§ IA·2026·001

NOTE DE RÉFÉRENCE · N°001 · 15 JUIN 2026 · RÉF. IA·2026·001

Accès aux modèles, contrôle export et souveraineté IA : les nouvelles dépendances de l'intelligence artificielle

Objet. Comprendre comment l'accès aux modèles, aux infrastructures et aux API d'intelligence artificielle est en train de devenir une question de souveraineté, de continuité d'activité et de gestion des risques pour les organisations européennes. Et cartographier les dépendances que les décideurs doivent désormais savoir identifier.

Par Stéphane Nachez, Président d'IntelligenceArtificielle.com

VERSION & MISES À JOUR

Note de référence – version 1.1 – 15 juin 2026 Cette note est mise à jour au fil des évolutions réglementaires, industrielles et géopolitiques. Dernière mise à jour : 15 juin 2026.

Principales évolutions depuis la v1.0 (14 juin) : publication par l'administration américaine (David Sacks) de sa version des faits, révélant un bras de fer documenté avec Anthropic ; mise en lumière de la dimension « accès chinois » comme motivation ; précision du déroulé (directive reçue le vendredi 12 juin, réactions et riposte le 13) ; statut de rétablissement et contexte d'introduction en Bourse d'Anthropic. **Synthèse de la couche.** Aucune chaîne entièrement européenne, du silicium à l'agent, n'existe à ce jour ; la rupture la plus dure reste matérielle. Mais, pour un grand nombre d'usages d'entreprise, une architecture souveraine *suffisante* est déjà assemblable aujourd'hui : un modèle Mistral en open weights ou un déploiement Pharia, sur cloud SecNumCloud (OVHcloud, Scaleway, Outscale), avec un index vectoriel européen. Les verrous résiduels (matériel et clauses contractuelles) sont à identifier explicitement.

Publié sur intelligenceartificielle.com · https://intelligenceartificielle.com/notes/acces-modeles-controle-export-souverainete-ia/

Résumé exécutif

- **Un précédent inédit.** Le vendredi 12 juin 2026, le Département du commerce américain a ordonné à Anthropic de suspendre l'accès à ses modèles les plus avancés, Fable 5 et Mythos 5, pour tout ressortissant étranger : sont visés aussi bien les employés non-américains d'Anthropic que les personnes présentes sur le sol américain. Pour se conformer, Anthropic a désactivé ces deux modèles **pour l'ensemble de ses clients**. C'est, à la connaissance des observateurs, la première fois que les États-Unis appliquent un contrôle à l'export à l'**accès à un grand modèle de langage**, et non plus seulement aux puces.
- **Un différend documenté, pas un simple coup de force** (**mise à jour 15 juin**). L'administration américaine (David Sacks) affirme avoir demandé à Anthropic de corriger une faille de Fable 5 ou de retirer le modèle, ce qu'Anthropic aurait refusé ; Anthropic conteste la gravité de la faille et dit n'avoir eu que 90 minutes. La motivation de fond invoquée, le risque d'**accès chinois** aux capacités cyber de Mythos, rattache l'épisode à la logique de contrôle export visant la Chine. Au 15 juin, l'accès n'est pas rétabli, et l'affaire pèse sur l'introduction en Bourse imminente d'Anthropic.

- **Le déclencheur est cyber, pas géopolitique au sens classique.** Anthropic indique que le gouvernement aurait eu connaissance d'une technique de « jailbreak » de Fable 5 liée à des capacités cyber. Après examen, l'entreprise estime les vulnérabilités mineures, déjà connues, et applicables à d'autres modèles publics comme GPT-5.5, qui ne sont, eux, pas visés.
- **Ce n'est pas un épisode isolé.** La décision intervient onze jours après l'Executive Order du 2 juin 2026 instaurant une revue gouvernementale (volontaire) des capacités *cyber* des modèles « frontier », et s'inscrit dans une trajectoire de durcissement entamée avec les contrôles sur les semi-conducteurs. L'accès aux modèles de pointe entre dans le champ des actifs stratégiques, au même titre que les puces, le cloud critique ou les technologies duales.
- **Une onde de choc politique transpartisane en Europe.** L'épisode a suscité des réactions convergentes de responsables français de sensibilités opposées (Bruno Retailleau, Benjamin Haddad, Édouard Philippe, Jordan Bardella), britanniques (Al Carns, Tom Tugendhat) et néerlandais (Geert Wilders), tous appelant à réduire la dépendance technologique et à accélérer l'investissement dans l'écosystème européen (réactions sourcées en §1).
- **La dépendance reste structurelle et multicouche.** Puces (NVIDIA/AMD), cloud hyperscale (AWS/Azure/GCP), modèles fermés et API propriétaires : aucune chaîne IA entièrement européenne n'existe aujourd'hui. La couche matérielle est le point de dépendance le plus dur. Aucun producteur européen de GPU IA n'opère à l'échelle industrielle.
- **Mais les alternatives européennes ont radicalement mûri.** L'image d'une Europe limitée à de « petits modèles » est dépassée : Mistral Large 3 (675 milliards de paramètres en architecture MoE) et Mistral Medium 3.5 (128 milliards, auto-hébergeable) constituent une offre ouverte compétitive à grande échelle, déployable et entièrement maîtrisable, adaptée à un large éventail d'usages d'entreprise ; Aleph Alpha (Pharia) propose des déploiements souverains air-gapped pour le secteur public ; OVHcloud, Scaleway et Outscale offrent du cloud certifié SecNumCloud. À noter toutefois que le haut du classement *open weights* mondial est aujourd'hui occupé par des modèles chinois (cf. §8), et qu'aucun modèle européen ne surpasse les tout meilleurs modèles propriétaires américains sur les tâches les plus complexes.
- **La souveraineté affichée n'est pas toujours la souveraineté réelle.** « Européen » se décline en couches (contrôle capitalistique, hébergement, matériel, financement) dont chacune peut être souveraine ou non. La fusion Cohere–Aleph Alpha place le « champion souverain » allemand sous contrôle majoritaire hors UE ; l'inclusion de S3NS (coentreprise Thales-Google) dans le marché de cloud souverain de la Commission a relancé le débat sur le *sovereignty washing*.
- **Le verrou matériel reste américain (et taïwanais), y compris pour Mistral.** Le « modèle souverain » français entraîne ses systèmes sur des puces NVIDIA (13 800 accélérateurs, campus de 1,4 GW co-financé avec NVIDIA et le fonds émirati MGX) et l'écosystème CUDA. L'Europe développe une souveraineté processeur (SiPearl, CPU « sans interrupteur d'arrêt ») mais pas encore d'accélérateur IA à l'échelle : ses propres supercalculateurs exascale greffent des GPU NVIDIA. Et au-dessus de NVIDIA plane un point de défaillance unique : la fabrication, concentrée chez TSMC. Un modèle déjà diffusé en open weights n'est pas « débranchable » à distance ; sa génération suivante dépend en revanche de l'accès continu au calcul.
- **La Chine a fait de l'open weights une arme de souveraineté.** Les modèles ouverts chinois (Qwen, DeepSeek, GLM) représentent désormais ~30 % des téléchargements mondiaux, devant les États-Unis. Cette stratégie, née des contrôles américains sur les puces, déplace la vraie ligne de fracture : non plus « américain contre européen », mais « **débranchable à distance contre maîtrisable localement** » : API fermée contre poids ouverts auto-hébergeables.

- **Le risque pour l'entreprise est opérationnel, pas idéologique.** Le bon cadrage n'est pas « faut-il quitter les fournisseurs américains ? » mais « quels usages critiques peuvent supporter une interruption, une restriction ou une modification unilatérale d'accès ? ». Les notions de RTO, RPO et lock-in, familières en gestion de continuité, s'appliquent désormais à la couche IA.
- **Cinq trajectoires sont à surveiller**, du durcissement généralisé des contrôles à une coopération internationale encadrée, en passant par la segmentation des modèles selon les profils d'utilisateurs et l'accélération de l'offre européenne. Le plus probable est une escalade graduelle couplée à une montée en puissance européenne.
- **Recommandation cardinale.** Cartographier ses dépendances IA critiques avant d'envisager toute substitution. La diversification, la réversibilité contractuelle et la capacité de repli sur des modèles auto-hébergeables sont les premiers leviers actionnables, bien avant tout débat sur l'indépendance complète.

---

1. Le fait déclencheur : l'affaire Anthropic

Le vendredi 12 juin 2026, Anthropic a annoncé avoir reçu du Département du commerce américain une directive d'export lui ordonnant de suspendre l'accès à ses deux modèles les plus avancés, **Fable 5 et Mythos 5, pour tout ressortissant étranger** (directive reçue à 17 h 21, heure de l'Est). La portée de la directive est large : elle vise non seulement les personnes situées hors des États-Unis, mais aussi tout ressortissant étranger présent sur le sol américain – y compris les employés non-citoyens d'Anthropic.

Face à cette portée, l'entreprise a estimé n'avoir d'autre choix que de **désactiver les deux modèles pour l'intégralité de ses clients**, américains compris, afin de garantir la conformité. Amazon Web Services, qui héberge ces modèles, a indiqué qu'Anthropic lui avait demandé d'en révoquer l'accès pour tous les utilisateurs, dans toutes les régions. Anthropic a précisé que l'accès à ses autres modèles, dont Claude Opus 4.8, n'était pas affecté, et a présenté ses excuses pour cette interruption en indiquant travailler à un rétablissement.

Un déclencheur de nature cyber. Le gouvernement n'aurait pas communiqué le détail de ses préoccupations de sécurité nationale. Anthropic indique cependant croire que les autorités avaient eu connaissance d'une méthode de « jailbreak » de Fable 5 – c'est-à-dire de contournement de ses garde-fous – exploitant des capacités d'analyse de code. Après examen d'une démonstration de cette technique, l'entreprise affirme avoir identifié un petit nombre de vulnérabilités déjà connues et relativement mineures, et estime que le même type de contournement pourrait produire des résultats comparables sur d'autres modèles publics, dont GPT-5.5 d'OpenAI, qui ne sont pas soumis à des contrôles équivalents.

Un précédent, pas une routine. Jusqu'à présent, les contrôles à l'export américains visaient les puces et les serveurs de calcul, principalement à destination de la Chine. C'est la première fois qu'une mesure de ce type frappe directement l'**accès à un grand modèle de langage** et, par ricochet, l'ensemble des clients internationaux d'un fournisseur. La mécanique juridique sous-jacente, la règle dite du *deemed export* (selon laquelle exposer une technologie contrôlée à un ressortissant étranger sur le sol américain équivaut à une exportation), explique pourquoi les propres salariés étrangers d'Anthropic se sont retrouvés exclus du modèle qu'ils avaient contribué à construire.

Un bras de fer documenté (état au 15 juin). L'épisode, d'abord présenté comme un acte gouvernemental unilatéral, s'est révélé être un standoff aux versions contradictoires. Le 13 juin, David Sacks, co-président du Conseil consultatif présidentiel pour la science et la technologie (et ancien « AI czar » de l'administration), a publié la version de l'exécutif : un partenaire de confiance d'Anthropic et du gouvernement (identifié par plusieurs médias comme Amazon) aurait découvert, lors de tests, un « jailbreak » des garde-fous séparant

Fable 5 (grand public) des capacités cyber non bridées de Mythos. L'administration aurait alors demandé à Dario Amodei de corriger la faille ou de retirer le modèle ; selon Sacks, « Dario a refusé », et le contrôle à l'export aurait été pris « à contrecœur ». Sacks ajoute que le rétablissement ne dépend que d'un correctif (« la balle est dans le camp d'Anthropic ») et reproche à l'entreprise une minimisation incompatible avec son image de laboratoire « safety-first ».

La version d'Anthropic diffère : l'entreprise soutient qu'un contournement étroit, déjà connu et présent dans des modèles rivaux, ne justifiait pas le retrait d'un modèle utilisé par des millions d'utilisateurs ; selon une source citée par *Fortune*, elle n'aurait disposé que de 90 minutes pour retirer le modèle, sans communication préalable d'une menace de sécurité nationale. Un fil de fond émerge des deux récits et de la couverture de *Semafor* : la décision serait liée à des craintes d'**accès chinois** aux capacités de Mythos, ce qui réinscrit l'épisode dans la logique de contrôle export visant la Chine, plutôt que dans une simple querelle de sécurité produit.

Cette controverse importe pour une lecture lucide : il ne s'agit ni d'un pur acte hostile, ni d'une politique froidement coordonnée, mais d'une **zone grise réglementaire** où s'entremêlent sécurité cyber, rivalité avec la Chine, et antécédents de friction (Anthropic avait été classée « risque chaîne d'approvisionnement » par le Pentagone en février 2026). Plusieurs analystes proches de l'administration ont par ailleurs jugé la mesure disproportionnée. Quoi qu'il en soit des responsabilités, l'enseignement pour un utilisateur ne change pas : un accès critique peut être suspendu sur décision étrangère, sans préavis utile.

Statut et enjeux. Au 15 juin, l'accès n'est pas rétabli ; les deux parties affichent leur volonté d'un retour rapide, conditionné à un correctif, et les marchés de prédiction situent une restauration plutôt début juillet. L'affaire intervient dans un contexte sensible : Anthropic a déposé début juin un dossier confidentiel d'introduction en Bourse (valorisation de l'ordre de 965 Md$), et la valeur de ses titres pré-IPO a reculé après l'annonce, le risque réglementaire s'invitant désormais dans le récit de la cotation.

L'onde de choc politique européenne. L'épisode a déclenché des réactions convergentes par-delà les clivages, compilées notamment par Euronews (13 juin 2026). En France :

- **Bruno Retailleau**, ancien ministre de l'Intérieur et candidat à l'élection présidentielle de 2027, y voit un signal d'alarme : selon lui, une nation qui dépend des autres pour sa technologie est une nation qu'on peut « débrancher du jour au lendemain ». Il cite Mistral, OVHcloud, Scaleway et ChapsVision comme des atouts français et appelle à « réarmer » la puissance technologique nationale.
- **Benjamin Haddad**, ministre délégué chargé de l'Europe, estime que l'Europe ne peut se contenter d'être un marché ouvert dépendant de technologies « conçues, financées et contrôlées » ailleurs, et plaide pour investir davantage et se doter des moyens de maîtriser ces technologies.
- **Édouard Philippe**, ancien Premier ministre et maire du Havre, qualifie l'IA d'infrastructure critique « aussi essentielle que l'électricité ou Internet », dont les modèles et la puissance de calcul, s'ils échappent à tout contrôle, peuvent être débranchés par d'autres.
- **Jordan Bardella**, président du Rassemblement national et député européen, appelle la France à accélérer son soutien « à la pépite Mistral AI et à tout l'écosystème ».

Hors de France, le constat est identique. Aux Pays-Bas, **Geert Wilders** (Parti pour la liberté, PVV) relie l'affaire à la souveraineté nationale et réclame le rétablissement du modèle. Au Royaume-Uni, le député **Al Carns**, ancien ministre des Forces armées, souligne que des chercheurs, entreprises et hôpitaux britanniques utilisaient le modèle désormais coupé – « ce n'est pas une histoire d'IA, c'est l'histoire de chaque industrie que nous dominions », tandis que **Tom Tugendhat**, ancien ministre de la Sécurité, résume que la souveraineté tient désormais « davantage au code qu'aux canons ». Cette convergence, de la gauche à la droite et d'un pays à l'autre, est le fait politique marquant de l'épisode, et la matière première d'un argumentaire que les décideurs européens vont devoir traduire en décisions concrètes.

2. Pourquoi ce n'est pas qu'un épisode Anthropic

L'affaire est spectaculaire, mais sa portée tient à ce qu'elle révèle d'une tendance de fond : **l'IA de pointe bascule dans la catégorie des actifs stratégiques**, soumis aux mêmes logiques de sécurité nationale que les semi-conducteurs ou les technologies duales. Trois signaux convergents le confirment.

Une trajectoire réglementaire qui s'étend des puces aux modèles. En janvier 2025, l'administration Biden avait publié le *Framework for Artificial Intelligence Diffusion* (« AI Diffusion Rule »), un dispositif ambitieux qui, pour la première fois, contrôlait à la fois l'export de puces de calcul avancées et celui de certains poids de modèles fermés (via une nouvelle classification, ECCN 4E091), tout en classant les pays du monde en trois tiers d'accès. Jugée trop large et nuisible à l'innovation, la règle a été abrogée par l'administration Trump le 13 mai 2025, deux jours avant son entrée en vigueur, au profit d'une approche recentrée sur les semi-conducteurs et la diligence des entreprises. L'épisode de juin 2026 montre que la question des poids de modèles, un temps écartée, est revenue par une autre voie.

Un cadre fédéral qui se referme sur les capacités cyber. Le 2 juin 2026, le président américain a signé l'Executive Order *« Promoting Advanced Artificial Intelligence Innovation and Security »*. Sans imposer de licence ni de pré-autorisation, il instaure un cadre **volontaire** par lequel les développeurs de modèles « frontier » sont invités à donner au gouvernement un accès anticipé (jusqu'à trente jours avant publication) pour une revue de sécurité, et crée un processus de *benchmarking classifié* des capacités cyber des modèles. La directive Anthropic, intervenue onze jours plus tard et précisément motivée par une préoccupation cyber, illustre la tension du moment : un cadre se voulant non contraignant, doublé d'une capacité d'action unilatérale brutale.

Une dépendance qui n'est plus théorique. Tant que l'accès aux modèles relevait du contrat commercial, la souveraineté restait un débat abstrait. L'épisode démontre concrètement qu'en l'absence de capacités de repli, une organisation européenne peut perdre du jour au lendemain l'accès à un outil critique, sur décision d'une administration étrangère, sans préavis et sans recours immédiat.

3. Chronologie

DATE	ÉVÉNEMENT
2024	Multiplication des contrôles américains sur l'export de puces IA avancées (NVIDIA/AMD), principalement vers la Chine.
13–15 janvier 2025	Publication par l'administration Biden de l'« AI Diffusion Rule » : contrôle des puces *et* de certains poids de modèles fermés, avec un cadre de pays à trois tiers. Conformité prévue pour le 15 mai 2025.
13 mai 2025	Le BIS (Département du commerce) abroge l'« AI Diffusion Rule » deux jours avant son entrée en vigueur ; les contrôles antérieurs sur les puces demeurent, une règle de remplacement est annoncée.
17 avril 2026	La Commission européenne attribue, via le dispositif Cloud III (DPS), un marché de cloud souverain de 180 M€ sur six ans à quatre consortiums (OVHcloud/CleverCloud/Post Telecom, STACKIT, Scaleway, et Proximus/S3NS) – évalués pour la première fois selon une grille de souveraineté mesurable (niveaux SEAL).
24 avril 2026	Annonce de la fusion Cohere (Canada) – Aleph Alpha (Allemagne), valorisation d'environ 20 Md$, investissement de 600 M$ du groupe Schwarz ; clôture attendue au second semestre 2026.

DATE	ÉVÉNEMENT
2 juin 2026	Signature de l'Executive Order « Promoting Advanced AI Innovation and Security » : revue gouvernementale *volontaire* des modèles frontier (accès anticipé jusqu'à 30 jours), benchmarking cyber classifié, sans régime de licence.
9 juin 2026	Anthropic met Fable 5 à disposition du grand public (version « Mythos-class » dotée de garde-fous) ; Mythos 5, non bridé, reste réservé à des partenaires sélectionnés.
10 juin 2026	La Commission européenne présente un paquet de mesures visant à renforcer les puces, l'IA et le cloud souverains européens. En parallèle, un « jailbreak » de Fable 5 circule publiquement.
11 juin 2026	OVH Groupe entre en négociations exclusives pour acquérir Gladia, spécialiste français du *speech-to-text* – troisième acquisition IA/sécurité du groupe en 2026 (après Seald et Dragon LLM).
12 juin 2026 (vendredi, 17 h 21 ET)	Directive du Département du commerce : Anthropic désactive Fable 5 et Mythos 5 pour **tous** ses utilisateurs afin de respecter l'interdiction d'accès aux ressortissants étrangers. Les autres modèles (dont Opus 4.8) ne sont pas affectés.
13 juin 2026 (samedi)	Vague de réactions politiques en Europe. David Sacks publie la version de l'administration (Anthropic aurait refusé de corriger ou retirer le modèle) ; la dimension « accès chinois » est mise en avant par la presse.
Depuis le 13 juin	Accès non rétabli ; rétablissement annoncé comme conditionné à un correctif. Contexte : Anthropic a déposé début juin un dossier confidentiel d'IPO (valorisation ~965 Md$), dont les titres pré-cotation ont reculé.

4. Cartographie des dépendances technologiques

Les organisations européennes sont exposées à plusieurs niveaux de dépendance, du plus matériel au plus contractuel. Identifier *où* l'on est exposé est le préalable à toute stratégie de réduction.

Puces et calcul (GPU). C'est le point de dépendance le plus dur. Les accélérateurs dominants proviennent de fabricants américains (NVIDIA, AMD). Aucun producteur européen de GPU IA n'opère à l'échelle industrielle, et même les supercalculateurs européens reposent sur du matériel américain. Les initiatives publiques (EuroHPC, projets de puces souveraines) existent mais ne pèsent pas encore sur le marché. Dépendance résiduelle : **quasi totale**.

Cloud et infrastructures (IaaS/PaaS). Les hyperscalers américains (AWS, Azure, Google Cloud) fournissent l'essentiel de la capacité. Des alternatives européennes solides existent – OVHcloud, Scaleway, Outscale (groupe Dassault Systèmes), ainsi que des offres certifiées SecNumCloud par l'ANSSI, mais avec moins de centres de données et une disponibilité GPU souvent plus contrainte, et en important elles-mêmes les puces. Dépendance résiduelle : **moyenne sur le cloud généraliste, forte sur le calcul GPU**.

Modèles de fondation et API. Les modèles propriétaires les plus capables (GPT-5.x, Claude Opus, Gemini) restent américains et s'accèdent via des API contrôlées par leurs éditeurs. C'est précisément la couche que l'affaire Anthropic vient de fragiliser : interruption possible, conditions modifiables unilatéralement, exposition à des décisions réglementaires étrangères. La nuance majeure (cf. section 6) est que l'offre *ouverte* et *auto-hébergeable* européenne y est désormais une alternative crédible pour de nombreux usages.

Données et RAG. Les architectures de *retrieval-augmented generation* reposent sur des bases vectorielles et des corpus. Des solutions open source européennes existent et sont matures (Qdrant, basée à Berlin ;

Weaviate, basée à Amsterdam), mais l'écosystème de services managés clés en main reste plus développé côté américain. Dépendance résiduelle : **modérée**, surtout liée à l'intégration interne nécessaire.

Agents et orchestration. Les frameworks d'orchestration les plus répandus sont des projets open source à gouvernance internationale. Il n'existe pas encore de solution européenne packagée de bout en bout pour orchestrer des agents complexes en production ; les entreprises assemblent des briques communautaires. Dépendance résiduelle : **structurelle mais peu sensible** (open source, donc non « débranchable » à distance).

Contrats et réversibilité. Les contrats des fournisseurs cloud et IA américains n'intègrent pas toujours de garanties de portabilité (export des index vectoriels, des configurations, des poids fine-tunés). Le verrouillage propriétaire (formats, dépendances d'API, coûts de migration) accroît la dépendance autant que la technologie elle-même.

Juridictions. Données et modèles hébergés hors UE relèvent de juridictions étrangères (CLOUD Act américain, régimes de contrôle export). Même sur un cloud « européen », la dépendance au matériel importé maintient un contrôle partiel. Les dispositifs SecNumCloud, GAIA-X et l'AI Act visent à renforcer la maîtrise, sans pouvoir aujourd'hui garantir qu'aucune couche n'échappe à toute ingérence extra-UE.

Compétences. L'Europe dispose de chercheurs et d'ingénieurs reconnus, mais d'un vivier de spécialistes IA plus restreint qu'aux États-Unis ou en Chine. Les projets de pointe mobilisent souvent une expertise rare, source de dépendance indirecte.

---

5. Grille de risques opérationnels

Le bon cadre d'analyse pour une entreprise n'est pas idéologique mais opérationnel : il s'agit d'appliquer à la couche IA les notions éprouvées de la gestion de continuité.

Interruption de service (**RTO élevé**). Si un modèle ou une API est coupé, le basculement vers une alternative – redéploiement d'un modèle auto-hébergeable, adaptation applicative, re-tests – prend du temps. Sans solution de repli préexistante, le *Recovery Time Objective* peut se compter en jours ou en semaines.

Perte de contexte (**RPO**). Prompts, historiques de dialogue, *embeddings* et logs sont parfois stockés côté fournisseur. Une coupure peut rendre obsolète le dernier point de sauvegarde et empêcher une reprise complète.

Verrouillage technologique (**lock-in**). L'usage intensif d'un modèle ou d'une plateforme unique (prompts optimisés, fine-tuning, API propriétaires) crée une adhérence forte. Changer de fournisseur impose souvent de refaire une partie des développements.

Continuité de conformité. L'arrêt d'un service traitant des données réglementées (santé, finance, secteur public) peut, en soi, créer une rupture de conformité – RGPD, exigences sectorielles, souveraineté des données – surtout en présence de législation extraterritoriale.

Dépendances cachées. Des sous-traitants (agences, éditeurs, startups) peuvent s'appuyer sur des API étrangères à l'insu du donneur d'ordre. Une coupure révèle ces dépendances en cascade et expose à un effet domino.

Risques contractuels et financiers. Pénalités liées à des SLA non tenues, surcoûts d'un fournisseur de substitution en urgence, clauses de force majeure incertaines en cas de décision gouvernementale étrangère : la chaîne de responsabilités (fondeur, cloud, éditeur, intégrateur) est difficile à démêler.

---

6. Alternatives européennes par couche – état réel 2026

Cette section est le cœur opérationnel de la note. Elle décrit l'offre européenne telle qu'elle existe à la date de publication, sans surévaluer ni sous-estimer sa maturité. Les niveaux sont qualitatifs.

Modèles de fondation

L'idée reçue d'une Europe cantonnée à de « petits modèles » est aujourd'hui fausse. **Mistral AI** (France) propose une gamme complète : *Mistral Large 3*, modèle phare en architecture *Mixture-of-Experts* de **675 milliards de paramètres** (41 milliards actifs, fenêtre de contexte de 256 k), publié fin 2025 ; *Mistral Medium 3.5*, modèle dense de **128 milliards de paramètres** auto-hébergeable sur aussi peu que quatre GPU, devenu le modèle par défaut de la plateforme Le Chat ; *Mistral Small 4* (119 milliards, MoE), unifiant raisonnement, multimodalité et code ; et la famille *Ministral 3* (14/8/3 milliards) pour l'embarqué. Plusieurs de ces modèles sont publiés en *open weights* (licences Apache 2.0 ou MIT modifiée), ce qui autorise un déploiement entièrement maîtrisé. Des organisations comme CMA CGM, La Banque Postale ou France Travail figurent parmi les utilisateurs annoncés.

Aleph Alpha (Allemagne, Heidelberg, fondée en 2019) a réorienté sa stratégie de l'API *Luminous* vers la plateforme **PhariaAI** et la famille de modèles **Pharia**, conçues pour des déploiements souverains : on-premise, VPC privé, *air-gapped*, avec exécution native sur le cloud STACKIT du groupe Schwarz. C'est une réponse de premier plan à la question « quel modèle puis-je faire tourner à l'intérieur de mon infrastructure ? ». **Point de vigilance majeur** : la fusion annoncée le 24 avril 2026 avec le canadien **Cohere** (entité combinée valorisée ~20 Md$, dont Cohere détiendrait ~90 %) place ce « champion souverain » allemand sous **contrôle majoritaire hors UE**, au point de heurter des clauses de souveraineté de certains contrats publics allemands exigeant un fournisseur « contrôlé en Europe ». Illustration concrète que la souveraineté d'un acteur n'est pas un acquis figé.

Maturité : **élevée** sur l'open weights auto-hébergeable (Mistral) et le déploiement souverain régulé (Aleph Alpha). Limites : aucun modèle européen propriétaire fermé ne surpasse les tout meilleurs modèles américains sur les tâches les plus complexes ; et, sur le terrain de l'open weights lui-même, le haut du classement mondial est aujourd'hui occupé par des modèles chinois (GLM, Qwen, Kimi, DeepSeek ; cf. §8), Mistral se situant un cran en dessous de cette frontière. L'écosystème d'inférence managée européen reste par ailleurs plus restreint.

Cloud et infrastructure

OVHcloud est le principal acteur européen, opérant un large parc de serveurs et une offre d'IA managée (OVHai). Le groupe mène une stratégie d'intégration verticale par acquisitions : *Seald* (chiffrement de bout en bout, janvier 2026), *Dragon LLM* (spécialisation de modèles pour industries régulées, mars 2026) et *Gladia* (*speech-to-text*, négociations exclusives annoncées le 11 juin 2026). **Scaleway** (groupe Iliad) et **Outscale** (groupe Dassault Systèmes) complètent l'offre, plusieurs services étant certifiés **SecNumCloud** par l'ANSSI.

La référence de marché est désormais le **cadre de souveraineté du cloud** de la Commission européenne, appliqué pour la première fois au marché Cloud III (180 M€ sur six ans, avril 2026). Il classe les fournisseurs sur une échelle mesurable (niveaux SEAL-0 à SEAL-4, huit objectifs pondérés, la chaîne d'approvisionnement comptant pour le poids le plus élevé) : OVHcloud, STACKIT et Scaleway y atteignent SEAL-3. Ce cadre offre la première grille de comparaison objective au-delà du marketing – utile, car les hyperscalers américains contrôlent encore environ 70 % du marché européen du cloud, contre ~15 % pour les fournisseurs européens.

Maturité : **élevée** pour le cloud généraliste et l'hébergement souverain ; **en consolidation** pour la disponibilité GPU à grande échelle. Limite : moins de centres et de capacité GPU que les hyperscalers, et puces importées.

Données, RAG et bases vectorielles

L'écosystème open source européen est mature : **Qdrant** (Berlin) et **Weaviate** (Amsterdam) sont des bases vectorielles de référence, conçues pour des déploiements maîtrisés et compatibles avec les exigences RGPD. Maturité : **élevée** en open source. Limite : moins de services *managés* clés en main que l'offre américaine, donc un effort d'intégration interne plus important.

Agents et orchestration

C'est la couche la moins consolidée côté européen. Des frameworks open source d'assistants conversationnels existent (par exemple **Rasa**, en Allemagne), mais il n'y a pas encore d'équivalent européen packagé de bout en bout pour orchestrer des agents complexes en production. Maturité : **émergente**. Atout paradoxal : reposant sur de l'open source, cette couche n'est pas « débranchable » à distance par un tiers.

Cadres juridiques et certifications

SecNumCloud (ANSSI) qualifie les hébergeurs pour les données sensibles ; **GAIA-X** vise une infrastructure fédérée européenne ; l'**AI Act** européen (règlement 2024/1689, entré en vigueur le 1er août 2024) introduit une classification des usages à risque qui orientera le choix des modèles et infrastructures. Son application est progressive : interdictions depuis février 2025, obligations sur les modèles à usage général depuis août 2025, et surtout entrée en application de l'essentiel des règles (systèmes à haut risque, transparence, début de l'enforcement) au **2 août 2026** (amendes jusqu'à 35 M€ ou 7 % du chiffre d'affaires mondial). Maturité : **en consolidation**. Limite : ces cadres garantissent la localisation et la gouvernance des données, mais ne neutralisent pas, à eux seuls, les lois à portée extraterritoriale.

> *Synthèse de la couche. Aucune chaîne entièrement européenne, du silicium à l'agent, n'existe à ce jour ; la rupture la plus dure reste matérielle. Mais, pour un grand nombre d'usages d'entreprise, une architecture souveraine suffisante est déjà assemblable aujourd'hui : un modèle Mistral en open weights ou un déploiement Pharia, sur cloud SecNumCloud (OVHcloud, Scaleway, Outscale), avec un index vectoriel européen. Les verrous résiduels (matériel et clauses contractuelles) sont à identifier explicitement.*

---

7. Qu'est-ce qui est réellement européen ? Les couches de la souveraineté

« Européen » n'est pas un statut binaire mais un empilement de couches, chacune pouvant être souveraine ou non indépendamment des autres. Une analyse rigoureuse en distingue au moins quatre : le contrôle capitalistique et juridique, l'infrastructure d'hébergement, le matériel de calcul, et le financement. Un acteur peut être souverain à un niveau et dépendant à un autre, et c'est précisément là que se logent les illusions.

La couche du contrôle capitalistique. La nationalité d'origine ne dit rien du contrôle réel. Deux cas survenus en 2026 l'illustrent :

- *Aleph Alpha → Cohere.* Le « champion souverain » allemand passe, via la fusion annoncée le 24 avril 2026, sous contrôle majoritaire (~90 %) du canadien Cohere, au point de heurter des clauses de contrats publics allemands exigeant un fournisseur « contrôlé en Europe ».

- *S3NS dans le cloud souverain de l'UE.* Le marché Cloud III de la Commission (avril 2026) a retenu un consortium incluant **S3NS, coentreprise de Thales et de Google Cloud**. Son inclusion dans un dispositif labellisé « souverain » a déclenché une polémique de *sovereignty washing* : CISPE, qui fédère 38 acteurs européens du cloud, y a vu un « but contre son camp ». Le cadre de la Commission lui-même n'a d'ailleurs accordé à ce consortium que le niveau SEAL-2 (souveraineté des données), contre SEAL-3 (résilience numérique) pour les consortiums purement européens (OVHcloud/CleverCloud, STACKIT, Scaleway).

À l'inverse, **Mistral illustre une souveraineté capitalistique préservée** : son actionnariat reste majoritairement européen, et la participation de Microsoft, souvent citée comme un signe de dépendance, se limite à 15 M€ investis en février 2024 (extension de Série A), soit une part minoritaire et non-contrôlante, largement diluée depuis. Microsoft n'a aucun levier de contrôle sur Mistral ; le partenariat Azure n'est qu'un canal de distribution parmi d'autres (Mistral est également présent sur AWS, Google Cloud et en auto-hébergement). Au niveau du contrôle, Mistral est donc nettement plus souverain qu'Aleph Alpha post-fusion, ce qui montre que la couche capitalistique se juge acteur par acteur, sans généralisation.

L'enseignement : la souveraineté se vérifie au registre du commerce et dans les clauses contractuelles, pas sur la plaque d'origine.

La couche du matériel : le verrou que personne ne contourne. C'est ici que se pose la question directe : *un acteur comme Mistral court-il un risque de blocage américain ?* La réponse honnête est : pas au niveau du modèle, mais oui au niveau du matériel. Mistral entraîne ses modèles sur des puces **NVIDIA** – 13 800 accélérateurs financés par 830 M$ de dette pour un centre de données près de Paris (opérationnel au T2 2026), et un campus de 1,4 GW co-développé avec NVIDIA, Bpifrance et le fonds émirati MGX (systèmes Grace Blackwell, horizon 2028). Tout le logiciel d'entraînement repose sur l'écosystème CUDA de NVIDIA.

Or NVIDIA est une entreprise américaine, et ses puces sont l'objet historique des contrôles à l'export américains. L'« AI Diffusion Rule » de janvier 2025 prévoyait explicitement un classement des pays en tiers conditionnant leur accès aux puces avancées – preuve que Washington a déjà envisagé de restreindre l'accès aux GPU par pays, alliés compris. Le risque de blocage d'un modèle européen ne prend donc pas la forme d'un « interrupteur » à distance comme pour Fable 5 : un modèle open weights déjà téléchargé ne se débranche pas. Il prend une forme plus lente mais plus structurelle : une contrainte sur l'approvisionnement en calcul. Un modèle déjà entraîné et diffusé en open weights est à l'abri d'une coupure ; sa *prochaine génération*, elle, dépend de l'accès continu aux puces.

Concrètement, les scénarios de pression américaine sur un acteur comme Mistral n'ont pas la forme d'un interrupteur logiciel, mais d'un étranglement de la chaîne d'approvisionnement : restriction des exportations de GPU avancés vers l'UE (le tiering de l'AI Diffusion Rule en était l'esquisse), conditions imposées à NVIDIA sur le support ou les mises à jour, ou – plus profond encore – contrainte sur la fabrication. Car au-delà de NVIDIA se cache un point de défaillance unique que peu d'acteurs européens contournent : **TSMC**. La quasi-totalité des puces IA avancées, NVIDIA comprise, est gravée à Taïwan ; c'est la dépendance la plus structurelle de toutes, et elle concerne aussi bien les acteurs américains que les rares projets européens.

L'état réel des alternatives matérielles. L'Europe construit une souveraineté *processeur*, mais pas encore une souveraineté *accélérateur IA* à l'échelle. Le français **SiPearl** conçoit des CPU haute performance « sans porte dérobée ni interrupteur d'arrêt » (Rhea1, puis Rhea2), retenus pour les premiers supercalculateurs exascale européens (JUPITER à Jülich, Alice Recoque/Jules Verne en France). Mais ce sont des CPU : dans JUPITER, ce sont encore des **GPU NVIDIA** qui sont greffés pour l'accélération IA. Des startups visent précisément l'accélérateur, **Axelera AI** (Pays-Bas/Italie, financée par l'EuroHPC, puce Titania visée pour 2028) et **Euclyd** (Pays-Bas, fondée par d'anciens d'ASML, architecture de calcul en mémoire revendiquant une efficacité d'inférence très supérieure), mais elles sont en phase pré-déploiement, à horizon 2027-2028. À court terme, aucune ne remplace NVIDIA pour l'entraînement de modèles de pointe.

La Chine, elle, a déjà bâti sa filière de repli sous la contrainte des sanctions : **Huawei Ascend** (910C visant 600 000 unités en 2026, gravé par SMIC, atteignant de l'ordre de 60 % des performances d'inférence d'un H100 NVIDIA selon les tests de DeepSeek), complétée par Alibaba (T-Head), Baidu (Kunlunxin) ou Hygon. L'écart avec le tout dernier NVIDIA reste large (la prochaine génération Ascend serait loin de la parité sur l'entraînement frontière), mais l'enseignement est stratégique : contrainte par l'embargo, la Chine a fait émerger une seconde source viable pour l'inférence à grande échelle. L'Europe, qui n'a pas subi le même choc, n'a pas (encore) produit cet effet.

La couche du financement. Les capitaux qui financent l'infrastructure européenne sont eux-mêmes internationaux : le campus de calcul de Mistral associe des fonds français (Bpifrance) et émiratis (MGX). Cela ne disqualifie pas la souveraineté, mais rappelle qu'elle se négocie aussi au niveau de l'actionnariat des infrastructures.

Synthèse. La bonne question n'est pas « est-ce européen ? » mais « à quelle couche, et avec quel degré de contrôle ? ». Le seul niveau où une souveraineté quasi complète est atteignable aujourd'hui est celui des **poids de modèles ouverts et auto-hébergés** : un modèle dont on détient les poids, exécuté sur sa propre infrastructure, ne peut être ni débranché ni modifié unilatéralement par un tiers. C'est la propriété qui a fait défaut aux utilisateurs de Fable 5, et c'est elle, plus que la nationalité du fournisseur, qui définit la résilience réelle.

---

8. La couche chinoise : les poids ouverts comme stratégie de souveraineté

Le débat européen oppose le plus souvent deux pôles, américain et européen, en négligeant un troisième acteur dont la stratégie reformule la question : la Chine.

Le basculement vers l'ouvert. Longtemps présentée comme un écosystème fermé, la Chine a fait de l'open weights le cœur de sa stratégie. Les familles Qwen (Alibaba), DeepSeek, GLM (Zhipu/Z.ai) et Kimi (Moonshot) sont publiées sous licences permissives (souvent Apache 2.0 ou MIT). Les modèles ouverts chinois représentent désormais de l'ordre de 30 % des téléchargements mondiaux de modèles – devant les États-Unis – et Qwen est devenu la plus grande famille open weights sur Hugging Face, avec plus de 100 000 modèles dérivés, dépassant le Llama de Meta. L'écart de capacité avec les meilleurs modèles propriétaires américains s'est resserré à une poignée de points.

Une stratégie née de la contrainte. Cette domination de l'ouvert est en partie une *réponse* aux contrôles américains. Privées d'accès aux puces NVIDIA haut de gamme depuis 2022, les équipes chinoises ont été contraintes d'innover sur l'efficacité logicielle – et certaines s'affranchissent désormais de NVIDIA : GLM-5 (744 milliards de paramètres) aurait été entraîné entièrement sur des puces Huawei Ascend. La contrainte d'hier est devenue un argument de souveraineté.

La logique des « deux boucles ». Diffuser des modèles ouverts sert un double objectif : réduire la dépendance chinoise aux technologies américaines, et installer une dépendance mondiale aux « rails » chinois. Des écosystèmes souverains s'y rallient déjà : la Malaisie a annoncé bâtir son IA souveraine sur DeepSeek, Singapour a retenu Qwen plutôt que Llama, et des entreprises américaines elles-mêmes (Airbnb pour son support client) utilisent Qwen pour son rapport coût/capacité.

Ce que cela change pour une organisation européenne. Paradoxalement, un modèle ouvert chinois est *insensible* au type de coupure qui vient de frapper les utilisateurs de Fable 5 : une fois les poids téléchargés et exécutés sur une infrastructure européenne, aucune administration, américaine ou chinoise, ne peut le débrancher à distance. Mais cela ne signifie pas qu'il soit sans risque : la dépendance se déplace du terrain géopolitique vers le terrain juridique et opérationnel, comme le détaille le point suivant.

Le test juridique : service hébergé contre poids auto-hébergés. C'est la distinction cardinale, souvent confondue. Recourir au **service hébergé** d'un acteur chinois (application ou API) expose à un risque réglementaire majeur : les données transitent vers la Chine, dont la loi PIPL reste subordonnée aux intérêts de l'État, et qui ne bénéficie d'aucune décision d'adéquation au titre du RGPD. Le précédent fait jurisprudence : en janvier 2025, l'autorité italienne (Garante) a bloqué le service DeepSeek pour traitement non conforme de données d'utilisateurs italiens (DeepSeek soutenant que le RGPD ne lui était pas applicable), des procédures parallèles s'ouvrant en Irlande, en France et en Belgique. À l'inverse, **télécharger et auto-héberger les poids ouverts** du même modèle sur une infrastructure européenne fait disparaître le problème de transfert : la donnée ne quitte jamais l'infrastructure de l'organisation. La nationalité du modèle ne détermine donc pas le risque RGPD ; c'est le *mode de consommation* qui le détermine.

Mais l'auto-hébergement ne neutralise pas tout. Trois risques subsistent et se reportent sur l'organisation déployante. D'abord l'**AI Act** : ses obligations sur les modèles à usage général et, à compter du 2 août 2026, sur les systèmes à haut risque s'appliquent quel que soit le pays d'origine du modèle, et c'est le déployeur qui en porte une partie de la charge (documentation, transparence, supervision humaine). Ensuite la **gouvernance du contenu** : les modèles chinois embarquent des restrictions codées en dur sur les sujets politiquement sensibles, qui peuvent biaiser des usages métier sans avertissement. Enfin la **traçabilité** : l'opacité des données d'entraînement complique toute garantie de conformité ou d'absence de contamination. Adopter un modèle chinois auto-hébergé, c'est donc échanger un risque de coupure géopolitique contre un ensemble différent de contraintes, à arbitrer usage par usage.

Le clivage utile n'est pas tant « américain contre européen » que « **débranchable à distance contre maîtrisable localement** », c'est-à-dire API fermée contre poids ouverts auto-hébergeables. Sur ce critère, l'offre ouverte, qu'elle soit française (Mistral), allemande (Aleph Alpha) ou chinoise (Qwen, DeepSeek), partage une propriété décisive que les API propriétaires américaines n'offrent pas : l'impossibilité, pour un tiers, de couper l'accès après déploiement.

---

9. Trois autres raccourcis à éviter

Au-delà des deux idées reçues déjà déconstruites (l'open weights ne garantit pas l'indépendance complète au niveau matériel ; un champion national peut basculer sous contrôle étranger), trois raccourcis méritent encore d'être écartés.

« **Hébergé en Europe = indépendance complète** ». Un cloud européen certifié protège la localisation et la gouvernance des données, mais reste tributaire de puces importées et, dans certains montages, d'une exposition à des législations extraterritoriales. La localisation est nécessaire, non suffisante.

« **Modèle américain = risque uniforme** ». Tous les usages ne portent pas le même risque. Un usage non critique, aisément substituable, n'appelle pas la même vigilance qu'un processus métier central bâti sur une seule API propriétaire. Le risque se mesure usage par usage, pas fournisseur par fournisseur.

« **Souveraineté = rupture avec les fournisseurs américains** ». L'objectif réaliste n'est pas l'autarcie mais la *maîtrise* : savoir où l'on est exposé, garder une capacité de repli sur les usages critiques, et négocier la réversibilité. La dépendance se cartographie et se réduit ; elle ne se supprime pas par décret.

---

10. Scénarios prospectifs

Ces cinq trajectoires ne sont pas exclusives ; la situation réelle en combinera plusieurs.

1. **Durcissement généralisé des contrôles.** Les États-Unis étendent les restrictions à d'autres modèles et d'autres pays ; la Chine renforce les siennes. L'Europe voit son accès aux dernières innovations américaines se restreindre, ce qui provoque un choc initial mais accélère massivement l'investissement et la recherche européens.
2. **Déconcentration et multi-fournisseurs.** Anticipant les coupures, entreprises et administrations généralisent les architectures multi-modèles (un modèle européen auto-hébergeable en secours d'une API américaine, plusieurs clouds). La résilience augmente, au prix d'une complexité opérationnelle accrue.
3. **Rattrapage européen.** Plans d'investissement massifs (HPC, puces, *AI factories*), consolidations transnationales et montée en gamme des modèles européens réduisent sensiblement la dépendance sur trois à cinq ans, soutenus par la commande publique.
4. **Fragmentation numérique.** Le marché se scinde en blocs régionaux aux standards divergents. Les entreprises opérant mondialement doivent dupliquer leurs chaînes IA par zone, ce qui renchérit l'innovation et complique l'interopérabilité.
5. **Coopération encadrée.** Sous l'impulsion d'enceintes internationales (G7, OCDE), les grandes puissances conviennent de cadres communs (revues de sécurité partagées, licences « de confiance »). Le choc s'atténue, au prix d'un encadrement plus lourd des lancements.

Le scénario central, à ce stade, combine une **escalade graduelle des contrôles** et une **accélération de l'offre européenne**, ce qui rend la cartographie des dépendances et les capacités de repli d'autant plus stratégiques.

---

11. Recommandations pour les décideurs

1. **Cartographier avant d'agir.** Dresser l'inventaire des usages IA en production (modèles, API, clouds, sous-traitants) et qualifier, pour chacun, la criticité métier et l'impact d'une coupure (RTO/RPO). C'est le préalable à toute décision.
2. **Établir un plan de continuité IA.** Pour chaque usage critique, définir une solution de repli, typiquement un modèle européen auto-hébergeable (Mistral en open weights, déploiement Pharia) avec index vectoriel redondant, et tester réellement le basculement.
3. **Distinguer le mode de consommation, pas seulement le fournisseur.** Le risque tient autant à *comment* un modèle est consommé qu'à *qui* le fournit : une API hébergée (où qu'elle soit) expose à la coupure et au transfert de données ; des poids ouverts auto-hébergés suppriment ces deux risques. Pour les usages critiques ou sensibles, privilégier l'auto-hébergement de poids ouverts, qu'ils soient européens ou non, sur la dépendance à une API distante.
4. **Diversifier.** Éviter la concentration sur un seul hyperscaler ou une seule API. Répartir les usages selon leur sensibilité et préparer les chemins de migration.
5. **Prioriser les solutions qualifiées pour les données sensibles.** Lors des appels d'offres, privilégier les hébergeurs certifiés (SecNumCloud) et évaluer sérieusement les modèles européens, sans en surestimer ni en sous-estimer la maturité.
6. **Renforcer la réversibilité contractuelle.** Exiger explicitement l'export périodique des index vectoriels, configurations et poids fine-tunés, et la coopération du prestataire en cas de résiliation forcée.

7. **Mettre en place une veille réglementaire.** Suivre les régimes d'export (BIS), les Executive Orders, l'AI Act et les dispositifs de cloud souverain, en associant DSI, RSSI et direction juridique pour anticiper tout changement brutal.
8. **Réduire la dépendance en compétences.** Internaliser progressivement l'expertise sur le déploiement et l'exploitation de modèles, pour ne pas reporter la dépendance technologique sur une dépendance de conseil.

Glossaire

Contrôles à l'export. Règles administratives soumettant à licence ou interdiction le transfert de certaines technologies à des destinations ou utilisateurs étrangers. Aux États-Unis, ils relèvent du *Bureau of Industry and Security* (BIS).

Deemed export. Principe selon lequel exposer une technologie contrôlée à un ressortissant étranger, **même sur le territoire national**, équivaut juridiquement à une exportation. C'est ce mécanisme qui a conduit à exclure les salariés étrangers d'Anthropic.

ECCN 4E091. Classification d'export créée par l'« AI Diffusion Rule » pour viser certains poids de modèles fermés : première tentative américaine de contrôler les modèles, et non plus seulement les puces.

Modèle « frontier ». Modèle d'IA parmi les plus avancés, au cœur des dispositifs de sécurité gouvernementaux (revues, benchmarking cyber).

Open weights. Modèle dont les poids sont publiés, autorisant un déploiement et un fonctionnement entièrement maîtrisés (y compris hors ligne). À distinguer d'un modèle « ouvert » au sens large.

PIPL (*Personal Information Protection Law*). Loi chinoise de protection des données personnelles (2021). À la différence du RGPD, elle reste subordonnée aux intérêts de l'État et à la sécurité nationale, et la Chine ne bénéficie d'aucune décision d'adéquation de l'UE, d'où le risque attaché au transfert de données vers un service hébergé en Chine.

SEAL (niveaux de souveraineté). Échelle (SEAL-0 à SEAL-4) du cadre de souveraineté du cloud de la Commission européenne, évaluant un fournisseur sur huit objectifs pondérés (la chaîne d'approvisionnement comptant pour le poids le plus élevé). Première grille de comparaison objective au-delà du marketing « cloud souverain ».

TSMC. Fondeur taïwanais qui grave la quasi-totalité des puces IA avancées mondiales (NVIDIA comprise) ; point de défaillance unique de la chaîne d'approvisionnement, en amont même de la dépendance à NVIDIA.

SecNumCloud. Qualification délivrée par l'ANSSI attestant qu'un hébergeur cloud répond aux exigences de sécurité et de souveraineté pour les données sensibles.

RTO (Recovery Time Objective). Délai maximal toléré pour rétablir un service après interruption – ici, le temps de basculer vers un modèle de substitution.

RPO (Recovery Point Objective). Volume de données maximal que l'on accepte de perdre – ici, la fraîcheur du dernier point de sauvegarde d'un système IA.

Lock-in. Dépendance à un fournisseur rendant la migration coûteuse (formats propriétaires, fine-tuning, API spécifiques, clauses contractuelles).

Souveraineté numérique. Capacité d'une organisation ou d'un État à maîtriser ses technologies et ses données sans dépendance critique à une puissance étrangère.

Sources principales

- Anthropic, communiqué et publication relatifs à la suspension de Fable 5 et Mythos 5 (directive reçue le 12 juin 2026, 17 h 21 ET) ; notice développeurs et statut AWS/Bedrock.
- Reuters, Fortune, *Time*, CNBC, CNN, Semafor, BleepingComputer, Tom's Hardware, American Banker, VentureBeat : couverture de la directive, de ses effets, de la riposte de l'administration et de la dimension « accès chinois » (12–15 juin 2026).
- David Sacks (Conseil consultatif présidentiel pour la science et la technologie) : publication, le 13 juin 2026, de la version de l'administration (refus allégué d'Anthropic de corriger ou retirer le modèle ; rétablissement conditionné à un correctif).
- Contexte financier : dépôt confidentiel d'un dossier d'introduction en Bourse par Anthropic (début juin 2026, valorisation ~965 Md$) et recul des titres pré-cotation après l'annonce (CNBC, TechCrunch).
- Euronews, *« Wake-up call: Europe reacts to Anthropic halting access to its Fable 5 and Mythos 5 AI models »* (13 juin 2026) – réactions attribuées de Bruno Retailleau, Benjamin Haddad, Édouard Philippe, Jordan Bardella, Geert Wilders, Al Carns, Tom Tugendhat.
- The White House, Executive Order *« Promoting Advanced Artificial Intelligence Innovation and Security »* (2 juin 2026) ; analyses Crowell & Moring, Latham & Watkins, Skadden, Council on Foreign Relations.
- BIS / Département du commerce américain : publication (janvier 2025) puis abrogation (13 mai 2025) de l'« AI Diffusion Rule » ; analyses WilmerHale, Morrison Foerster, Kirkland & Ellis, Akin.
- Mistral AI : annonces et documentation des modèles Large 3, Medium 3.5, Small 4 et Ministral 3 (décembre 2025 – mars 2026) ; financement (830 M$ de dette) pour l'acquisition de 13 800 puces NVIDIA et projet de campus de calcul de 1,4 GW avec NVIDIA, Bpifrance et MGX ; investissement minoritaire de Microsoft (15 M€, février 2024) et examen par la Commission européenne.
- Aleph Alpha / Cohere : annonce de fusion (24 avril 2026) ; documentation PhariaAI et Pharia ; couverture du contrôle capitalistique.
- OVH Groupe : communiqués d'acquisition Seald (janvier 2026), Dragon LLM (mars 2026) et Gladia (négociations exclusives, 11 juin 2026).
- Commission européenne : cadre de souveraineté du cloud et marché Cloud III (180 M€, attribution avril 2026 à quatre consortiums, niveaux SEAL) ; couverture de la controverse S3NS (CISPE).
- Écosystème open weights chinois : études MIT / Hugging Face et analyses (Stanford HAI, USCC, MIT Technology Review, CIGI) sur la part des modèles ouverts chinois (Qwen, DeepSeek, GLM, Kimi) dans les téléchargements mondiaux et leur lien avec les contrôles export américains.
- Encadrement juridique des modèles chinois : décision du Garante italien bloquant le service DeepSeek (janvier 2025) ; procédures parallèles (Irlande, France, Belgique) ; analyses comparatives RGPD / PIPL.
- Matériel et puces : Huawei (feuille de route Ascend 910C/950/960/970, production via SMIC) ; SiPearl (CPU Rhea1/Rhea2 « backdoor-free », supercalculateurs JUPITER et Alice Recoque) ; Axelera AI et Euclyd (financements EuroHPC / capital-risque) ; concentration de la fabrication chez TSMC.
- AI Act européen (règlement 2024/1689) : calendrier d'application officiel de la Commission ; ANSSI (SecNumCloud) ; initiative GAIA-X.

Cette note inaugure une série de publications de référence d'IntelligenceArtificielle.com consacrées aux transformations structurelles du marché de l'IA, à destination des décideurs économiques, publics et institutionnels. Elle est appelée à être enrichie : toute annonce, source ou évolution pertinente peut être signalée pour mise à jour.