France Digitale

FDécryptage
Avril 2024

Des puces aux applications, l'Europe peut-elle être une *puissance* de l'IA générative ?

Sommaire

L'édito

L'émergence de l'IA générative, apparue au grand public en fin d'année 2022 interroge : quelles sont les implications sociétales de cette technologie ? Quel impact aura-t-elle sur les individus et l'environnement ? Comment les entreprises en tireront-elles de la valeur ?

Plus largement, **l'Europe a-t-elle les moyens matériels, financiers et technologiques de gagner en souveraineté grâce à l'IA générative ?**

C'est à cette dernière question que nous avons voulu répondre. En effet, derrière les annonces des derniers mois - valorisations records, partenariats stratégiques parfois controversés, investissements massifs des États et des entreprises - la vraie interrogation porte sur **les perspectives économiques et géopolitiques** de l'IA générative.

Alors que la plupart de ces annonces proviennent d'acteurs technologiques américains déjà dominants sur le marché du numérique, l'IA générative va-t-elle renforcer la position de ces acteurs - que l'Europe essaie déjà de contrôler au travers du Digital Markets Act - ou de nouveaux concurrents, potentiellement européens, auront-ils les moyens d'émerger et de prospérer ?

Pour répondre à cette question, nous avons interrogé une quarantaine de startups européennes et fonds de capital-risque actifs dans le secteur de l'IA générative, et élargi notre champ d'investigation pour **obtenir une vue d'ensemble de la chaîne de valeur de l'IA générative dans le monde**. Voici nos conclusions :

1. **Les entreprises européennes excellent à différents niveaux de la chaîne de valeur de l'IA générative** : de l'assemblage de puces à la conception de modèles de fondation, l'hébergement de données ou le développement d'applications hautement spécialisées.

2. Toutefois, **l'Europe n'a pas encore les moyens matériels, financiers, technologiques et humains d'une totale indépendance de l'ensemble des acteurs de cette chaîne de valeur**. Il n'existe pas aujourd'hui de chaîne de valeur strictement européenne, de la même manière qu'il n'existe pas non plus de chaîne de valeur strictement américaine ou chinoise.

Le marché est mondialisé. Cependant, l'interdépendance des acteurs européens avec les acteurs extra-européens peut être progressivement réduite.

L'édito

3. **Comment réduire cette interdépendance ?** En renforçant à chaque couche de la chaîne de valeur l'échelon européen.

Cela nécessite :

- d'investir massivement sur l'ensemble des couches de la chaîne de valeur ;
- de sourcer davantage de matériaux rares - quitte à capitaliser sur de nouveaux gisements, comme celui déjà présent dans nos téléphones, nouvelles "mines urbaines" ;
- de doper la commande publique et privée en privilégiant ouvertement la commande européenne ;
- de former et d'attirer en Europe des ingénieurs spécialisés.

4. **Mais aussi d'avoir une vision politique et réglementaire adaptée au contexte économique et financier** : toute réglementation visant à encadrer les règles concurrentielles, les investissements étrangers ou les acquisitions prédatrices devra appréhender les effets collatéraux sur les startups et leurs investisseurs européens, et tenir compte des interdépendances, des structures de coûts et de la complexité des cycles d'innovation rapides.

Faisons ensemble de l'Europe une puissance de l'IA générative !

Maya NOËL
Directrice générale
France Digitale

Vous ne parlez pas IA ?

On vous a concocté un lexique pour (enfin) comprendre tous les termes techniques liés à l'intelligence artificielle. Rendez-vous ici.

Le résumé du résumé

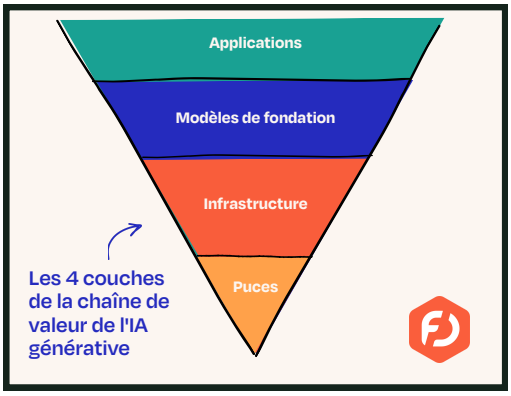
Si vous ne devez lire qu'une page, c'est celle là !

On a décrypté toute la chaîne de valeur de l'IA générative, pour identifier les principaux acteurs, les interdépendances et les mécanismes à l'œuvre sur ce marché.

On en a tiré un résumé des principaux enjeux et une infographie.

- **La chaîne de valeur de l'IA générative est complexe**, englobant à la fois des éléments matériels (*hardware*) et des logiciels (*software*). Chaque segment et sous-segment de cette chaîne représentent un marché à part entière, tout en étant inextricablement liés aux autres, créant un réseau d'interdépendances entre les acteurs.

- **L'écosystème de l'IA générative repose sur quatre couches fondamentales** : les puces (logiciels de conception, matières premières, machines de haute précision, fabrication), l'infrastructure (centres de données, réseaux, logiciels et services, y compris les services de distribution), les modèles de fondation et les applications.

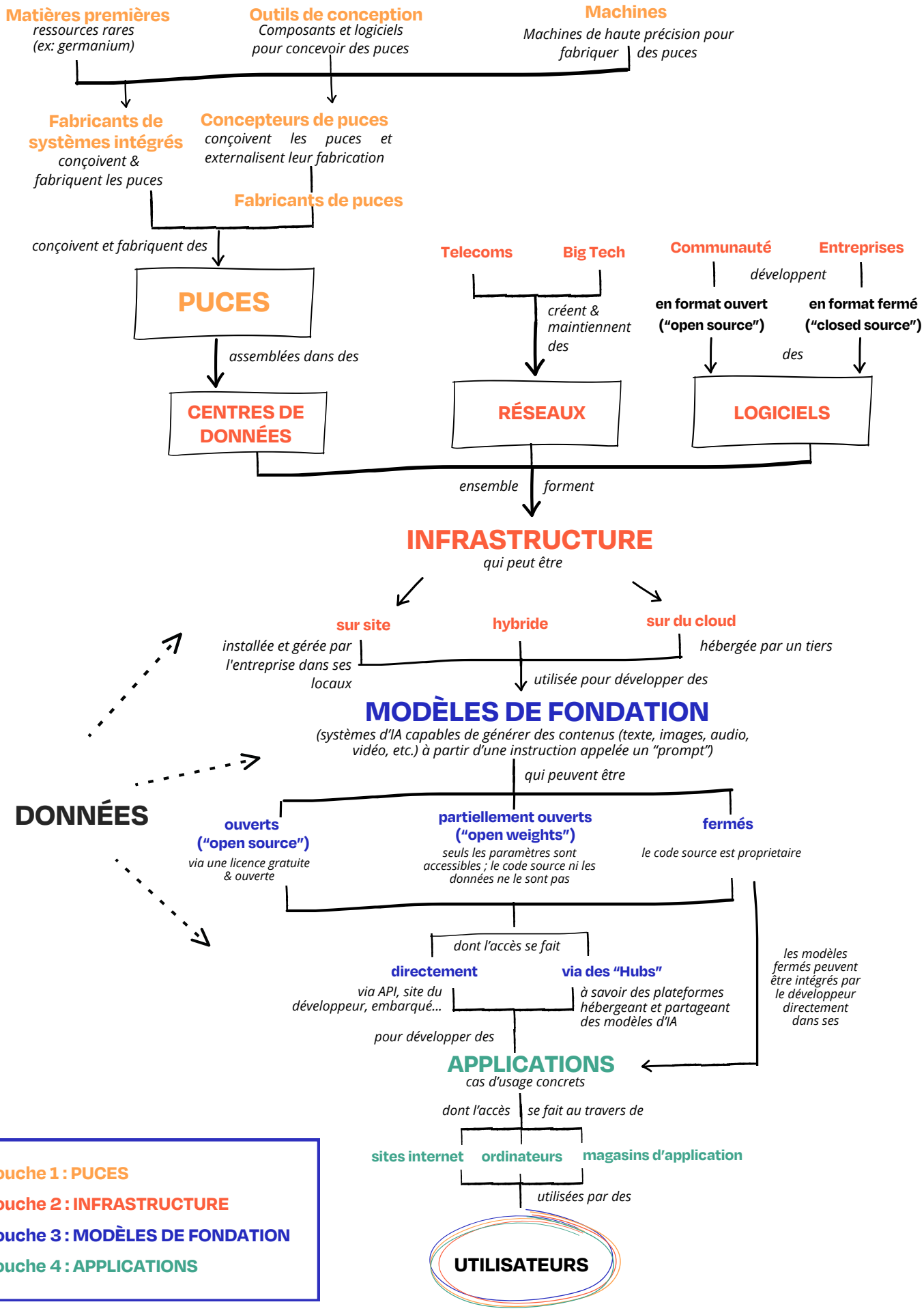


- Si les modèles de fondation d'IA générative, comme OpenAI ou Mistral, suscitent un vif enthousiasme auprès du grand public, **la valeur économique dans la chaîne de valeur réside principalement dans les puces et l'infrastructure**. Les applications spécialisées peuvent également avoir une valeur économique importante (dans le domaine de la santé ou de la finance par exemple), parce qu'elles répondent à des cas d'usage concrets pour les utilisateurs.

- Des acteurs majeurs (tels qu'Amazon, Google, Microsoft et, dans une moindre mesure, Nvidia) se distinguent par leur **structure intégrée verticalement**, à savoir une présence à différents niveaux de la chaîne de valeur par le biais de partenariats stratégiques ou d'investissements. Cela leur confère **un avantage stratégique indéniable dans la capture de valeur** avec deux conséquences principales :
  - **D'une part, les partenariats prennent une place importante dans le marché de l'IA générative**, tant entre petites et grandes entreprises qu'entre grandes entreprises. Cette tendance entraîne souvent une dynamique de coopétition, où la collaboration et la concurrence se côtoient et s'entremêlent.
  - **D'autre part, les entreprises verticalement intégrées ne sont pas de simples concurrentes** des plus petites entreprises ; elles jouent également un rôle crucial en tant que fournisseurs d'infrastructure et d'accès à de nouveaux marchés.

- Pour apprécier **les comportements anti-concurrentiels** sur la chaîne de valeur de l'IA générative, le diable est dans les détails. Des comportements anti-concurrentiels peuvent émerger d'une accumulation de pratiques de marché apparemment légitimes prises individuellement, mais devenant anti-concurrentielles lorsqu'elles sont appliquées de manière systématique par des entreprises dominant leur marché ou profitant de la dépendance de leurs partenaires. Si, à ce jour, aucune pratique de marché ouvertement abusive n'a été observée, **l'existence d'entreprises dominant le marché et disposant d'une structure verticalement intégrée n'est pas sans risque**. Nous avons identifié les risques à chaque couche de la chaîne de valeur.

Comprendre la chaîne de valeur de l'IA générative en 1 infographie



Couche 1 : PUCES
Couche 2 : INFRASTRUCTURE
Couche 3 : MODÈLES DE FONDATION
Couche 4 : APPLICATIONS

Couche n°1

Les puces, au cœur de l'IA Générative

Les principales entreprises derrière les puces de l'IA Générative

Matières premières
ressources rares
(silicium, terres rares, germanium, gallium, etc.)

Outils & Services
pour la conception des puces

Machines
machines de haute précision pour fabriquer des puces
ASML (NL)

Semi-conducteurs (circuits préconçus)
***Arm (UK)**, Cadence (US), Synopsis (US), **Siemens EDA (DE)**, **Imagination Technologies (UK)**, eMemory (TW), **Dolphin Designs (FR)**, Lattice Seminc. (US), Rambus (US), Ceva (US)*

Logiciels de conception assistée par ordinateur (CAO)
*Cadence (US), Synopsis (US), **Siemens EDA (DE), Defacto Tech (FR)***

Fabricants de systèmes intégrés
conçoivent et fabriquent les puces
Intel (US), Siemens (DE), Samsung (KO), SK Hynix Inc. (KO)

Concepteurs de puces
conçoivent les puces et externalisent leur fabrication
*AMD (US), Broadcom (US), IBM (US), Mythic (US), Nvidia (US), Qualcomm (US), Xilinx/AMD (US), Cerebras (US), Habana (IL), Cambricon (CN), **Flex.ai (FR), Sipearl (FR), Graphcore (UK)***

Amazon (US), Apple (US), Google (US), Meta (US), Microsoft (US), Alibaba (CN)

Fabricants de puces
TSMC (TW), UMC Group (TW), PSMC (TW), VIS (TW), SMIC (CN), HUAHONG (CN), Samsung (KO), DB Hitek (KO), Global Foundries (US), Intel (US) Tower Semiconductor (IL)

PUCES
qui servent à

entrainer un système d'IA
c'est-à-dire créer un modèle d'IA à partir d'un ensemble de données

faire de l'inférence de système d'IA
c'est-à-dire produire un contenu à partir d'un modèle d'IA pré-entrainé sur des données

il y a d'ailleurs différents types de puces

GPUs
Nvidia (US), AMD (US), Intel (US)

ASICs / FPUs
*Amazon (US), Google (US), Apple (US), Meta (US), Microsoft (US), Broadcom (US), IBM (US), Mythic (US), Qualcomm (US), Xilinix/AMD (US), Cerebras (US), Habana (IL), Alibaba (CN), Cambricon (CN), Samsung (KR), **Graphcore (UK), Flex.ai (FR)***

CPUs
*Intel (US), AMD (US), Apple (US), Qualcomm (US), **Sipearl (FR)***

Comment ça marche ?

Les puces - autrement appelées “circuits intégrés” - sont les composantes électroniques essentielles qui forment les systèmes informatiques. Concrètement, une puce est un ensemble de circuits électroniques gravés sur un petit disque de silicium (appelé "wafer"), puis programmée pour exécuter certaines tâches. En pratique, les puces fournissent la puissance de calcul et les fonctions de mémoire suffisantes pour développer et déployer des logiciels, y compris l'IA générative.

FDécodeur - Pourquoi les systèmes d'IA générative ont-ils besoin de puissance de calcul ? Les systèmes d'IA générative ont besoin de puissance de calcul pour effectuer deux fonctions principales :

(1) **l’entraînement :** il s'agit du processus de conception et de développement d'un modèle d'IA générative en utilisant des architectures d'apprentissage automatique (“Machine Learning”) et d'apprentissage profond (“Deep Learning”) pour analyser et apprendre à partir d'un ensemble de données apportées. *Exemple : la conception de GPT-3 par OpenAI repose sur l'utilisation d'une architecture d'apprentissage profond spécifique appelée "Transformer", ainsi que sur l'analyse des vastes données disponibles sur Internet.*

(2) **l’inférence :** il s'agit du processus par lequel un modèle d'IA générative pré-entraîné utilise des données externes pour produire du contenu inédit. *Example : quand on fournit à un chatbot pré-entraîné un contrat, et qu’on lui pose des questions sur son contenu, le chatbot analysera le contrat et fournira des réponses à nos questions.*

Deux techniques permettent en outre d'**affiner** un modèle pour une tâche précise et obtenir des réponses adaptées à un cas d'utilisation particulier.

- Une première technique s’appelle le “**fine tuning”**. Elle consiste à poursuivre le processus d'entraînement du modèle en utilisant un ensemble de données spécifiques. L'objectif est d'améliorer les performances du modèle dans des domaines ou pour des tâches particulières. *Exemple : entraîner un modèle de reconnaissance d'image pour identifier des races de chiens spécifiques en utilisant un ensemble de photos de chiens.*

- La seconde technique est le “***Retrieval Augmented Generation*” ou RAG**. Cette technique permet à l'IA de récupérer des informations à partir d'un ensemble de données externes. Pour ce faire, elle combine des fonctions de recherche pour identifier des documents pertinents et un modèle de langage pour résumer ces documents ou répondre à des questions basées sur l'information récupérée. *Exemple: Intégrer un modèle de langage (LLM) dans le système d'information d'une entreprise. Un chatbot basé sur cet LLM pourra alors fournir des réponses basées sur les informations internes de l'entreprise.*

La fabrication de puces repose sur trois éléments indispensables :

- des **matières premières**, principalement du silicium, mais aussi une grande quantité d'eau et d'autres éléments (germanium, gallium, phosphore, phosphure d'indium, bore et certaines terres rares), qui constituent les fondations physiques des puces ;

- des **outils de conception électronique**, comme des logiciels et des services spécialisés, qui permettent de concevoir les puces ; et

- des **machines de haute précision**, qui permettent de réaliser des gravures de circuits intégrés extrêmement précises.

Ces éléments convergent dans des usines spécialisées (appelées "**fonderies**"), responsables de la fabrication des puces, pour être finalement assemblés et commercialisés.

Toutes les puces ne sont pas équivalentes, et certaines exigent des techniques de production plus sophistiquées que d'autres. Plus la gravure des circuits est fine et précise, plus la puce est puissante. Or, seules les puces les plus performantes, issues de techniques de production de pointe, sont capables d'exécuter les calculs complexes nécessaires aux systèmes d'IA générative.

FDécodeur - une puce adaptée à chaque usage

S'il existe une grande diversité de puces, toutes ne sont pas adaptées à l'IA générative. Trois architectures se distinguent par leur aptitude à exceller dans ce domaine :

- **Les Unités de Traitement Graphique ("GPUs")** peuvent effectuer plusieurs calculs simultanément, ce qu'on appelle le "traitement parallèle". Initialement conçues pour le traitement d'images, notamment dans l'industrie des jeux vidéo, les GPUs se sont révélées très efficaces pour l'entraînement des systèmes d'IA générative. Elles peuvent également charger et exécuter rapidement des modèles d'IA lourds pour l'inférence grâce à leurs importantes capacités de mémoire vive et l'utilisation de milliers de cœurs de traitement. *Exemple : la carte Nvidia A100 80GB - avec un prix unitaire d'environ 21 000 €.*

- **Les Unités de Traitement Central ("CPUs")** offrent une puissance de calcul supérieure à celle des GPUs, mais leur capacité en calcul parallèle est moindre en raison de leur utilisation de la mémoire vive et de leur nombre limité de cœurs de traitement. Par conséquent, les CPUs sont moins efficaces pour l'entraînement des modèles d'IA générative, mais elles peuvent être employées seules ou avec des GPUs pour des tâches d'inférence spécialisées. Elles présentent l'avantage d'être moins onéreux et de consommer moins d'énergie que les GPUs, mais leur vitesse d'inférence est approximativement trois fois plus lente. *Exemple : le processeur Intel Xeon Scalable.*

- **Les Circuits Intégrés Spécifiques à une Application ("ASICs") et les Unités en Virgule Flottante ("FPUs")** sont utilisés pour leur rapidité et leur efficacité énergétique. Cependant, ils sont limités à des tâches spécifiques de l'entraînement et l'inférence des modèles. *Exemple : Tensor Processing Unit (TPU) de Google, plus rapide que les GPUs dans l'entraînement.*

Quels sont les risques concurrentiels sur le marché des puces ?

Dans cette couche de la chaîne de valeur, nous observons une concentration du marché, mais également l'apparition d'entraves à la concurrence.

La domination de la Chine sur les matières premières : une menace pour l'industrie des puces?

Le silicium, matière première essentielle pour la fabrication des puces, est largement disponible. Cependant, son extraction est dominée par la Chine, qui contrôle aujourd'hui 70% du marché. Ce pays détient également une position de force sur d'autres matériaux critiques pour la fabrication de semi-conducteurs, détenant 80% de la production de germanium et de gallium. En ce qui concerne les terres rares, la Chine dispose de certains des plus importants gisements minéraux au monde, lui permettant de représenter 60% de la production mondiale, comparativement à 15% pour les États-Unis. Actuellement, aucun site d'extraction de terres rares n'est opérationnel en Europe ; bien qu'un gisement significatif ait été découvert en Suède l'année dernière, il faudrait encore 10 à 15 ans pour lancer les opérations d'extraction. En ce qui concerne le traitement des terres rares, la Chine détient 85% du marché mondial : non seulement elle a consacré des années au développement de ses capacités minières et de traitement, mais elle a par ailleurs constitué d'importantes réserves stratégiques de métaux indispensables à la fabrication des dispositifs numériques.

Cette concentration confère à la Chine une position dominante sur le marché, exposant ainsi les entreprises en aval à divers risques, notamment **des restrictions à l'exportation susceptibles de provoquer des hausses de prix artificielles**. De telles mesures pourraient être mises en œuvre **soit par la Chine de manière individuelle, soit de manière collective à travers de l'alliance des BRICS** (Brésil, Russie, Inde, Chine, Afrique du Sud). La seconde option pourrait être particulièrement préoccupante, car le groupe détiendrait alors 72% des réserves mondiales de terres rares. Les restrictions à l'exportation représenteraient une forme d'**abus de pouvoir envers les partenaires commerciaux dépendants** et pourraient entraîner des perturbations majeures dans la chaîne de valeur. Il convient de rester particulièrement vigilant face aux évolutions commerciales et géopolitiques dans ce domaine.

Historique des restrictions chinoises à l'exportation de matières premières

La Chine a déjà utilisé des restrictions à l'exportation de matières premières critiques comme arme géopolitique. En 2023, la Chine a limité ses exportations de germanium et de gallium, deux matériaux essentiels à la fabrication de puces. Cette décision a été prise en réponse aux mesures protectionnistes américaines sur les puces. Ce n'est pas la première fois que la Chine utilise ces techniques. En 2010, le pays a réduit de 40% ses exportations de certaines terres rares et imposé un embargo total au Japon, suite à des tensions politiques entre les deux pays. L'augmentation conséquente du prix des terres rares a été condamnée par l'Organisation mondiale du commerce (OMC) comme une restriction commerciale injustifiée.

Un leader européen dans le segment de la machinerie à haute précision

Le segment de la gravure des puces les plus avancées (avec des nœuds de circuit inférieurs à 14 nm) est actuellement dominé par une seule entreprise : la néerlandaise Advanced Semiconductor Material Lithography (ASML). ASML est l'unique fabricant d'équipements nécessaires à la production de ces puces de pointe. Si Canon, un concurrent japonais, a annoncé son intention d'entrer sur ce marché, il n'est pas encore opérationnel.

La conception de puces : un marché essentiel et sous-estimé

Le marché des outils de conception de puces se compose de deux segments principaux :

- Les fournisseurs de propriété intellectuelle (IP), qui proposent des circuits pré-conçus que les clients peuvent adapter à leurs besoins : le segment des fournisseurs d'IP est dominé par l'entreprise britannique ARM dont la part de marché s'approche des 41%, mais d'autres acteurs européens, américains et asiatiques sont également présents ;

- les systèmes CAO (Conception Assistée par Ordinateur), qui sont des outils et logiciels pour concevoir des circuits. Le segment des CAO est dominé par trois entreprises (Cadence, Siemens EDA et Synopsys) qui capturent 75% de la part de marché. Le marché présente également une concentration géographique importante, avec deux fournisseurs basés aux États-Unis, Cadence et Synopsys, qui détiennent ensemble 62% de la part de marché. Il est important de souligner que ces trois entreprises sont également **intégrées verticalement**, étant présentes dans le segment de la propriété intellectuelle (notamment Cadence et Synopsys), ce qui crée un **risque de verrouillage** pour les utilisateurs.

Géants du cloud contre titans des puces : une histoire de coopétition

Dans le **segment de la conception de puces**, nous distinguons principalement deux types d'acteurs :

(1) les fabricants de dispositifs intégrés (IDM), qui sont responsables à la fois de la conception et de la fabrication des puces,

(2) les entreprises sans usine (dites "fabless"), qui conçoivent les puces, mais externalisent leur fabrication à des fonderies spécialisées. Au sein des entreprises sans usine, on identifie deux sous-segments : celles qui conçoivent des puces spécifiques à un dispositif ou à une tâche, et celles qui conçoivent des puces à usage général.

- Dans le sous-segment des puces spécifiques à un dispositif ou à une tâche, nous trouvons des fabricants de dispositifs, mais aussi des **entreprises intégrées verticalement proposant des services cloud et logiciels**. Les principaux acteurs sont américains (notamment Amazon, Apple, Google, Meta et Microsoft), mais aussi chinois (Alibaba). Ils ont investi dans ce marché pour assurer leur approvisionnement en puces en développant des puces spécifiques adaptées à leurs appareils et/ou services, comme les TPU de Google pour le cloud computing, qui ont été développés depuis 2018.

- Le sous-segment des puces à usage général est également dominé par des sociétés américaines, mais qui ne fabriquent pas toutes des types de puces compatibles avec l'IA générative. La conception des CPUs est largement dominée par Intel qui détient une part de marché de 71 %, mais de nouveaux concurrents tels qu'AMD, AWS et Ampere émergent et remettent en question sa position. En revanche, le marché des GPUs est beaucoup plus concentré, avec **Nvidia contrôlant 84 % du marché total et 92 % du marché des GPUs destinés au cloud computing**. Ses deux principaux concurrents, AMD et Intel, rencontrent des difficultés pour s'imposer, bien que selon les observateurs du marché et les startups interrogées dans le cadre de notre recherche, AMD est en meilleure position. Cependant, même si AMD émerge en tant que concurrent solide sur le marché des GPUs cloud, il est largement prévu que Nvidia conserve sa domination sur le marché pendant un certain temps, exposant ainsi les entreprises en aval à plusieurs risques.

Bien que le marché des GPUs soit dominé par Nvidia, il n'y a eu aucune preuve concrète de pratiques monopolistiques à ce jour. Cependant, le contexte actuel présente un risque potentiel pour de telles pratiques, telles que **des fixations des prix, des restrictions de la production, des conditions contractuelles déloyales ou des comportements discriminatoires**. Malgré ces risques, les startups interrogées ayant des relations directes avec Nvidia semblent entretenir des relations mutuellement bénéfiques, **bien que fragiles**. Elles entretiennent toutes une relation client-fournisseur "privilégiée", souvent complétée par un partenariat (par exemple, un accord de revente). Cependant, elles sont conscientes du déséquilibre du rapport de force avec Nvidia, ce qui pourrait entraîner l'interruption unilatérale du partenariat ou la cessation du traitement préférentiel. Par exemple, en cas de pénuries d'approvisionnement, il est probable que les clients les plus importants soient favorisés.

Parmi 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. Cette relation complexe peut être qualifiée de "**coopétition**", illustrant la collaboration entre concurrents. L'absence d'exclusivité dans ces partenariats permet de maximiser la disponibilité des GPUs pour les utilisateurs. Cette stratégie pourrait s'avérer utile pour contrer d'éventuelles accusations de comportement anticoncurrentiel de la part de Nvidia ou d'autres acteurs du marché.

FDécodeur : qu'est-ce que la coopétition ?

La coopétition est une alliance stratégique entre des organisations concurrentes. Cette pratique est particulièrement fréquente dans les secteurs du logiciel et du matériel informatique, où les entreprises cherchent à tirer des avantages mutuels de projets spécifiques. (source)

La fabrication de puces : un marché exposé aux risques géopolitiques

Le marché de la fabrication de puces avancées (de 3 à 7 nm) se caractérise par une forte concentration. En dehors des fabricants de puces intégrés (IDM), **TSMC (Taïwan) domine largement le marché avec une part de marché de 90 %**, suivi par Samsung Foundry (Corée du Sud). D'autres acteurs importants dans la production de puces avancées (de 7 à 28 nm) incluent SMIC en Chine et Global Foundries aux États-Unis. Bien qu'il n'y ait apparemment pas d'accord d'exclusivité entre les entreprises *fabless* et les fabricants de puces, cette concentration du marché présente plusieurs enjeux. En plus des préoccupations concernant l'abus de position dominante, la concentration géographique de la fabrication de puces en Asie de l'Est engendre des risques géopolitiques. Les tensions ou les conflits entre ces pays (Chine et Taïwan, Chine et États-Unis, Corée du Nord et Corée du Sud) pourraient entraîner des perturbations dans l'approvisionnement en technologies critiques.

Face à la concentration du marché des fonderies de puces avancées, **le secteur privé réagit, et des acteurs existants, comme Intel, se lancent dans le marché des fonderies**. Ainsi, l'entreprise américaine Intel entend devenir un concurrent direct de TSMC d'ici 2030, alors que pendant des décennies, elle s'est limitée à la fabrication de puces pour sa propre utilisation. Désormais, l'entreprise envisage de scinder ses activités en créant une division baptisée Intel Foundry Services, dédiée à la production de circuits intégrés avancés pour d'autres fournisseurs *fabless*. Cette initiative ouvre Intel à des fournisseurs qu'elle considérait auparavant comme des concurrents, comme Microsoft, Nvidia, Qualcomm, Google, et même AMD. Avec cette nouvelle stratégie, le géant américain adopte un modèle similaire à celui de Samsung, qui fabrique des puces pour ses propres besoins ainsi que pour des fournisseurs fabless.

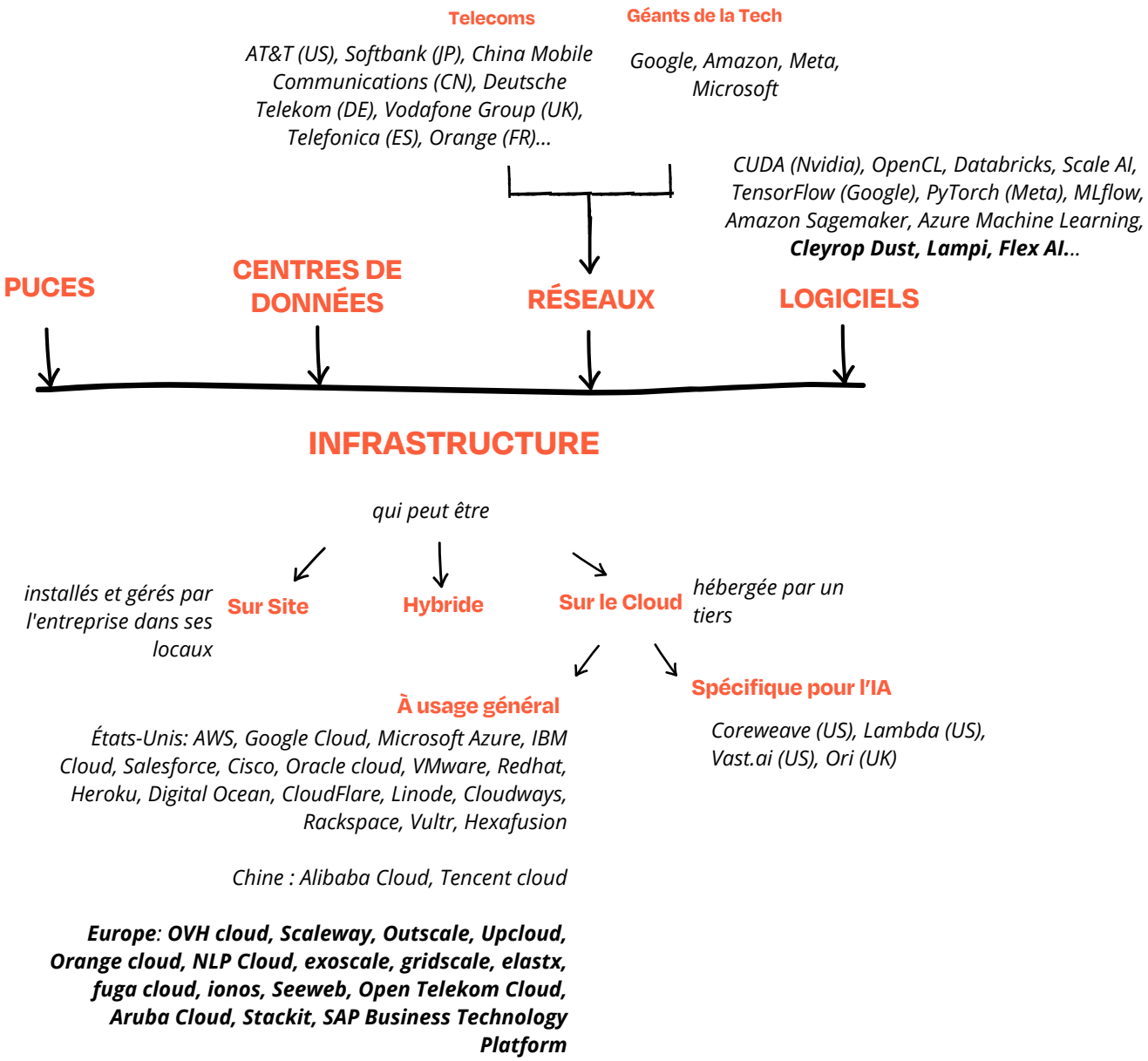
Outre les acteurs privés, **les pouvoirs publics s'impliquent dans la lutte contre la concentration du marché des fonderies de puces avancées.** Ainsi, en 2022, les États-Unis ont lancé le fonds CHIPS for America, doté de 53 milliards de dollars. L'objectif de ce fonds est de stimuler la construction de fonderies avancées sur le territoire américain et des acteurs tels que TSMC, Intel et d'autres fabricants de puces sont susceptibles de bénéficier de ces subventions. L'annonce du fonds a eu un effet catalyseur, attirant 166 milliards de dollars d'investissements privés supplémentaires dans le secteur des fonderies.

L'Europe se mobilise aussi : le Chips Act, adopté en juillet 2023, veut mobiliser plus de 43 milliards d'euros d'investissements publics et privés en collaboration entre États et partenaires internationaux. Ainsi, TSMC, Intel et Global Foundries (avec STMicroelectronics), ont annoncé la construction de "mégafabs" en France, Allemagne et Italie.

Couche n°2

L’infrastructure avec l'omniprésence des hyperscalers

Les principales entreprises de la couche “infrastructure” de l’IA générative



Comment ça marche ?

L'infrastructure de l’intelligence artificielle désigne l'ensemble des ressources matérielles et logicielles nécessaires pour créer, entraîner et exécuter des modèles d'IA.

- **ressources matérielles (“hardware”)**: il s’agit de tous les composants physiques et électroniques (tels que les puces, les serveurs, les centres de données, les réseaux, etc.), nécessaires au stockage des données, à l'entraînement des algorithmes et au déploiement des systèmes d'IA.

- **les logiciels (“software”) :** il s’agit des instructions informatiques intangibles utilisées pour développer et déployer des systèmes d'IA, tels que des logiciels de programmation de puces, des frameworks pour le Deep Learning et Machine Learning, des bibliothèques de développement, des outils de gestion et d'analyse des données, entre autres.

L'infrastructure est utilisée pour **collecter, stocker et gérer les données.** Elle joue un rôle crucial dans le développement et le déploiement des modèles d'IA générative. Le choix de l'infrastructure dépend des besoins spécifiques de chaque projet d'IA. Deux approches principales se distinguent :

- **l'infrastructure sur site,** à savoir celle qui est installée et gérée par l'entreprise sur son site ou dans ses locaux. Cette option offre un contrôle total sur les actifs physiques et numériques, notamment les données. Cependant, elle nécessite un investissement important en matériel, logiciel et personnel pour assurer la gestion et la maintenance des centres de données et des réseaux.

- **l'infrastructure basée sur le cloud (public ou privé) :** elle est alors hébergée par un tiers, le fournisseur de cloud. Cette option offre une capacité d'évolution et une flexibilité importante, sans avoir à gérer le fardeau de la gestion d'un centre de données.

Tous les fournisseurs de cloud ne se situent pas au même niveau en termes d'échelle et de services offerts. On peut les classer en deux catégories principales :

- **Les hyperscalers** sont des acteurs mondiaux tels qu'Amazon Web Services (AWS), Microsoft Azure et Google Cloud Platform (GCP). Ils fournissent une vaste gamme de services (PaaS et SaaS) ainsi qu'une infrastructure massive capable de développer et de maintenir de grands modèles de fondation.
- **Les fournisseurs de cloud de niveau intermédiaire,** proposent une infrastructure de taille plus réduite et peuvent se concentrer sur des services spécifiques (tels que le cloud privé, le cloud souverain, ou encore le cloud destiné à une verticale, entre autres).

La majorité des startups et des scale-ups interrogées pour cette étude privilégient les services de cloud public. Parmi elles, plusieurs adoptent une stratégie dite "**multi-cloud**", utilisant simultanément plusieurs fournisseurs.

Certaines adoptent cette approche de manière opportuniste, en profitant de crédits cloud disponibles (sortes d'offres gratuites), tandis que **d'autres le font délibérément pour des raisons de performances et de confidentialité**. En effet, il est généralement reconnu par les entreprises interrogées que les géants du cloud américains sont actuellement les fournisseurs de services cloud les plus performants sur le marché. Cependant, leurs obligations envers la législation extraterritoriale américaine sur l'accès aux données (telles que le CLOUD Act et la Section 702 de la FISA) soulèvent des inquiétudes parmi les clients européens des start-ups, dès lors qu'elles opèrent dans le secteur privé ou public et traitent des données sensibles (par exemple, les données financières ou de santé). En réponse, les start-ups ont adopté diverses stratégies pour atténuer ce risque.

Ainsi, un nombre considérable de start-ups adoptent une stratégie multi-cloud, **combinant des fournisseurs américains et européens**, en fonction des besoins en matière de confidentialité de leurs clients. Certaines proposent même la possibilité de **déployer leurs logiciels sur l'infrastructure du client**, que ce soit sur site ou dans un cloud privé. Une minorité préfère s'appuyer sur un **cloud privé** ou possède sa **propre infrastructure sur site**, souvent complétée par l'utilisation d'un cloud public pour les tests, la recherche ou d'autres activités non stratégiques.

FDécodeur : un cloud adapté à chaque besoin

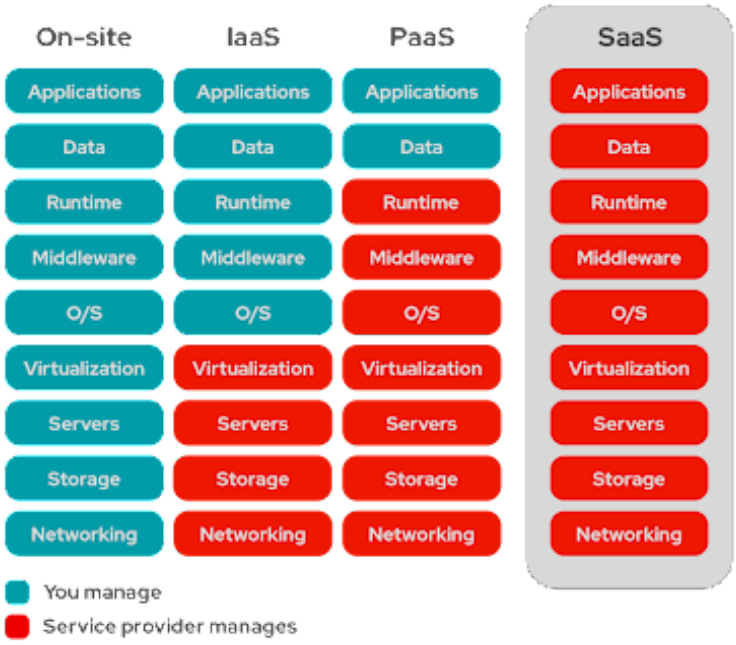
Aujourd'hui, les fournisseurs de cloud proposent aux entreprises trois niveaux de services, couvrant l'ensemble du spectre, depuis l'infrastructure jusqu'au développement et déploiement d'applications :

- **l'Infrastructure en tant que service (ou 'IaaS")** fournit une partie ou la totalité de l'infrastructure de base sous forme virtualisée. Les utilisateurs assument la responsabilité de la gestion de leur propre système d'exploitation, de leurs applications, de leurs données et de leurs configurations.

- la **Plateforme en tant que service ("PaaS")** fournit aux utilisateurs une plateforme de développement et d'exécution pour créer et déployer leurs propres applications.

- le **Logiciel en tant que service ("SaaS")** fournit des applications logicielles hébergées sur le cloud et accessibles aux utilisateurs via Internet, sur une base d'abonnement.

Le choix du modèle de service de cloud dépend des besoins spécifiques de l'entreprise, de ses compétences techniques et de ses objectifs. Par exemple, le niveau IaaS offre à l'utilisateur plus de contrôle sur les données et l'infrastructure, tandis que le SaaS délègue plus de contrôle au tiers.



Le niveau de contrôle de l'infrastructure par l'utilisateur dans les services IaaS, Paas et SaaS
source: RedHat

Certains géants du cloud, tels qu'Azure de Microsoft, ainsi que certains fournisseurs de niveau intermédiaire comme OVHcloud et Scaleway, offrent des solutions IaaS et PaaS dans le cadre de leur offre de **cloud privé**. Ces services permettent aux clients de bénéficier de l'**utilisation exclusive** d'une portion spécifique de l'infrastructure du fournisseur.

Quels sont les risques concurrentiels sur le marché de l'infrastructure ?

La domination des interfaces de programmation de puces les plus utilisées

Les interfaces de programmation des puces, qui constituent le système d'exploitation des puces, sont le premier niveau d'infrastructure où émergent des risques concurrentiels. Tout comme l'emprise de Microsoft Windows sur les ordinateurs personnels dans les années 1990 et au début des années 2000, il existe un risque que certains modèles de programmation des puces dominent le marché et induisent un enfermement des utilisateurs. Le modèle de programmation des puces CUDA de Nvidia illustre ce phénomène. Introduit en 2007, il est devenu la norme de facto pour l'accélération GPU dans le domaine de l'IA. Le succès de cette interface de Nvidia s'explique par sa facilité d'utilisation, son investissement précoce dans des outils, des bibliothèques et des frameworks complémentaires d'IA, ainsi que par ses collaborations avec des universités et des entreprises technologiques. Cependant, en raison de sa nature propriétaire, les bases de code développées sur CUDA demeurent souvent difficiles à migrer vers d'autres modèles de programmation, ce qui entraîne un verrouillage pour les utilisateurs.

Pour réduire la dépendance envers CUDA, des concurrents tels qu'Intel investissent dans des modèles de programmation alternatifs. De plus, des frameworks open source de plus en plus disponibles et interopérables entre les modèles de programmation des puces sont en développement, et des langages neutres pour les GPU font leur apparition. Les GPU d'AMD, qui ne sont pas liés à CUDA, gagnent également en popularité. Cependant, un nouveau risque émerge. **Si Nvidia ne parvient pas à s'adapter à cette nouvelle réalité plus concurrentielle, elle pourrait recourir à des pratiques anticoncurrentielles, à l'instar de Microsoft au début des années 1990** pour maintenir son monopole sur les marchés des systèmes d'exploitation et des navigateurs. **Les autorités de la concurrence doivent donc rester vigilantes quant à l'ouverture de ce marché :** les GPU de Nvidia doivent demeurer compatibles avec des modèles de programmation autres que CUDA, la migration et l'interopérabilité entre les modèles doivent être renforcées, et des alternatives à CUDA doivent avoir une chance équitable de se développer.

Les fournisseurs de puces se lancent au marché du cloud

Face à l'entrée des hyperscalers sur le marché des puces pour réduire leur dépendance à Nvidia, Nvidia lui-même se positionne à son tour sur le marché du cloud pour diversifier sa clientèle. Récemment, elle a amorcé des investissements dans des fournisseurs de cloud spécialisés dans l'IA, tels que Coreweave, par le biais de sa branche d'investissement d'entreprise. 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.

Un verrouillage concurrentiel dans le marché du cloud

Le marché du cloud public est largement dominé par quelques grands acteurs américains, les *hyperscalers*. Ces entreprises, telles qu'Amazon Web Services (AWS), Microsoft Azure et Google Cloud Platform (GCP), détiennent une part de marché collective de 65 % au niveau mondial. En France, cette domination est encore plus prononcée, avec ces trois acteurs accaparant 71 % du marché. Cette concentration excessive soulève des inquiétudes quant à l'équité de la concurrence sur plusieurs fronts. Comme évoqué dans la Couche 1, les *hyperscalers* bénéficient d'accords préférentiels pour l'accès à des ressources informatiques puissantes, comme les GPU. En plus de ces partenariats, ils mettent en place des pratiques commerciales et techniques visant à renforcer artificiellement leur domination sur le marché.

Parmi les pratiques commerciales mises en place par les *hyperscalers* pour verrouiller leurs utilisateurs professionnels, deux se distinguent particulièrement : les crédits cloud et les frais de sortie.

Les crédits cloud, allocations de services gratuits offertes pendant une durée limitée, s'avèrent indéniablement précieux pour les startups, surtout en phase de démarrage. Cependant, sur le long terme, ils peuvent induire une dépendance vis-à-vis d'un fournisseur cloud spécifique. **Les *hyperscalers*, disposant de ressources financières colossales par rapport aux acteurs de taille intermédiaire, sont mieux placés pour proposer des crédits cloud sur des périodes prolongées.** Cette situation, fausse la concurrence et désavantage les petits acteurs. Fait notable, les startups interrogées ont observé **une augmentation significative des offres de crédits cloud ces derniers mois,** dans le sillage de l'engouement pour l'IA. AWS, Azure et GCP se livrent une lutte acharnée pour attirer le plus grand nombre possible de startups.

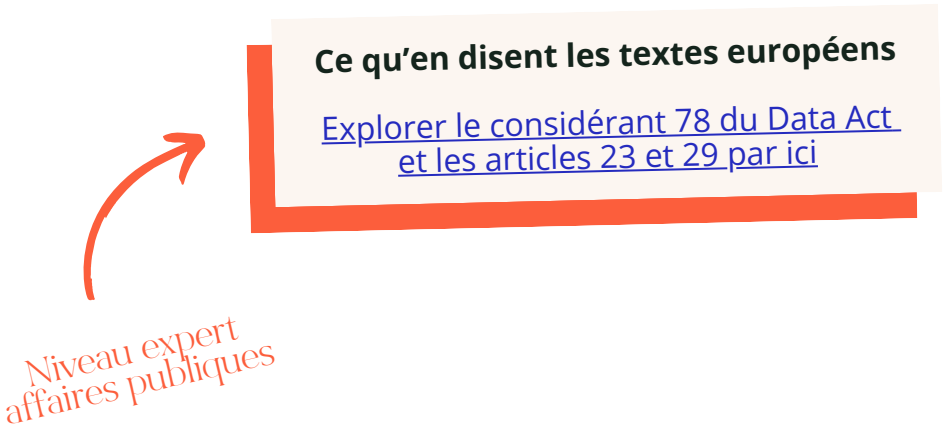
Les frais de sortie, sorte de pénalité facturée pour le transfert de données d'un fournisseur de cloud à un autre, ne sont pas associés aux coûts réels supportés par le fournisseur et constituent par conséquent une pratique abusive. Si tous les fournisseurs de cloud ne recourent pas à ces frais, il n'en demeure pas moins qu'ils figurent souvent dans les clauses contractuelles des principaux *hyperscalers*.

Parmi les pratiques techniques mises en place par les hyperscalers pour verrouiller leurs utilisateurs professionnels, le manque d'intéropérabilité et la restriction d'accès à certains logiciels sont particulièrement prégnants. Les *hyperscalers* peuvent ainsi refuser l'intéropérabilité avec d'autres fournisseurs de cloud, ce qui signifie que les applications et les données ne peuvent pas facilement migrer d'une plateforme à l'autre. Ils peuvent également restreindre l'accès à certaines API, limitant ainsi la capacité des développeurs à créer des applications portables, et enfin, restreindre l'accès à certains logiciels en cas de résiliation de contrat.

Si le règlement européen sur les données (Data Act) vise à limiter les pratiques de verrouillage des *hyperscalers*, **une période de risque subsiste jusqu'à son entrée en vigueur complète*.** Les *hyperscalers* n'ont aucun intérêt à abandonner par anticipation les pratiques lucratives de verrouillage des consommateurs, telles que les crédits cloud et les frais de sortie, alors même que ces pratiques sont déjà qualifiées et prouvées : Google Cloud Platform a annoncé la fin des frais de transfert de cloud, mais a ensuite précisé que cela ne concernait qu'une sélection de clients. AWS a également supprimé les frais de sortie, mais pour certains clients uniquement.

D'autres problèmes non réglementés par le Data Act subsistent. Notamment, **la durée maximale légitime des crédits cloud et les restrictions abusives à l'accès aux logiciels**. Ces lacunes de la législation européenne ouvrent la voie à une réglementation nationale fragmentée, susceptible de créer des distorsions au sein du marché unique et de permettre aux *hyperscalers* de poursuivre des pratiques abusives.

**Certaines dispositions du Data Act ne seront applicables qu'en septembre 2025 et d'autres en janvier 2027.*

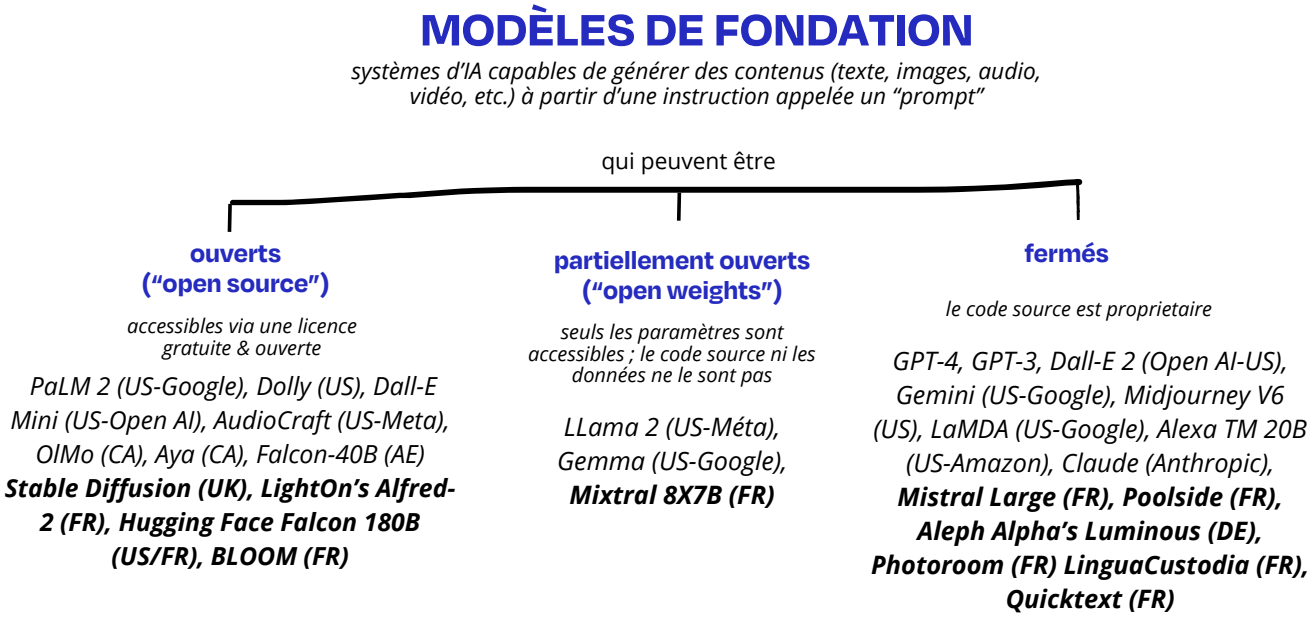


Couche n°3

Les modèles de fondation, un marché avec de fortes barrières à l'entrée

La troisième couche de l'IA générative - les modèles de fondation, un marché avec de fortes barrières à l'entrée

Les principales entreprises productrices de modèles de fondation



Comment ça marche ?

Les modèles de fondation dans l'IA générative sont des **systèmes d'IA capables de produire de nouveaux contenus ou données**, comme du texte, des images, de l'audio ou de la vidéo, à partir d'un ensemble de données et d'instructions fournies par l'utilisateur (appelées "prompt"). Un modèle d'IA est entraîné sur un ensemble de données spécifiques. Cet apprentissage lui permet d'identifier les schémas et les relations entre les données, et de créer du contenu similaire à l'original. Une fois entraîné, le modèle peut être appliqué à n'importe quel nouvel ensemble de données. Il utilise ses connaissances acquises pour générer du contenu inédit original (on parle alors d'inférence).

Les modèles de fondation se distinguent par leur taille et leur échelle. On peut distinguer deux grandes catégories :

- **les modèles à usage général (les "general purpose models")**, comme les Large Language Models (LLMs) tels que GPT-3. Ils sont caractérisés par leur grande taille et leur capacité à traiter de gros volumes de données. Ils sont entraînés sur des ensembles de données massifs et appliquent jusqu'à des milliards de paramètres. De nature généraliste, ils sont capables de produire une large gamme de contenus avec un degré de précision variable sur divers sujets ;

- **les modèles spécialisés**, qui sont de plus petite taille et se focalisent sur des tâches spécifiques. Ils sont entraînés sur des ensembles de données plus restreints, mais sont plus précis que les modèles à usage général.

Ces modèles sont plus ou moins ouverts ; on parle alors de modèles développés en “open source”, “open weights” et “closed source” :

- **Les modèles “open source”** sont mis à la disposition du public sous une licence libre et ouverte qui permet l'accès, l'utilisation, la modification et la distribution du modèle. Les modèles open source rendent l'architecture du modèle, le code d'entraînement, les paramètres ("weights"), les données d'entraînement et les informations sur l'utilisation du modèle publiquement disponibles. Cette transparence permet aux utilisateurs de modifier et d'adapter le modèle. *Exemple : OLMo d'Allen.ai*

- **Les modèles “open weights”** sont accessibles au public, mais avec un niveau d'ouverture moindre par rapport aux modèles open source complets. En effet, ces modèles partagent publiquement leurs paramètres ("weights"), l'architecture du modèle et des informations sur son utilisation. Cependant, contrairement aux modèles open source complets, les données d'entraînement et le code source utilisé pour entraîner le modèle ne sont pas divulgués. *Exemple : 7B de Mistral, LLama de Meta.*

- **Les modèles “closed source”** reposent sur un code source propriétaire qui n'est accessible qu'après le paiement d'une licence, d'un abonnement ou d'une autre forme de compensation financière. Ils peuvent être utilisés et intégrés par des tiers dans leurs logiciels et/ou applications via une API (*exemple : GPT4 d'OpenAI*) ou intégrés exclusivement par un développeur dans son propre logiciel, application, services, etc. (*exemple: le chatbot “Velma” de Quicktext ou l’application de Photoroom*).

Les développeurs ont le choix **entre créer leurs propres modèles d'IA générative ou utiliser des modèles existants** (open source, open weights, closed source). La création d'un modèle d'IA générative requiert principalement trois éléments : la puissance de calcul, les données et le talent. En revanche, les modèles existants sont accessibles soit directement via une API, soit indirectement via un hub de modèles. **Les hubs de modèles sont des plateformes spécialisées qui partagent et hébergent des modèles d’IA.**

Hubs de Modèles

plateformes hébergeant et partageant des modèles

Hugging Face (FR/US)
Github (US-Microsoft)
PyTorch Hub (US-Meta),
Qualcomm AI Hub (US),
TensorFlow Hub (Google, US)
Amazon Bedrock (US)

Des barrières à l'entrée élevées

Nous avons demandé à 30 startups s'il était intéressant qu'elles développent, en propre, des modèles de fondation à usage général. 28 d'entre elles n'en sont pas convaincues, considérant ce projet trop coûteux - tant financièrement que techniquement - et incertain en termes de retour sur investissement.

Il y a en effet déjà une grande disponibilité de solutions propriétaires prêtes à l'emploi ainsi que des modèles personnalisables en "open source" et "open weight" sur le marché. Les quatre investisseurs interrogés ne sont également pas convaincus de l'opportunité de financer une pluralité de modèles de fondation à usage général, pour les mêmes raisons - et notamment l'intense besoin d'investissement initial en capital. Seuls quelques modèles - les premiers développés - attirent et concentrent les capitaux.

Le phénomène est différent s'agissant des modèles d'IA générative spécialisés, plus petits en termes de paramètres et de données requises, donc moins coûteux à produire, et moins gourmands en ressources - notamment énergétiques - que les modèles à usage général. Le coût à l'entrée étant moins élevé d'un point de vue financier, ces modèles sont plus faciles à développer pour les startups, et plus attractifs pour des investisseurs privés.

Les startups et les investisseurs ont exprimé une nette préférence pour les modèles d'IA générative spécialisés. Cependant, seules trois startups interrogées (LinguaCustodia, Photoroom et Quicktext) développent leur modèle de base spécialisé et l'intègrent dans leur produit. Les autres startups interrogées préfèrent intégrer l'IA générative dans leur offre en s'appuyant sur des modèles disponibles, en les adaptant et les améliorant à l'aide de techniques telles que le fine tuning ou le RAG.

Pourquoi le développement de modèles d'IA générative est-il si compliqué ?

Les startups évoquent les raisons suivantes, en sus de l'accès au financement :

- **L'accès à la puissance de calcul :** il existe une corrélation directe entre la taille d'un modèle de fondation et la puissance de calcul nécessaire. Pour illustrer cela, les chercheurs estiment que la formation d'un grand modèle de langage (LLM) similaire à GPT-3 a nécessité environ 1 024 GPU NVIDIA A100 pendant 34 jours. Bien qu'il n'y ait pas de données officielles sur la puissance de calcul totale utilisée pour former GPT-4 - dont la taille est nettement plus importante que GPT-3 -, nous savons qu'OpenAI a utilisé le supercalculateur de 10 000 GPU de Microsoft Azure, et Intel estime que le coût de développer de GPT-4 pourrait se chiffrer en milliards de dollars.

- **L'accès aux ressources énergétiques** : une utilisation massive de GPU entraîne par ailleurs des coûts d'électricité importants. Par exemple, la formation de ChatGPT-3 a nécessité approximativement 1 283 MWh, ce qui équivaut à la consommation énergétique moyenne d'environ 274 foyers français pendant une année entière. L'inférence nécessite également une intensité énergétique forte. Ainsi, le coût énergétique d'une seule requête sur ChatGPT est estimé à 2,9 Wh. Lorsqu'il est multiplié par le volume de requêtes par jour enregistré au début de 2023, cela équivaut à 564 MWh par jour et 206 GWh par an, soit la consommation d'électricité annuelle d'un pays comme la République centrafricaine. Ce besoin de ressources énergétiques des modèles les plus grands limite l'accessibilité des modèles généraux à quelques entreprises. Cependant, les modèles spécialisés ou le fine-tuning des grands modèles nécessitent beaucoup moins de ressources pour la formation, l'utilisation et l'adaptation, les rendant plus attractifs pour la communauté de l'IA générative.

- **L'accès aux données** : l'accès à une quantité suffisante de données de haute qualité pour entraîner des modèles d'IA est crucial pour les développeurs de modèles d'IA générative - qu'ils soient à usage général ou spécialisé. Les données constituent, avec la puissance de calcul, la matière première de l'IA générative. A titre d'exemple, les grands modèles de langage (LLMs) nécessitent des quantités massives de données pour leur entraînement (plusieurs trillions de jetons de données). Il existe quatre principales catégories de données d'entraînement :

  - **les données commerciales :** il s'agit de contenus de haute qualité (tels que des livres, de la musique, des journaux, des publications scientifiques, etc.). L'accès à ces données est contrôlé par le biais du droit d'auteur, et encadré par des modalités d'accès techniques et contractuelles. *Exemple: Shutterstock propose des APIs, des licences et des abonnements pour accéder à des millions d'images, vidéos, textes, musiques, effets sonores.*

  - **les données utilisateur :** il s'agit d'informations sur l'utilisation d'un certain logiciel ou plateforme (tels que les données d'interaction, de préférences, etc.). *Exemple: Meta a accès aux préférences publicitaires des utilisateurs de Facebook.*

  - **les données en accès libre :** il s'agit de données accessibles, traitables et réutilisables librement. Les startups interrogées pour cette étude soulignent que les données en accès libre actuellement disponibles sont insuffisantes pour être les seules sources d'entraînement des modèles d'IA. *Exemple: les textes législatifs de l'UE mis en ligne sont des données en libre accès.*

  L'accès aux données est un défi complexe pour les entreprises. D'une part, le cadre juridique de l'accès aux données - commerciales, personnelles et en accès libre - est complexe et fragmenté, tant au niveau européen qu'au niveau international. D'autre part, les détenteurs des données utilisateur et des données commerciales sont peu nombreux, le marché est très concentré, et ils ont une position déjà consolidée, créant un risque concurrentiel potentiellement important pour les nouveaux entrants. Nous y reviendrons en page 31.

- **Le nettoyage et la notation des données :** Avant de pouvoir être utilisées pour entraîner ou affiner un système d'IA, les données nécessitent un processus de préparation minutieux par des logiciels et du personnel qualifié. Elles doivent d'abord être nettoyées et étiquetées afin de les rendre compréhensibles et de minimiser autant que possible les biais. Le processus d'entraînement et d'amélioration d'un modèle d'IA ne s'arrête pas à la notation initiale des données. Selon les startups interrogées, le processus de notation des données, pour perfectionner un modèle d'IA spécialisé, peut prendre jusqu'à six mois. Des itérations supplémentaires sont souvent nécessaires pour affiner le modèle et optimiser ses performances. Ces itérations, utilisant des techniques comme l'apprentissage par renforcement, prolongent encore la durée totale du processus de développement.

- **L’accès au talents :** Les professionnels qualifiés dans le domaine de l'intelligence artificielle, notamment les ingénieurs capables de préparer et développer des modèles de fondation, sont rares. Par conséquent, ils sont non seulement coûteux, mais aussi difficiles à recruter. Il y a donc une course à l’attractivité, favorisant les entreprises qui peuvent payer des salaires élevés et proposer des avantages sociaux importants. Au-delà de cet aspect financier, un moyen de rétention des salariés est de pouvoir offrir aux développeurs un écosystème complet de travail (comme le proposent AWS ou GCP) et des systèmes de reconnaissance professionnelle tels que les certifications Google Developer Expert ou les distinctions Github Star.

Quels sont les risques concurrentiels sur le marché des modèles de fondation ?

Contrairement au marché des puces et de l'infrastructure, qui existe depuis des décennies, le marché des modèles d'IA est encore très récent (à titre d'exemple, l'entité commerciale de Open AI existe depuis 2019). Bien qu'il soit trop tôt pour affirmer avec certitude qu'il existe déjà des pratiques anticoncurrentielles sur ce marché, nous avons identifié les risques potentiels.

La complexité des partenariats stratégiques entre géants technologiques et modèles de fondation

L'année 2023 a vu fleurir de nombreux partenariats stratégiques des géants technologiques établies et des pépites de l'IA générative. **Ces partenariats n'ont pas tous le même objet** : il peut s'agir d'investissement en capital, d'accès aux puces et/ou à l'infrastructure (sous forme de clauses d'accès privilégié ou d'accords d'exclusivité), ou d'accès favorisé à un marché de distribution (plateforme ou services des grandes entreprises).

Nous avons résumé ci-dessous les principaux partenariats stratégiques publics au 1er mars 2024 :

Géant technologique	Cocontractant	Nature du partenariat stratégique
Microsoft (US)	OpenAI (US)	• Microsoft investit 13 milliards de dollars (participation de 49 %. Source. • OpenAI accède à un supercalculateur Azure dédié (10 000 GPU). Source. • OpenAI est intégré dans la plateforme de distribution d'Azure. Source.
Microsoft (US)	Mistral (FR)	• Microsoft investit 15 millions de dollars sous forme d'obligations convertibles (participation minoritaire). • Mistral accède à l'infrastructure Azure. • Disponibilité de Mistral dans la market place d'Azure. Source.
Google (US)	Anthropic (US)	• Google investit 2 milliards de dollars américains (participation de 10%). Source. • Disponibilité d'Anthropic via Google Cloud Platform. • Anthropic utilisera les puces Google TPU v5e pour l'inférence en IA. Source.
Amazon (US)	Anthropic (US)	• Amazon investit 4 milliards de dollars américains (participation minoritaire). Source. • Anthropic utilisera les puces AWS. • Disponibilité d'Anthropic sur AWS Amazon Bedrock. Source.
Intel (US)	Stability.ai (US)	• Intel investit 50 millions de dollars américains. • Stability.ai utilisera exclusivement les puces Intel. Source.

Pour l'ensemble des startups interrogées, ces partenariats stratégiques sont une opportunité économique, à condition qu'ils ne soient pas exclusifs - tant au niveau des startups que des géants technologiques. Cela signifie en pratique :

- **veiller à ce que toutes les startups aient une chance égale d'être présentes sur des plateformes** comme celle d'Azure. A ce titre, l'ouverture de l'accès d'Azure aux modèles d'OpenAI et de Mistral est saluée par les startups, car elle illustre le caractère non exclusif des partenariats établis par Azure.

- **garantir aux startups la liberté de nouer des partenariats supplémentaires avec d'autres canaux de distribution ou d'autres fournisseurs de cloud.** Une analyse approfondie sera en outre nécessaire pour déterminer si les fournisseurs de cloud constituent des contrôleurs d'accès (ou "gatekeepers") tels que définis dans le Digital Markets Act, comme le préconise le rapport du Parlement européen de 2023 sur la concurrence dans l'IA générative.

Il est essentiel d'appréhender l'impact à long terme de ces partenariats stratégiques sur la concurrence. Si, à court terme, ils 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 spécifique 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.

Les acquisitions de startups par des géants technologiques

Les acquisitions font partie du cycle de financement classique des startups, dont la structure est composée à la fois d'investissement en capital risque lors des phases d'amorçage et de croissance, puis d'options de sorties (telles que les acquisitions, l'introduction en bourse ou la consolidation - LBO) pour permettre aux investisseurs de récupérer leur mise.

Dans un secteur comme l'intelligence artificielle, où les investissements sont rapidement élevés compte tenu du coût d'entrée dans la technologie, seuls quelques grands acteurs concentrent les capacités financières et l'appétit nécessaires pour mener des acquisitions, et sont localisés principalement aux Etats-Unis et en Asie.

Aussi, toute acquisition devra être analysée avec un impératif d'équilibre entre deux enjeux:

- d'un côté, le **risque de voir une acquisition par un géant technologique consolider sa position dominante** sur le marché en renforçant son intégration verticale, et

- de l'autre côté, la **nécessité pour les startups de disposer de sorties attrayantes** pour rembourser leurs investisseurs - lesquels réinvestiront ensuite dans d'autres entreprises en Europe.

Face à ce double enjeu, la seule vraie option pour l'Europe est de renforcer massivement les capacités d'investissement des fonds privés et publics en Europe, pour offrir des **alternatives viables aux acquisitions par les grands acteurs technologiques étrangers.**

Toutefois, ne nous méprenons pas sur l'alternative "strictement" européenne : dans un marché mondialisé, il est normal et recommandé d'avoir des investisseurs de tous les continents - sauf exception géopolitique - pour élargir son influence économique.

Le verrouillage par l'intégration verticale

Le marché des modèles de fondation ne se limite pas aux nouvelles startups : des acteurs déjà bien établis, comme Google et Meta, y jouent également un rôle actif. Ces entreprises se caractérisent par une intégration verticale complète, s'étendant des puces à l'infrastructure, en passant par les modèles fondamentaux et les applications. Cette **intégration verticale n'est pas sans risque** :

- les entreprises qui adoptent une stratégie d'intégration verticale pourraient jouir d'un avantage concurrentiel déloyal en **exploitant les données utilisateurs** collectées via leurs différents produits, et les utiliser pour mieux vendre leurs modèles IA ;

- ces entreprises pourraient être tentées de **privilégier leurs propres modèles d'IA** par rapport à ceux de leurs concurrents ("auto-préférence");

- enfin, elles pourraient être tentées de **limiter stratégiquement l'accès à leurs modèles d'IA à leurs concurrents** sur des marchés en aval ou adjacents.

Pour prévenir les effets néfastes de la limitation stratégique, il est crucial d'appliquer le principe FRAND (Équitables, Raisonnables et Non Discriminatoires) du Digital Markets Act à l'accès aux modèles d'IA générative.

Des canaux de distribution à maintenir ouverts : le cas des hubs de modèles

Aujourd'hui, les développeurs disposent de plusieurs canaux pour accéder aux modèles de fondation : le site Web du fournisseur, les plateformes de distribution des fournisseurs de cloud et les hubs tiers. Cette diversité doit être préservée pour garantir le choix de l'utilisateur. **Les plateformes de distribution et les hubs de modèles ne doivent pas évoluer vers des environnements fermés**, comme cela a pu être le cas avec les magasins d'application dans l'écosystème de la téléphonie mobile. Les hubs de modèle ne doivent pas devenir les prochains contrôleurs d'accès ("gatekeepers") comme défini dans le DMA.

En effet, l'utilisation de modèles prêts à l'emploi fournis par des fournisseurs cloud dominantes au travers de leurs propres plateformes de distribution comporte certains risques, notamment l'auto-préférence et le verrouillage des utilisateurs. Ainsi, le fournisseur de cloud pourrait privilégier son modèle propriétaire par rapport aux alternatives tierces, et même limiter la migration vers un concurrent. Si la plateforme cloud contrôle tout, des données d'entrée à l'architecture du modèle, en passant par le code et les poids, elle pourrait techniquement entraver l'interopérabilité ou l'exportation du modèle par l'utilisateur.

Il conviendra donc de vérifier que :

- les **conditions** imposées par les plateformes pour la distribution de modèles d'IA ne sont **pas abusives**
- il y a un **partage juste de la valeur** créée entre les plateformes et les développeurs de modèles tiers; et
- **aucun accord d'exclusivité** n'existe entre les plateformes de distribution, les hubs de modèles et les créateurs de modèles.

Il est par ailleurs important d'encourager l'existence de hubs de modèles tiers (par exemple, celui développé par la startup franco-américaine Hugging Face) pour prévenir les pratiques abusives mentionnées précédemment et fournir des alternatives significatives aux places de marché gérées par des entreprises dominantes (comme Amazon Bedrock).

Retrouvez en annexe ce que dit le DMA

en particulier les articles 6(4) et 6(5)

Niveau expert affaires publiques

Les risques d'entraves sur l'accès aux données

Le marché des données est marqué par plusieurs risques d'entraves à la concurrence, dues à un cadre juridique fragmenté, une concentration du marché et la présence d'acteurs consolidés**.**

→ **S'agissant du cadre juridique**, le principal frein concurrentiel réside dans la multiplication des réglementations relatives aux données et au droit d'auteur dans le monde. S'agissant du **droit d'auteur**, nous avons résumé ici quelques exemples :

- Au sein de l'**Union européenne**, la directive 2019/790 sur le droit d'auteur autorise l'exploration automatisée de données ("data mining"). Cependant, les détenteurs de données choisissent souvent de se retirer du champ d'application de la directive, limitant ainsi l'accès à des ensembles de données précieux.
- Aux **États-Unis**, la doctrine du "fair use" (usage équitable) crée une exception au droit d'auteur pour l'utilisation de données à des fins spécifiques. Cependant, son application pour entraîner des modèles d'IA générative a conduit à de nombreux litiges non encore résolus, remettant en question sa légitimité dans ce contexte.
- Le **Japon** dispose d'une large exception au droit d'auteur permettant l'entraînement de modèles d'IA générative à des fins commerciales et non commerciales. Source.

La **protection des données personnelles** fait elle aussi l'objet de réglementations distinctes dans le monde. L'Union européenne se distingue par son cadre juridique, le RGPD, considéré comme le plus protecteur au niveau mondial. Mais son interprétation par les différentes autorités de protection des données à l'échelon national peut être stricte voire en opposition, ce qui place les entreprises européennes dans une situation de désavantage concurrentiel.

→ S'agissant des données des utilisateurs, le marché est concentré autour d'un petit nombre de grandes entreprises technologiques, telles qu'Amazon, Google et Meta. Ces acteurs contrôlent des plateformes majeures destinées aux consommateurs et agissent comme des contrôleurs d'accès ("gatekeepers") pour les autres entreprises souhaitant interagir avec ces utilisateurs. Si le Digital Markets Act vise à atténuer cette position dominante, les startups interrogées dans le cadre de cette étude rapportent que les contrôleurs d'accès continuent d'utiliser diverses stratégies pour limiter l'accès des tiers aux données de leurs utilisateurs. Ces stratégies incluent par exemple l'ajout de nouvelles fonctionnalités complexes, l'arrêt du support technique sur les technologies clés utilisées par les concurrents, ou encore l'utilisation abusive de prétextes comme la protection des données personnelles ou la sécurité pour justifier, d'un point de vue juridique, le blocage de l'accès aux concurrents à cette plateforme de distribution.

→ S'agissant des données commerciales, ce marché est largement contrôlé par des acteurs bien établis et divers : grandes entreprises technologiques telles Google dans le domaine des livres numérisés, mais aussi par des éditeurs comme Axel Springer ou encore un vaste réseau d'intermédiaires détenant des droits dans les industries créatives. Les startups qui entrent sur le marché de l'IA générative sont confrontées à plusieurs pratiques contractuelles anticoncurrentielles de la part des détenteurs de droits, notamment :

- **des termes contractuels abusifs :** 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.

- **des partenariats exclusifs :** le recours à des clauses d'exclusivité et de priorité par les grands acteurs pourrait verrouiller les fournisseurs de données, limitant ainsi les opportunités des concurrents. Par exemple, cette pratique pourrait prendre la forme de contrats à prix élevé qui excluent *de facto* les plus petits concurrents avec moins de ressources financières.

Bien que les startups et les investisseurs consultés reconnaissent l'utilité des partenariats, ils soulignent que ces derniers ne constituent pas une solution pérenne. En effet, ils estiment qu'une **mise à jour des règles du droit d'auteur à l'ère de l'intelligence artificielle est nécessaire, parallèlement à la généralisation de clauses contractuelles équitables, raisonnables et non discriminatoires (FRAND) dans les accords d'accès aux données.** Cette approche est en ligne avec les principes énoncés dans le Data Act.

Retrouvez en annexe ce que dit le Data Act

Et en particulier l'article 13

Les clauses abusives dans les contrats de travail

Le recrutement est l'un des trois principaux freins à la croissance des startups depuis de nombreuses années en France. La pénurie de profils qualifiés et la concurrence des grandes entreprises technologiques sont citées comme les obstacles les plus récurrents à cet égard. La complexité administrative de l'embauche transfrontalière au sein de l'UE aggrave encore ce problème. Cela rend le recrutement de talents étrangers plus difficile et coûteux.

Bien qu'aucun membre de France Digitale n'ait signalé de barrières anti-concurrentielles directement liée au recrutement, il est important de noter que les grandes entreprises technologiques américaines font l'objet d'une attention accrue pour deux pratiques : l'insertion de clauses de non-concurrence dans les contrats de travail, et la conclusion d'accords de non-débauchage.

Une **clause de non-concurrence** est une clause classique du contrat de travail qui empêche un employé d'exercer, dans ses futures fonctions, une profession ou une activité similaire et directement concurrente à celle de son précédent employeur (Source). Bien qu'autorisées sous certaines conditions aux États-Unis et en Europe, **des études universitaires ont démontré que les clauses de non-concurrence ont été utilisées de manière anti-concurrentielle par les grandes entreprises technologiques aux États-Unis** (source), ce qui a incité la Federal Trade Commission à proposer une règle sur les clauses de non-concurrence en 2023. Si elle venait à être adoptée aux Etats-Unis, cette règle interdirait les clauses de non-concurrence à moins que le nouvel employeur ne détienne au moins 25% du précédent employeur. D'autres clauses restrictives d'emploi, telles que les accords de non-divulgation et les accords de non-sollicitation de clients, demeureraient légaux.

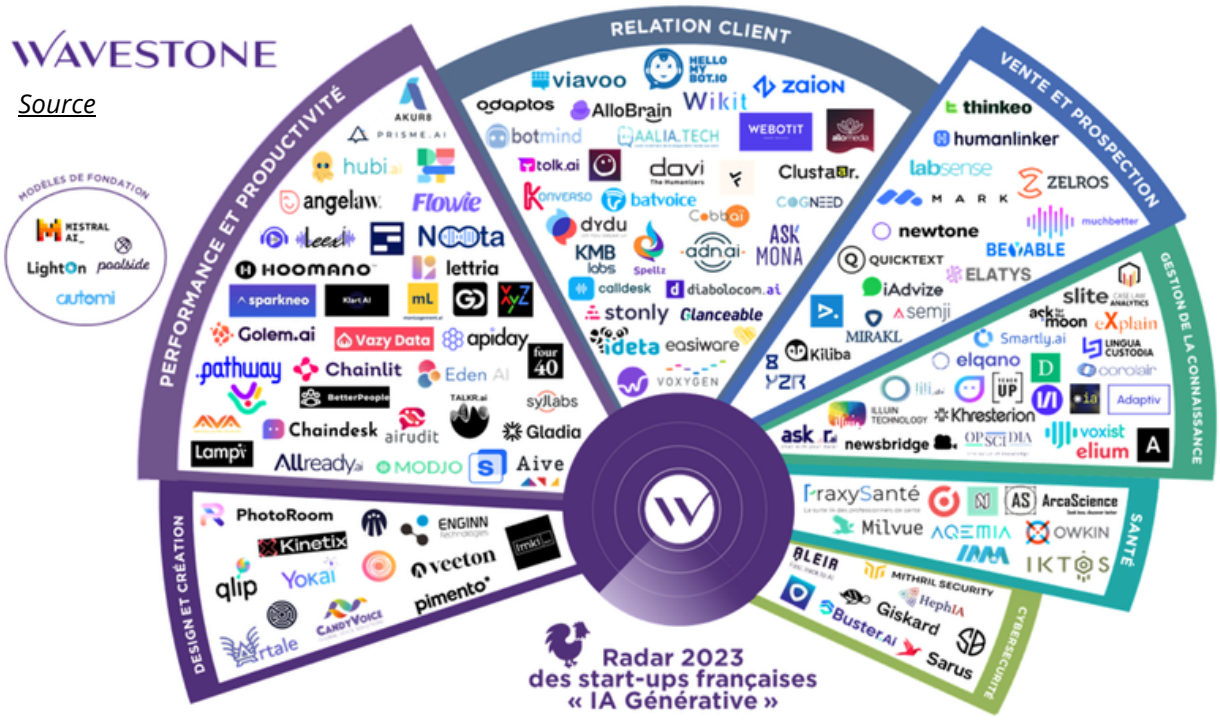
Un **accord de non-débauchage** est une entente entre plusieurs entreprises visant à s'abstenir mutuellement de recruter les employés des entreprises signataires de l'accord. En Europe et aux États-Unis, les accords de non-débauchage sont généralement illégaux, sauf exception pour les coentreprises. Des cas d'accords de non-débauchage dans le secteur technologique ont été recensés, notamment aux États-Unis. En 2010, le Département de la Justice américain a porté plainte contre Apple, Adobe et Google pour ce motif, mais l'affaire s'est soldée par un règlement à l'amiable (source). À ce jour, aucun cas d'accord de non-débauchage n'a été détecté entre entreprises technologiques en Europe. Néanmoins, une surveillance accrue de ce domaine est recommandée, en particulier en tenant compte des retours d'expérience des startups.

Couche n°4

Les applications, un marchè dynamique dans l'ombre des gèants technologiques

La quatrième couche de l'IA générative - les applications, un marché dynamique dans l'ombre des géants technologiques

Le mapping des startups françaises de l'IA générative



Comment ça marche ?

Les développeurs d'applications en IA générative jouent un rôle crucial dans le déploiement de cas d'usage pour les utilisateurs finaux. À travers des applications telles que les chatbots, le remplissage automatique de documents et la correction d'images, les modèles d'IA générative trouvent des utilisations concrètes. On distingue deux types de développeurs d'applications en IA génératives :

- **ceux qui développent des applications de bout en bout** : ils développent leurs propres modèles d'IA personnalisés pour répondre aux besoins spécifiques de leurs propres applications ;

- **les intégrateurs de modèles tiers** : ils utilisent des modèles d'IA pré-entraînés développés par des tiers pour enrichir les fonctionnalités de leurs applications. Ils peuvent intégrer le modèle d'IA sous sa forme originale ou l'adapter en utilisant diverses techniques (fine tuning, RAG, etc.).

Dans les deux cas, il peut s'agit **d'applications à usage général** (comme ChatGPT qui peut répondre à des questions sur n'importe quel sujet) ou **d'applications spécialisées** (comme le chatbot Velma de Quicktext, qui ne peut répondre qu'aux requêtes liées aux hôtels).

Quels sont les risques concurrentiels sur le marché des applications ?

Le stade de la chaîne d'approvisionnement en IA générative où la concurrence est la plus intense est celui des applications. C'est là que se concentrent le plus grand nombre d'acteurs, comme en témoigne le dynamisme des startups dans ce domaine. En France, par exemple, le nombre de startups qui proposent des applications en IA générative a connu une croissance fulgurante, passant de 86 en janvier 2023 à plus de 135 en janvier 2024.

Un consensus se dégage parmi les startups et les investisseurs du secteur : la valeur principale réside dans les applications hautement spécialisées plutôt que dans les applications à usage général (compte tenu de la domination d'acteurs déjà bien établis sur le marché des applications généralistes). En effet, à ce stade de la chaîne de valeur de l'IA générative, les startups sont confrontées à un double problème de concurrence face aux acteurs établis: le développement d'applications et la distribution d'applications.

Le risque d'auto-préférence ("self-preferencing")

En matière de développement d'applications, **les acteurs intégrés verticalement, tels que Google et Microsoft, représentent une concurrence directe pour les startups, en particulier pour les applications à usage général.**

Google, par exemple, développe son propre chatbot, Gemini, tandis que Microsoft propose l'assistant IA Copilot. Le risque est que ces entreprises intègrent ces applications dans leurs services existants (les suites 365 et Google, respectivement), conférant ainsi un avantage à leurs propres produits.

Cette pratique, connue sous le nom d'auto-préférence ("self-preferencing"), est interdite par le Digital Markets Act (DMA). Pour garantir la concurrence sur ce marché et permettre aux services alternatifs d'émerger, **l'installation et l'intégration d'applications tierces sur les plateformes systémiques doivent toujours être autorisées et valorisées sur les plateformes de distribution au même niveau que les produits des distributeurs**. Afin d'évaluer si des services comme Google Suite et Microsoft 365 constituent des contrôleurs d'accès ("gatekeepers") au sens du DMA, une enquête approfondie peut être nécessaire.

Retrouvez en annexe ce qu'en dit le DMA

Et en particulier les articles 6 (4), 6 (5), 6 (6).

Niveau expert affaires publiques

La relation ambiguë avec les distributeurs d'applications

En matière de distribution d'applications, les startups évoluent dans une relation à double tranchant avec les acteurs établis.

D'un côté, les startups tirent profit et cherchent activement à être référencées sur les plateformes de ces acteurs, comme les fournisseurs d'infrastructure. Cela leur permet d'accéder à une large base d'utilisateurs potentiels. Des partenariats tels que celui entre Mistral et Microsoft Azure ou Voxist et OVHcloud sont donc très appréciés par les startups car ils constituent un excellent moyen d'acquérir de nouveaux clients.

De l'autre, cela pose des questions sur le rôle des acteurs établis en tant que portes d'entrée pour l'acquisition de clients par les startups. Cette situation pourrait les amener à être considérés comme des contrôleurs d'accès ("gatekeepers") au sens du DMA. De plus, des questions subsistent concernant les services tiers référencés et leur intégration aux plateformes des acteurs établis. À ce jour, les startups collaborent souvent via des partenariats bilatéraux. Ces partenariats doivent rester non exclusifs et être négociés et rédigés selon les conditions FRAND.

En ce qui concerne la distribution via les magasins d'applications, il est nécessaire de maintenir sa régulation dans le cadre du DMA afin de préserver l'équité.

Lexique

Accord de non-débauchage : accord entre des entreprises pour ne pas débaucher les employés de l'autre.

Apprentissage profond (“deep learning”) : sous-catégorie de l'apprentissage automatique utilisant des réseaux neuronaux artificiels à plusieurs couches pour extraire des caractéristiques et apprendre des données.

Auto-préférence : comportement anticoncurrentiel dans lequel une entreprise priorise ses propres produits ou services par rapport à ceux d'un tiers.

Circuits Intégrés Spécifiques à une Application (“ASIC”) : type de circuit intégré conçu pour une tâche ou un cas d'utilisation spécifique.

Coeur (“Core”) : unité à l'intérieur de la puce capable d'exécuter des instructions et d'effectuer des calculs. L'utilisation de plusieurs cœurs facilite le traitement parallèle, permettant ainsi à la puce d'exécuter plusieurs tâches simultanément.

Conception assistée par ordinateur (CAO) : outils logiciels pour concevoir des systèmes électroniques, y compris des circuits intégrés.

Contrôleur d'accès (“gatekeeper”) : entreprise fournissant des services de plateforme critiques.

Coopétition : alliance stratégique entre des concurrents pour des bénéfices mutuels, souvent dans le domaine des logiciels et du matériel.

Entrainement : processus de création d'un modèle d'IA générative en appliquant des architectures d'apprentissage automatique et d'apprentissage profond à un ensemble de données apportées.

Fonderie : usine hautement spécialisée où sont fabriqués les circuits intégrés.

Fine-Tuning : processus qui consiste à affiner un modèle pré-entraîné sur une tâche spécifique pour améliorer ses performances pour un usage particulier.

Fournisseur de propriété intellectuelle : entreprise proposant une propriété intellectuelle qui peut être intégrée dans des conceptions de puce ou des composants.

Hyperscaler : entreprise offrant une infrastructure massive capable d'entraîner des modèles à usage général et une gamme complète de services cloud (IaaS, PaaS et SaaS).

Inférence : processus par lequel un modèle d'IA pré-entraîné génère du nouveau contenu basé sur un ensemble de données externes.

Intégration verticale : stratégie où une entreprise contrôle plusieurs étapes de la chaîne de valeur d'un produit ou service.

Interface de programmation d'une puce : système d'exploitation des puces, logiciel utilisé pour programmer et contrôler leur comportement.

Plateforme de distribution : plateforme où les utilisateurs peuvent acheter, vendre ou louer divers produits et services.

Mémoire Vidéo à Accès Aléatoire (VRAM) : type de mémoire spécifiquement conçu pour une utilisation dans les GPU optimisés pour le transfert de données à haute vitesse.

Prompt : entrée ou instruction fournie à un système d'IA pour générer une réponse ou effectuer une tâche.

Retrieval-Augmented-Generation (RAG) : technique permettant à l'IA de récupérer des informations à partir d'un ensemble de données externes.

Unités Centrales de Traitement ("CPU") : type de circuits intégrés chargés d'exécuter des instructions et d'effectuer des calculs.

Unités en virgule flottante ("FPU") : type de circuit intégré spécialisé dans les opérations mathématiques complexes, principalement utilisé pour entraîner des modèles d'IA.

Unité de traitement graphique ("GPU") : type de circuit intégré avec plusieurs cœurs, conçu pour gérer plusieurs processus de calcul simultanément (« traitement parallèle »). Principalement utilisé pour entraîner des modèles d'IA.

Vente liée ("Bundling") : pratique consistant à vendre plusieurs produits et/ou services dans un seul ensemble ou package.

Annexe - Ce qu'en disent les textes européens

Data Act

Considérant 78. Les crédits cloud ne doivent pas verrouiller les utilisateurs. [...] Les clients bénéficiant d'offres gratuites devraient également bénéficier des dispositions relatives au changement de fournisseur prévues par le présent règlement, de sorte que ces offres n'entraînent pas un effet de verrouillage pour les clients.

Article 13. Clauses contractuelles abusives imposées unilatéralement à une autre entreprise
1. Une clause contractuelle concernant l'accès aux données et l'utilisation des données ou la responsabilité et les voies de recours en cas de violation ou d'extinction d'obligations liées aux données qu'une entreprise a imposée unilatéralement à une autre entreprise ne lie pas cette dernière entreprise si elle est abusive.

2. Une clause contractuelle qui reflète des dispositions impératives du droit de l'Union ou des dispositions du droit de l'Union qui s'appliqueraient si les clauses contractuelles ne réglaient pas la question n'est pas considérée comme étant abusive.

3. Une clause contractuelle est abusive si elle est d'une nature telle que son utilisation s'écarte manifestement des bonnes pratiques commerciales en matière d'accès aux données et d'utilisation des données, contrairement à la bonne foi et à un usage loyal.

4. En particulier, aux fins du paragraphe 3, une clause contractuelle est abusive si elle a pour objet ou pour effet:
- a) d'exclure ou de limiter la responsabilité de la partie qui a unilatéralement imposé la clause en cas d'actes intentionnels ou de négligence grave;
- b) d'exclure les voies de recours dont dispose la partie à laquelle la clause a été unilatéralement imposée en cas d'inexécution d'obligations contractuelles ou la responsabilité de la partie qui a unilatéralement imposé la clause en cas de manquement à ces obligations;
- c) de donner à la partie qui a unilatéralement imposé la clause le droit exclusif de déterminer si les données fournies sont conformes au contrat ou d'interpréter toute clause contractuelle.

5. Aux fins du paragraphe 3, une clause contractuelle est présumée être abusive si elle a pour objet ou pour effet:
- a) de limiter de manière inappropriée les voies de recours en cas d'inexécution des obligations contractuelles ou la responsabilité en cas de manquement à ces obligations, ou d'étendre la responsabilité de l'entreprise à laquelle la clause a été imposée unilatéralement;
- b) de permettre à la partie qui a imposé unilatéralement la clause d'avoir accès aux données de l'autre partie contractante et de les utiliser d'une manière qui porte gravement atteinte aux intérêts légitimes de l'autre partie contractante, en particulier lorsque ces données contiennent des données commercialement sensibles ou sont protégées par des secrets d'affaires ou des droits de propriété intellectuelle;

- c) d'empêcher la partie à laquelle la clause a été imposée unilatéralement d'utiliser les données qu'elle a fournies ou générées pendant la durée du contrat, ou de limiter l'utilisation de ces données dans la mesure où cette partie n'est pas autorisée à utiliser ou à enregistrer ces données, à y accéder ou à les contrôler ou à en exploiter la valeur de manière adéquate;
- d) d'empêcher la partie à laquelle la clause a été imposée unilatéralement de résilier l'accord dans un délai raisonnable;
- e) d'empêcher la partie à laquelle la clause a été imposée unilatéralement d'obtenir une copie des données qu'elle a fournies ou générées pendant la durée du contrat ou dans un délai raisonnable après la résiliation de celui-ci;
- f) de permettre à la partie qui a imposé unilatéralement la clause de résilier le contrat dans un délai excessivement court, compte tenu des possibilités dont l'autre partie contractante dispose raisonnablement pour se tourner vers un service alternatif et comparable et du préjudice financier causé par cette résiliation, sauf s'il existe des motifs sérieux de le faire;
- g) de permettre à la partie qui a imposé unilatéralement la clause de modifier substantiellement le prix indiqué dans le contrat ou toute autre condition de fond liée à la nature, au format, à la qualité ou à la quantité des données à partager, lorsqu'aucun motif valable ou aucun droit pour l'autre partie de résilier le contrat dans le cas d'une telle modification n'est stipulé dans le contrat.

Article 23. Suppression des obstacles à un changement de fournisseur effectif.

Les fournisseurs de services de traitement de données prennent les mesures prévues aux articles 25, 26, 27, 29 et 30 afin de permettre aux clients de changer de fournisseur pour passer à un service de traitement de données, couvrant le même type de service, qui est fourni par un fournisseur de services de traitement de données différent, ou passer à une infrastructure TIC sur site, ou, le cas échéant, recourir simultanément à plusieurs fournisseurs de services de traitement de données. En particulier, les fournisseurs de services de traitement de données n'imposent pas d'obstacles et suppriment les obstacles précommerciaux, commerciaux, techniques, contractuels et organisationnels, qui freinent les clients dans les démarches suivantes:

- a) la résiliation, après le préavis maximal et l'achèvement avec succès du processus de changement de fournisseur, conformément à l'article 25, du contrat portant sur le service de traitement de données;
- b) la conclusion de nouveaux contrats avec un fournisseur de services de traitement de données différent couvrant le même
- c) le portage des données exportables et des actifs numériques du client vers un fournisseur de services de traitement de données différent ou vers une infrastructure TIC sur site, y compris après avoir bénéficié d'une offre gratuite;
- d) conformément à l'article 24, la réalisation de l'équivalence fonctionnelle lors de l'utilisation du nouveau service de traitement de données dans l'environnement TIC d'un fournisseur de services de traitement de données différent couvrant le même type de service;
- e) le découplage, lorsqu'il est techniquement possible, des services de traitement de données visés à l'article 30, paragraphe 1, des autres services de traitement de données fournis par le fournisseur de services de traitement de données.

Article 29. Suppression progressive des frais de changement de fournisseur.

1. À compter du 12 janvier 2027, les fournisseurs de services de traitement de données ne peuvent imposer aucun frais de changement de fournisseur au client pour le processus de changement de fournisseur.
2. À compter du 11 janvier 2024 et jusqu'au 12 janvier 2027, les fournisseurs de services de traitement de données peuvent imposer des frais de changement de fournisseur réduits au client, pour le processus de changement de fournisseur.
3. Les frais de changement de fournisseur réduits visés au paragraphe 2 ne dépassent pas les coûts supportés par le fournisseur de services de traitement de données qui sont directement liés au processus de changement de fournisseur concerné.

Digital Markets Act

Art 6 (4) Accès à un logiciel tiers.
Le contrôleur d'accès autorise et permet techniquement l'installation et l'utilisation effective d'applications logicielles ou de boutiques d'applications logicielles de tiers utilisant ou interopérant avec son système d'exploitation, et permet l'accès à ces applications logicielles ou boutiques d'applications logicielles par des moyens autres que les services de plateforme essentiels concernés du contrôleur d'accès. Le cas échéant, le contrôleur d'accès n'empêche pas une application logicielle ou boutique d'application logicielle de tiers téléchargée d'inviter les utilisateurs finaux à choisir s'ils souhaitent utiliser par défaut ladite application logicielle ou boutique d'application logicielle téléchargée. Le contrôleur d'accès permet techniquement aux utilisateurs finaux qui choisissent d'utiliser par défaut ladite application logicielle ou boutique d'application logicielle téléchargée de procéder facilement à ce changement.

Rien n'empêche le contrôleur d'accès de prendre, dans la mesure où elles ne vont pas au-delà de ce qui est strictement nécessaire et proportionné, des mesures visant à éviter que les applications logicielles ou les boutiques d'applications logicielles de tiers ne compromettent l'intégrité du matériel informatique ou du système d'exploitation qu'il fournit, à condition que ces mesures soient dûment justifiées par le contrôleur d'accès.

En outre, rien n'empêche le contrôleur d'accès d'appliquer, dans la mesure où elles ne vont pas au-delà de ce qui est strictement nécessaire et proportionné, des mesures et des paramètres autres que les paramètres par défaut permettant aux utilisateurs finaux de protéger efficacement la sécurité en ce qui concerne les applications logicielles ou les boutiques d'applications logicielles de tiers, à condition que ces mesures et paramètres autres que les paramètres par défaut soient dûment justifiés par le contrôleur d'accès.

Art 6 (5) Traitement préférentiel.
Le contrôleur d'accès n'accorde pas, en matière de classement ainsi que pour l'indexation et l'exploration qui y sont liées, un traitement plus favorable aux services et produits proposés par le contrôleur d'accès lui-même qu'aux services ou produits similaires d'un tiers. Le contrôleur d'accès applique des conditions transparentes, équitables et non discriminatoires à ce classement.

Art 6 (6) Migration vers un autre logiciel.
Le contrôleur d'accès ne restreint pas techniquement ou d'une autre manière la capacité des utilisateurs finaux de changer d'applications logicielles et de services qui sont accessibles en utilisant les services de plateforme essentiels du contrôleur d'accès et de s'y abonner, y compris en ce qui concerne le choix des services d'accès à l'internet pour les utilisateurs finaux.

En savoir plus

FRANCE DIGITALE

Fondée en 2012, France Digitale est la plus grande association de startups en Europe, avec plus de 2000 membres startups et investisseurs français du numérique.

Nos missions ?

- Faire émerger des champions du numérique en Europe ;
- Porter la voix et fédérer celles et ceux qui innovent pour changer la face du monde ;
- Créer des ponts et des opportunités de business entre tous les acteurs de l'innovation qu'ils soient grands groupes, décideurs publics, investisseurs, entrepreneurs ou salariés des startups et scale-ups.

Comment ?

- En connectant les acteurs de l'écosystème lors de rencontres de qualité, de rendez-vous d'affaires ou même d'auditions institutionnelles ;
- En faisant gagner du temps aux entrepreneurs grâce au partage d'expériences, de bonnes pratiques et des meilleurs outils ;
- En lançant des campagnes de sensibilisation auprès des politiques et du grand public (pour que même notre famille comprenne ce qu'on fait !).

Pour suivre nos actualités, abonnez-vous à notre newsletter.

Les thématiques IA et affaires publiques vous intéressent ?
Adhérez à France Digitale et rejoignez nos actions !

Chaque année, France Digitale c'est :

- 50+ événements : des FDTour pour lever des fonds partout en France, des C-level days pour fédérer les professionnels des startups et scalups ou encore le FDDay, grande réunion annuelle pour tout l'écosystème ;
- 1,000+ rencontres entre des startups et des investisseurs et 450+ rencontres avec des grands groupes et des décideurs publics ;
- +20 mappings et décryptages des sujets de fond et réglementaires, et même une newsletter avec le meilleur de l'actu tech pour toujours garder un temps d'avance.

L'objectif pour les années à venir ?
Accompagner la croissance de l'écosystème au service de l'Homme et de la planète pour faire de l'Europe le leader de l'innovation responsable.

France Digitale
Jouons collectif

Méthodologie

Cette étude a été réalisée sur la base de recherches documentaires et d'entretiens qualitatifs exclusifs avec une quarantaine d'entreprises, à savoir 30 startups et scale-ups (dont deux concepteurs de puces), 4 fonds de capital-risque et 2 fournisseurs de cloud européens.

Liste des entreprises interrogées :

Alpha Intelligence Capital, Buster.ai, Case Law Analytics, Cleyrop, Criteo, Doctrine, Dust, Elaia, Emocio.hr, Explain, Flex.ai, Flowie, Giskard, Glanceable, Gleamer, Golem.AI, Hugging Face, IRIS, Lampi.AI, La Forge, LinguaCustodia, Malt, Mistral, OVHcloud, PhotoRoom, :probabl, Quicktext, ReciTAL, Scaleway, Serena, SiPearl, Stonly, Suzan.ai, Voxist, Welcome to the Jungle et XXII

Auteurs

Marianne TORDEUX BITKER

> Directrice Affaires Publiques chez France Digitale

Agata Hidalgo

> Responsable Affaires Publiques chez France Digitale

Gaël Gutierrez

> Policy Analyst chez France Digitale

ap@francedigitale.org