La virtualité supposée de l'IA face au mur du réel

Comment 200 milliards US$ de GPU révèlent les contraintes physiques de l'intelligence artificielle

Jérôme Vetillard

VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products

Compliance by Design | PhD ENS Ulm | MIT Sloan | IE Business School & Brown University

Édition révisée — Février 2026

Résumé exécutif

Entre 2022 et 2025, les hyperscalers ont investi plus de 200 milliards de dollars en infrastructures GPU pour l'IA générative — une stratégie d'options réelles rationnelle face à l'incertitude radicale sur la taille du marché. Deux ans plus tard, l'adoption en entreprise reste marginale (moins de 10 % des entreprises américaines, 74 % sans valeur tangible extraite), créant un déséquilibre entre capacité installée et usage réel.

Pour amortir cette surcapacité, l'industrie pousse vers des usages intensifs (agents autonomes persistants), déplaçant le goulot d'étranglement des GPU vers l'énergie, l'eau et les données d'entraînement. L'IA entre dans un régime d'industrie lourde où l'avantage concurrentiel se joue autant sur les contrats énergétiques et les droits d'eau que sur l'excellence algorithmique. La question finale n'est plus technique mais politique : quelle part de nos ressources physiques finies nos sociétés accepteront-elles d'allouer à l'IA, et au détriment de quels autres usages ?

I. Introduction : le paradoxe de l'abondance contrainte

Nous parlons de l'intelligence artificielle comme d'une technologie immatérielle. Des algorithmes, des modèles, des « tokens » générés dans l'abstraction du cloud. Cette représentation domine les discours stratégiques, les analyses de marché, les prévisions d'adoption. Elle devient pourtant de plus en plus trompeuse.

Derrière chaque requête à un modèle de langage, chaque image générée, chaque ligne de code assistée, se cache une réalité physique : des dizaines de milliers de GPU alignés dans des datacenters, une consommation d'énergie électrique comparable à celle d'une ville moyenne, des volumes d'eau significatifs pour le refroidissement, des raccordements à des réseaux électriques déjà sous tension. L'IA générative n'est plus seulement une industrie du « logiciel augmenté » : elle tend structurellement vers une industrie lourde.

Entre 2022 et 2025, les hyperscalers (Microsoft, Google, Meta, Amazon) se sont engagés dans une course à l'investissement sans précédent, principalement en infrastructures de calcul. Cette accumulation n'était ni irrationnelle ni accidentelle : face à une incertitude radicale sur les usages et l'ampleur du marché, sécuriser la capacité avant les concurrents constituait une option stratégique défendable. Le coût d'un excès temporaire paraissait inférieur au risque d'être contraint au moment où le marché basculerait.

Deux ans plus tard, un décalage se précise : l'usage en entreprise progresse, mais reste loin d'un déploiement massif et homogène. Les retours sur investissement sont diffus, et l'usage reste fréquemment sporadique, centré sur des interactions humaines non persistantes. La surcapacité initialement assumée devient alors un déséquilibre économique : une infrastructure installée très importante, mais imparfaitement monétisée.

C'est cette tension qui pousse aujourd'hui l'industrie vers des usages plus intensifs : agents persistants, orchestrations multi-modèles, workloads continus. Ces paradigmes augmentent drastiquement le taux d'utilisation des machines, mais déplacent le facteur limitant. Le goulot d'étranglement n'est plus prioritairement le nombre de GPU disponibles. Il tend à devenir la capacité à mobiliser durablement des gigawatts raccordables, des volumes d'eau, des données d'entraînement de qualité, des permis, et une acceptabilité sociale suffisante.

Ce basculement redéfinit les règles du jeu concurrentiel : l'avantage glisse des seuls laboratoires vers les contrats énergétiques long terme, l'ingénierie d'infrastructure, les choix de localisation et les compromis territoriaux. L'IA entre dans un régime où ses possibilités techniques se heurtent à des contraintes thermodynamiques et à la gouvernance de ressources physiques finies.

Cette transformation n'est pas une simple crise de croissance : c'est un changement de régime industriel. Elle pose une question que nous n'avons pas encore affrontée de front : l'intelligence artificielle sera-t-elle limitée par notre créativité algorithmique, ou par les contraintes thermodynamiques et politiques de la planète qui la supporte ?

II. 2022–2024 : l'accumulation préemptive

Lorsque ChatGPT franchit le million d'utilisateurs en cinq jours, en novembre 2022, rien n'est encore stabilisé : ni les usages finaux, ni la vitesse d'adoption, ni la taille réelle du marché. L'IA générative peut devenir un outil de niche comme une plateforme universelle. Cette incertitude radicale aurait pu conduire à l'attentisme. Elle produit l'inverse : une mobilisation de capital sans précédent.

Entre 2022 et 2025, les hyperscalers engagent collectivement plus de 200 milliards de dollars dans les infrastructures de calcul. Microsoft annonce 50 milliards sur quatre ans, Meta 37 milliards pour la seule année 2024, Amazon 75 milliards dédiés aux datacenters IA. Ces montants dépassent largement les investissements observés lors de précédentes vagues technologiques (cloud, mobile, big data).

Cette accumulation obéit à une logique économique bien identifiée : celle des options réelles. Dans un contexte d'incertitude radicale, construire une capacité excédentaire revient à acheter une option stratégique. Le coût de cette option (des GPU temporairement sous-utilisés) est jugé inférieur au risque d'être exclu d'un marché émergent à rendements croissants faute de capacité disponible au moment critique. Microsoft ne construit pas pour la demande de 2023, mais pour celle, hypothétique et massive, de 2026–2027. La surcapacité n'est pas subie : elle est assumée.

Cette stratégie se heurte immédiatement à une contrainte physique : le silicium. Produire un GPU NVIDIA H100 requiert des procédés de gravure en 4–5 nm, maîtrisés quasi exclusivement par TSMC. Les délais de livraison atteignent 18 à 24 mois, rendant l'offre structurellement inélastique à court terme. Les GPU deviennent alors des actifs stratégiques. Les contrats avec NVIDIA impliquent prépaiements, engagements de volume et, dans certains cas, priorités d'allocation. Microsoft sécurise l'accès à plusieurs centaines de milliers de H100 ; Meta développe ses propres puces (MTIA) ; Amazon déploie Trainium et Inferentia ; Google exploite ses TPU depuis 2016.

En 2023–2024, les GPU font l'objet d'une véritable course à l'armement. Leur valeur n'est plus seulement fonctionnelle, mais positionnelle. La Chine, mise à l'écart par le protectionnisme américain, se dote de son propre « projet Manhattan » pour réaliser des gravures en dessous de 5 nm.

Cette dynamique produit une concentration extrême des moyens de calcul. Fin 2024, Microsoft viserait environ 1,8 million de GPU, tandis que Stanford University dispose d'environ 300 GPU seulement (déclarations de Russell Wald, Stanford HAI). Même les initiatives académiques ambitieuses restent marginales : le superpod Marlowe de Stanford (248 H100), le cluster du Kempner Institute à Harvard (384 H100), ou les 32 H100 disponibles au MIT en 2024, restent trois à quatre ordres de grandeur sous les capacités d'un hyperscaler individuel. Lors d'une audition au Sénat américain en novembre 2024, Russell Wald résumait la situation sans détour : « L'ensemble des universités américaines ne pourrait pas construire une version de ChatGPT aujourd'hui. »

Les publications de pointe migrent progressivement des laboratoires universitaires vers les équipes internes des hyperscalers. L'IA devient une science expérimentale réservée aux acteurs disposant de plusieurs milliards de dollars de CapEx. Cette accumulation préemptive s'inscrit pleinement dans la théorie des marchés à rendements croissants (Arthur, 1989) : effets de réseau, coûts de changement élevés, économies d'échelle, dynamiques de winner-takes-most. Satya Nadella le formule explicitement en 2023 : « We're not going to be caught short on compute. »

Cependant, à mesure que cette capacité GPU se déploie, une contrainte nouvelle apparaît : la mémoire. Les architectures modernes sont désormais memory-bound plutôt que compute-bound. Un H100 embarque 80 Go de HBM3 ; un GB200 promet 192 Go, mais les modèles de nouvelle génération requièrent plusieurs téraoctets de mémoire agrégée. L'industrie HBM — dominée par SK Hynix, Samsung et Micron — peine à suivre : délais de 8 à 12 mois, hausse des prix estimée à +300 % entre 2022 et 2024.

Fin 2024, l'industrie dispose de plusieurs millions de GPU avancés, représentant une puissance agrégée supérieure à 100 exaflops, largement au-delà de l'usage immédiat mesurable. C'est précisément le résultat recherché ex ante. Mais cette rationalité stratégique va se heurter, ex post, à une réalité plus complexe.

III. 2024–2025 : le piège de la sous-consommation

3.1. L'écart d'adoption : données convergentes

Deux ans après le lancement de ChatGPT, un décalage structurel apparaît entre la capacité de calcul installée et l'usage réel. Les données convergent : l'adoption de l'IA générative en entreprise reste marginale, fragmentée et largement non rentabilisée.

Selon le US Census Bureau, seules 9,2 % des entreprises américaines déclaraient utiliser l'IA au deuxième trimestre 2025, contre 5,7 % fin 2024 — soit une progression réelle, mais très inférieure aux projections initiales. Une étude BCG (2024) portant sur 1 000 dirigeants montre que 74 % des entreprises n'extraient aucune valeur tangible de leurs initiatives IA. McKinsey confirme : plus de 80 % des organisations ne rapportent aucun impact matériel de l'IA générative sur leurs résultats financiers. Seuls 6 % des répondants — que McKinsey qualifie de « high performers » — déclarent attribuer à l'IA un effet supérieur à 5 % sur l'EBIT. Cette concentration pose question : ces organisations sont précisément celles qui ont le plus massivement investi et qui seraient les plus exposées à une dévaluation boursière en cas d'annonce négative sur leur retour IA. Le biais déclaratif ne peut être écarté.

Le problème n'est pas seulement le déploiement, mais l'usage effectif. Une étude du MIT indique que 95 % des projets pilotes d'IA générative échouent à produire des retours mesurables, non pour des raisons techniques, mais faute de passage à l'échelle opérationnelle. L'usage demeure sporadique : là où les projections anticipaient plus de 100 requêtes mensuelles par utilisateur actif, la réalité observée se situe entre 8 et 12 requêtes. Même les copilots largement diffusés (Microsoft 365 Copilot, Google Workspace AI) peinent à transformer l'utilité perçue en gains économiques mesurables.

Plus révélateur encore : certaines organisations, dont Microsoft lui-même, ont intégré l'adoption de Copilot dans les métriques d'évaluation de performance de leurs employés. Des indicateurs tels que le « taux d'engagement Copilot » ou les « actions Copilot par utilisateur » figurent désormais parmi les KPI suivis par les managers. Cette pratique — parfois qualifiée de « dogfooding institutionnel » — révèle l'écart entre le narratif d'adoption

organique et la réalité d'une diffusion par incitation hiérarchique. L'usage contraint ne garantit ni appropriation réelle, ni génération de valeur.

3.2. Anatomie des freins

Cette sous-performance n'a rien de mystérieux. McKinsey montre que près de 70 % des obstacles sont humains et organisationnels, 20 % technologiques, et seulement 10 % algorithmiques, alors même que ces derniers concentrent l'essentiel de l'attention et des ressources.

Les freins techniques (20 %) tiennent à l'intégration avec des systèmes hérités et à la qualité des données, souvent fragmentées et peu gouvernées. L'IA générative ne peut être dissociée de la maturité de la gouvernance des données ; de même, toute ambition MLOps présuppose une culture DevOps préalablement établie. Selon Curt Jacobsen de McKinsey, 30 à 50 % du temps des équipes innovation est consommé à assurer la conformité réglementaire ou à attendre que les politiques organisationnelles évoluent.

Les freins organisationnels (70 %) sont plus profonds. La résistance au changement n'est pas irrationnelle : elle reflète l'incertitude sur l'impact futur des emplois, l'absence de formation adéquate, et le déficit de vision stratégique. Une enquête Gallup (2024) révèle que seuls 15 % des employés américains déclarent que leur entreprise a communiqué une stratégie IA claire. Moins de 30 % des PDG sponsorisent directement l'agenda IA selon McKinsey. Les freins juridiques s'intensifient avec le durcissement des régulations : RGPD, AI Act adopté en 2024, questions de responsabilité et d'explicabilité. Les freins économiques apparaissent ex post : le TCO réel d'une IA en production (inférence, fine-tuning, monitoring, conformité) s'avère 10 à 20 fois supérieur aux coûts des POC initiaux.

3.3. Tentatives de stimulation de la demande et pivot agentique

Face à cette sous-consommation, l'industrie tente d'activer la demande par plusieurs leviers : prolifération de use cases marketing pour démontrer l'universalité de la technologie ; bundling agressif (Microsoft intègre Copilot dans toutes ses suites enterprise, Google fait de même avec Workspace AI) ; guerre des prix (OpenAI réduit le prix de GPT-3.5 de plus de 90 % entre 2023 et 2024) ; et pivot massif vers l'agentic AI à partir du quatrième trimestre 2024.

Ce dernier pivot n'est pas accidentel. Comme le formule explicitement McKinsey dans son rapport de juin 2025 : les copilots horizontaux n'ont pas généré de valeur à l'échelle ; les acteurs se tournent vers des agents autonomes, intégrés dans des processus métiers verticaux, capables de fonctionner en continu. Pourtant, une question s'impose : pourquoi des organisations incapables de rentabiliser des usages simples d'IA générative parviendraient-elles soudainement à extraire de la valeur de systèmes autrement plus complexes à déployer, gouverner et sécuriser ?

3.4. Déséquilibre économique structurel et risque financier

Le résultat est un déséquilibre économique majeur. Des millions de GPU sont installés, mais une fraction seulement est effectivement utilisée de manière productive. Les taux d'utilisation réels des clusters GPU se situent entre 15 % et 30 % selon les analyses sectorielles, bien en-deçà des 60 à 80 % nécessaires à la rentabilité de tels investissements.

Ce déséquilibre soulève une question financière que l'industrie évite encore de poser frontalement : les 200 milliards de dollars investés constituent-ils une allocation stratégique rationnelle ou les prémices d'une bulle spéculative ? Plusieurs indicateurs appellent à la vigilance. Le ratio entre le CapEx engagé et les revenus IA directement attribuables reste défavorable : pour chaque dollar investi en infrastructure, les revenus spécifiquement IA ne génèrent qu'une fraction de retour, le reste étant comptabilisé dans les revenus cloud généraux. La duration des actifs GPU est problématique : le cycle d'obsolescence technologique (18 à 24 mois entre générations) est significativement plus court que l'amortissement comptable (5 à 7 ans), créant un risque de dépréciation accélérée. Enfin, le risque de rupture architecturale n'est pas négligeable : si des architectures alternatives (SSM de type Mamba, modèles de diffusion pour le texte, architectures hybrides) réduisaient significativement les besoins en GPU optimisés pour l'attention, des vagues entières de hardware deviendraient des stranded assets.

Le précédent le plus éclairant est celui de la surcapacité en fibre optique de 2000–2001. À la fin des années 1990, les opérateurs télécoms avaient massivement investi dans des réseaux de fibre, convaincus que la demande de bande passante croîtrait exponentiellement. La demande a effectivement crû, mais avec un décalage temporel de plusieurs années par rapport aux projections. Les entreprises les plus exposées (WorldCom, Global Crossing, 360networks) ont fait faillite. L'infrastructure a finalement été utilisée, mais par d'autres acteurs, à d'autres conditions, après une destruction massive de valeur actionnariale. L'analogie n'est pas parfaite — les hyperscalers actuels sont financièrement plus résilients que les télécoms de 2000 — mais le pattern structurel (surcapacité préemptive → décalage d'adoption → pression sur les valorisations) est analogue et mérite attention.

Cette pression pousse les hyperscalers vers une course au volume : baisse des prix, bundling, et invention de nouveaux paradigmes d'usage plus intensifs. Les agents autonomes fonctionnant 24/7 consomment un ratio estimé de 50 à 100 fois plus de ressources qu'une requête ponctuelle (estimation fondée sur le rapport des duty cycles : un agent persistant maintient un contexte mémoire et exécute des boucles d'inférence en continu pendant des heures ou des jours, contre quelques secondes pour une requête interactive ; le ratio reflète principalement la différence de temps d'occupation des ressources GPU et mémoire HBM, et non une augmentation proportionnelle du compute brut par inférence unitaire). En cherchant à rentabiliser la surcapacité de calcul via l'intensification de l'usage, l'industrie déplace mécaniquement le goulot d'étranglement.

IV. 2025–2027 : le basculement vers les contraintes physiques

L'intensification des usages de l'IA ne se heurte plus principalement à la disponibilité des GPU, mais à la capacité de soutenir leur fonctionnement continu. Le facteur limitant se déplace vers des intrants physiques dont l'extension est lente, coûteuse et politiquement contrainte. Ce basculement marque un changement de régime.

4.1. Énergie : le mur des gigawatts

Les ordres de grandeur suffisent à mesurer l'ampleur du phénomène. Un datacenter hyperscale consomme typiquement 150 à 300 MW en charge continue, soit l'équivalent de la consommation d'une ville de 100 000 à 200 000 habitants. Un cluster de 10 000 GPU H100 requiert environ 20 MW de puissance permanente, hors refroidissement.

Les besoins énergétiques de l'entraînement des grands modèles restent opaques, mais les estimations convergent vers des dizaines de mégawatts mobilisés sur plusieurs mois, représentant des volumes de l'ordre de 100 GWh pour les modèles de génération GPT-4. Selon l'IEA, les datacenters ont consommé environ 460 TWh en 2022. Les projections à horizon 2030 se situent entre 1 000 et 1 300 TWh, dont 40 à 50 % attribuables à l'IA.

Cette demande se concentre géographiquement, créant des tensions locales sévères. En Irlande, les datacenters représentaient environ 20 % de la consommation électrique en 2023, avec des projections proches de 30 % à horizon 2030. En Virginie du Nord, les capacités approchent de la saturation. Singapour et les Pays-Bas ont imposé des moratoires sur les nouveaux projets. Ajouter de la capacité électrique prend du temps : 3 à 5 ans pour une centrale à gaz, 10 à 15 ans pour le nucléaire traditionnel.

Face à ces contraintes, l'industrie se tourne vers le nucléaire. Microsoft a signé un accord de long terme avec Constellation Energy pour la remise en service du réacteur Unit 1 de Three Mile Island (≈ 835 MW). Une course aux SMR (Small Modular Reactors) s'engage : AWS vise plus de 5 GW d'ici 2039, Google environ 500 MW via Kairos Power, Oracle conçoit des datacenters alimentés par SMR, et TerraPower (Bill Gates) développe le réacteur Natrium (345 MW) — un réacteur à neutrons rapides utilisant le sodium fondu comme caloporteur primaire, dans la lignée des filières Phénix, Superphénix et Astrid en France.

Les SMR offrent des avantages théoriques (modularité, délais réduits, proximité de la charge), mais introduisent des risques systémiques nouveaux. Beaucoup reposent sur de l'uranium HALEU (jusqu'à 19,75 %), dont le taux d'enrichissement est proche du seuil militaire. La dissémination potentielle de centaines de réacteurs complique la surveillance internationale et fragilise les régimes actuels de non-prolifération.

4.2. Eau : la contrainte invisible

L'eau constitue l'autre goulot d'étranglement critique, souvent absent des discours publics. Selon la taille, le climat et la technologie de refroidissement, un datacenter hyperscale peut consommer plusieurs millions de litres d'eau par jour. Microsoft a déclaré une hausse de 34 % de sa consommation d'eau entre 2021 et 2022, Google de 20 %, largement attribuées à l'extension de leurs infrastructures IA.

Selon le World Resources Institute, environ 40 % des datacenters mondiaux sont situés dans des zones de stress hydrique moyen à élevé. Cette contrainte est géographique et saisonnière, alimentant des conflits d'usage déjà visibles en Arizona, en Uruguay ou en Espagne. Les alternatives technologiques (refroidissement à air, immersion cooling, localisation nordique) impliquent des compromis sévères en termes de coûts, d'énergie ou de latence.

4.3. Données d'entraînement : le quatrième mur

Au-delà de l'énergie et de l'eau, une contrainte matérielle émerge qui relève de la même logique de finitude : l'épuisement des données d'entraînement de haute qualité. Les scaling laws qui gouvernent l'amélioration des modèles de fondation reposent sur trois intrants : le compute, les paramètres et les données. Les deux premiers sont extensibles par l'investissement ; le troisième se révèle fini.

Villalobos et al. (2022) estiment que les stocks de texte de haute qualité disponibles sur le web public pourraient être épuisés entre 2026 et 2032, selon les hypothèses de croissance des modèles. Muennighoff et al. (2023) parviennent à des conclusions convergentes et montrent que la réutilisation répétée des mêmes données (epochs multiples) produit des rendements décroissants au-delà d'un seuil relativement bas.

La substitution par des données synthétiques — générées par les modèles eux-mêmes — pose un problème fondamental. Shumailov et al. (2023) démontrent que l'entraînement récursif sur des données synthétiques provoque une dégénérescence progressive des modèles (« model collapse ») : les distributions se contractent, les queues de distribution disparaissent, et le modèle converge vers une représentation appauvrie du monde. Ce phénomène est analogue à la consanguinité génétique : chaque génération perd de la diversité informationnelle.

Cette contrainte renforce l'argument central de cet article : l'IA se heurte non pas à une seule limite physique mais à un faisceau convergent de contraintes matérielles — énergie, eau, silicium, données — qui encadrent sa croissance de tous côtés. La raison fondamentale est la même dans chaque cas : ces ressources sont finies, leur extension est lente, et leur substitution introduit des compromis sévères.

4.4. Le verrouillage géographique

L'IA devient géographiquement déterminée par la convergence de ces contraintes physiques : énergie bas carbone abondante et pilotable, bassins hydrologiques pérennes, et

connectivité aux dorsales internet avec latence acceptable. Ces conditions sont rarement réunies simultanément, créant un verrouillage territorial durable. Certaines régions disposent d'avantages relatifs (pays nordiques, Québec, Nouvelle-Zélande, France), mais ceux-ci restent partiels. En France, les épisodes de canicule et de faible débit des fleuves ont déjà conduit à des réductions de puissance nucléaire, révélant la dépendance croisée entre énergie et eau.

Technologie réputée immatérielle, l'IA devient l'une des industries les plus matériellement contraintes du XXIe siècle.

V. Dynamique systémique : l'effet rebond à l'échelle industrielle

5.1. Le paradoxe de Jevons revisité

En 1865, William Stanley Jevons observait dans The Coal Question que l'amélioration de l'efficience des machines à vapeur n'avait pas réduit la consommation totale de charbon. En abaissant le coût énergétique par unité de travail, l'efficience avait élargi le champ des usages possibles, entraînant une augmentation nette de la demande globale. Ce mécanisme — désigné comme backfire dans la littérature sur l'effet rebond — ne se manifeste pleinement que dans des contextes où la demande est peu saturée et fortement sensible aux coûts marginaux.

L'IA générative réunit précisément ces conditions. Les progrès algorithmiques et matériels ont réduit drastiquement le coût unitaire du calcul. Les modèles de génération GPT-4 produisent un token pour une fraction (souvent estimée autour d'un dixième) de l'énergie requise par GPT-3. Les générations récentes de processeurs spécialisés affichent des gains de performance par watt de l'ordre de ×2 à ×3 par rapport aux générations précédentes.

Ces gains sont réels. Pourtant, la consommation énergétique totale de l'écosystème IA augmente rapidement. La raison est structurelle : chaque amélioration d'efficience élargit le périmètre des usages économiquement viables dans un contexte où la demande n'est ni saturée ni contrainte par des budgets énergétiques explicites. Les volumes d'usage ont été multipliés par un facteur estimé entre 50 et 100 entre 2022 et 2025, dépassant largement les gains unitaires d'efficience.

Cette dynamique rappelle la loi de Wirth en informatique : « le logiciel ralentit plus vite que le matériel n'accélère ». Dans le contexte de l'IA, on peut proposer une heuristique équivalente : les modèles consomment plus vite que les puces n'économisent. Il ne s'agit pas d'une loi physique, mais d'une régularité observée dans un régime où l'efficience libère de nouveaux degrés de liberté plutôt que de réduire la consommation totale.

5.2. L'agentic AI comme catalyseur de l'effet rebond

L'émergence de systèmes dits agentiques — agents autonomes capables d'orchestrer des tâches complexes, de maintenir un contexte étendu et de fonctionner en continu — agit comme un catalyseur de cette dynamique. La différence énergétique entre les paradigmes est structurelle. Une requête ponctuelle mobilise des ressources pendant quelques secondes. Un agent autonome opère sur des horizons longs : contextes persistants en mémoire, interrogations régulières de systèmes externes, coordination de plusieurs modèles spécialisés, boucles de rétroaction continues.

À infrastructure GPU installée constante, le passage à des agents persistants augmente le taux d'utilisation effectif (duty cycle) des ressources de 15–30 % (usage humain ponctuel) à 60–80 % (agents 24/7). Cette intensification se traduit par une hausse de la consommation énergétique absolue de ×3 à ×5, non parce que les modèles deviennent moins efficients, mais parce qu'ils sont mobilisés en permanence.

Le paradoxe est temporel. À court terme, l'agentic AI contribue à résoudre un problème économique : la rentabilisation d'infrastructures déjà financées. À moyen terme, elle accélère la collision avec des contraintes physiques irréductibles. L'IA ne se heurte pas à ses limites malgré ses gains d'efficience, mais à cause de la manière dont ces gains élargissent le champ des usages possibles.

VI. Recomposition de l'avantage compétitif

6.1. Évolution des facteurs de succès

Au début des années 2020, l'avantage compétitif en IA reposait principalement sur des actifs immatériels : talent scientifique, innovations architecturales, accès à de vastes jeux de données propriétaires. Cinq ans plus tard, ces facteurs demeurent nécessaires, mais leur poids relatif s'est redistribué. La diffusion rapide des connaissances via l'open source, les pré-publications et la mobilité des talents réduit la durabilité des avantages purement algorithmiques. Les modèles de fondation convergent progressivement en performance sur un large spectre de tâches.

En parallèle, des facteurs jusqu'alors secondaires gagnent en importance : accès à une énergie bas carbone abondante, sécurisation de ressources hydriques, proximité des infrastructures de production électrique, acceptabilité territoriale des projets, et — de manière croissante — accès à des corpus de données propriétaires de haute qualité dans des domaines verticaux où les données publiques sont insuffisantes. Ces dimensions relèvent davantage de la logique des industries lourdes que de celle du logiciel.

6.2. L'IA comme industrie lourde

Cette redistribution rapproche structurellement l'IA d'industries historiquement contraintes par l'accès à l'énergie et aux ressources naturelles. L'aluminium s'est concentré autour de l'hydroélectricité au début du XX^{e} siècle ; la chimie lourde près des ports, de l'eau et de

l'électricité ; les semi-conducteurs sont depuis longtemps water-intensive et géographiquement concentrés.

L'intensité énergétique de l'IA reste difficile à mesurer précisément, en raison de l'hétérogénéité des architectures et des usages. Néanmoins, des comparaisons en ordre de grandeur suggèrent une intensité de 0,40–0,60 kWh par dollar de valeur ajoutée pour l'IA générative (note méthodologique : cette estimation intègre l'inférence, l'entraînement et le refroidissement rapportés au chiffre d'affaires directement attribuable à l'IA ; elle exclut l'amortissement du hardware et les coûts de construction des datacenters), contre environ 0,05 pour les services financiers, 0,08 pour le software traditionnel, 0,80 pour la sidérurgie et 1,20 pour l'aluminium. L'IA ne rejoint pas encore l'électrométallurgie, mais s'en rapproche structurellement.

6.3. Nouveaux déterminants de l'avantage concurrentiel

Dans ce nouveau régime, les sources d'avantage concurrentiel durable se recomposent : contrats énergétiques long terme (PPA bas carbone sur 10 à 20 ans comme actif stratégique) ; proximité des sources de production (énergie pilotable et abondante) ; droits d'eau sécurisés (avantage difficilement réplicable dans les régions en stress hydrique) ; acceptabilité politique et sociale ; et maîtrise de la chaîne énergétique via une intégration verticale partielle (nucléaire avancé, SMR, partenariats long terme).

VII. Géopolitique de l'IA sous contraintes physiques

7.1. Fragmentation des chaînes de valeur du calcul avancé

La production de processeurs de calcul avancé repose sur une chaîne de valeur hautement fragmentée et asymétriquement distribuée. La conception reste majoritairement occidentale, tandis que la fabrication des puces logiques avancées est quasi monopolistique : Taïwan concentre plus de 90 % des capacités mondiales en dessous de 7 nm. La mémoire HBM dépend de quelques acteurs concentrés en Asie de l'Est. La lithographie EUV demeure un monopole industriel européen (ASML).

Cette architecture crée des vulnérabilités systémiques majeures. Taïwan constitue un point de défaillance unique : toute perturbation durable bloquerait l'essentiel de la production mondiale de puces avancées. Au-delà des considérations de droit international et de bastion démocratique, une intervention militaire de la Chine à Taïwan perturberait très fortement cette nouvelle économie liée à l'IA.

7.2. L'asymétrie Chine/Occident face aux contraintes physiques

La Chine a engagé depuis 2022 un effort massif d'autonomisation technologique, parfois qualifié de « projet Manhattan du silicium ». Des progrès réels ont été observés (gravure 7

nm par contournement DUV chez SMIC), mais au prix de rendements faibles, de surcoûts énergétiques et d'une scalabilité incertaine.

Or, l'hypothèse implicite que les contraintes physiques s'appliquent symétriquement à tous les acteurs mérite d'être interrogée. La Chine dispose d'avantages structurels significatifs sur plusieurs des dimensions identifiées dans cet article. En matière de déploiement d'infrastructures énergétiques, la vitesse d'exécution chinoise est sans équivalent en Occident : entre 2020 et 2024, la Chine a mis en service plus de capacité solaire que le reste du monde réuni, tout en maintenant un programme nucléaire (150 réacteurs planifiés ou en construction) qui dépasse de loin les ambitions européennes et américaines combinées. En matière d'acceptabilité sociale et territoriale, le modèle de gouvernance chinois élimine les délais liés aux oppositions locales, aux permis environnementaux et aux consultations publiques qui rallongent de plusieurs années les projets en Europe et en Amérique du Nord.

La stratégie chinoise ne se limite pas au contournement lithographique. Elle inclut des investissements massifs dans des architectures alternatives — chiplets, wafer-level integration, compute-in-memory — qui pourraient rendre partiellement obsolète la course aux nanomètres. Deux scénarios se dessinent : soit une convergence progressive vers les nœuds avancés neutralisant l'arme technologique occidentale, soit un contournement par des voies architecturales différentes. Dans les deux cas, l'avantage relatif de la Chine sur les contraintes physiques (déploiement rapide d'énergie, capacité d'exécution, tolérance politique aux externalités) pourrait compenser partiellement son retard sur le silicium avancé.

7.3. Uranium HALEU et nouvelle diplomatie nucléaire

L'essor des SMR et des réacteurs avancés, envisagés pour alimenter les infrastructures IA, introduit une dépendance critique à l'uranium HALEU (5–19,75 %). Le marché actuel est extrêmement concentré : la Russie assure l'essentiel de la production mondiale. Dans le contexte de la guerre en Ukraine, dépendre de la Russie pour la fourniture d'uranium HALEU reviendrait à financer le conflit en troquant le pétrole par l'uranium. Le HALEU pose également un problème de gouvernance internationale : son niveau d'enrichissement est à proximité du seuil militaire, réduisant drastiquement le temps de bascule vers un usage militaire en cas de détournement.

7.4. Câbles sous-marins et souveraineté numérique

La matérialité de l'internet repose presque exclusivement sur les câbles sous-marins. Depuis une décennie, leur propriété bascule des opérateurs télécoms vers les hyperscalers. Les routes sont concentrées autour de quelques goulots géographiques (Atlantique Nord, Méditerranée orientale, détroits asiatiques), exposés aux sabotages et aux pressions géopolitiques. Pour l'IA, toute fragmentation durable des réseaux conduirait à une régionalisation forcée des modèles, accentuant les asymétries entre blocs.

VIII. Contre-forces et zones d'incertitude

Les contraintes énergétiques et hydriques ne constituent pas un horizon strictement indépassable. Plusieurs dynamiques sont susceptibles d'infléchir la trajectoire actuelle. Leur portée demeure toutefois incertaine, ce qui impose de nuancer toute lecture déterministe, sans invalider les tendances structurelles identifiées.

8.1. Innovations d'efficience : gains réels, portée systémique limitée

Les progrès en efficience algorithmique et matérielle sont réels. La sparsification, la quantization (FP32 → INT8/INT4), les architectures Mixture-of-Experts et la distillation de modèles réduisent significativement les coûts d'inférence et la consommation mémoire, parfois d'un ordre de grandeur. Sur le plan matériel, des trajectoires prometteuses émergent (neuromorphique, photonique, in-memory computing). Hooker (2021) montre que la charge computationnelle nécessaire pour atteindre une performance donnée sur ImageNet a été divisée par 10 entre 2012 et 2022. Toutefois, ces améliorations ont été accompagnées d'une croissance rapide du volume total d'usages, confirmant la prédominance de l'effet rebond. En outre, la plupart de ces innovations impliquent des délais de déploiement industriels longs (5 à 10+ ans).

8.2. Inférence décentralisée et edge computing

Une trajectoire alternative émerge, jusqu'ici peu intégrée dans les analyses de consommation énergétique : l'inférence décentralisée sur des modèles distillés ou quantifiés, exécutée sur des appareils locaux. Apple Intelligence, les modèles Llama on-device, et le déploiement de NPU (Neural Processing Units) dans les processeurs grand public (Qualcomm Snapdragon X, Apple M4, Intel Meteor Lake) dessinent un scénario où une fraction significative de l'inférence migre vers le edge, réduisant la pression sur les datacenters centralisés.

Ce scénario ne résout pas le problème de l'entraînement (qui reste centralisé et hyper-intensif), ni celui des agents persistants (qui requièrent des contextes mémoire étendus et une coordination multi-modèles incompatibles avec les capacités actuelles des NPU). Il pourrait néanmoins modifier significativement le profil énergétique de l'inférence interactive — qui représente la majorité de la consommation opérationnelle actuelle — en redistribuant la charge vers des milliards de dispositifs dont le coût énergétique unitaire est négligeable. L'ampleur de ce transfert dépendra de la vitesse de convergence entre les capacités des modèles on-device et les attentes utilisateurs en matière de qualité de réponse.

8.3. Compute scheduling et orchestration énergétique intelligente

Une contre-force technologique déjà en déploiement mérite attention : le compute scheduling intelligent, qui aligne les workloads de calcul sur la disponibilité temporelle d'énergie décarbonée. Google a pionnerié cette approche avec son système de carbon-aware computing, qui déplace les workloads d'entraînement et de batch processing vers les heures et les régions où l'électricité disponible est la plus décarbonée.

Cette approche réduit l'empreinte carbone marginale sans réduire la consommation énergétique absolue. Elle est inefficace pour les workloads à faible latence (inférence temps réel) mais pertinente pour l'entraînement, le fine-tuning et les tâches batch — qui représentent une fraction substantielle de la consommation totale. Son efficacité dépend toutefois du mix énergétique local : dans des systèmes majoritairement fossiles, déplacer un workload de la nuit au jour (pour capter le solaire) peut n'avoir qu'un impact marginal si la production de base reste carbone-intensive.

8.4. Énergies renouvelables dédiées : efficacité conditionnelle

Les hyperscalers investissent massivement dans les énergies renouvelables. Microsoft a contracté plus de 10 GW de capacités à horizon 2030 ; Google vise une alimentation 24/7 en énergie décarbonée. Leur efficacité réelle dépend toutefois étroitement du mix énergétique local. Dans des systèmes électriques encore majoritairement fossiles (comme plusieurs réseaux régionaux aux États-Unis, dont le mix oscille entre 60 % et 80 % d'énergies fossiles), l'achat d'électricité dite « verte » ne correspond pas nécessairement à une consommation physique d'énergie décarbonée additionnelle. Les Power Purchase Agreements et certificats d'origine opèrent surtout une réallocation comptable de la production existante.

En l'absence d'ajout simultané de capacités bas carbone pilotables (nucléaire, hydroélectricité, géothermie) ou de solutions de stockage longue durée, l'augmentation de la demande liée aux datacenters se traduit mécaniquement par une hausse de la production fossile ailleurs dans le système. Les stratégies de « verdissement » peuvent ainsi réduire l'empreinte déclarée sans diminuer — voire en augmentant — les émissions globales à l'échelle du réseau.

8.5. Régulations émergentes et positionnement européen

Les cadres réglementaires évoluent rapidement, mais de manière fragmentée. L'Union européenne débat d'une taxe carbone sur le compute ; la Californie envisage des obligations de transparence sur la water footprint ; Singapour impose des normes strictes de PUE ; l'Irlande plafonne la part des datacenters dans la consommation électrique nationale.

Le positionnement européen mérite une attention particulière. L'initiative Gaia-X, les projets de cloud souverain français (S3NS, Bleu), et les réflexions autour d'un « cloud de confiance » européen représentent une tentative institutionnelle de répondre aux asymétries identifiées dans cet article. Toutefois, leur efficacité reste incertaine. La fragmentation des initiatives nationales, les délais de mise en œuvre, et l'absence de capacités de gravure avancée en

Europe limitent la portée de ces stratégies. L'Europe dispose en revanche d'atouts réels sur les contraintes physiques : un mix électrique parmi les moins carbonés au monde (France, Suède, Finlande), une expérience réglementaire avancée (AI Act, RGPD), et un savoir-faire industriel en nucléaire. La question est de savoir si ces avantages structurels seront mobilisés à l'échelle et à la vitesse requises.

Le risque de carbon leakage est réel : certaines juridictions peuvent maintenir des régulations permissives pour attirer les investissements. Aucune contre-force prise isolément ne constitue une solution unique. Leur combinaison pourrait repousser l'échéance des contraintes physiques, sans les supprimer.

IX. Implications stratégiques

9.1. Pour les entreprises technologiques

À court terme (2025–2026), la priorité stratégique consiste à sécuriser l'accès à l'énergie avant une tension accrue des marchés. Les contrats énergétiques long terme deviennent des actifs stratégiques au même titre que les brevets ou les talents. La diversification géographique multi-sites entre juridictions complémentaires renforce la résilience. À moyen terme (2027–2030), l'intégration verticale partielle dans la chaîne énergétique (fusion, SMR, TerraPower) marque une rupture culturelle pour un secteur historiquement centré sur le logiciel. Enfin, les arbitrages coût-latence-empreinte devront être explicitement formalisés : les workloads tolérant des latences élevées déportés vers des sites énergétiquement avantageux, l'inférence temps réel restant proche des utilisateurs finaux.

9.2. Pour les régulateurs

Les régulateurs disposent d'une fenêtre d'action limitée, située approximativement entre 2025 et 2027, avant que les investissements ne produisent des effets de verrouillage durables. Les leviers disponibles incluent les normes d'efficience contraignantes (plafonds de PUE, obligations de réutilisation de la chaleur fatale), les marchés carbone sectoriels appliqués au compute, la planification territoriale (zoning, quotas régionaux), et la transparence obligatoire (reporting standardisé et audité des consommations). Leur efficacité dépend de leur coordination internationale : sans harmonisation minimale, les politiques nationales risquent de se neutraliser.

9.3. Pour les territoires

Les territoires disposant d'avantages comparatifs font face à un choix stratégique : valoriser ces actifs pour attirer des investissements massifs, ou limiter leur exposition à une industrie énergivore et potentiellement volatile. Les opportunités sont réelles (milliards d'investissement, emplois qualifiés, recettes fiscales). Les risques sont substantiels (dépendance mono-industrielle, stranded assets, conflits d'usage).

Les controverses autour des méga-bassines agricoles en France constituent un précédent éclairant. Bien que reposant sur une logique hydrologique rationnelle, ces infrastructures ont cristallisé une opposition sociale intense, non sur leur faisabilité technique, mais sur la perception d'un accaparement privatif d'une ressource commune. Les datacenters IA s'inscrivent dans une dynamique comparable. Sans mécanismes explicites de contreparties locales, de gouvernance et de compensation environnementale, l'acceptabilité sociale devient un facteur limitant dominant, indépendamment de la pertinence technique du projet.

X. Conclusion — la question thermodynamique (et donc politique)

L'intelligence artificielle a franchi un seuil : ce qui était présenté comme une industrie du « logiciel augmenté » se comporte désormais comme une industrie électro- et hydro-intensive, territorialisée, contrainte par des ressources physiques et donc politiquement disputées.

Ce changement de régime met au jour une contradiction centrale. D'un côté, les acteurs dominants promeuvent une démocratisation de l'IA afin d'élargir la base d'usages, d'accélérer l'adoption et d'amortir des investissements massifs déjà engagés. De l'autre, la matérialité des contraintes réintroduit la gestion du coût d'opportunité. Lorsque l'électricité, l'eau, les données et la capacité réseau deviennent des facteurs limitants à coût marginal croissant, tous les usages ne se valent plus. Mobiliser un mégawatt pour la recherche scientifique, la santé, la modélisation climatique ou des infrastructures critiques n'a pas le même rendement social que mobiliser le même mégawatt pour des usages récréatifs ou faiblement productifs.

La tension devient alors systémique. La logique de démocratisation maximale — nécessaire au ROI industriel — entre en conflit avec une logique d'allocation efficiente de ressources rares. Autrement dit, la question n'est plus seulement qui peut accéder aux modèles, mais quels usages justifient, au regard de leur valeur collective, le coût marginal réel des ressources qu'ils mobilisent.

La question finale n'est donc plus seulement algorithmique. Elle est thermodynamique, puis immédiatement politique : quelle part d'électricité, d'eau, de données, de réseau et d'acceptabilité sociale nos sociétés accepteront-elles d'allouer à l'IA — et au détriment de quels autres usages ? La réponse se jouera dans les arbitrages d'infrastructure, la gouvernance des externalités, et la hiérarchisation collective des usages légitimes du compute. L'IA n'échappe pas aux lois de la physique : elle oblige à décider explicitement comment nous choisissons de les payer.

Références

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. The Economic Journal, 99(394), 116–131.

Jevons, W. S. (1865). The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal Mines. Macmillan.

Hooker, S. (2021). The hardware lottery. Communications of the ACM, 64(12), 58–65.

Villalobos, P., et al. (2022). Will we run out of data? An analysis of the limits of scaling datasets in machine learning. arXiv:2211.04325.

Muennighoff, N., et al. (2023). Scaling data-constrained language models. Advances in Neural Information Processing Systems (NeurIPS).

Shumailov, I., et al. (2023). The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493.

International Energy Agency (2024). Electricity 2024: Analysis and Forecast to 2026. IEA Publications.

World Resources Institute (2023). Aqueduct Water Risk Atlas. WRI Global.

Boston Consulting Group (2024). From Potential to Profit with GenAI: Bridging the Gap. BCG Henderson Institute.

McKinsey Global Institute (2025). The State of AI: How Organizations Are Rewiring to Capture Value. McKinsey & Company.

Gallup (2024). The Real State of AI Adoption in American Workplaces. Gallup Inc.

US Census Bureau (2025). Business Trends and Outlook Survey: Artificial Intelligence Use. BTOS Q2 2025.

Stanford HAI (2024). Artificial Intelligence Index Report 2024. Stanford University.

European Parliament (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).

Sorrell, S. (2009). Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. Energy Policy, 37(4), 1456–1469.

Kaack, L. H., et al. (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12, 518–527.