui correspondent selon les mêmes conglomérats technologiques aux données partagées avec des opérateurs autres que ceux opérant les écosystèmes fermés, d'avec la donnée ou le contenu faisant l'objet d'une protection au titre du droit d'auteur ou du droit voisin, ou du droit des bases de données issu de la Directive CE n° 96/9, ou même de la donnée personnelle ou non personnelle au sens du RGPD et de la Directive e-Privacy, ou de la donnée commercialement sensible pouvant être protégée au titre du secret des affaires* ».

202. C'est la raison pour laquelle des groupes emblématiques de producteurs de contenus ont entamé des discussions avec les acteurs majeurs de l'IA générative : soit en trouvant des accords comme Axel Springer et Le Monde l'ont fait récemment (voir *supra*), soit en lançant des procédures judiciaires, comme le New York Times l'a fait vis-à-vis d'OpenAI. Toutefois, les sommes en jeu pour accéder à ces contenus limitent pour l'instant la capacité d'acteurs de taille plus modeste à conclure des accords similaires, comme le montre par exemple l'accord conclu entre OpenAI et NewsCorp pour un montant de près de 250 millions de dollars (soit plus de 230 millions d'euros) sur 5 ans. Ainsi, en l'état des informations dont dispose l'Autorité, seule OpenAI a été en mesure de conclure de tels accords avec des éditeurs de presse à ce jour.

203. Par ailleurs, les données issues des services des grandes entreprises du numérique ne sont pas aisément accessibles pour les développeurs, sauf à enfreindre les règles d'utilisation applicables à ces services. À la suite de rumeurs concernant l'entraînement de Sora (le modèle d'OpenAI capable de produire des vidéos à la suite de requêtes textuelles) sur les vidéos issues de YouTube, le PDG de YouTube a ainsi indiqué que si tel était le cas, « *cela constituerait une violation claire des politiques applicables à la plateforme* »[129]. Microsoft aurait également menacé de refuser l'accès aux données issues de son moteur de recherche (Bing) à ses concurrents sous licence si ceux-ci se fondaient sur ces données pour entraîner leurs outils d'IA générative[130].

204. La plupart des géants du numérique considèrent toutefois que le caractère privilégié de leur accès aux données peut être nuancé dans la mesure où les développeurs auraient accès à d'importantes quantités de données sur Internet et dans des jeux de données publiquement accessibles.

205. Or, s'il existe en effet plusieurs jeux de données publiques disponibles (voir *supra* paragraphes 150 et suivants) les grands conglomérats technologiques gardent toutefois un avantage indéniable en matière d'accès aux données. En effet, les modèles de fondation de ces grandes entreprises ont non seulement été entrainés sur ces jeux de données publiques mais également sur des données propriétaires qui ne sont pas accessibles dans les mêmes conditions aux tiers. Cet accès aux données propriétaires est d'autant plus important que

[129] The Verge, OpenAI training Sora on YouTube videos would violate the platform's rules, 4 avril 2024.

[130] Reuters, Microsoft threatens to restrict data from rival AI search tools- Bloomberg News, 27 mars 2023.

l'accès aux données publiques devient progressivement plus difficile. La conclusion d'accords entre développeurs de modèles et fournisseurs de contenus montre également que les données publiques ne sont pas nécessairement suffisantes. Par ailleurs, même lorsque l'accès aux données est comparable, ces entreprises contrôlent également la puissance de calcul nécessaire, bénéficient d'un meilleur accès à l'expertise spécialisée et ont acquis une vaste expérience dans la collecte, l'étiquetage (voir glossaire) et l'analyse des données.

206. Il est vrai que certaines entreprises de taille plus modeste peuvent avoir un avantage lorsqu'il s'agit de spécialiser un modèle déjà entrainé pour un secteur spécifique et qu'elles ont accès aux données spécifiques de ce secteur. Cela peut être le cas par exemple d'une société pharmaceutique qui pourrait avoir accès à des données d'essais cliniques propriétaires qu'elle pourrait utiliser pour former ou affiner un modèle de fondation utilisé dans le secteur médical. Cependant, les grandes entreprises du numérique bénéficient aussi d'un accès privilégié à des ensembles de données personnalisées et spécialisées dans de nombreux domaines, tels que la santé, la finance et les transports. C'est le cas par exemple de Google pour les données de cartographie et de santé (grâce à Fitbit) et pour Amazon, qui a accès à des données sensibles de santé grâce à sa récente acquisition de One Medical[131]. Dans le secteur de la finance, l'avis « Fintech » de l'Autorité[132] avait aussi révélé qu'Amazon, Apple, Google ou Meta étaient susceptibles d'accéder à des données financières dans le cadre du développement de solutions de paiement.

207. Au-delà de l'accès par des entreprises à des données spécifiques à un secteur, il convient également de rappeler que différentes activités sont plus ou moins productrices de données. Certaines entreprises peuvent ainsi être conduites à traiter de très grandes quantités de données compte tenu de la nature de leurs activités (par exemple, les banques) ou parce qu'elles ont pour objet même l'intermédiation de données.

c) La capacité d'attirer les meilleurs talents

208. L'Autorité a rappelé *supra* la nécessité de disposer de compétences techniques pointues afin de pouvoir développer des modèles de fondation performants. Or, les grandes entreprises du numérique disposent de nombreux atouts pour attirer les meilleurs talents.

209. Leurs capacités financières leur permettent d'offrir à ces profils des salaires attractifs, des actions et stock-options et des avantages sociaux importants. Par ailleurs, ces entreprises sont en mesure d'offrir des perspectives de travail intéressantes à ces talents, compte tenu de leur réputation en matière d'innovation, leur positionnement global et la profondeur de leur catalogue de services. Par exemple, comme relevé *supra*, la plupart de ces grandes entreprises disposent en interne de laboratoires de recherche très performants leur permettant d'offrir des conditions de travail très confortables. Le nouveau centre de recherche de Google pour l'IA à Paris est ainsi doté d'un budget de 300 millions d'euros et devrait rassembler plus de 300 chercheurs disposant d'un accès à des outils d'IA avancés[133]. Enfin,

---

[131] Open markets institute, AI in the Public Interest: Confronting the Monopoly Threat, novembre 2023.

[132] Avis n° 21-A-05 du 29 avril 2021 portant sur le secteur des nouvelles technologies appliquées aux activités de paiement, paragraphes 356 à 358.

[133] La Tribune, IA : avec son nouveau centre de recherche à Paris, Google entend former 100.000 professionnels, 15 février 2024.

si la réputation de ces entreprises a pu souffrir de scandales tels que Cambridge Analytica[134], les géants du numérique comme Google, Microsoft et Apple figurent au classement des entreprises préférées des jeunes cadres diplômés en 2023, malgré une baisse de leur position depuis 2022.[135]

210. Au-delà de la capacité d'attirer les meilleurs talents, ces grandes entreprises peuvent également nouer des partenariats avec de jeunes pousses très innovantes, ce qui leur permet d'avoir également accès à la meilleure expertise possible (voir *supra*).

211. Ces éléments sont confirmés par plusieurs acteurs du secteur. Selon l'un d'eux, « [l]*es grands acteurs du numérique, principalement américains, bénéficient d'atouts structurants pour attirer les meilleurs talents : la notoriété d'entreprises innovantes, des projets stimulants, des moyens et des outils quasiment illimités, et des niveaux de rémunération très importants. Elles proposent donc un cadre de travail général particulièrement attractif pour les talents* ».

2. LES AVANTAGES LIES A L'INTEGRATION VERTICALE ET CONGLOMERALE DES GRANDES ENTREPRISES TECHNOLOGIQUES

a) Économies d'échelle, de gamme et effets de réseaux

212. Les grandes entreprises technologiques intégrées verticalement ou de manière conglomérale ont accès aux financements, aux talents, aux données et à la puissance de calcul en raison de leurs activités dans des marchés distincts mais liés au secteur de l'IA générative.

213. Le secteur est caractérisé par d'importants coûts fixes occasionnés par l'entraînement initial d'un modèle de fondation (compte tenu notamment de l'acquisition des ressources de calcul ou des éventuels contrats d'acquisitions de données), ce qui donne lieu à **des économies d'échelle,** les acteurs cherchant à amortir ces coûts sur le plus grand nombre d'utilisateurs possible. De ce fait, un acteur déjà établi, disposant d'importantes capacités de production et d'une base d'utilisateurs, sera avantagé par rapport à des acteurs plus modestes. Par ailleurs, au niveau même des centres de données, l'augmentation de l'activité peut se traduire par un agrandissement de la taille du centre, ce qui conduit à un accroissement des coûts fixes mais permet également de diminuer les coûts unitaires grâce aux gains en termes d'énergie (baisse du coût de refroidissement notamment), de main-d'œuvre ou de sécurité par exemple.

214. Les potentielles économies d'échelle (du fait de coûts fixes importants) semblent moindres à l'aval dès lors que les nouveaux entrants ont la possibilité d'utiliser des modèles ouverts et que la puissance de calcul nécessaire pour l'inférence est dépendante du nombre d'utilisateurs.

215. Les produits d'IA générative peuvent également se caractériser par **des économies de gamme**. Les économies de gamme se matérialisent lorsqu'une entreprise peut accroître sa production en produisant des biens distincts à partir des mêmes facteurs. Or, une fois développé, un modèle de fondation peut servir à une grande variété d'applications, les coûts de réglage fin étant modestes par rapport au développement initial du modèle. Ainsi, le

---

[134] CNBC, Facebook has struggled to hire talent since the Cambridge Analytica scandal, according to recruiters who worked there, 16 mai 2019.

[135] Les Echos Start, Classement des boîtes préférées des jeunes cadres : l'industrie progresse, les GAFAM en baisse, 23 juin 2023.

modèle généraliste Gemini de Google a permis l'entraînement de plusieurs modèles spécialisés, comme MedGemini dans la santé.

216. Par ailleurs, les grandes entreprises du numérique sont particulièrement bien placées pour exploiter leur accès aux infrastructures de calcul et aux données ainsi que leurs connaissances techniques pour le lancement et le développement de modèles, ce que confirme l'Organisation de coopération et de développement économiques (OCDE) dans une note récente : « *les données pourraient constituer une économie de gamme si le fait d'opérer sur des marchés adjacents permet aux entreprises de capturer des données qui améliorent leur capacité à développer de meilleurs modèles génératifs d'IA. Il peut également y avoir des synergies si le personnel travaillant dans des domaines connexes peut travailler au développement de l'IA* »[136] (traduction libre).

217. Ces éléments conduisent plusieurs acteurs du secteur à considérer qu'il est nécessaire pour les développeurs de modèles d'atteindre une taille critique sur la distribution des modèles, à l'aval du marché, pour amortir les coûts d'investissements initiaux très élevés.

218. Le secteur de l'IA générative est également caractérisé par **la présence d'effets de réseaux**. Un service d'IA générative peut améliorer ses performances au fur et à mesure de son utilisation puisque les données de retours des utilisateurs permettent d'affiner les modèles. Ainsi, plusieurs acteurs indiquent que « *l'accès à une large base d'utilisateurs peut créer un cercle vertueux : plus la base de clients est importante, plus il est possible d'améliorer les modèles et donc d'attirer de nouveaux utilisateurs. Les entreprises ou les utilisateurs finaux pourraient être incités à choisir le modèle de fondation qui a déjà une présence significative sur le marché, en partant du principe qu'il est le plus efficace* ». Les géants du numérique disposant de grandes bases d'utilisateurs sont plus à même de profiter de ces effets de réseaux, ce qui pourrait dans le futur constituer une barrière à l'entrée.

219. Les activités existantes de ces entreprises peuvent enfin contribuer à **financer de nouvelles initiatives en matière d'IA**, dès lors qu'elles sont aussi présentes dans des secteurs d'activité à forte marge, comme le *cloud* par exemple. Dans le même temps, les développeurs de modèles doivent continuer de lever des capitaux et produire des revenus pour continuer à financer leurs activités de recherche et développement et leurs déploiements de modèles.

220. À l'aval, en étant directement impliquées dans toutes les étapes de la chaîne de valeur, ces entreprises peuvent également **mieux comprendre les besoins du marché** et ajuster plus rapidement leurs produits pour répondre aux besoins du client. C'est ce que confirme un acteur du secteur : « *En développant des outils d'IA générative intégrés aux services cloud à destination des entreprises, les hyperscalers ont la capacité de fournir à leurs services d'IA générative des informations hautement qualifiées, pouvant être utilisées, par exemple, pour le réglage fin des modèles, basé sur une compréhension très poussée du comportement de leurs clients en termes d'usages numériques, de leurs besoins en termes de services informatiques et de gestion des données, et ce, souvent à l'échelle mondiale pour des clients multinationaux. Ces informations hautement qualifiées apparaissent clairement comme un facteur différenciant difficile à reproduire et parfaitement adaptée à l'activité d'analyse des besoins pouvant être couverts par l'IA, érigeant au fil du temps une barrière à l'entrée significative* ». Il est possible, par exemple, d'imaginer une situation hypothétique où un grand acteur du numérique, tirant parti de la puissance d'un modèle d'IA et des données ainsi que des métadonnées recueillies par l'exploitation d'un réseau social, pourrait concevoir des applications spécialisées utilisant l'IA dans le secteur des ressources humaines.

---

[136] OCDE, Artificial intelligence, data and competition - Background Note, 6 mai 2024, page 27.

221. Dans un futur plus ou moins proche, les entreprises du secteur pourraient être incitées à conclure des accords visant à faciliter leur accès à l'énergie, ce qui pourrait diminuer leurs coûts, notamment pour les grands acteurs. Le partenariat de Microsoft avec l'entreprise Hélion, spécialisée dans la fusion nucléaire, peut être un signe avant-coureur de cette tendance[137].

b) La mise en place progressive d'écosystèmes

222. L'Autorité constate également que ces entreprises commencent à intégrer les outils d'IA générative **dans leurs écosystèmes de produits et de services**.

223. Dans son avis n° 23-A-08, l'Autorité avait constaté que certains fournisseurs de services constituaient des écosystèmes *cloud* : « *L'analyse du fonctionnement du secteur ainsi que le positionnement des différents acteurs tend à montrer que certains fournisseurs constituent des écosystèmes cloud, c'est-à-dire un ensemble de services intégrés auxquels peuvent accéder les clients, comprenant les services propriétaires du fournisseur, mais également, au travers généralement de places de marché, un ensemble de services de développeurs tiers, conçus pour fonctionner dans cet écosystème. Le secteur pourrait ainsi se structurer autour d'une concurrence entre écosystèmes cloud* » (paragraphe 252).

224. La même logique semble être à l'œuvre dans le secteur de l'IA générative.

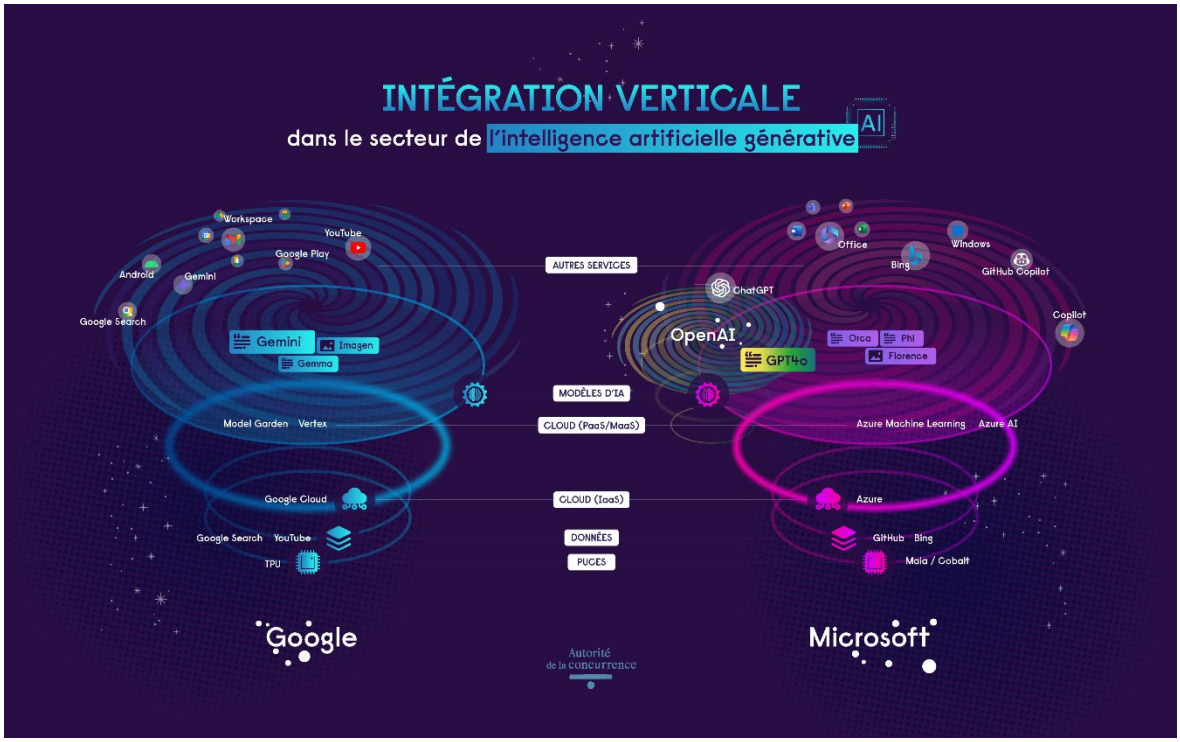
225. **Les services d'IA générative sont de plus en plus intégrés aux services situés sur des marchés distincts mais connexes sur lesquels les grandes entreprises du numérique ont un fort pouvoir de marché**. Ainsi, Microsoft déploie ses propres modèles et ceux de son partenaire OpenAI dans la fonction « Copilot » afin d'améliorer la fonctionnalité de recherche de Microsoft Bing et propose avec « Copilot for Microsoft 365 », un assistant IA conçu pour fonctionner avec l'offre Microsoft 365 (incluant notamment Word, Excel, Outlook, Teams et PowerPoint). De même, Google utilise Gemini afin d'améliorer son moteur de recherche (AI Overview) et commence à proposer le service « Gemini for workspace » afin de faciliter la rédaction dans Gmail et Docs ainsi qu'un générateur d'images créatives dans Slides.

226. Par ailleurs, **les places de marché MaaS de ces grandes entreprises permettent d'accéder à des modèles d'IA générative, propriétaires ou de tiers conçus pour fonctionner dans leur écosystème.** À titre d'exemple, Model Garden de Google propose plus de 130 modèles de fondation comprenant des modèles propriétaires comme Gemini, des modèles ouverts comme Llama 2 de Meta et des modèles tiers comme Claude 3 d'Anthropic.

227. Il apparaît que Microsoft et Google en particulier bénéficient d'avantages significatifs dans le secteur de l'IA générative. Ils contrôlent en effet la majorité des accès aux intrants nécessaires pour développer des modèles de fondation et sont de surcroît intégrés verticalement sur la chaîne de valeur, ce qui leur permet de contrôler le développement de leurs modèles de fondation propres et de ceux des tiers. Leurs activités sur d'autres marchés numériques leur permettent notamment de bénéficier des investissements nécessaires et de mettre progressivement en place des écosystèmes. La figure 8 présente l'écosystème de Google et Microsoft tout au long de la chaîne de valeur de l'IA générative.

[137] Hélion Energy, Announcing Helion's fusion power purchase agreement with Microsoft.

Figure n° 8 : intégration verticale de Google et Microsoft dans la chaîne de valeur de l'IA générative



Source : Autorité de la concurrence.
OpenAI est inclus dans la chaîne de valeur afférente à Microsoft car la plupart des services de Microsoft en aval (Copilot, Bing, etc.) se fondent sur les modèles d'OpenAI GPT-4 ou GPT-4o.

228. Les grands acteurs du numérique ont donc un accès privilégié à l'ensemble des intrants nécessaires au développement de modèles de fondation. Au-delà de leur puissance financière leur permettant de conclure notamment de nombreux partenariats, ils ont également un accès facilité à la puissance de calcul nécessaire, un accès privilégié aux données et la capacité d'attirer les meilleurs talents. Tous ces facteurs réunis constituent un obstacle considérable à l'entrée et à l'expansion de concurrents sur le marché, et les avantages dont bénéficient les grandes entreprises, peuvent entraîner des risques pour la concurrence.

C. LES RISQUES CONCURRENTIELS A L'AMONT DE LA CHAINE DE VALEUR

229. L'instruction a révélé différentes pratiques mises en œuvre ou susceptibles d'être mises en œuvre dans ce secteur qui pourraient restreindre la concurrence.

1. LES PRATIQUES SUSCEPTIBLES D'ETRE SANCTIONNEES EN DROIT DES PRATIQUES ANTICONCURRENTIELLES

a) Observations préliminaires

230. Préalablement à la caractérisation d'éventuels abus de position dominante, la question du pouvoir de marché de certaines entreprises appelle plusieurs observations.

231. À ce stade précoce du développement et de la mise en œuvre de l'IA générative, nous ne disposons pas aujourd'hui d'informations suffisantes pour déterminer avec précisions, dans le cadre de cet avis, les marchés pertinents et évaluer les parts de marché des différents acteurs impliqués. En effet, les caractéristiques des modèles varient fortement et les voies d'accès au marché à l'aval ainsi que la commercialisation de ces modèles sont en cours de développement. Certaines contributions identifient néanmoins « *deux marchés principaux, le marché primaire pour la conception et la pré-formation de modèles fondamentaux primaires et le marché secondaire pour le développement de modèles soumis à une spécialisation ou à un réglage fin pour répondre à des objectifs prédéterminés* ». D'autres acteurs s'interrogent sur l'impact de l'intégration de l'IA générative dans la définition des marchés pertinents : « *En ce qui concerne l'analyse des marchés pertinents, l'intégration de l'IA générative dans la commercialisation d'un produit ou d'un service doit-elle être évaluée comme une innovation qui ajoute une nouvelle fonctionnalité aux offres existantes, ou les positionne-t-elle sur un marché distinct par rapport aux mêmes produits ou services non dotés de l'IA générative ? De même, un système d'IA générative intégré au cœur des fonctionnalités d'un produit ou d'un service d'une entreprise doit-il être considéré comme un composant supplémentaire appartenant à un marché distinct de celui du produit et du service auquel il est rattaché ?* ».

232. Il convient de relever que, en dehors des parts de marché, d'autres critères sont pris en compte par les autorités de concurrence pour apprécier le pouvoir de marché dont pourrait disposer une entreprise dans des secteurs a priori dynamiques, comme par exemple, l'existence de barrières à l'entrée ou à l'expansion.

233. S'il semble prématuré de tirer des conclusions définitives sur la définition des marchés pertinents et le pouvoir de marché de certains acteurs, il convient néanmoins de rester vigilant car l'accès de certaines entreprises à des intrants clés ainsi que les avantages liés à leur intégration verticale et conglomérale créent les conditions d'une forte concentration à leur profit et renforcent leur puissance sur des marchés distincts mais liés, ou connexes, tels que les logiciels de productivité, les moteurs de recherche ou la publicité en ligne. **Dans certains cas, l'analyse concurrentielle pourra donc utilement reposer sur la constitution ou le renforcement d'écosystèmes plutôt que sur une analyse marché par marché.**

234. Par ailleurs, **il ne semble pas que le secteur de l'IA générative remette en question la pertinence des outils et concepts classiques du droit de la concurrence.**

235. En ce sens, le recours aux outils traditionnels du droit de la concurrence **tels que le droit des ententes** (voir *infra*) **et surtout l'abus de position dominante** conserve toute son efficacité et peut se justifier. En effet, il ressort d'une pratique décisionnelle et d'une jurisprudence constantes qu'une entreprise en position dominante sur un marché donné peut se voir reprocher un abus dont les effets affectent d'autres marchés dès lors que le marché sur lequel l'entreprise détient une position dominante et ceux sur lesquels l'abus déploie ses effets revêtent un caractère de connexité suffisant et qu'il existe des circonstances particulières justifiant l'application des règles prohibant les abus de position dominante[138].

236. Par ailleurs, au sein de ces écosystèmes, une pratique abusive peut avoir pour objectif d'évincer certains concurrents de l'opérateur dominant mais elle peut également exploiter la

[138] Arrêts de la CJCE, 14 novembre 1996, Tetra Pak International SA/Commission, aff. C-333/94 P, Rec. 1996 I-05951, point 27 ; CJUE, 17 février 2011, Konkurrensverket/TeliaSonera Sverige AB, précité, point 86 ; décision n° 22-D-20 du 15 novembre 2022 relative à des pratiques mises en œuvre dans le secteur des solutions de gestion de la paie des intermittents du spectacle, paragraphes 89 et suivants.

fragilité d'autres opérateurs de l'écosystème en leur imposant des prix plus élevés, des restrictions contractuelles (comme des contraintes de mono-hébergement ou des clauses d'exclusivité) ou en contraignant directement ou indirectement leur utilisation des services (comme des données auxquelles l'accès est restreint de manière asymétrique ou une extraction excessive des données).

237. Le droit des ententes a aussi toute sa part, comme il sera vu *infra*.

238. Enfin, d'autres outils juridiques pourraient être mobilisés. C'est le cas par exemple de **l'abus de dépendance économique**, qui ne s'apprécie pas par rapport à la position d'une entreprise sur un marché donné mais au regard des spécificités de la relation commerciale qu'elle entretient avec ses partenaires, situés en amont ou en aval et qui permet de s'affranchir de la condition de position dominante. Ainsi que l'a déjà souligné l'Autorité dans son avis n° 23-A-08 précité ainsi que dans son étude de juin 2020 consacrée au commerce en ligne, l'abus de dépendance économique peut permettre d'appréhender les pratiques contractuelles abusives des opérateurs du numérique et des places de marché[139].

239. Le **droit des pratiques restrictives de concurrence** dont la mise en œuvre relève, principalement, de la compétence de la Direction générale de la concurrence, de la consommation et de la répression des fraudes (ci-après « DGCCRF ») et des juridictions commerciales, fournit aussi une arme efficace pour sanctionner les pratiques déloyales. Ces outils ont fait l'objet d'une présentation détaillée dans l'avis n° 23-A-08 précité de l'Autorité (paragraphes 562 et suivants), auquel le présent avis renvoie.

240. L'Autorité s'attachera ainsi à présenter les risques concurrentiels présents à chaque strate de l'amont de la chaîne de valeur. Elle évoquera également, de manière plus globale, les risques résultant de la présence d'acteurs importants sur plusieurs marchés liés ou connexes.

b) Des risques d'abus à l'amont de la chaîne de valeur

Plusieurs risques d'abus au niveau des composants informatiques

241. Il ressort des développements précédents (voir *supra* paragraphes 124 et suivants) qu'un seul acteur, Nvidia, semble avoir une position prépondérante dans le secteur des composants informatiques nécessaires à l'entraînement des modèles de fondation.

242. Compte tenu des contraintes fortes liées à l'approvisionnement des processeurs graphiques et la concentration du secteur, plusieurs acteurs craignent que ce contexte favorise la mise en œuvre de pratiques potentiellement anticoncurrentielles :

- l'étude récente de France Digitale[140], réalisée notamment sur la base d'entretiens avec une quarantaine d'entreprises du secteur, fait ainsi état de risques potentiels tels que des **fixations des prix, des restrictions de l'approvisionnement, des conditions contractuelles déloyales ou des comportements discriminatoires** ;

---

[139] Signalons à cet égard que, conscientes des profonds déséquilibres caractérisant les marchés du numérique, des autorités nationales de concurrence, telles que l'autorité italienne, ont intégré des dispositions relatives aux abus de dépendance. Ainsi, en droit italien, depuis le 1er novembre 2022, existe une présomption de dépendance économique dans les relations commerciales entre les plateformes numériques et les entreprises qui utilisent leurs services d'intermédiation si la plateforme joue un rôle déterminant pour atteindre les utilisateurs finaux ou les fournisseurs, notamment en termes d'effets de réseau ou de disponibilité des données.

[140] France Digitale, « Des puces aux applications, l'Europe peut-elle être une puissance de l'IA générative ? », avril 2024, page 13.

- par ailleurs, des préoccupations relatives à la dépendance du secteur envers le logiciel de programmation de puces **CUDA** de Nvidia, seul environnement parfaitement compatible avec les GPU devenues incontournables pour le calcul accéléré, ont été exprimées ;
- les récentes annonces d'investissements de Nvidia dans des fournisseurs de services *cloud* spécialisés dans l'IA, tels que Coreweave, suscitent également les inquiétudes de fournisseurs de *cloud* généralistes. Selon France Digitale, « *ces fournisseurs spécialisés bénéficient d'un partenariat privilégié avec Nvidia, leur permettant ainsi d'offrir un accès aux GPU à des tarifs jusqu'à 80% moins élevés que ceux proposés par les fournisseurs de cloud généralistes. Cette situation pourrait engendrer une concurrence déloyale en matière de tarification avec les fournisseurs de cloud généralistes, notamment les acteurs de milieu de gamme qui ne disposent pas des ressources financières des hyperscalers* ».

243. Le **risque de concentration de la puissance de calcul à terme au profit des grands acteurs du numérique constitue un autre sujet de préoccupation**. En effet, comme exposé *supra* aux paragraphes 189 et suivants, leur accès préférentiel aux GPU de Nvidia, le développement de leurs propres accélérateurs d'IA et leurs prises de participations au sein d'entreprises innovantes du secteur pourraient réduire progressivement la concurrence sur l'accès à la puissance de calcul[141].

244. Le secteur des cartes graphiques, **qui a fait l'objet d'une opération de visite et saisie inopinée en septembre 2023**, est attentivement scruté par les services d'instruction de l'Autorité[142].

Des risques de verrouillage par les grands fournisseurs de services cloud

245. Les fournisseurs de services *cloud* en place jouent un rôle important dans le développement de nouvelles technologies d'IA car ils fournissent de grandes quantités de ressources informatiques, nécessaires notamment aux développeurs de modèle de langage. Cependant, la position que ces entreprises occupent en tant que fournisseurs d'un intrant essentiel pour les technologies d'IA crée le risque qu'elles en abusent pour étendre leur pouvoir de marché et diminuer la concurrence.

246. **L'Autorité constate que plusieurs pratiques de verrouillage financier et technique, déjà identifiées lors de l'avis n° 23-A-08 sur le *cloud*, semblent perdurer et même s'intensifier afin d'attirer le plus grand nombre de *start-ups* actives dans le secteur de l'IA générative.**

247. Tout d'abord, des offres de **crédits *cloud*** particulièrement élevées sont proposées notamment à destination des entreprises innovantes du secteur.

[141] Stucke, Maurice E. and Ezrachi, Ariel, Antitrust & AI Supply Chains, 11 mars 2024, page 18.

[142] Communiqué de l'Autorité, Le rapporteur général de l'Autorité de la concurrence indique qu'une opération de visite et saisie inopinée a été réalisée dans le secteur des cartes graphiques, 27 septembre 2023.

Pour mémoire, les crédits *cloud* sont des offres d'essai qui prennent la forme d'allocations de services proposées par un fournisseur, octroyant un accès gratuit à un client dans un délai défini. Dans son avis précité sur le secteur du *cloud*, l'Autorité avait considéré que les programmes de crédits à destination exclusive de certaines cibles, notamment celles qui concentrent un fort potentiel d'innovation telles que les *start-ups*, les développeurs, les chercheurs ou les étudiants devaient faire l'objet d'une attention particulière. L'Autorité avait ainsi indiqué que « [l]*es montants élevés parfois proposés, le vaste écosystème d'entreprises qu'ils concernent, leur durée de validité et les risques de verrouillage décrits plus haut les distinguent significativement des essais gratuits qui peuvent être traditionnellement observés dans d'autres industries, et soulèvent des doutes quant à la capacité de tous les fournisseurs de services cloud à les répliquer* »[143].

248. Une stratégie similaire est ici développée, notamment par les *hyperscalers,* à destination de jeunes entreprises actives dans le secteur de l'IA générative. À titre d'exemple, Google Cloud offre « *aux start-ups d'IA* » éligibles au programme « *Cloud Google for Startups* »[144] jusqu'à 350 000 dollars (environ 325 000 euros) sur deux ans pour l'utilisation de Google Cloud et Firebase (une plateforme de Google permettant de développer rapidement des applications), soit 150 000 dollars (environ 140 000 euros) de plus par rapport aux jeunes entreprises actives dans d'autres secteurs[145]. Des programmes similaires avec des primes pour les *start-ups* IA sont également proposés par Amazon[146] et Microsoft[147].

249. Le montant de ces crédits peut aussi être augmenté sous la condition que les *start-ups* en bénéficiant utilisent les nouveaux produits de ces fournisseurs pour l'IA. Par exemple, les entreprises sélectionnées dans le cadre du « AWS Activate »[148], un programme réservé aux *start-ups,* peuvent bénéficier d'un maximum de 100 000 dollars (soit environ 92 000 euros) de crédits promotionnels AWS pour commencer à créer des modèles. Par ailleurs, les *start-ups* qui s'appuient sur les accélérateurs d'IA d'AWS (AWS Trainium et AWS Inferentia) peuvent être éligibles à un montant de 300 000 dollars (soit environ 280 000 euros) de crédits supplémentaires.

250. Ces offres sont particulièrement attractives pour ces entreprises, car elles leur permettent d'avoir accès gratuitement aux services *cloud*, nécessaires pour l'entraînement, la

---

[143] Voir l'avis de l'Autorité n° 23-A-08 précité, paragraphe 423.

[144] Google Cloud offre à ses clients des réductions et des crédits pour les produits de calcul et autres produits (y compris l'assistance aux *start-ups* axées sur l'IA) via le programme Cloud Google for startups.

[145] Le site de Google précise « *Bénéficiez de l'aide d'experts dédiés aux start-up, d'une prise en charge de vos coûts Google Cloud et Firebase allant jusqu'à 200 000 $ (jusqu'à 350 000 $ pour les start-up d'IA) pendant deux ans, d'une formation technique, d'une assistance commerciale et d'offres valables sur divers produits et services Google* ».

[146] Le programme intitulé « AWS Generative AI Accelerator » proposait en 2023 un programme mondial conçu pour aider dix *start-ups* actives dans le domaine de l'IA générative à réaliser leur potentiel. Les participants admissibles pouvaient recevoir jusqu'à 300 000 dollars (soit environ 280 000 euros) de crédits AWS. Amazon a récemment annoncé élargir son offre de crédits pour couvrir les frais des *start-ups* utilisant les principaux modèles d'IA (comme Anthropic, Mistral ou Cohere) en indiquant « *C'est un autre cadeau que nous faisons à l'écosystème des startups, en échange de ce que nous espérons, c'est-à-dire que les startups continuent à choisir AWS en première intention* » (traduction libre).

[147] Microsoft soutient les *start-ups* actives dans le secteur de l'IA générative par le biais de son programme « hub des créateurs ». Celui-ci offre jusqu'à 150 000 dollars (soit 140 000 euros) de crédits Azure aux entreprises éligibles.

[148] https://aws.amazon.com/fr/startups/generative-ai/.

spécialisation ou le déploiement de leurs solutions. Néanmoins, compte tenu de l'ampleur des coûts nécessaires pour l'entraînement ou le réglage fin de modèles d'IA, cette pratique a pour effet d'encourager les utilisateurs à choisir les services de ces *hyperscalers* en raison du montant des crédits *cloud* qui leur sont offerts et non pas seulement parce que ceux-ci répondraient le mieux à leurs besoins à long terme. Ces crédits pourraient ainsi avoir pour effet de verrouiller ces entreprises au sein des écosystèmes des *hyperscalers*, dans un contexte de freins techniques et tarifaires à la migration.

251. D'autres pratiques ont été identifiées au-delà des crédits *cloud* comme **des pratiques de verrouillage technique**. Les *hyperscalers* proposeraient en effet des solutions propriétaires (par exemple, des services d'apprentissage automatique automatisés, voir glossaire) pour les entreprises souhaitant créer ou régler leurs modèles plus facilement. Cependant, lorsque le modèle final est créé, les utilisateurs n'auraient pas accès au modèle lui-même, mais seulement à la possibilité de l'utiliser ou de le déployer à partir de l'infrastructure du fournisseur de services *cloud*. Cette pratique verrouillerait ainsi l'utilisateur qui, s'il voulait changer de fournisseur de services *cloud*, devrait recréer son modèle d'IA à partir de zéro, ce dernier ne pouvant être transféré chez un autre fournisseur.

252. Outre qu'elles pourraient être qualifiées en droit de la concurrence, notamment d'abus de position dominante, certaines de ces pratiques sont également encadrées par la loi SREN ou par le règlement européen sur les données (« *Data Act* »).

<u>La loi SREN</u> :

La loi SREN repose sur trois axes principaux : la protection des citoyens, des plus jeunes, et des entreprises et collectivités. Ce dernier volet a notamment pour objet d'anticiper la mise en œuvre du règlement européen sur les données et de limiter dans le temps la pratique des crédits *cloud*.

Dans son avis n° 23-A-05 sur le projet de loi[149], l'Autorité avait souligné que, compte tenu du contexte réglementaire européen dans lequel le projet de loi s'insérait, il convenait de s'assurer de la bonne articulation des mesures envisagées avec le futur cadre européen, afin de ne pas pénaliser les acteurs opérant sur le marché français.

L'article 26 de la loi SREN prévoit ainsi que « *les avoirs d'informatique en nuage* » (les crédits *cloud*) sont limités à un an et ne peuvent être assortis d'une clause d'exclusivité. Ce même article interdit la pratique commerciale déloyale visant à subordonner la vente d'un produit ou d'un service à la conclusion concomitante d'un contrat de fourniture de services *cloud*. La loi prévoit enfin la remise d'un rapport par l'Autorité sur la pratique d'autopréférence (en anglais « *self-preferencing* ») ainsi que les éventuelles améliorations procédurales ou législatives nécessaires.

L'article 27 interdit les frais de transfert de données (*egress fees*) supérieurs aux coûts facturés lorsqu'un client souhaite transférer ses données vers ses propres infrastructures ou vers les infrastructures d'un autre fournisseur au sein d'une architecture multi-*cloud*. L'article 28 prévoit pour sa part notamment des exigences d'interopérabilité des services *cloud* et de portabilité des données. Ces deux articles sont assortis d'une clause d'extinction au 12 janvier 2027 afin d'assurer leur compatibilité avec les dispositions du Data Act.

[149] Avis de l'Autorité n° 23-A-05 du 20 avril 2023 concernant le projet de loi visant à sécuriser et réguler l'espace numérique.

Des préoccupations de concurrence concernant l'accès aux données

253. Plusieurs préoccupations de concurrence sont susceptibles d'être soulevées aussi bien à l'amont qu'à l'aval de la chaîne de valeur, notamment en ce qui concerne l'accès aux données.

254. En effet, les développeurs ont besoin de données massives et générales lors de la phase d'entraînement et de données spécialisées lors de la phase de réglage fin des modèles de fondation. À l'étape de l'inférence, les agents d'IA conversationnels doivent accéder aux données nécessaires pour répondre aux requêtes des utilisateurs.

255. Or, les entreprises innovantes du secteur pourraient être confrontées à des **pratiques de refus d'accès ou d'accès discriminatoire aux données** de la part d'entreprises disposant d'un accès significatif aux données. France Digitale indique à cet effet : « *à titre d'exemple, une entreprise disposant d'un accès significatif aux données, comme un index web ou un moteur de recherche, pourrait refuser ou restreindre l'accès aux données sous son contrôle. De même, ces acteurs pourraient favoriser les développeurs avec lesquels ils ont établi un partenariat (par exemple, pour la fourniture de services cloud ou de plateforme), ou privilégier leurs propres services internes. De plus, les entreprises dominant le marché pourraient contraindre leurs partenaires contractuels à ne pas fournir leurs données à des développeurs d'IA concurrents. Par exemple, elles pourraient imposer des restrictions au web scraping ou accorder des droits exclusifs d'utilisation des données en échange de services publicitaires, de référencement web ou de services cloud. Enfin, les grands acteurs pourraient proposer des services ou des technologies (comme des droits d'inférence) en échange de données, rendant ainsi les discussions avec d'autres acteurs moins attrayantes pour les détenteurs de droits* »[150].

256. De ce point de vue, **l'Autorité sera vigilante aux données rendues accessibles aux grands acteurs du numérique dans le cadre de leurs partenariats noués avec des entreprises**. En effet, un partenariat conclu entre une grande entreprise du numérique et une entreprise industrielle française pour le réglage fin de ses modèles sur ses données propriétaires ne lui donnerait pas accès au même volume de données qu'un partenariat noué par exemple avec une entreprise active dans l'intermédiation de données. Ce dernier serait examiné avec attention par les services de l'Autorité.

257. Le refus d'accès aux données pourrait prendre des formes plus subtiles. Des opérateurs puissants pourraient ainsi chercher à acquérir ou à consolider une position dominante dans le secteur de l'IA générative en proposant le versement de rémunérations importantes aux créateurs de contenus, en particulier pour **exclure des acteurs concurrents** moins établis ou de potentiels entrants. Ainsi, selon certaines parties prenantes, la rémunération élevée pourrait être compensée par un pouvoir de marché accru du fait de la marginalisation ou de l'exclusion des acteurs moins établis[151]. Une rémunération élevée des créateurs de contenus pourrait donc, potentiellement, constituer un abus de position dominante.

258. Cependant, **les données numériques constituent économiquement un bien non rival** : en d'autres termes, vendre des données à un acteur ne limite a priori pas la capacité de vendre

[150] Étude France Digitale précitée, page 32.

[151] Voir la littérature économique sur les incitations à augmenter les coûts des concurrents, notamment : Salop et Scheffman (1983), « *Raising rivals' costs* », American Economic Review ; Krattenmaker et Salop (1986), « *Anti-competitive foreclosure : raising rivals' cost to achieve power over price* » the Yale Law Journal ; Salop (2017), « *The raising rivals' cost foreclosure paradigm, conditional pricing practices, and the flawed incremental price-cost test* », Antitrust Law Journal.

les mêmes données à un autre acteur, le cas échéant à un prix différent. **La question de l'incitation** des fournisseurs de contenus à conclure de tels accords différenciés se pose néanmoins, et à la connaissance de l'Autorité aucun éditeur de presse n'a, à ce jour, signé d'accord avec plusieurs développeurs de modèles et à des prix différents.

259. Par ailleurs, de **telles pratiques, combinées à des clauses d'exclusivité,** pourraient renforcer d'éventuelles préoccupations de concurrence. De telles clauses, mises en œuvre par des opérateurs puissants, seraient en effet susceptibles d'empêcher leurs concurrents d'accéder aux données dans les mêmes conditions. Ces accords seraient ainsi susceptibles de verrouiller les fournisseurs de données, limitant ainsi les opportunités des concurrents.

260. Certains acteurs s'inquiètent également de l'existence d'un avantage à être le premier à nouer des partenariats, via l'ajout dans les contrats entre les développeurs de modèles et les fournisseurs de contenus de clause échangeant des données contre des services *cloud*. Ainsi, un acteur indique que « *les grands acteurs pourraient proposer des services ou des technologies (comme des droits d'inférence) en échange de données, rendant ainsi les discussions avec d'autres acteurs moins attrayantes pour ces fournisseurs de données* ».

261. L'accès **aux données des utilisateurs** constitue également un enjeu majeur, comme le relève la Commission pour l'IA : « *force est de constater que beaucoup de données intéressantes pour entraîner des IA ont un caractère personnel. Dans la santé, bien entendu, mais pas uniquement. Même l'IA générative, a priori plus intéressée par les données culturelles, peut en avoir besoin pour développer une capacité d'interaction spécifique. Dans l'éducation, entraîner un modèle capable d'interagir de façon crédible et pertinente avec un élève nécessitera probablement un entraînement sur des données de dialogue entre élèves et enseignants, qui sont des données personnelles* »[152]. En effet, plusieurs acteurs rapportent que les grandes entreprises du secteur continuent d'utiliser diverses stratégies pour limiter l'accès des tiers aux données de leurs utilisateurs, en faisant un usage abusif des règles juridiques comme la protection des données personnelles ou encore des préoccupations de sécurité.

262. L'interaction entre le droit de la concurrence et la protection des données personnelles fait l'objet d'une attention particulière de l'Autorité, comme le montre sa récente déclaration conjointe avec la CNIL sur le thème « *concurrence et données personnelles : une ambition commune* ». Celle-ci a permis de rappeler les modalités de prise en compte, par l'Autorité, du paramètre de concurrence relatif aux « données personnelles » et a également confirmé que « *certaines politiques de confidentialité posent la question de l'éventuelle utilisation d'arguments relatifs à la protection de la vie privée à des fins anticoncurrentielles* » [153]. L'Autorité est particulièrement attentive à ce que la mise en œuvre du règlement général sur la protection des données (ci-après « RGPD ») par les grands acteurs du numérique ne crée pas de risque de comportement d'exclusion ou d'autopréférence.

263. L'arrêt Meta de la Cour de justice de l'Union européenne (ci-après la « Cour de justice ») a d'ailleurs confirmé[154] qu'une autorité nationale de concurrence peut constater une violation du RGPD et la qualifier d'abus de position dominante, précisant les modalités de coopération avec les autorités de contrôle du règlement sur la protection des données personnelles.

---

[152] Rapport de la Commission pour l'IA précité, page 100.

[153] Déclaration conjointe de l'Autorité de la concurrence et de la CNIL, « Protection de données et concurrence : une ambition commune », 12 décembre 2023, page 7.

[154] CJUE, 4 juillet 2023, Meta Platforms Inc. e.a. contre Bundeskartellamt, C-252/21.

264. Il convient enfin de relever que les éditeurs expriment de grandes préoccupations liées à l'exploitation de leurs contenus par les fournisseurs de modèles de fondation, **sans l'autorisation des détenteurs de droits**. La décision récente de l'Autorité dans l'affaire des « droits voisins » a ainsi établi que Google avait utilisé, aux fins d'entraînement de son modèle de fondation Gemini (anciennement Bard), des contenus des éditeurs et agences de presse, sans les avertir et sans leur offrir la possibilité effective d'exercer leur droit de retrait[155]. Si cette question soulève des questions de respect des droits de propriété intellectuelle qui vont au-delà du champ d'étude du présent avis[156], le droit de la concurrence pourrait, sur le principe, appréhender ces questions sur le fondement d'une atteinte à la loyauté de la transaction, par exemple, et donc, de l'abus d'exploitation. L'Autorité rappelle à cet égard qu'elle sanctionne les comportements, qui, sous couvert de protection des droits de propriété intellectuelle, constituent en réalité des pratiques anticoncurrentielles, car elles vont au-delà de ce qui est nécessaire pour cette légitime protection[157].

Une vigilance particulière s'impose sur les risques liés à l'accès à une main-d'œuvre qualifiée

265. Si plusieurs entreprises ont développé des solutions innovantes d'IA générative et disposent donc d'un savoir-faire important dans ce secteur, les grandes entreprises du numérique pourraient mettre en œuvre des pratiques susceptibles de limiter ou d'empêcher le libre mouvement du personnel qualifié, et donc la concurrence associée.

266. Pour l'IA générative, comme pour le reste du secteur du numérique, les ressources humaines constituent en effet un bien particulièrement rare et disputé par les entreprises, en recherche perpétuelle de talents et de moyens pour les fidéliser[158].

---

[155] Décision n° 24-D-03 du 15 mars 2024 relative au respect des engagements figurant dans la décision de l'Autorité de la concurrence n° 22-D-13 du 21 juin 2022 relative à des pratiques mises en œuvre par Google dans le secteur de la presse. L'Autorité a notamment considéré que l'absence de toute information donnée par Google sur les utilisations faites des contenus des éditeurs et agences de presse par son service Bard constituait un manquement à l'obligation de transparence résultant de l'engagement 1. Google a également enfreint l'engagement n° 6 en liant l'utilisation des contenus des éditeurs et agences de presse par son service d'IA à l'affichage des contenus protégés sur des services comme Search, Discover et Actualités. La question de savoir si l'utilisation de publications de presse dans le cadre d'un service d'IA relève de la protection au titre de la règlementation des droits voisins n'a pas été tranchée. Google n'a pas contesté les pratiques reprochées et a sollicité le bénéfice de la procédure de transaction. L'Autorité a prononcé une sanction totale de 250 millions d'euros à l'encontre des sociétés Alphabet Inc, Google LLC, Google Ireland Ltd et Google France et présenté une série de mesures correctives visant à répondre aux préoccupations identifiées.

[156] Plusieurs éditeurs de contenu ont indiqué ne jamais avoir été informés ou avoir autorisé l'usage de leurs contenus pour cette finalité ni avoir obtenu de rémunération tandis que certains développeurs de modèles se prévaudraient de l'article L. 122-5-3 III du code de la propriété intellectuelle prévoyant que des copies ou reproductions numériques d'œuvres auxquelles il a été accédé de manière licite peuvent être réalisées en vue de fouilles de textes et de données menées à bien par toute personne, quelle que soit la finalité de la fouille, sauf si l'auteur s'y est opposé de manière appropriée, notamment par des procédés lisibles par machine pour les contenus mis à la disposition du public en ligne. Une mission a été confiée au Conseil supérieur de la propriété littéraire et artistique aux fins d'examiner les mécanismes juridiques envisageables afin de garantir la juste rémunération des ayants droit et analyser les enjeux économiques sous-jacents à l'accès aux données protégés lorsque celles-ci sont utilisées par les IA.

[157] Décision n° 23-D-14 du 20 décembre 2023 relative à des pratiques mises en œuvre dans les secteurs des consoles statiques de jeux vidéo de huitième génération et des accessoires de contrôle compatibles avec la console PlayStation 4.

[158] Voir, par exemple, https://www.rhmatin.com/formation/digital-learning/enjeux-rh-et-formation-aux-ia-generatives-quelle-echelle-faut-il-atteindre-en-france.html et plus largement, le bilan 2022 et perspectives

267. En France et plus globalement en Europe, les moyens juridiques auxquels ont recours les entreprises sont d'abord strictement encadrés par le droit civil. En effet, au niveau national, si l'article L. 1121-1 du code du travail affirme le principe de la liberté du travail (« [n]*ul ne peut apporter aux droits des personnes et aux libertés individuelles et collectives de restrictions qui ne seraient pas justifiées par la nature de la tâche à accomplir ni proportionnées au but recherché* »), certaines clauses peuvent être insérées par les entreprises dans les contrats de travail de leurs employés, d'une part, et dans les contrats entre entreprises, d'autre part, afin de restreindre ou empêcher la mobilité de leur personnel. Ces stipulations font l'objet d'un contrôle étroit par les juridictions.

268. C'est le cas, par exemple, des clauses suivantes :

- **une clause de non-concurrence** est une clause du contrat de travail par laquelle le salarié s'engage à ne pas exercer, pendant une période déterminée à partir de la cessation de la relation de travail, une activité concurrente à celle de son employeur, pour son propre compte ou celui d'un autre employeur[159]. Pour qu'elle soit valide, la clause de non-concurrence doit répondre à plusieurs critères cumulatifs comme le fait d'être justifiée par les intérêts légitimes de l'entreprise mais aussi être limitée dans le temps et l'espace. Elle doit également viser une activité précise et prévoir une contrepartie financière pour le travailleur à qui elle peut être opposée[160] ;
- **une clause de non-sollicitation** de personnel est une clause entre entreprises par laquelle le bénéficiaire interdit au débiteur de solliciter et/ou d'embaucher son personnel, sous peine d'indemnisations généralement assises sur les salaires mensuels des travailleurs concernés. Ces clauses sont souvent utilisées lorsqu'un prestataire de services met du personnel à la disposition d'une entreprise. Elles sont particulièrement utilisées dans le secteur du numérique, pour la fourniture de produits ou solutions numériques déployés dans les entreprises clientes, et visent à permettre aux fournisseurs de garder leurs techniciens, experts ou « consultants ». La Cour de cassation estime que ce type de clause n'est ni une variante, ni une précision de la clause de non-concurrence[161].

269. En droit de la concurrence, les pratiques mises en œuvre sur les marchés du travail font également l'objet d'une vigilance particulière des autorités de contrôle. Au-delà des accords de fixation des salaires entre entreprises, les accords de non-débauchage (*no-poach*) peuvent également constituer des pratiques anticoncurrentielles prohibées.

270. C'est ainsi qu'en 2010 aux États-Unis, une action a été engagée par le ministère de la justice à l'encontre des entreprises Adobe, Apple, Google, Intel, Intuit et Pixar pour s'être interdit réciproquement de solliciter les salariés les plus qualifiés[162].

2023 de Numeum, une organisation professionnelle représentant l'écosystème numérique en France, faisant état d'une « *pénurie de talents formés à l'ensemble des compétences nécessaires pour déployer les dernières innovations technologiques* ».

[159] https://www.dalloz.fr/documentation/Document?id=DZ%2FOASIS%2F000185.

[160] Voir notamment Cassation, soc., 10 juillet 2002, 00-45.135.

[161] Voir notamment Cass. com., 31 janv. 2012, n° 11-11.071, P+B, SAS Capp invest immo c/ Sté Socorpi.

[162] Antitrust Division U.S. V. Adobe Systems, Inc., Apple Inc., Google Inc., Intel Corporation, Intuit, Inc., And Pixar. Les accords conclus entre Apple et Google, Apple et Adobe, Apple et Pixar et Google et Intel empêchaient les entreprises de solliciter directement (« *no cold call agreements* ») les employés de l'autre partie. Un accord entre Google et Intuit empêchait Google de solliciter directement les employés d'Intuit. Ces accords étaient vus comme éliminant ainsi une forme importante de concurrence pour attirer des employés hautement qualifiés et diminuant globalement la concurrence au détriment des employés concernés. L'action,

271. En France, dans une affaire concernant les revêtements de sols[163], l'Autorité a, par ailleurs, examiné et sanctionné, parmi d'autres pratiques anticoncurrentielles, des accords tacites de non-agression ou « *gentleman*['s] *agreement* » entre concurrents visant notamment à interdire le démarchage de leurs salariés respectifs. Les services d'instruction de l'Autorité ont par ailleurs annoncé avoir notifié des griefs d'entente sur le marché du travail à plusieurs entreprises des secteurs de l'ingénierie, du conseil en technologies et des services informatiques, en novembre 2023[164].

272. Selon la Commission européenne[165], les accords de non-débauchage, comme les accords de fixation de salaire, sont susceptibles d'être qualifiés d'accords restrictifs de concurrence par objet, et sont prohibés par l'article 101, paragraphe 1, du Traité sur le fonctionnement de l'Union européenne. Elle précise que des effets proconcurrentiels de ces accords sont possibles, mais incertains : ces effets doivent être démontrés et significatifs, tandis qu'il peut exister des moyens moins restrictifs ou plus respectueux des droits et libertés des employés de les obtenir, en prenant notamment pour référence les clauses de non-concurrence, si elles sont conformes aux législations nationales.

273. Sur ces dernières clauses, les États-Unis semblent adopter une approche différente puisque la FTC aux États-Unis vient d'interdire la majorité des clauses de non-concurrence[166]. Les clauses de non-concurrence existantes vis-à-vis des cadres supérieurs restent toutefois en vigueur.

274. Un sujet de préoccupation additionnel consiste dans **le recrutement, par les géants du numérique, d'équipes entières** (comme le montre le recrutement par Microsoft d'une grande partie des 70 employés de la *start-up* Inflection) **ou d'employés stratégiques** de développeurs de modèles (comme le bref recrutement par Microsoft du fondateur d'OpenAI après son licenciement, avant qu'il ne soit finalement réintégré dans la société). Si cette pratique peut être examinée sous l'angle du contrôle des concentrations (voir *infra* paragraphes 295 et suivants), elle peut également s'analyser en une tentative d'exclusion de concurrents du secteur. En effet, la raréfaction des talents pourrait empêcher les développeurs de modèles de fondation d'entraîner des modèles performants, susceptibles de concurrencer ceux des géants du secteur. De telles pratiques pourraient être appréhendées en droit des pratiques anticoncurrentielles par la prohibition des abus de position dominante.

275. Au demeurant, une voie de droit spécifique et éprouvée est prévue parallèlement en droit commun national : l'action en concurrence déloyale permet de sanctionner sur le fondement du droit commun de la responsabilité civile extracontractuelle les actes de désorganisation, de parasitisme[167] et de dénigrement pouvant résulter de démarches d'embauches massives, ou ciblées sur des travailleurs-clés, afin d'obtenir réparation du préjudice subi. De telles

---

engagée sur le fondement de la section 1 du *Sherman Act*, s'est terminée par un accord transactionnel qui a mis fin aux poursuites.

[163] Décision n° 17-D-20 du 18 octobre 2017 relative à des pratiques mises en œuvre dans le secteur des revêtements de sols résilients.

[164] Autorité de la Concurrence, Le rapporteur général de l'Autorité de la concurrence a notifié des griefs d'entente sur les marchés du travail à plusieurs entreprises des secteurs de l'ingénierie, du conseil en technologies et des services informatiques, 23 novembre 2023.

[165] https://competition-policy.ec.europa.eu/document/download/adb27d8b-3dd8-4202-958d-198cf0740ce3_en.

[166] FTC, FTC Announces Rule Banning Noncompetes, 23 avril 2024.

[167] Voir notamment Cass. com., 5 février 1991, n° 88-16.214.

démarches peuvent ainsi être considérées comme fautives lorsqu'elles ont lieu dans des conditions déloyales et entrainent une désorganisation de l'entreprise ciblée[168].

276. Compte tenu de l'extrême rapidité des évolutions technologiques dans le domaine de l'IA générative, il est déterminant pour les entreprises, et notamment les développeurs de modèles, de pouvoir recruter et conserver des talents nouvellement formés et/ou maitrisant les toutes dernières technologies. Si l'adoption des clauses précitées peut poursuivre un objectif légitime (protection de l'investissement dans la formation des salariés, protection contre la concurrence déloyale abusive, protection du savoir-faire et de secrets industriels), des voies de droit existent et les grandes entreprises du numérique doivent s'interdire les pratiques déloyales, anticoncurrentielles ou préjudiciables aux employés, afin de ne pas dissuader l'entrée de nouvelles entreprises innovantes ou limiter de manière indue ou abusive la mobilité des travailleurs vers leurs concurrents. Ces pratiques et restrictions pourraient avoir pour conséquence une concentration des talents au sein d'un nombre limité d'entreprises pouvant entrainer des distorsions de concurrence importantes sur le marché du travail, préjudiciables aux employés concernés et nuisant *in fine* aux consommateurs.

277. Ainsi, s'il ressort de l'instruction du présent avis que de telles restrictions ne semblent pas, pour l'instant, soulever de préoccupations particulières des parties prenantes, l'Autorité estime qu'une vigilance s'impose sur ces questions.

Les modèles en accès libre peuvent entrainer des risques concurrentiels

278. Si les modèles en accès libre peuvent permettre d'abaisser les barrières à l'entrée (voir *supra* paragraphes 179 et suivants), ils peuvent également susciter des préoccupations de concurrence. En effet, dans certains cas, les conditions d'accès et de réutilisation des modèles ou de certains de leurs composants peuvent conduire au verrouillage des utilisateurs.

279. L'affaire « Google Android » [169] illustre le cas de restrictions anticoncurrentielles dans le secteur des logiciels libres.

Arrêt du Tribunal de l'Union européenne dans l'affaire Google et Alphabet c/ Commission (Google Android)

Google avait empêché des fabricants d'appareils d'utiliser une autre version d'Android (son système d'exploitation *open source*) non approuvée par elle (les « fourches » Android, c'est-à-dire un nouveau logiciel créé à partir du code source d'un logiciel existant). Pour pouvoir préinstaller sur leurs appareils les applications propriétaires de Google, y compris Play Store et Google Search, les fabricants devaient s'engager à ne développer ou vendre aucun appareil fonctionnant sous une fourche Android.

Le Tribunal de l'Union européenne a considéré que la pratique en cause avait conduit au renforcement de la position dominante de Google sur le marché des services de recherche générale, tout en constituant un frein à l'innovation, dans la mesure où elle avait limité la diversité des offres accessibles aux utilisateurs. Cette pratique avait notamment conduit la Commission européenne à sanctionner Google sur le fondement de l'abus de position dominante.

[168] Voir notamment Cass. com, 8 juillet 2020, n° 18-17.169, Cass. com., 9 mars 1999 n° 97-12.009.

[169] Voir le cas de la Commission européenne AT.40099 — Google Android et l'arrêt du Tribunal dans l'affaire T-604/18, Google et Alphabet/Commission (Google Android) (pourvoi en cours).

280. Des restrictions de concurrence similaires peuvent également exister dans le secteur de l'IA générative. L'Autorité perçoit deux types de risques associés aux modèles *open source*.

281. En premier lieu, des risques sont présents dès la mise à disposition des modèles. Les développeurs de modèles peuvent ainsi prévoir l'interdiction de développer des modèles qui concurrencent directement leurs propres modèles, imposer des limites à leur exploitation commerciale ou à la conception de produits ou services concurrents. C'est le cas par exemple du modèle Llama 2 de Meta imposant l'octroi d'une licence supplémentaire si son usage dépasse plus de 700 millions d'utilisateurs[170].

282. En second lieu, certaines entreprises peuvent initialement adopter une approche ouverte de l'IA générative afin d'étendre leur pouvoir de marché[171] en verrouillant les entreprises utilisatrices. Un acteur met ainsi en garde contre ce risque : « *les bénéfices concurrentiels de l'open source supposent que les producteurs des modèles n'en restreignent pas l'accès par la suite en exploitant la dépendance des utilisateurs ("lock in")* ».

c) Les risques liés à la présence d'entreprises sur plusieurs marchés distincts

283. Plus globalement, l'intégration verticale de certains acteurs du numérique, et leur écosystème de services, sont susceptibles de donner lieu à plusieurs pratiques abusives.

284. À l'amont, les développeurs de modèles pourraient se voir opposer **un refus (ou des limites) d'accès à des puces ou des données nécessaires pour entraîner des modèles de fondation concurrents.** Par exemple, les développeurs pourraient être lésés par des accords qui permettraient à un fournisseur d'infrastructure *cloud* d'avoir un accès exclusif à des données clés nécessaires pour entraîner des LLM ou de monopoliser les puces nécessaires pour les développer et les exécuter. Cette pratique pourrait avoir pour effet d'entraîner des retards ou de conduire à la mise en place de modèles moins ambitieux, nuisant ainsi au maintien d'une concurrence effective sur le marché.

285. Plusieurs acteurs s'inquiètent également **des accords d'exclusivité** entre fournisseurs de services *cloud* et développeurs de modèles de fondation. Selon eux, ces accords viseraient en effet à s'assurer que les développeurs dépendent exclusivement de ces fournisseurs pour l'accès aux services *cloud* nécessaires et pour la distribution aux clients et seraient ainsi susceptibles d'avoir **un impact sur l'innovation** et la concurrence entre fournisseurs, surtout lorsqu'un modèle particulier occupe une position significative sur le marché.

286. Ces effets de verrouillage sont encore aggravés lorsqu'ils sont combinés à d'autres mesures leur conférant une influence sur le développeur de modèles (comme des prises de participations élevées).

287. D'autres risques découlent de l'utilisation en aval des modèles d'IA générative :

- **ventes liées** : les entreprises détenant des positions prééminentes ou dominantes sur des marchés connexes pourraient lier la vente de ces produits ou services à celles de leurs propres solutions d'IA. Cela peut être le cas de pratiques ou d'accords liant différents produits entre eux, par exemple en intégrant une solution d'IA générative directement dans une offre logicielle (comme le déploiement de Copilot par Microsoft, son assistant

[170] L'Usine Digitale, Meta lance Llama 2, un grand modèle de langage open source et gratuit même pour une utilisation commerciale, 19 juillet 2023.

[171] Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI, 17 août 2023.

d'IA alimenté par le modèle GPT-4 d'OpenAI) ou dans des moteurs de recherche (comme c'est le cas avec l'introduction récente du « *AI Overview* » de Google qui dispense l'utilisateur de consulter les sites sources pour obtenir une information). **Les intégrations d'outils d'IA générative sur certains supports, comme les *smartphones,* suscitent également des inquiétudes.** Ces supports peuvent faire partie de l'écosystème de ces grands acteurs (Google a récemment étendu la disponibilité de son application Gemini à davantage de smartphones Android) ou appartenir à d'autres fabricants (Samsung et Google Cloud ont ainsi annoncé une collaboration à long terme pour rendre la technologie d'IA générative de Google Cloud accessible aux utilisateurs de Samsung dans le monde entier[172]). Ce type de pratiques pourrait consolider durablement le secteur de l'IA générative autour d'entreprises numériques déjà dominantes ;

- les concurrents en aval pourraient également être lésés par des pratiques **d'autopréférence de la part d'acteurs verticalement intégrés**, affectant la capacité des développeurs de modèles non intégrés verticalement à les concurrencer. Cela peut être le cas par exemple de pratiques consistant à exploiter les données utilisateurs collectées via leurs différents produits et à les utiliser pour améliorer la performance de leurs modèles d'IA ou des stratégies de verrouillage des modèles de langages ayant pour effet de limiter la concurrence en aval.

288. Sur l'ensemble de la chaîne de valeur, les parties prenantes observent qu'elles disposent de très peu de marge de manœuvre sur la négociation des prix et des conditions contractuelles pour utiliser les modèles d'IA de ces *hyperscalers*, car elles subissent le même traitement que pour les autres produits et services proposés par les mêmes *hyperscalers*.

289. L'ensemble de ces comportements pourraient permettre à certaines entreprises d'utiliser leur pouvoir de marché au détriment d'acteurs alternatifs, restreignant les choix offerts aux utilisateurs et l'incitation à développer des solutions alternatives.

2. LES PRISES DE PARTICIPATIONS MINORITAIRES ET PARTENARIATS DES GEANTS DU NUMERIQUE PEUVENT EGALEMENT SOULEVER DES PREOCCUPATIONS CONCURRENTIELLES

290. Dans un secteur comme l'IA, où les investissements sont très élevés compte tenu du coût d'accès aux intrants, seuls quelques grands acteurs concentrent les capacités financières pour conclure des accords de jeunes entreprises innovantes ou prendre des participations en leur sein (voir *supra*). Les investissements et les partenariats entre acteurs du secteur ne sont pas condamnables en soi. Ils peuvent en effet permettre aux *start-ups* de bénéficier des ressources financières et technologiques des grandes entreprises du numérique et, ainsi, favoriser l'innovation. Pour l'acquéreur, ces investissements permettent de se diversifier ou d'avoir accès à des technologies innovantes de nature à améliorer la qualité de ses services. Ils sont donc nécessaires au développement du secteur de l'IA.

291. Ils présentent néanmoins des risques concurrentiels qui appellent une vigilance particulière des autorités de concurrence. Or, les participations minoritaires, en ce qu'elles ne confèrent généralement pas de contrôle, ne sont que rarement examinées *ex ante* dans le cadre du contrôle des concentrations. Elles peuvent néanmoins être appréhendées *ex post* sous l'angle du droit des pratiques anticoncurrentielles.

---

[172] Communiqué de presse de Samsung, Samsung et Google Cloud unissent leurs forces avec l'annonce de l'IA générative sur la série Samsung Galaxy S24, 17 janvier 2024.

a) La nécessité d'une vigilance particulière dans le secteur de l'IA générative

292. Certaines prises de participations nécessitent une vigilance accrue. En effet, alors que ces opérations sont susceptibles d'entraîner de nombreux risques concurrentiels, les autorités de concurrence ne disposent pas d'informations claires sur les modalités de ces accords.

293. Ces participations, même minoritaires, peuvent avoir un impact concurrentiel sur le secteur :

- en effet, si le détenteur de ce type de participations est un concurrent de l'entreprise cible, il peut alors avoir des droits sur ses revenus, ce qui peut entraîner **un affaiblissement de l'intensité concurrentielle** entre les deux entités dans la mesure où les revenus de l'entreprise cible participent également aux revenus de l'acquéreur. Ces partenariats peuvent également éliminer l'un des concurrents si celui-ci intègre simplement les modèles de fondation du partenaire dans ses produits au lieu de développer ses propres modèles ;
- **des effets verticaux** peuvent aussi être constatés sur ces marchés puisqu'une entreprise cliente détentrice d'une participation minoritaire peut être incitée à s'approvisionner auprès de l'entreprise cible, entrainant de ce fait un désavantage dans la concurrence entre l'entreprise cible et ses concurrents sur le marché amont. Par ailleurs, ces accords entre fournisseurs de services *cloud* et développeurs de modèles, surtout lorsqu'ils comportent une clause d'exclusivité, peuvent renforcer le pouvoir de marché du fournisseur ;
- l'acquisition de participations minoritaires peut également entraîner des effets coordonnés en **renforçant la transparence du marché** dès lors que l'acquéreur est susceptible d'acquérir des informations commercialement sensibles (comme des plans d'affaires, des données actualisées sur les prix, des informations liées à l'infrastructure des modèles tels que les hyperparamètres, les traitements appliqués pour nettoyer les données, ou l'utilisation par les concurrents du modèle de langage du développeur). Un acteur du secteur s'inquiète ainsi que des partenariats entre un fournisseur d'infrastructure *cloud* et un développeur de grands modèles de langage permettent l'accès à des informations sensibles: « [e]*n l'absence de garanties robustes, les partenariats entre un fournisseur d'infrastructure cloud et un développeur de grands modèles de langage pourraient donner au fournisseur d'infrastructure cloud un accès à des informations sensibles sur le plan concurrentiel concernant les concurrents qui utilisent les modèles de langage du développeur, y compris, par exemple, des informations concernant les relations de ses concurrents et l'utilisation du fournisseur de modèle, les prompts et les réponses générés par le modèle, ou les feuilles de route de produits (y compris les innovations prévues liées aux fonctionnalités d'IA)* ». L'acquéreur pourra ainsi anticiper le comportement concurrentiel de l'entreprise cible et réagir en conséquence ;
- ces effets coordonnés peuvent être notamment renforcés **lorsqu'une entreprise investit dans plusieurs entreprises concurrentes** (comme c'est le cas de Microsoft avec les développeurs de modèles OpenAI et Mistral AI) **ou lorsque plusieurs entreprises puissantes du secteur du numérique investissent dans une même entreprise cible** (par exemple, les investissements d'Amazon et Google dans Anthropic). Comme l'indique un acteur du secteur : « *lorsqu'un acteur important détient des participations dans plusieurs entreprises concurrentes, cela peut en fonction des droits dont ils disposent créer des conflits d'intérêts et potentiellement renforcer le verrouillage auprès de certains fournisseurs. Cela peut limiter la concurrence en donnant à cet acteur un avantage concurrentiel déloyal, en restreignant l'accès à des technologies ou des*

ressources clés, ou en influençant les décisions stratégiques des entreprises dans lesquelles il détient des participations. Cela peut également conduire à une concentration du pouvoir et à une réduction de la concurrence, qui sera préjudiciable à l'innovation et aux consommateurs » ;

- enfin, le **recrutement** de nombreux employés d'Inflection par Microsoft a suscité des discussions sur la possibilité de considérer cette pratique comme une concentration[173].

294. L'instruction menée par les services de l'Autorité a confirmé que ces partenariats suscitent l'inquiétude de nombreux acteurs du secteur :

- pour un acteur : « [l]*es participations de ces dernières* [les GAFAM] - *même minoritaires - dans les technologies émergentes conduisent de facto à des verrouillages de marché, la domination dans le marché amont profitant largement au marché aval. Ici encore, la position de Microsoft ainsi que ses prises de participation sur de multiples marchés connexes conduisent à fausser le rapport de force.* (…) *la position de force de Microsoft lui permet de protéger sa position dans le domaine de l'IA générative, domaine sur lequel il s'appuie pour renforcer sa position à la fois sur le marché des solutions collaboratives ou sur le marché des moteurs de recherche, pénalisant ainsi les acteurs minoritaires à la fois sur le marché du search et sur celui de l'IA générative* » ;
- un autre acteur du secteur considère que « *si, à court terme, ces accords présentent l'avantage de faciliter l'accès à des ressources essentielles et à des canaux de distribution pour les startups, ces dernières pourraient se retrouver dépendantes des puces et/ou de l'infrastructure de l'entreprise partenaire. De plus, la plateforme cloud de l'entreprise partenaire pourrait tenter de s'imposer comme le canal de distribution exclusif pour certains modèles de fondation. Cela pourrait se produire, si, par exemple, les participations minoritaires actuelles des entreprises établies deviennent majoritaires ou sont transformées en acquisitions à part entière* ».

b) Le contrôle des concentrations permet le contrôle *ex ante* de certaines prises de participations

Ces opérations sont soumises à autorisation préalable si elles confèrent aux investisseurs un contrôle de fait et dépassent les seuils communautaires et nationaux de notification

295. Une prise de participation minoritaire est soumise à autorisation préalable au titre du contrôle des concentrations si elle confère un changement durable de contrôle au sens du paragraphe 1 de l'article 3 du règlement CE n° 139/2004 sur le contrôle des concentrations, c'est-à-dire la capacité d'exercer une **influence déterminante** sur la stratégie de la cible. Le droit français reprend cette logique à l'article L. 430-1 du code de commerce et précise la notion de contrôle. Le III de cet article dispose que « *le contrôle découle des droits, contrats ou autres moyens qui confèrent, seuls ou conjointement et compte tenu des circonstances de fait ou de droit, la possibilité d'exercer une influence déterminante sur l'activité d'une entreprise* ». Une fois le contrôle établi, le chiffre d'affaires de l'entreprise cible doit dépasser les seuils communautaires et nationaux de notification.

296. Selon les lignes directrices de l'Autorité relatives au contrôle des concentrations, une participation minoritaire peut ainsi permettre à un actionnaire d'exercer une influence

[173] Mlex, Microsoft's AI hires resemble 2017 case evading merger veto, Germany's Mundt says, 9 avril 2024.

déterminante si elle est assortie de droits qui excèdent ce qui est normalement consenti à des actionnaires minoritaires afin de protéger leurs intérêts financiers ou si ces droits, examinés selon la méthode du faisceau d'indices, sont de nature à démontrer l'existence d'une influence déterminante[174]. Des droits spéciaux conférant une part déterminante dans les décisions de l'entreprise (comme des droits de veto[175]), des pactes spécifiques d'actionnaires ou la possibilité de nommer certains responsables au sein des organes dirigeants de l'entreprise peuvent ainsi octroyer un contrôle sur l'entreprise cible, au sens des règles sur les concentrations. Exceptionnellement, une entreprise peut disposer d'une influence déterminante sans aucune participation au capital[176].

297. **Au-delà de la question de l'influence déterminante, l'Autorité peut également tenir compte des liens économiques entre les entreprises et des situations de contrôle de fait comme, par exemple, le fait d'être le principal voire le seul actionnaire actif, soit sur son secteur, soit dans des secteurs connexes, alors que les autres actionnaires sont des investisseurs financiers par exemple, ou l'existence de relations commerciales privilégiées comme des contrats commerciaux exclusifs** (paragraphe 48 des lignes directrices de l'Autorité précitées).

298. En l'espèce, au-delà de l'étendue de la participation au capital, une attention particulière pourrait être portée à l'influence particulière de ces grandes entreprises du numérique, ce qui les distingue d'autres profils d'investisseurs tels que les fonds de capital risque ou les organismes publics. Par ailleurs, les accords d'exclusivité portant sur la fourniture de services *cloud* ou sur les voies de commercialisation des produits et services de l'entreprise cible pourraient également être examinés afin de déterminer si l'acquéreur a une influence déterminante sur la stratégie de la cible.

299. Aux termes des I et II de l'article L. 430-8 du code de commerce, l'absence de notification, tout comme la réalisation anticipée de l'opération, sont, chacune, sanctionnées par une amende qui peut aller jusqu'à 5 % du chiffre d'affaires de l'entreprise responsable de la notification. Les entreprises concernées doivent donc être vigilantes en cas de modification du capital ou des accords conclus dans le cadre de ces prises de participation et partenariats.

En dessous des seuils de notification, une prise de participation peut également faire l'objet d'un examen par les autorités de concurrence

300. Le renouvellement de la doctrine appliquée par la Commission aux renvois au titre de l'article 22 du règlement n° 139/2004 apporte une réponse adéquate pour examiner les opérations qui ne sont pas de dimension communautaire et échappent au contrôle des autorités nationales de concurrence en vertu de leur droit national, en dépit des effets dommageables qu'elles peuvent avoir sur la concurrence. La Cour de justice est appelée à se prononcer sur la validité de cette interprétation et de cette application de l'article 22. Dans ses conclusions du 21 mars 2024 relatives à l'affaire Illumina/Grail, l'avocat général Emiliou

---

[174] Voir paragraphes 35 et suivants des lignes directrices de l'Autorité.

[175] Décision n° 11-DCC-10 du 25 janvier 2011 relative à la prise de contrôle exclusif du groupe Parisot par Windhurst Industries et de la prise de contrôle conjoint de Windhurst Industries par le FSI.

[176] À titre d'exemple, dans son avis n° 91-A-09 du 15 octobre 1991, le Conseil de la concurrence a considéré que la détention par Gillette d'obligations convertibles en actions d'Eemland, qui détenait la marque Wilkinson, ainsi que l'existence d'accords permettant à Gillette d'influencer la politique commerciale de la marque Wilkinson, bien que ne conférant aucun droit de vote, conféraient à Gillette une influence déterminante sur Eemland et devaient s'analyser comme une concentration.

a considéré que l'article 22 n'habilite pas la Commission à accueillir de tels renvois[177]. Si la Cour de justice devait suivre ces conclusions, cette question devrait faire l'objet d'une nouvelle réflexion.

301. Il convient par ailleurs de rappeler que, par un arrêt Towercast du 16 mars 2023, la Cour de justice a considéré que le règlement sur les concentrations ne s'opposait pas à ce qu'une opération de concentration, dépourvue de dimension communautaire, se situant sous les seuils juridictionnels de contrôle *ex ante* prévus par le droit national et n'ayant pas fait l'objet d'un renvoi sur le fondement de l'article 22 de ce règlement « *soit analysée par une autorité de concurrence d'un État membre comme étant constitutive d'un abus de position dominante prohibé à l'article 102 TFUE au regard de la structure de la concurrence sur un marché de dimension nationale.* »[178]

Ces participations peuvent également être examinées en marge d'une opération de concentration

302. Dans des cas spécifiques, même si elles ne remplissent pas le critère de l'influence déterminante, ces participations peuvent être examinées dans l'analyse des effets d'une opération de concentration.

303. En 2022, l'Autorité a ainsi considéré pour la première fois qu'une prise de participation minoritaire non contrôlante concomitante à une prise de contrôle exclusif était susceptible de porter atteinte à la concurrence. Dans le cadre du rachat de la société Bio Pôle Antilles par le groupe Inovie[179], cette dernière avait informé l'Autorité de son intention de procéder à l'acquisition d'une participation minoritaire dans le capital de la société Synergiebio, unique concurrent privé de Bio Pôle Antilles en Guadeloupe et à Saint-Martin. L'Autorité a conclu que cette prise de participation ne conférerait pas à Inovie de droits de contrôle lui permettant d'exercer une influence déterminante sur Synergibio. Elle a toutefois considéré que, compte tenu de son caractère suffisamment certain, ce projet de prise de participation minoritaire non contrôlante devait être pris en compte dans l'analyse des effets de l'opération au niveau local. Dans ces conditions, et afin d'obtenir l'autorisation de concentration de sa première opération avec Bio Pôle, Inovie s'est engagée à renoncer à toute prise de participation dans le capital de la société Synergibio, pour une durée de dix ans.

304. Par ailleurs, la prise de participation minoritaire examinée peut aussi être préexistante à l'opération de concentration. Ainsi, dans l'affaire Carrefour/Promodes[180], la Commission avait relevé que Carrefour détenait une participation de 42 % au capital de la société GMB qui contrôlait la société Cora. Afin de répondre aux préoccupations concurrentielles de la Commission sur la capacité de Cora et Casino à assurer un contrepoids à Carrefour/Promodes, la société Carrefour s'était ainsi engagée à céder sa participation dans la société GMB.

---

[177] Conclusions de l'avocat général dans les affaires jointes C-611/22 P | Illumina/Commission et C-625/22 P | Grail/Commission et Illumina, paragraphe 265.

[178] Arrêt de la Cour de justice du 16 mars 2023, Towercast, C-449/21, paragraphe 53.

[179] Voir la décision de l'Autorité n° 22-DCC-35 du 27 avril 2022 relative à la prise de contrôle exclusif de la société Bio Pôle Antilles par le groupe Inovie.

[180] Affaire n° COMP/M. 1684 - Carrefour/Promodes.

Un manque de transparence sur ces participations et partenariats

305. Malgré les nombreux risques concurrentiels soulignés *supra*, **les autorités ne disposent pas toujours d'informations** susceptibles de déterminer si ces accords peuvent nuire à la concurrence et donc aux consommateurs.

306. Dès 2014[181], la Commission s'était prononcée en faveur d'un système d'information obligatoire dans les cas où l'opération créerait un « *lien significatif d'un point de vue concurrentiel* ». Afin de garantir la sécurité juridique aux parties, seules les opérations remplissant les critères cumulatifs suivants étaient concernées : « *acquisition d'une participation minoritaire dans un concurrent ou une entreprise liée verticalement* (…) *et le lien concurrentiel est considéré comme significatif lorsque la participation minoritaire est (1) d'environ 20 % ou (2) comprise entre 5 % et environ 20 % mais assortie de facteurs supplémentaires tels que des droits assurant à l'acquéreur une minorité de blocage de fait, un siège au sein du conseil d'administration ou un accès à des informations commercialement sensibles de la cible* »[182]. Cette initiative avait finalement été écartée par la mandature suivante.

307. Or, dans le secteur de l'IA générative, certains acquéreurs minoritaires peuvent jouer un rôle plus décisif que ne le laisse entrevoir l'intitulé de certains partenariats, comme l'a montré l'implication de Microsoft dans les changements de gouvernance de la société OpenAI au mois de novembre 2023, malgré l'absence de pouvoir décisionnel officiel au sein de la société. Les récentes déclarations du président-directeur général de Microsoft semblent confirmer que ce partenariat donne des droits importants à Microsoft.[183] Ces éléments ont conduit le Président du Bundeskartellamt, à se demander si certains de ces accords ne constituaient pas en réalité des « *concentrations déguisées* »[184].

308. Ces inquiétudes sont partagées par les autorités de concurrence dans le monde, comme le montrent les investigations en cours de la Commission européenne[185] et de la CMA au Royaume-Uni[186] sur les investissements de Microsoft au sein de la société OpenAI, les

---

[181] Commission européenne, Livre blanc vers un contrôle plus efficace des concentrations dans l'UE, 9 juillet 2014.

[182] Livre blanc précité, paragraphe 47.

[183] Discutant de l'investissement de Microsoft au sein d'OpenAI dans un article de presse, M. A…, président-directeur général de Microsoft, a indiqué « *Si OpenAI disparaissait demain, je ne voudrais pas qu'un de nos clients s'en inquiète, très honnêtement, parce que nous avons tous les droits pour poursuivre l'innovation. Non seulement pour servir le produit, mais aussi pour faire ce que nous faisions nous-mêmes en partenariat. Nous avons le personnel, nous avons l'informatique, nous avons les données, nous avons tout. Mais en même temps, je me suis engagé dans le partenariat avec OpenAI et c'est ce que je leur ai dit* (…). *Et aussi, ce n'est pas une question d'autonomie, n'est-ce pas ? Nous sommes là-dedans. Nous sommes en dessous d'eux, au-dessus d'eux, autour d'eux »* (traduction libre).

[184] Selon Andreas Mundt : « *Les partenariats entre les grandes entreprises technologiques et les startups spécialisées dans le développement de l'IA, telles que OpenAI, devraient alerter les autorités de la concurrence sur le fait que les accords de coopération pourraient être des concentrations déguisées* » (traduction libre) (Mlex, Watch out for AI cooperation agreements that are really mergers, Germany's Mundt warns, 21 septembre 2023).

[185] Communiqué de presse de la Commission européenne du 9 janvier 2024 : « la Commission européenne vérifie si l'investissement de Microsoft dans OpenAI est susceptible de faire l'objet d'un examen au regard du règlement de l'UE sur les concentrations ».

[186] Le 8 décembre 2023, la CMA a ouvert une enquête sur le partenariat entre Microsoft et OpenAI sur le fondement de contrôle des concentrations.

enquêtes engagées début 2024 par la FTC aux États-Unis à l'encontre des sociétés Alphabet, Amazon, Anthropic, Microsoft et OpenAI[187] et l'appel à contributions lancé par la CMA en avril 2024 sur ce sujet[188]. La coopération entre Microsoft et OpenAI a également été examinée par le Bundeskartellamt en 2023. Si l'autorité allemande a conclu que la coopération en question ne relevait pas du contrôle des concentrations national compte tenu de l'absence d'un lien suffisant entre l'opération et le territoire allemand, elle a néanmoins confirmé que l'influence de Microsoft sur la société OpenAI constituait une concentration au sens de la loi.

c) Ces participations peuvent être appréhendées par le droit des pratiques anticoncurrentielles

309. Dans l'hypothèse où ces participations minoritaires n'entraîneraient pas de contrôle de l'acquéreur sur l'entreprise cible (et ne rempliraient donc pas les critères du droit des concentrations), ces participations peuvent être appréhendées *ex-post* sous l'angle du droit des pratiques anticoncurrentielles, sur le fondement du droit des ententes ou de l'abus de position dominante notamment.

310. Cette possibilité a été confirmée par la Cour de justice dans une affaire ancienne. La Cour de justice s'était ainsi interrogée, dans l'affaire Philip Morris[189], sur le point de savoir si, et le cas échéant, dans quelles conditions, la prise d'une participation minoritaire dans le capital d'une entreprise concurrente pouvait constituer une violation des articles 85 et 86 du Traité (devenus articles 101 et 102 TFUE). Après avoir rappelé que le fait, pour une entreprise, de prendre une participation dans le capital d'une entreprise concurrente ne constitue pas en soi un comportement restrictif de concurrence sur le fondement du droit des ententes, la Cour de justice avait indiqué « [qu']*une telle prise de participation peut néanmoins constituer un moyen apte à influer sur le comportement commercial des entreprises en cause, de manière à restreindre ou à fausser le jeu de la concurrence sur le marché où ces deux entreprises déploient leurs activités commerciales* » (paragraphe 37). Selon la Cour de justice, tel serait notamment le cas « *si l'entreprise qui investit obtient un contrôle de droit ou de fait sur le comportement commercial de l'autre entreprise ou si l'accord prévoit une coopération commerciale entre les deux entreprises* […] *Tel peut également être le cas si l'accord réserve à l'entreprise qui investit la possibilité de renforcer , à un stade ultérieur, sa position en prenant le contrôle effectif de l'autre entreprise* » (paragraphes 38-39). La Cour de justice avait ensuite indiqué la nécessité de faire preuve d'une vigilance particulière, en examinant notamment « *si un accord qui, à première vue, ne prévoit qu'un investissement passif dans une entreprise concurrente, ne vise pas, en réalité, à prendre le contrôle de cette entreprise, le cas échéant à un stade ultérieur, ou à instaurer une coopération entre les entreprises en vue d'un partage du marché* » (paragraphe 45). Elle avait finalement rejeté l'ensemble du recours.

311. Ces prises de participation pourraient donc potentiellement être examinées sous l'angle de l'abus de position dominante, éventuellement collective.

---

[187] FTC Launches Inquiry into Generative AI Investments and Partnerships, 25 janvier 2024.

[188] CMA seeks views on AI partnerships and other arrangements, 24 avril 2024.

[189] Cour de Justice, 17 novembre 1987, British-American Tobacco Company Ltd et R.J. Reynolds Industries Inc. c. Commission, affaires jointes 142 et 156/84, paragraphe 37 et suivants.

312. Pour démontrer l'existence d'une position dominante collective, il convient d'établir que les entreprises « *ont, ensemble, notamment en raison des facteurs de corrélation existant entre elles, le pouvoir d'adopter une même ligne d'action sur le marché et d'agir dans une mesure appréciable indépendamment des autres concurrents, de leur clientèle et, finalement, des consommateurs* »[190], ce qui peut ressortir de l'examen des liens ou facteurs de corrélation juridiques existant entre les entreprises ou de l'examen de la structure du marché selon les critères dégagés par le Tribunal dans l'arrêt Airtours[191]. Ainsi, l'existence de liens structurels entre des entreprises tels que des liens en capital ou des accords formalisés entre elles, d'une part, et l'adoption d'une ligne commune d'action sur le marché, d'autre part, suffisent à démontrer l'existence d'une position de dominance collective. Les prises de participations de plusieurs entreprises au sein de la même cible pourraient ainsi être examinées sur ce fondement. L'Autorité rappelle à cet égard qu'une position dominante collective n'est pas répréhensible en soi, seul un abus de cette position l'étant.

313. Les accords entre entreprises pourraient également relever du droit des ententes, par exemple si ces accords ont pour objectif une répartition des marchés ou favorisent la transparence du marché.

314. L'Autorité a ainsi récemment mis en application la jurisprudence Towercast précitée en examinant si une opération de concentration se situant sous les seuils de contrôle était constitutive d'une pratique anticoncurrentielle contraire au TFUE, en l'espèce d'une entente contraire à l'article 101[192] (voir encadré ci-dessous).

Décision n° 24-D-05 du 2 mai 2024 relative à des pratiques mises en œuvre dans le secteur de l'équarrissage

Dans leur notification de griefs, les services d'instruction reprochaient à Akiolis, Saria et Verdannet l'élaboration et la mise en œuvre d'une entente de répartition géographique de marché, finalement réalisée à travers des cessions croisées de fonds de commerce.

Dans cette affaire, l'Autorité a notamment analysé si les opérations de concentration, qui n'avaient pas fait l'objet de notification *ex ante* au titre du contrôle européen ou national des concentrations, étaient susceptibles, à elles seules, de constituer une entente anticoncurrentielle contraire aux articles 101 TFUE et L. 420-1 du code de commerce. Les mises en cause estimaient que la jurisprudence Towercast ne portait que sur l'applicabilité de l'article 102 TFUE et n'était donc pas transposable à l'article 101 TFUE. Elles soutenaient également que l'application du droit des ententes à une concentration nécessitait d'identifier une pratique anticoncurrentielle détachable de l'opération de concentration.

L'Autorité a considéré que, « *conformément à la jurisprudence de la Cour de justice, une opération de concentration qui est « dépourvue de dimension communautaire, au sens de l'article 1er de ce règlement* [concentrations]*, située en dessous des seuils de contrôle ex ante obligatoire prévus par le droit national et n'ayant pas donné lieu à un renvoi à la Commission en application de l'article 22 dudit règlement » est susceptible de faire l'objet d'un contrôle a posteriori sur le fondement des articles 101 TFUE et L. 420-1 du code de*

[190] Voir, par exemple, la décision de l'Autorité n° 20-D-11 du 9 septembre 2020 relative à des pratiques mises en œuvre dans le secteur du traitement de la dégénérescence maculaire liée à l'âge (DMLA) et décision n° 12-D-06 du 26 janvier 2012 relative à des pratiques mises en œuvre dans le secteur des agrégats et des marchés aval à Saint-Pierre-et-Miquelon.

[191] TPICE, T-342/99, Airtours c/Commission, 6 juin 2002, point 62.

[192] Décision n° 24-D-05 du 02 mai 2024 relative à des pratiques mises en œuvre dans le secteur de l'équarrissage.

commerce ». C'est la première fois que l'Autorité examine, sous l'angle du droit des ententes, des opérations de concentration situées sous les seuils nationaux de notification. Un non-lieu a finalement été prononcé.

3. LES RISQUES DE COLLUSION ENTRE ENTREPRISES DU SECTEUR

315. L'utilisation de l'IA générative pourrait potentiellement avoir des conséquences sur la potentielle mise en œuvre de pratiques concertées[193].

316. La quasi-totalité des parties prenantes interrogées lors de la consultation publique de l'Autorité n'ont pas fait état d'inquiétudes spécifiques sur cette question. Par ailleurs, la majorité des préoccupations concernent l'aval de la chaîne de valeur si bien qu'elles se situent en dehors du champ d'étude de l'Autorité dans le cadre de cet avis. Il convient néanmoins de citer l'exemple de la société Samsung, qui a récemment interdit à ses employés d'utiliser des outils d'IA générative tels que ChatGPT après avoir découvert que des membres du personnel avaient téléchargé du code sensible sur la plateforme, ce qui pouvait conduire à la divulgation de ces informations à d'autres utilisateurs[194]. Cela justifie donc une attention des autorités de concurrence sur ces questions.

317. L'étude conjointe de l'Autorité et du Bundeskartellamt sur le thème « *algorithmes et concurrence* », à laquelle le présent avis renvoie[195], aborde de manière spécifique les enjeux de collusion entre algorithmes et le cadre juridique susceptible de s'appliquer à ces questions. Or, pour mémoire, la majorité des modèles de langage actuels sont développés en utilisant le même algorithme d'apprentissage profond intitulé *Transformers.* Les conclusions de cette étude peuvent donc être étendues aux algorithmes d'IA générative et plusieurs situations pourraient entraîner des risques concurrentiels, notamment :

- les algorithmes d'IA générative peuvent être un moyen de soutenir ou de faciliter des pratiques anticoncurrentielles préexistantes (comme une entente) ;
- la collusion peut être fondée sur un algorithme entre concurrents impliquant un tiers « *hub and spoke* ». Cette situation fait référence au cas dans lequel un tiers comme un consultant externe ou un développeur de logiciels fournit aux concurrents le même algorithme ou des algorithmes coordonnés, sans qu'il y ait de communication directe entre les différents concurrents ;
- la collusion peut être induite par l'utilisation parallèle d'algorithmes individuels distincts ou par le recours à des algorithmes d'apprentissage automatique. Dans cette dernière situation, les algorithmes peuvent apprendre par eux-mêmes à converger vers un équilibre collusif.

---

[193] Selon la jurisprudence, les pratiques concertées se distinguent des accords entre entreprises « *dans le dessein d'appréhender sous les interdictions de cet article* [article 101 du TFUE] *une forme de coordination entre entreprises qui, sans avoir été poussée jusqu'à la réalisation d'une convention proprement dite, substitue sciemment une coopération pratique entre elles aux risques de la concurrence* » **(**Arrêt de la Cour du 14 juillet 1972. - Imperial Chemical Industries Ltd. contre Commission des Communautés européennes. - Affaire 48-69, paragraphe 64**).**

[194] Bloomberg, Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak, 2 mai 2023, cité par Carugati, C. (2023) 'Competition in generative artificial intelligence foundation models', Working Paper 14/2023, Bruegel.

[195] Étude de l'Autorité et du Bundeskartellamt, Algorithmes et concurrence, 6 novembre 2019.

318. Un acteur du secteur résume ainsi les questions, relatives notamment à la mise en œuvre de la responsabilité, susceptibles de se poser : « *Comment évaluer les risques de collusion introduits par l'utilisation croisée de l'IA générative par des entreprises d'un même marché, notamment en termes de transparence et d'échange d'informations sensibles ? Comment définir la chaîne de responsabilité associée en matière de collusion lorsque la prise de décision de l'humain est progressivement remplacée par une IA ?* ».

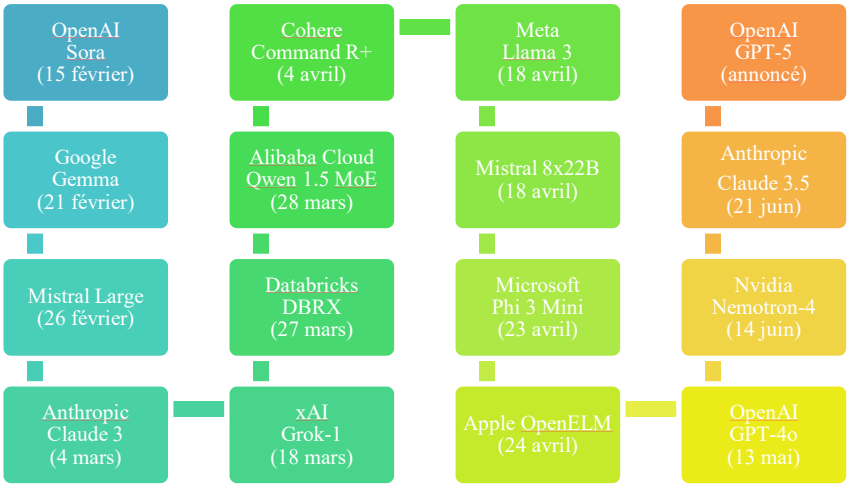
III. Perspectives et recommandations

A. LE SECTEUR DE L'IA GENERATIVE EST LOIN D'AVOIR ATTEINT SON POTENTIEL

319. Le secteur de l'IA générative est en pleine croissance. Moins de deux ans après le lancement de ChatGPT, de nombreux acteurs établis ont investi dans ce domaine et une multitude de jeunes entreprises sont apparues pour accélérer la recherche et permettre l'utilisation de cette technologie innovante par la majorité des entreprises et des consommateurs. OpenAI a ainsi dépassé, début 2024, la barre des deux milliards de dollars de chiffre d'affaires (1,85 milliard d'euros), dont la majeure partie a été obtenue depuis décembre 2022.

320. De nombreux modèles d'IA génératives ont été annoncés au cours du premier semestre 2024 (l'instruction de cet avis a ainsi vu l'annonce de nombreux modèles, Mistral Large par Mistral AI, Claude 3 par Anthropic et Llama 3 par Meta), venant souligner le dynamisme et la volatilité de ce marché). **La course à l'innovation** et au développement de nouveaux modèles d'IA générative devrait se poursuivre sur au moins deux axes : la taille des modèles (plus un modèle est grand, plus il est performant) et l'optimisation des modèles à taille constante.

Figure n° 9 : principaux modèles publiés entre le 8 février 2024 et le 24 juin 2024



Source : Autorité de la concurrence

321. Bien que les plus grands modèles semblent avoir les meilleures performances, ils ne sont pas adaptés à tous les cas d'usage. De nombreux acteurs préfèrent donc des modèles plus petits

et moins coûteux. La question de la portabilité des modèles d'IA générative sur des supports moins puissants et sans processeurs graphiques se pose également et devrait entraîner une concurrence entre les acteurs. Apple ou Samsung ont déjà annoncé l'intégration future d'outils d'IA générative sur leurs téléphones portables.

322. La taille des modèles influe également sur **l'impact environnemental de l'IA** générative. Bien qu'il soit difficile d'estimer le coût supplémentaire induit par l'usage de l'IA générative, il n'en reste pas moins certain que cette technologie va entraîner, au moins à court terme, une augmentation de la consommation énergétique. Des estimations indiquent par exemple qu'un moteur de recherche utilisant l'IA serait dix fois plus consommateur d'énergie qu'un moteur de recherche sans IA. La réduction de l'impact énergétique est donc un autre axe d'innovation possible, sachant que certains acteurs du secteur ont annoncé vouloir atteindre une empreinte environnementale neutre en carbone en 2030[196], ce qui va les inciter à innover pour réduire leurs coûts. Cette innovation peut être de nature technologique, aux différentes étapes de la création des modèles, de l'entraînement à l'inférence, porter sur les modes d'utilisation de l'IA générative par les utilisateurs finaux, par exemple pour réduire la fréquence des requêtes, ou porter sur l'accès à l'énergie des acteurs du secteur. Comme évoqué ci-avant (paragraphe 221), il conviendra de veiller à ce que la montée en puissance de l'enjeu énergétique ne crée pas de nouvelles barrières à l'entrée pour des raisons de coût ou en cas d'intégration verticale de certains acteurs.

323. Le développement de nouvelles formes de calcul tel que l'informatique quantique ou l'informatique en périphérie (en anglais « *edge computing* ») pourrait permettre d'accélérer le développement du secteur de l'IA générative ainsi que son adoption par les utilisateurs, mais pourrait également renforcer les risques concurrentiels identifiés dans cet avis si ces nouvelles formes de calcul sont contrôlées par les grands acteurs, directement ou au travers de partenariats.

324. L'Autorité observe également une tendance à la « plateformisation » dans le secteur de l'IA générative. Ainsi, OpenAI propose l'ajout de *plugins* sur ChatGPT, via le GPTStore, une place de marché sur laquelle tous les développeurs et entreprises peuvent proposer leur *plugin* spécifique. D'un autre côté, HuggingFace s'impose comme la place de marché de référence pour la publication de données et de modèles d'IA générative ouverts. Les fournisseurs de services *cloud* et notamment les *hyperscalers* mettent à disposition de leurs clients des places de marché MaaS permettant de faciliter l'accès aux principaux modèles d'IA générative. Ces places de marché *cloud* apparaissent comme des points de passage obligatoire pour les développeurs de modèles qui souhaitent atteindre les consommateurs ou les entreprises utilisatrices d'IA.

325. Un des principaux enjeux pour le bon développement de la concurrence dans le secteur de l'IA générative réside dans la diffusion de ressources ouvertes. Or, comme évoqué supra, l'*open source* dans le secteur de l'IA générative représente des situations différentes, chaque acteur ayant ses propres caractéristiques et besoins. Si le secteur bénéficiait de critères plus précis pour qualifier l'ouverture d'un modèle, cela permettrait aux acteurs qui le souhaitent de faire valoir cette qualité comme un avantage concurrentiel.

326. Les entreprises du secteur, et notamment les *start-ups*, doivent également s'installer de manière pérenne ce qui nécessite un modèle de rémunération permettant aux acteurs de rentabiliser leurs investissements initiaux très importants, tout en continuant à se développer.

---

[196] Les Échos, Comment l'IA plombe le bilan carbone de Microsoft, 16 mai 2024.

Cette problématique est d'autant plus pertinente pour les entreprises tournées vers *l'open source*, comme le souligne le PEReN : « *les logiciels open source sont par essence diffusables librement, donc a priori gratuitement, ce qui rend complexes les modèles économiques reposant sur la seule vente du logiciel* »[197].

327. Les analyses et risques identifiés par cet avis concernent l'amont de la chaîne de valeur de l'IA générative à modèle technologique constant, c'est-à-dire fondé sur les grands modèles de langage. Une nouvelle étape technologique qui permettrait de dépasser ces modèles nécessiterait de renouveler l'analyse concurrentielle.

B. RECOMMANDATIONS

328. Les conséquences potentielles de l'IA générative pour la productivité des services publics et des entreprises, pour l'organisation du travail et au-delà, pour des priorités essentielles de politique publique telles que l'éducation ou la santé, plaident pour un cadre réglementaire qui favorise l'adoption de l'IA par les ménages et les entreprises[198] et permette une diversité de cas d'usage et de modèles, tout en assurant un contrôle vigilant des risques dans des domaines comme la sécurité nationale, le respect de la vie privée et de la propriété intellectuelle et artistique. Au vu des développements qui précèdent, l'Autorité estime que les recommandations qui suivent seraient de nature à favoriser la dynamique concurrentielle du secteur.

329. Ces recommandations, pour la plupart, ne nécessitent pas d'initiative législative au niveau français ou européen.

1. DES PROPOSITIONS, A DROIT CONSTANT, VISANT A RENDRE PLUS EFFICACE LE CADRE REGLEMENTAIRE APPLICABLE AU SECTEUR

330. Comme exposé ci-avant (voir paragraphes 104 et suivants), de nombreuses réglementations se sont succédé au niveau mondial, européen et français ces dernières années. Toutes ne sont pas encore mises en application[199] et d'autres sont encore en projet[200]. Compte tenu de la rapidité avec laquelle le secteur évolue, ces réglementations devront être pleinement mises

---

[197] PEReN, Éclairage sur…n°7 – Open source et IA : des synergies à repenser ?, 3 avril 2024.

[198] Rapport de la Commission IA précité.

[199] Par exemple, la plupart des dispositions du règlement sur l'IA (en attente de publication au *Journal officiel*) ne seront applicables qu'à compter de l'année 2026, à l'exception de certaines dispositions particulières. Par ailleurs, la suppression des frais de changement de fournisseur de services *cloud* (appelés « *egress fees* »), prévue dans le cadre du règlement sur les données, sera effective à compter du 12 janvier 2027.

[200] Par exemple, la Commission a publié une proposition relative à l'adaptation des règles en matière de responsabilité civile extracontractuelle au domaine de l'IA le 28 septembre 2022, toujours en discussion. L'objectif de la proposition est de « *garanti*[r] *aux victimes de dommages causés par l'IA une protection équivalente à celle des victimes de dommages causés par les produits de manière générale. Elle réduit également l'insécurité juridique qui plane sur l'éventuelle exposition de la responsabilité des entreprises qui développent ou utilisent l'IA et évite l'apparition, dans les règles nationales en matière de responsabilité civile, d'adaptations fragmentées spécifiques à l'IA* ».

en œuvre et leur impact évalué afin d'éviter des effets négatifs sur l'innovation et la concurrence.

331. À droit constant, des améliorations pourraient cependant être envisagées. En effet, certaines réglementations sont apparues avant l'émergence de l'IA générative si bien que ses effets ne peuvent être totalement appréhendés par la législation.

332. C'est le cas du DMA par exemple. En effet, les obligations du DMA ne peuvent s'appliquer qu'aux services de plateformes essentiels des contrôleurs d'accès qui sont visés à l'article 2 du DMA, parmi lesquels ne figurent pas les MaaS.

333. Toutefois, il ne peut être exclu que certains services MaaS, en raison des caractéristiques décrites aux paragraphes 140 et suivants, puissent rentrer dans une des catégories énumérées à l'article 2, notamment dans celle des «*services d'informatique en nuage*»[201], sous réserve de l'interprétation de la Commission.

334. Dans cette hypothèse, les entreprises fournissant des services MaaS pourraient être désignées comme contrôleur d'accès pour ce qui concerne ces services, par le biais d'une désignation au titre du paragraphe 4 de l'article 3 du DMA en cas de franchissement des seuils quantitatifs ou, si ces seuils ne sont pas franchis, dans le cadre d'une désignation sur le fondement de critères qualitatifs à l'issue d'une enquête de marché prévue au paragraphe 8 de l'article 3 et à l'article 17 du même règlement. Au cours de cette analyse, il conviendra néanmoins de vérifier que les services MaaS servent réellement d'interface entre les entreprises utilisatrices de leurs services et les utilisateurs finaux[202].

Proposition n° 1 : la Commission devrait porter une attention particulière au développement des services MaaS pour évaluer la possibilité de désigner les entreprises fournissant de tels services en tant que contrôleurs d'accès.

335. Comme indiqué par l'Autorité dans son avis n° 23-A-08 précité du 29 juin 2023, les dispositions du Data Act, particulièrement en matière d'interopérabilité des données, auront un effet globalement positif sur la concurrence dans le secteur du *cloud*. Néanmoins, d'autres risques concurrentiels subsistent comme ceux liés aux avoirs d'informatique en nuage (ou « crédits *cloud* »), qui sont appréhendés en France par les II à V de l'article L. 442-12 du code de commerce (créés par l'article 26 de la loi SREN), mais pas au niveau européen. Il y a lieu de rappeler que le II de l'article L. 442-12 du code de commerce limite la durée des avoirs d'informatique en nuage et interdit qu'ils soient assortis de conditions d'exclusivité, sous peine d'une amende prononcée par le ministre de l'économie. Le IV du même article interdit de subordonner la vente d'un produit ou d'un service à la conclusion concomitante d'un contrat de fourniture de services d'informatique en nuage, dès lors que celle-ci constitue une pratique commerciale déloyale, cette interdiction étant également assortie d'une amende administrative. Enfin, selon le V du même article, « [l]*'Autorité de la concurrence peut, soit d'office, soit à la demande du ministre chargé du numérique ou de toute personne morale concernée, se saisir de tout signalement effectué vis-à-vis des pratiques d'autopréférence. Elle les sanctionne ou adopte toute mesure nécessaire, le cas échéant, sur le fondement des titres II et VI du présent livre. L'Autorité de la concurrence*

[201] À ce jour, aucune entreprise fournissant des services d'informatique en nuage n'a été désignée par la Commission en tant que contrôleur d'accès pour ce qui concerne ces services.

[202] S'il était au contraire considéré que les services MaaS et autres services en amont de la chaîne de valeur de l'IA n'étaient pas visés par l'article 2, la Commission pourrait lancer une enquête de marché sur le fondement de l'article 19, afin d'élargir la liste de l'article 2, ce qui nécessiterait alors une révision du règlement.

dispose, pour la mise en œuvre de ces dispositions, des pouvoirs qui lui sont reconnus au titre V du présent livre ».

336. L'Autorité recommande que, dans l'application de ces nouvelles dispositions, une attention particulière soit accordée par la DGCCRF aux pratiques concernant plus spécifiquement le domaine de l'IA. L'Autorité, pour sa part, s'engage à faire preuve de vigilance, s'agissant des pratiques d'autopréférence, visées au titre du V ci-dessus.

> **Proposition n° 2** : dans la mise en œuvre des dispositions de la loi SREN sur les avoirs d'informatique en nuage, la DGCCRF pourrait accorder une attention particulière à l'utilisation de ces avoirs dans le domaine de l'IA.

337. Il conviendra enfin d'être vigilant sur les effets du règlement sur l'IA (**AI Act**, voir les paragraphes 104 et suivants) sur la dynamique concurrentielle du secteur. En effet, celui-ci soumet les fournisseurs de systèmes d'IA générative à de multiples obligations réglementaires, qui requièrent la mobilisation d'importants moyens financiers, humains et techniques susceptibles de freiner l'émergence ou l'expansion d'opérateurs de taille plus modeste. Il conviendrait par exemple de s'assurer que les grands acteurs n'utilisent pas certaines dispositions du règlement sur l'IA pour assoir davantage leur pouvoir de marché. Par ailleurs, il conviendra d'être vigilant sur les exemptions du règlement liées aux modèles *open source*, dans la mesure où ceux-ci peuvent faire référence à des situations très variables (voir paragraphes 179 et suivants).

> **Proposition n° 3** : le futur Bureau de l'IA, mis en place à l'article 64 du règlement sur l'IA, et l'autorité nationale compétente en France, qui sera désignée en application de l'article 70 du règlement sur l'IA, devront s'assurer d'une part que la mise en œuvre du règlement ne freine pas l'émergence ou l'expansion d'opérateurs de taille plus modeste, et d'autre part que les plus grands acteurs du secteur ne détournent pas le texte à leur avantage.

338. Enfin, une coordination internationale est nécessaire, compte tenu des initiatives en cours en France, en Europe et dans le reste du monde, afin de s'assurer qu'elles ne créent pas de distorsions et de surcoûts pour les entreprises. Le sommet sur l'IA qu'accueillera la France au mois de février 2025 constituera l'occasion de renforcer la gouvernance mondiale de l'IA.

2. MOBILISER LES OUTILS DU DROIT DE LA CONCURRENCE ET DU DROIT DES PRATIQUES RESTRICTIVES DE CONCURRENCE

339. Face aux risques identifiés dans les développements précédents, la mobilisation des outils de concurrence aura un rôle essentiel pour prévenir l'émergence ou la consolidation de positions dominantes ou des ententes qui affecteraient la dynamique concurrentielle du secteur.

340. En effet, il est important que les autorités de concurrence restent engagées et attentives à ce qu'aucun acteur ne soit en mesure de verrouiller l'accès à des intrants essentiels pour le développement de l'IA générative tout en donnant aux marchés la possibilité et l'incitation de continuer à se développer et à innover. En ce sens, les services de l'Autorité ont déjà mis en œuvre plusieurs initiatives, en instruisant l'avis sur le fonctionnement concurrentiel du secteur du *cloud* ou bien encore en analysant, dans l'affaire « Google droits voisins », le rôle des données des éditeurs de presse au stade du « *grounding* ».

341. L'Autorité restera vigilante sur les évolutions du secteur, notamment la situation des cartes graphiques (à la suite de l'opération de visite et saisie conduite en 2023 dans le secteur), les accords des géants du numérique avec des fournisseurs de contenus (y compris les données sensibles comme les données financières) et les risques concurrentiels liés au déploiement des modèles sur des marchés distincts.

342. L'Autorité dispose d'outils qui lui permettent d'agir rapidement et de manière efficace. Les mêmes instruments sont disponibles **au niveau européen** même si les conditions d'application peuvent être différentes[203].

343. Face à une situation nécessitant une intervention dans l'urgence, l'Autorité peut être amenée à prononcer des **mesures conservatoires** en attendant de décider du fond du dossier, en cas d'atteinte grave et immédiate aux intérêts d'un secteur économique ou d'une entreprise. Les pratiques d'une entreprise en position dominante peuvent donner lieu au prononcé de mesures conservatoires car pendant la phase de structuration des marchés nouveaux, il est nécessaire d'éviter que l'opérateur prenne une avance technologique trop importante sur ses concurrents[204] ou que la structure oligopolistique du marché[205] se renforce. Cette procédure a permis de rendre des décisions dans des délais restreints au cours des dernières années, souvent inférieurs à six mois. Dans l'attente de la pleine effectivité des derniers textes réglementaires, le recours à des mesures conservatoires permettant de préserver les conditions de concurrence sur ce secteur peut apparaître particulièrement pertinent.

344. L'Autorité peut aussi décider de **transiger** avec la ou les entreprises mises en cause afin d'accélérer le délai de traitement d'une affaire. Cette procédure a été notamment mise en œuvre sur le marché des serveurs publicitaires pour éditeurs de sites en ligne et applications mobile[206].

345. L'Autorité peut également imposer une **sanction pécuniaire** et/ou des **injonctions comportementales ou structurelles** visant à faire cesser les pratiques en cause ou à contraindre l'entreprise concernée à modifier ses comportements. Ces injonctions peuvent prendre différentes formes, allant de l'injonction à négocier de bonne foi comme dans le cadre de l'affaire des « droits voisins » précitée, à la modification des règles de fonctionnement d'une plateforme[207] ou encore à la cessation de toute discrimination[208].

346. Elle peut opter, si l'affaire s'y prête, pour une solution négociée qui consiste à rendre obligatoires des **engagements** structurels et/ou comportementaux proposés par l'entreprise,

---

[203] Au niveau européen, le standard de preuve en matière de mesures conservatoires est plus contraignant, ce qui a entrainé un usage plus restreint des mesures conservatoires par la Commission.

[204] Conseil de la concurrence, décision n° 00-MC-01 du 18 février 2000 relative à une demande de mesures conservatoires présentée par la société 9 Télécom Réseau.

[205] Décision n° 23-MC-01 du 4 mai 2023 relative à une demande de mesures conservatoires de la société Adloox.

[206] Décision n° 21-D-11 du 7 juin 2021 relative à des pratiques mises en œuvre dans le secteur de la publicité sur Internet.

[207] Décision n° 19-D-26 du 19 décembre 2019 relative à des pratiques mises en œuvre dans le secteur de la publicité en ligne liée aux recherches.

[208] Décision n° 14-D-06 du 08 juillet 2014 relative à des pratiques mises en œuvre par la société Cegedim dans le secteur des bases de données d'informations médicales.

lorsque ces derniers permettent de remédier aux préoccupations de concurrence[209]. Cette procédure permet de régler rapidement certaines situations très en amont. Elle évite aussi à l'Autorité la lourdeur d'une instruction contentieuse et lui permet de libérer des ressources pour d'autres affaires. L'Autorité a ainsi accepté les engagements de Google créant un cadre de négociation et de partage des informations nécessaires à une évaluation transparente de la rémunération des droits voisins[210] ou les engagements de Meta afin de mettre un terme à des pratiques susceptibles de soulever des préoccupations de concurrence sur le marché français de la publicité en ligne non liée aux recherches[211].

347. L'Autorité dispose donc déjà d'une boite à outils qui lui permet d'agir efficacement en fonction des pratiques en cause, le cas échéant en utilisant ces instruments isolément, simultanément, ou de manière séquentielle, pour autant que ses ressources soient suffisantes.

348. **Les pratiques restrictives de concurrence**, qui relèvent principalement du ressort de la DGCCRF et des juridictions commerciales, peuvent également constituer une réponse adaptée aux risques observés dans le secteur. En effet, les dispositions relatives à ces pratiques (titre IV du livre IV du code de commerce) ont été appliquées à l'économie numérique et ont permis de condamner certaines pratiques contractuelles des plateformes numériques au cours des dernières années (voir paragraphes 612 et suivants de l'avis n° 23-A-08 de l'Autorité précité). Pour la mise en œuvre de ces dispositions également, les moyens nécessaires doivent être mis à la disposition de l'administration et des juridictions.

Proposition n° 4 : les autorités chargées de la régulation concurrentielle des marchés devront demeurer vigilantes dans le secteur de l'IA générative et mobiliser, si nécessaire, l'ensemble des outils à leur disposition pour agir de manière rapide et efficace.

3. ASSURER UN ACCES A LA PUISSANCE DE CALCUL POUR ENCOURAGER L'INNOVATION

349. L'accès à la puissance de calcul est indispensable pour permettre le développement de la recherche et l'émergence de nouvelles entreprises dans le secteur de l'IA générative.

350. De nombreuses initiatives ont lieu dans le marché des composants informatiques pour le calcul, et notamment en Europe avec SiPearl, une *start-up* ayant notamment pour objectif d'équiper les supercalculateurs européens avec ses CPU. Bien que cette initiative ne permette pas de fournir une solution directe pour le secteur spécifique de l'IA générative, qui a besoin

[209] D'après l'étude de l'Autorité sur les engagements comportementaux du 17 janvier 2020, « *cette dichotomie repose, traditionnellement, sur les effets produits par les engagements, les premiers modifiant directement et par eux-mêmes la structure des marchés (le nombre, la qualité ou le périmètre des opérateurs actifs sur un marché) quand les seconds se limitent à réguler les comportements de leurs souscripteurs. Ainsi, lorsque les engagements imposent une cession d'actif(s) ou une rupture de liens contractuels afin de permettre le maintien d'une offre indépendante sur le marché, ils sont considérés comme « structurels ». S'ils contraignent le comportement commercial ou stratégique d'une entreprise, ils sont alors qualifiés de « comportementaux* » (page 27).

[210] Décision n° 22-D-13 du 21 juin 2022 relative à des pratiques mises en œuvre par Google dans le secteur de la presse.

[211] Décision n° 22-D-12 du 16 juin 2022 relative à des pratiques mises en œuvre dans le secteur de la publicité sur Internet.

de puces plus spécialisées que les CPU, elle démontre la capacité d'initiatives européennes à répondre à des besoins industriels.

351. Les parties prenantes consultées considèrent que l'accès à la puissance de calcul est un sujet concurrentiel, en termes de délai (dans un marché très dynamique) et de coût, mais qui devrait perdre de son acuité avec l'émergence d'une concurrence accrue sur le marché des composants informatiques pour l'IA.

352. L'Autorité souligne l'importance de la disponibilité de ressources de calcul publiques, via des supercalculateurs, accessibles gratuitement pour les acteurs en contrepartie d'une contribution à la science ouverte. Plusieurs acteurs publics ont d'ailleurs appelé récemment au renforcement de la puissance de calcul européenne[212].

Proposition n° 5 : poursuivre les investissements dans le développement des supercalculateurs au niveau européen, pour permettre au plus grand nombre d'acteurs d'accéder à la puissance de calcul.

353. Au vu du dynamisme du marché et de l'exigence de puces de dernière génération pour l'entraînement des modèles, ces supercalculateurs nécessitent des investissements permanents pour rester une alternative viable pour l'entraînement de modèles d'IA générative et/ou le réglage fin de modèles pré-entrainés. Ces investissements pourraient être financés, au moins en partie, par des acteurs privés qui utilisent les ressources de calcul.

Proposition n° 6 : le gouvernement et/ou les sociétés assurant la gestion des supercalculateurs pourraient engager une réflexion afin de proposer un cadre ouvert et non-discriminatoire permettant à des acteurs privés d'utiliser les ressources des supercalculateurs publics contre rémunération, tout en conservant la priorité aux recherches notamment académiques.

354. Au vu de l'augmentation de la demande pour les ressources des supercalculateurs, il y a lieu d'insister sur le caractère ouvert des modèles d'IA entrainés sur les supercalculateurs publics, et d'accorder la priorité aux projets ayant la plus grande stratégie d'ouverture, notamment au regard des critères de l'*open source* indiqués dans le cadre de l'AI Act[213].

Proposition n° 7 : en lien notamment avec l'AI Act, fixer des critères d'ouverture des modèles d'IA génératives entrainés sur des supercalculateurs publics.

[212] Voir le rapport de la Commission de l'IA précité, le discours du 22 mai 2024 d'Emmanuel Macron en marge de l'ouverture du salon Vivatech, ou le post de Thierry Breton, Commissaire européen au marché intérieur, sur l'IA en Europe en septembre 2023.

[213] L'article 53(2) du projet d'AI Act à la date du 13 juin 2024 indique à cet effet : « *les obligations énoncées au paragraphe 1, points a) et b), ne s'appliquent pas aux fournisseurs de modèles d'IA qui sont diffusés sous une licence libre et gratuite permettant l'utilisation, la modification et la distribution du modèle, et dont les paramètres, y compris les poids, l'architecture du modèle et les informations sur l'utilisation du modèle sont mis à la disposition du public.* (…) » (traduction libre).

4. SUR LE MARCHE DES DONNEES, ASSURER UN EQUILIBRE ENTRE JUSTE REMUNERATION DES AYANTS DROIT ET ACCES DES DEVELOPPEURS DE MODELES AUX DONNEES NECESSAIRES POUR INNOVER, EN PRENANT EN COMPTE LA DIVERSITE DES CAS D'USAGE DES DONNEES

355. L'Autorité a pu constater les inquiétudes tant des développeurs de modèles d'IA quant à l'accès aux données nécessaires pour entrainer et utiliser leurs modèles que des ayants droit, comme les éditeurs et agences de presse, quant au respect de leurs droits. Il importe de respecter un équilibre entre ces deux considérations pour assurer la soutenabilité d'un modèle fondé, à technologie constante, sur l'utilisation d'une quantité toujours plus grande de données.

356. Il y a lieu de noter que les usages des données diffèrent selon le stade auquel elles sont utilisées au sein de la chaîne de valeur, qu'il s'agisse en amont de l'entraînement du modèle ou à l'aval avec des techniques telle que le « *grounding* » permettant d'améliorer le contenu créé par le modèle à l'aide de connaissance externe, comme les articles de presse. Ainsi, les accords entre ayants droit et développeurs devraient refléter l'importance relative de ces données pour les développeurs suivant les cas d'usage, et préciser le cadre dans lequel ces données peuvent être utilisées.

357. Par exemple, les données des éditeurs de presse sont essentielles pour les acteurs qui mettent en place des agents conversationnels à destination des utilisateurs avec du « *grounding* » et ont donc une forte valeur économique dans ce cas précis. Toutefois, pour l'entraînement de modèles d'IA générative, leur importance marginale est relative dans le volume considérable des données nécessaires, qui accorde en outre une place importante à la description de faits et aux raisonnements logiques, par exemple issus d'encyclopédies ou d'articles scientifiques.

358. Il ressort dès lors de l'analyse de l'Autorité que la valeur des données semble liée, en l'état actuel de la technologie :

- à l'étape de l'entrainement, au volume et à la valeur descriptive de données ;
- à l'étape du réglage fin, à la spécificité des données. Des données sectorielles auront ainsi plus de valeur pour un modèle voulant se spécialiser sur le secteur en question (par exemple, des données de santé) ;
- à l'étape de l'inférence, notamment pour le RAG ou le *grounding*, à la pertinence et à l'actualité des données, c'est-à-dire la capacité d'être en mesure d'apporter l'information manquante qu'un modèle n'aura pas intégrée lors de sa phase d'entraînement.

359. Par ailleurs, les coûts de transaction sont un point d'attention qu'il ne faut pas négliger dans le cas des données utilisées pour l'entraînement. En effet, si la plupart du temps, les coûts de transaction sont négligeables par rapport aux prix d'acquisition, ils pourraient s'avérer prohibitifs dans le cas des données d'entraînement si un développeur de modèle devait contracter individuellement avec chaque acteur dont il veut utiliser les données. À cet égard, l'Autorité observe que des propositions émergent de certains analystes, comme la mise en place de licences collectives ou l'octroi d'une « sphère de sécurité » (en anglais « *safe harbor* ») qui protégerait certains fournisseurs de modèles de toute responsabilité juridique, à condition qu'ils respectent certaines normes de transparence et d'éthique[214].

[214] TechCrunch, OpenAI's deals with publishers could spell trouble for rivals, 13 mars 2024.

Proposition n° 8 : les autorités publiques, notamment dans le cadre de la mission confiée par la ministre de la culture au Conseil supérieur de la propriété littéraire et artistique, pourraient inciter les ayants droit à tenir compte de la valeur économique des données selon les cas d'usage (en introduisant par exemple des prix différenciés), et à proposer des offres groupées réduisant les coûts de transaction, ceci afin de garantir les capacités d'innovation des développeurs de modèles

360. L'Autorité estime également que l'ouverture des données participe efficacement à l'animation concurrentielle du secteur en abaissant les barrières à l'entrée et en réduisant l'incertitude vis-à-vis de l'accès aux données (voir *supra*). La sphère publique, lorsque cela est possible, doit jouer un rôle moteur dans le prolongement de l'ouverture des données publiques de l'administration via le portail data.gouv.fr et de l'appel à projet concernant les communs numériques évoqué *supra*. Par exemple, l'INA ou la Bibliothèque nationale de France possèdent des ensembles de données massifs, qui pourraient être mis à disposition de développeurs de modèles, dans des conditions à définir. La sphère privée peut également contribuer, notamment en fournissant des données spécifiques, par exemple sectorielles, qui permettent le réglage fin de modèles.

361. Ces initiatives peuvent également permettre d'assurer une meilleure représentation de la langue et de la culture française (et européenne) parmi les modèles d'IA générative, où l'anglais prédomine actuellement. Ceci permettra d'améliorer leurs performances dans ces langues, tout en permettant une meilleure prise en compte de la diversité culturelle, au bénéfice de l'innovation et des utilisateurs finaux.

Proposition n° 9 : faciliter la mise à disposition des données de la sphère publique et privée pour l'entraînement ou le réglage fin de modèles d'IA générative, et encourager les initiatives publiques ou privées visant à diffuser les données francophones, qu'il s'agisse de textes, d'images ou de vidéos.

5. UNE MEILLEURE TRANSPARENCE SUR LES PRISES DE PARTICIPATIONS DES GEANTS DU NUMERIQUE DANS LES ENTREPRISES INNOVANTES DU SECTEUR PARAIT JUSTIFIEE

362. L'Autorité considère que, dans l'attente de la décision de la Cour de justice sur l'article 22, le cadre juridique actuel permet d'appréhender la plupart des préoccupations de concurrence concernant les accords entre entreprises, qu'il s'agisse du droit du contrôle des concentrations ou du droit des pratiques anticoncurrentielles. Néanmoins, l'Autorité considère qu'à droit constant il convient d'assurer une meilleure transparence des participations minoritaires non contrôlantes dans le secteur.

363. Bien que le DMA ne vise pas directement les services d'IA générative parmi les services de plateforme essentiels, son article 14 a un champ d'application large puisqu'il s'applique à tout projet de concentration lorsque les entités qui fusionnent ou la cible de la concentration fournissent des services de plateforme essentiels ou tout autre service du secteur numérique ou permettant la collecte de données. Conformément à la pratique actuelle en matière de contrôle des concentrations, le document d'information relatif à l'article 14[215] pourrait comprendre une obligation de communication d'informations relatives aux participations

[215] Le modèle relatif à l'article 14 et datant du 27 octobre 2023 est disponible ici.

minoritaires acquises dans le même secteur que la cible. Ainsi, un contrôleur d'accès informant la Commission d'une concentration dans le secteur de l'IA générative informerait également de toute participation minoritaire acquise dans le même secteur que sa cible.

364. Cette proposition serait sans préjudice du contrôle *ex post* de ces prises de participations minoritaires non contrôlantes en application des règles du droit de la concurrence relatives à l'abus de position dominante ou aux ententes.

> **Proposition n° 10** : À l'occasion de l'obligation d'information des concentrations prévue à l'article 14 du DMA, la Commission pourrait également demander, dans le modèle relatif à l'article 14 du règlement sur les marchés numériques, des informations sur les participations minoritaires détenues dans le même secteur d'activité que la cible.

365. Une autre possibilité, plus contraignante, et allant au-delà du seul secteur de l'IA générative, serait de modifier l'article 14 du règlement sur les marchés numériques afin que la Commission soit systématiquement informée de ces prises de participations.

366. Cette obligation d'information pourrait comporter des conditions similaires à celles envisagées par la Commission en 2014 (voir *supra*, paragraphe 306), ce qui permettrait de limiter les obligations d'information aux opérations potentiellement problématiques et d'éviter de faire peser une charge administrative disproportionnée sur les entreprises et la Commission. Elle pourrait également comporter des demandes d'informations complémentaires comme les éventuels accords d'exclusivité entre les parties. Par ailleurs, les informations reçues au titre de l'article 14 pourraient être utiles pour l'application des règles relatives à l'abus de position dominante ou le droit des ententes.

Conclusion

367. L'Autorité a procédé à un examen attentif du secteur de l'IA générative, en se concentrant particulièrement sur l'amont de la chaîne de valeur, susceptible de soulever davantage de risques concurrentiels.

368. Au-delà de l'analyse poussée du secteur et de l'identification des intrants nécessaires au développement de modèles de fondation, l'Autorité s'est prononcée sur des questions particulièrement structurantes pour l'avenir, comme les pratiques susceptibles d'être mises en œuvre sur les marchés du travail, les enjeux de la rémunération des contenus ou l'appréciation concurrentielle des participations minoritaires des grands acteurs au sein de jeunes entreprises innovantes.

369. L'Autorité, qui s'exprime à titre consultatif et non contentieux, ne se prononce pas sur la licéité des pratiques précitées. Néanmoins, les risques concurrentiels exposés dans le présent avis seront suivis avec attention par ses services, notamment en ce qui concerne les pratiques limitant abusivement l'accès à des intrants indispensables, les partenariats conclus par des entreprises numériques déjà dominantes assortis ou pas de rapports d'exclusivité et des pratiques de ventes liées ou groupées susceptibles de consolider durablement le secteur de l'IA générative autour de ces entreprises, sans préjudice des pratiques à l'aval de la chaîne de valeur, qui ne font pas l'objet du présent avis. Cette vigilance est indispensable pour contribuer au développement d'une IA ouverte, respectueuse des droits, dans laquelle les acteurs de taille modeste ont une chance de réussir et où les entreprises et les utilisateurs ont accès à des modèles variés et innovants.

Délibéré sur le rapport oral de Mme Elodie Vandenhende et M. Quentin Deltour, rapporteurs, et l'intervention de M. Yann Guthmann, chef du service de l'économie numérique, par M. Benoît Cœuré, président, Mme Fabienne Siredey-Garnier, Mme Irène Luc et M. Thibaud Vergé, vice-présidents et Mme Valérie Bros, Mme Julie Burguburu, Mme Catherine Prieto et M. Jérôme Pouyet, membres.

La chargée de séance,	Le président,
Caroline Orsel	Benoît Cœuré

Glossaire

- **Accélérateur d'IA :** circuit intégré, conçu et optimisé pour les charges de travail en intelligence artificielle.
- **API / Interface de programmation d'application** (en anglais « *Application Programming Interface* ») **:** interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités (définition issue du glossaire de la CNIL).
- **Apprentissage automatique** (en anglais « *machine learning* » ou « *ML* ») **:** champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'« apprendre » à partir de données, via des modèles mathématiques (définition issue du glossaire de la CNIL).
- **Apprentissage profond** (en anglais « *deep learning* ») **:** procédé d'apprentissage automatique utilisant des réseaux de neurones possédant plusieurs couches de neurones cachées. Ces algorithmes possédant de très nombreux paramètres, ils demandent un nombre très important de données afin d'être entrainés (définition issue du glossaire de la CNIL).
- **Cadriciel (« *framework* ») :** ensemble cohérent de composants logiciels servant à créer les fondations et l'architecture d'un logiciel.
- **Common Crawl :** organisation à but non lucratif fondée en 2007 aux États-Unis, le Common Crawl a pour mission de fournir gratuitement des archives d'Internet. Depuis 2008, une centaine de *crawl* d'Internet ont été réalisés.
- **CPU** (« *Central Processing Unit* ») **:** processeurs centraux.
- **Crawl :** collection automatique par un robot du contenu d'une page web.
- **CUDA (« *Compute Unified Design Architecture* ») :** logiciel propriétaire développé par la société Nvidia pour permettre la programmation sur ses propres GPU.
- **Données d'entraînement :** jeu de données (texte, sons, images, listes, etc.) utilisé lors de la phase d'entraînement / d'apprentissage : le système s'entraine sur ces données pour effectuer la tâche attendue de lui (définition issue du glossaire de la CNIL).
- **Données synthétiques :** données artificielles générées à partir de données originales et d'un modèle entrainé à reproduire les caractéristiques et la structure des données originales.
- **Entraînement (ou apprentissage) :** processus de l'apprentissage automatique pendant lequel le système d'intelligence artificielle construit un modèle à partir de données (définition issue du glossaire de la CNIL).
- **Étiquetage des données :** processus consistant à identifier des données brutes (images, fichiers texte, vidéos, etc.) et à ajouter une ou plusieurs étiquettes informatives et pertinentes pour apporter du contexte qui va renseigner le modèle d'apprentissage.
- **FLOPS (« *Floating-point Operations Per Second* ») :** nombre d'opérations en virgule flottante par seconde, unité de mesure de la puissance de calcul.
- **GPT (« *Generative Pretrained Transformers* ») :** réseau de neurones pré-entrainés d'architecture Transformers.

- **GPU / Processeur graphique :** processeur composé de nombreux cœurs spécialisés, permettant d'assurer les fonctions de calcul d'images de manière parallélisée. On le trouve généralement sur les cartes graphiques (définition issue de l'avis n° 23-A-08 de l'Autorité).
- **Hyperparamètre :** variable régissant le processus d'entraînement en lui-même et déterminée par le développeur. Il peut s'agir du nombre de nœuds par couche, de la taille des couches cachées du réseau de neurones, de l'initialisation des poids, du coefficient d'apprentissage, de la fonction d'activation, du nombre de fois où chaque donnée sera utilisée pendant l'entraînement, etc.
- **IaaS (« Infrastructure as a Service ») :** service d'informatique en nuage consistant à fournir le traitement, le stockage, les réseaux et d'autres ressources informatiques fondamentales dans lesquelles le consommateur peut déployer et exécuter les logiciels de son choix (définition issue de l'avis n° 23-A-08 de l'Autorité précité, page 198).
- **Inférence :** processus par lequel un modèle entrainé est utilisé pour effectuer des prédictions sur de nouvelles données, après sa phase d'apprentissage. Dans le cadre de l'intelligence artificielle générative, cela correspond à la production de contenu.
- **Intelligence artificielle :** tout outil utilisé par une machine afin de « *reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité* » (définition issue du Parlement européen).
- **Intelligence artificielle générative :** branche de l'intelligence artificielle qui vise à créer de nouveaux contenus (textes, images, vidéos, audio, etc.).
- **LLM / Grand modèle de langage** (en anglais « *Large Langage Model* ») **:** modèle d'IA générative de texte comprenant un grand nombre de paramètres.
- **Lora / Adaptation de rang faible** (en anglais « *Low Rank Adaptation* ») **:** technique de réglage fin introduite par une équipe de chercheurs de Microsoft en 2021, et réduisant les besoins en puissance de calcul.
- **MaaS (« Model as a Service ») :** plateforme ou place de marché *cloud* permettant aux développeurs d'accéder à plusieurs modèles de fondation par le biais d'une interface de programmation commune.
- **Modèle :** le modèle d'IA est la construction mathématique générant une déduction ou une prédiction à partir de données d'entrée. Le modèle est estimé à partir de données lors de la phase d'entrainement du système d'IA (définition issue du glossaire de la CNIL).
- **Modèle de fondation :** modèle d'IA de grande taille, entrainé sur un vaste ensemble de données et pouvant être adapté pour différentes tâches à l'aval.
- **MoE / combinaison d'experts** (en anglais « *Mixture of Experts* ») **:** architecture de modèle d'IA divisée en plusieurs sous-ensembles de réseaux de neurones appelés experts et spécialisés sur une tâche spécifique, ainsi qu'un routeur qui détermine quel expert doit être utilisé pour répondre à une requête.
- ***Open source* :** logiciel dont le code source est à la disposition du grand public. Le développement de ces « logiciels libres » implique un effort de collaboration où les programmeurs améliorent ensemble le code source et partagent les changements au sein d'une communauté (définition issue de l'avis de n° 14-A-18 de l'Autorité).
- ***Open-weights* :** désigne les modèles de fondation dont les poids ont été rendus disponibles à tous.

- **PaaS (« *Platform as a Service* ») :** service d'informatique en nuage consistant à déployer sur une infrastructure *cloud* des applications créées ou acquises par le consommateur à l'aide de langages de programmation, de bibliothèques, de services et d'outils pris en charge par le fournisseur (définition issue de l'avis n° 23-A-08 de l'Autorité précité, page 198).
- **PEReN (Pôle d'Expertise et de Régulation du Numérique) :** service à compétence nationale placé sous l'autorité conjointe des ministres chargés de l'économie, de la culture et du numérique.
- **Poids / Paramètre :** dans un réseau de neurones, un poids est un coefficient de puissance de la connexion entre deux neurones, qui s'ajuste pendant toute la phase d'entraînement (définition issue du guide ANSSI, Recommandations de sécurité pour un système d'IA générative, 29 avril 2024).
- **RAG / Génération augmentée de récupération (en anglais « *Retrieval Augmented Generation* ») :** la génération augmentée par récupération est une technique permettant d'améliorer la précision et la fiabilité des modèles génératifs d'IA à l'aide de données provenant de sources externes.
- **Réglage fin :** technique consistant à spécialiser un modèle d'IA pré-entrainé pour accomplir une tâche spécifique. Cela consiste généralement à entrainer le modèle dans son ensemble, ou seulement certaines couches d'un réseau de neurones, pour un faible nombre d'itérations sur un ensemble de données spécifiques correspondant à la tâche visée (définition issue du glossaire de la CNIL).
- **Réseaux de neurones :** dans le domaine de l'intelligence artificielle, un réseau de neurones est un ensemble organisé de neurones artificiels interconnectés permettant la résolution de problèmes complexes tels que la vision par ordinateur ou le traitement du langage naturel (définition issue du glossaire de la CNIL).
- **RLHF / Apprentissage par renforcement et rétroaction humain** (en anglais « *Reinforcement Learning from Human Feedback* ») **:** approche d'apprentissage par renforcement qui utilise les commentaires et les évaluations d'utilisateurs humains pour guider l'apprentissage d'un modèle d'intelligence artificielle. Ce type d'apprentissage est utilisé dans les générateurs de texte fondés sur les grands modèles de langue (définition issue du glossaire de la CNIL).
- **SaaS (« *Software as a Service* ») :** service d'informatique en nuage consistant à offrir la capacité au consommateur d'utiliser les applications du fournisseur exécutées sur une infrastructure *cloud* (définition issue de l'avis n° 23-A-08 de l'Autorité précité, page 199).
- **Services d'apprentissage automatique automatisés** (en anglais « *automated machine learning* » ou « autoML ») **:** génération automatique de modèles d'apprentissage optimisés, permettant son utilisation par des utilisateurs non experts.
- **SLM** (« *Small Language Models* ») **:** petits modèles de langage.
- **Supercalculateur :** un très grand ordinateur, réunissant plusieurs dizaines de milliers de processeurs, et capable de réaliser un très grand nombre d'opérations de calcul ou de traitement de données simultanées (définition issue du CEA).
- **TAL (traitement automatique du langage naturel),** en anglais « *Natural Language Processing* » ou « *NLP* » **:** domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle (définition issue du glossaire de la CNIL).

- **Token :** série de quelques lettres qui ne forment pas toujours des mots complets.
- **TPU** (en anglais « *Tensor Processing Units* ») **:** processeurs tensoriels.
- **Vision par ordinateur** (en anglais, « *computer vision* ») **:** branche de l'IA dont le principal but est de permettre à une machine d'analyser et traiter une ou plusieurs images ou vidéos prises par un système d'acquisition (définition issue du glossaire de la CNIL).