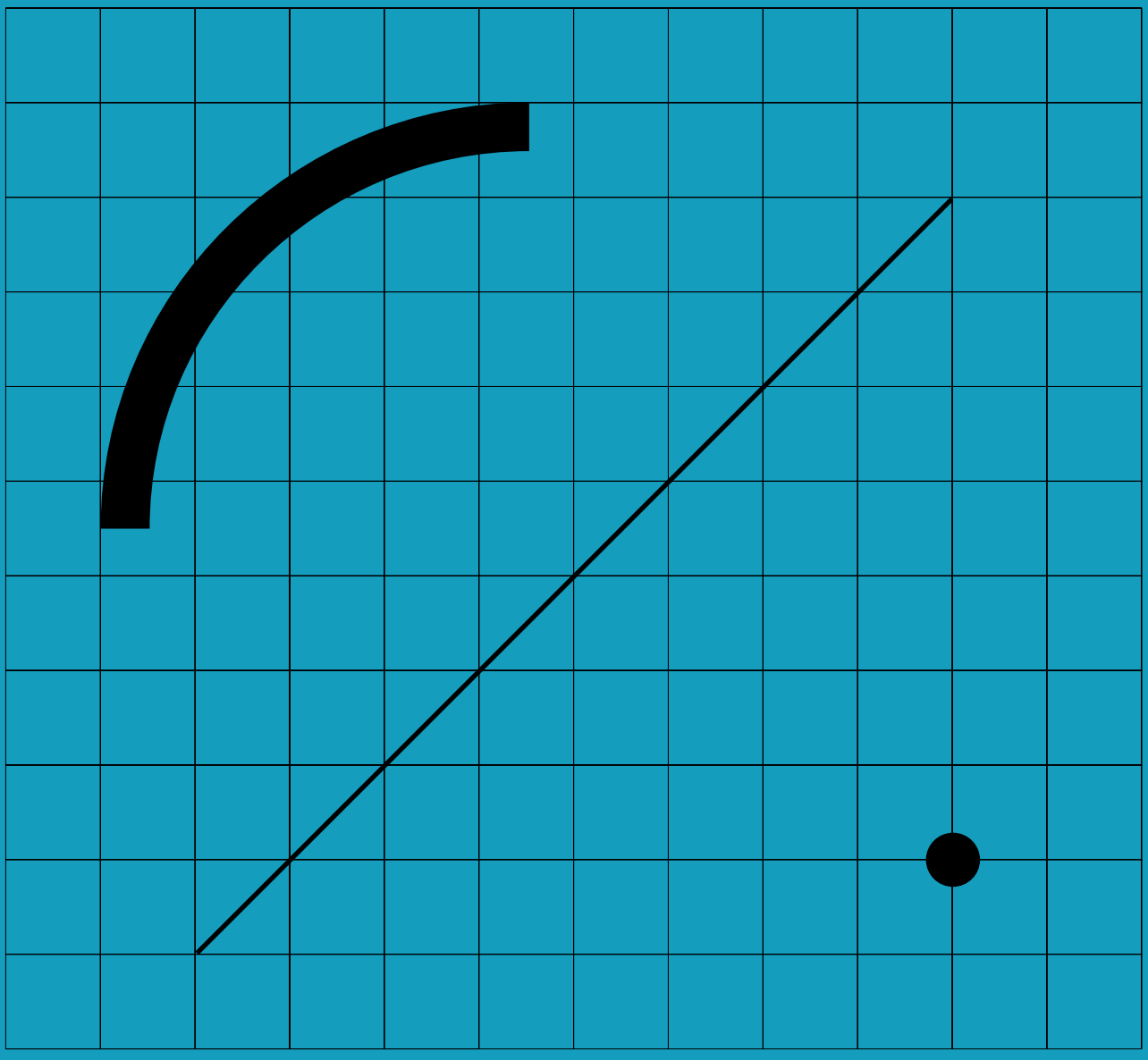
Financer les infrastructures pour une IA européenne compétitive



NOTE POUR L'ACTION - FÉVRIER 2025

groupe d'études geopolitiques

groupe d'études geopolitiques

Financer les infrastructures pour une IA européenne compétitive

AUTEURS

RAPHAËL DOAN • HAUT FONCTIONNAIRE ET ESSAYISTE. ANCIEN ÉLÈVE DE L'ÉCOLE NORMALE SUPÉRIEURE ET DE L'ÉCOLE NATIONALE D'ADMINISTRATION, AGRÉGÉ DE LETTRES CLASSIQUES, IL A NOTAMMENT PUBLIÉ QUAND ROME INVENTAIT LE POPULISME (2019, CERF), LE RÊVE DE L'ASSIMILATION (2021, PASSÉS COMPOSÉS), ET SI ROME N'AVAIT PAS CHUTÉ (2023, PASSÉS COMPOSÉS).

ANTOINE LEVY • ASSISTANT PROFESSOR AT UC BERKELEY HAAS SCHOOL OF BUSINESS.

VICTOR STORCHAN • SCIENTIFIC DIRECTOR OF "POWER OF AI" SERIES PUBLISHED BY LE GRAND CONTINENT. VICTOR RECENTLY HELD THE POSITIONS OF AI RESEARCH LEAD AT MOZILLA AI AND VP AI/ML AT JP MORGAN CHASE.

CITATION
RAPHAËL DOAN, ANTOINE LEVY, VICTOR STORCHAN,FINANCER LES INFRASTRUCTURES POUR UNE IA EUROPÉENNE COMPÉTITIVE, NOTE POUR L'ACTION, GROUPE D'ÉTUDES GÉOPOLITIQUES, 10 FÉVRIER 2025.

En bref

L'intelligence artificielle (IA) devient une infrastructure critique pour l'économie mondiale, comparable à l'électricité ou à Internet. D'ici 2030, la majorité des tâches cognitives, industrielles et administratives seront augmentées ou automatisées par l'IA, ce qui aura un impact profond sur la productivité et la compétitivité économique. Sans un investissement massif et coordonné dans ses infrastructures, l'Europe risque de dépendre technologiquement des États-Unis et de la Chine, ce qui menacerait son modèle démocratique, sa souveraineté et sa compétitivité. D'ici 2030, la liberté économique et technologique aura un prix : celui des GPU.

LES INFRASTRUCTURES DE L'IA, SOCLE FONDAMENTAL DE L'AVENIR DE L'EUROPE

- **Une transformation économique en cours** : d'ici 2030-2035, l'IA et les modèles de langage (LLM) deviendront omniprésents dans tous les secteurs (industrie, services, santé, finance, éducation). L'accès à la puissance de calcul sera un facteur de production aussi essentiel que le charbon au XIXe siècle.
- **Un décrochage économique préoccupant** : depuis 2000, la productivité européenne progresse deux fois moins vite que celle des États-Unis. Sans infrastructures propres, l'Europe ne bénéficiera pas des gains de productivité massifs apportés par l'IA.
- **Un risque de dépendance stratégique** : actuellement, 70 % de la puissance de calcul mondiale pour l'IA est détenue par les Etats-Unis dont 80% par les *hyperscalers* américains. L'Europe ne représente que 4 % de la capacité mondiale et souffre de coûts énergétiques industriels 1,5 à 3 fois plus élevés que ceux des États-Unis.

SCALABILITÉ DES MODÈLES D'IA : IMPLICATIONS MATÉRIELLES ET ÉNERGÉTIQUES

- Schématiquement, le développement des modèles d'IA comporte deux grandes phases: l'entraînement (la phase d'apprentissage à partir des données) et l'inférence (utilisation du modèle pour générer des réponses et accomplir des tâches).
- **À capacité constante, le coût d'entraînement d'un modèle d'IA diminue avec le temps** (d'un facteur proche de 4 chaque année) du fait des gains de performance du hardware (améliore la capacité de calcul par dollars) et efficacité algorithmique (qui réduit le nombre d'opérations à effectuer pour entraîner le modèle). En introduisant entre autres des innovations aux niveaux de l'architecture des modèles, des méthodes d'entraînement ou de l'optimisation des vitesses d'interconnexion entre GPUs, DeepSeek a pu trouver ces gains d'efficacité.
- **Les coûts d'entraînement des modèles augmentent d'un facteur d'environ 2.4 chaque année.** Les entreprises développant l'IA à la frontière technologique ne vont pas se mettre à dépenser moins pour l'entraînement de leurs modèles. Ainsi, le CapEx des GAFA relatifs aux dépenses pour les data centers et la puissance de calcul ont dépassé les 100 milliards en 2024, avec une augmentation de plus de 35 % par rapport à l'année précédente.

LA TAILLE ET LE COÛT DES INFRASTRUCTURES DE L'IA EN EUROPE ET EN FRANCE

- L'Europe représente aujourd'hui seulement 4 % de la puissance de calcul mondiale déployée pour l'IA.
- Pour la France, un objectif minimal serait de sécuriser sur son territoire une capacité de calcul dédiée à l'IA équivalente à 10 % de celle des États-Unis, reflétant ainsi son poids relatif dans le PIB américain – soit environ 5-6 GW à horizon 2028.
- Si elle se fixait l'objectif de représenter 16 % – en proportion de son poids dans l'économie mondiale – de la puissance de calcul globale en IA à horizon 2030, elle devrait porter à 20 GW sa puissance

énergétique dédiée à l'IA.
- Cela chiffrerait l'objectif français à 250-300 milliards d'euros d'investissements (soit plus de deux fois le montant de 109 milliards annoncé par le président Emmanuel Macron le 9 février) et l'objectif européen à 600-850 milliards.

LE FINANCEMENT DES INFRASTRUCTURES DE L'IA

- Dans le contexte d'une série de contraintes budgétaires au niveau national une double perspective s'impose : activer la capacité d'emprunt collective de l'Union via la réallocation des fonds non utilisés de NextGenEU et l'émission de nouvelles obligations pan-européennes pour financer les infrastructures IA et énergétiques.
- Du côté des investissements privés, **la modulation des primes de risque régulatoires associées au secteur de l'IA et au financement des infrastructures** pour les assureurs et les banques permettrait de mobiliser davantage de ressources financières.
- La création de «fonds IA» dédiés, accessibles aux épargnants européens sans limite de montant et éligibles aux produits d'investissement défiscalisés (à l'image du PEA en France), pourrait être étudiée.
- Du point de vue des entreprises, **la généralisation et l'harmonisation par le haut à l'échelle de l'Union des crédits d'impôts innovation et recherche** permettrait d'assurer la continuité fiscale du traitement des investissements en clusters entre pays de la zone.

LA NÉCESSITÉ D'UNE SIMPLIFICATION RÉGLEMENTAIRE

- **Un cadre législatif inadapté** : aujourd'hui, il faut au moins cinq ans pour installer un data center en France en raison de lourdeurs administratives et de délais de raccordement électrique.
- **Une simplification à accélérer** :
    - Le projet de loi de simplification de la vie économique prévoit que les centres de données de dimension industrielle soient qualifiés de projet d'intérêt national majeur, permettant d'accélérer certaines procédures. Déposé au Parlement en avril 2024, ce texte n'a pas encore été voté par les deux chambres.
    - Cette démarche trouve un écho outre-Manche, dans le cadre des «AI Growth zones» britanniques, qui vise à réduire les obstacles réglementaires à la construction des data centers.
- En France, la puissance électrique nucléaire, décarbonée, ne pourra devenir un atout pour l'essor de l'IA que si son accès est priorisé et le prix de son raccordement maîtrisé.

En 2012, la révolution du Deep Learning s'est accélérée avec la publication des performances d'AlexNet, l'un des premiers modèles entraînés alors sur deux GPU NVIDIA GTX 580. Treize ans plus tard, les infrastructures indispensables au développement de l'IA impliquent des coûts d'équipement (CapEx) sans précédent, que ni le secteur privé ni la puissance publique ne peuvent assumer seuls.

Il est urgent pour l'Union européenne de sécuriser la puissance de calcul, l'énergie nécessaire et l'écosystème industriel de la donnée pour soutenir des objectifs économiques et stratégiques ambitieux en matière d'IA pour les dix prochaines années.

Pour la France, cela pourrait impliquer, d'une part, de se doter d'une puissance de calcul sur son sol capable de soutenir au moins cinq acteurs nationaux ou européens développant des modèles de fondation à la frontière technologique tout en répondant aux besoins d'usage a minima des secteurs stratégiques de son économie. Conjointement, il est essentiel de garantir une fourniture énergétique adaptée et de structurer un écosystème industriel robuste autour de la donnée.

POURQUOI INVESTIR : LES INFRASTRUCTURES DE L'IA, SOCLE FONDAMENTAL DE L'AVENIR DE L'EUROPE

Quel que soit le résultat final de l'initiative Stargate, l'annonce faite par OpenAI et ses associés à la Maison Blanche révèle l'état d'esprit dans lequel se place l'industrie américaine et l'État fédéral américain : les 500 milliards de dollars d'investissements privés annoncés dépassent les 30 milliards du projet Manhattan ou les 250 milliards du programme Apollo, exprimés en dollars actuels. Ce n'est pas qu'un effet de surenchère : l'intérêt économique de moyen terme de l'IA est beaucoup plus grand que celui de la course à l'espace des années 1960. Si l'Europe doit faire l'effort de s'y atteler sérieusement, ce n'est ni pour suivre une mode, ni pour courir après une politique de prestige : c'est une question de puissance tout à fait réelle.

L'économie de 2030-2035 sera structurellement transformée par l'omniprésence des modèles de langage (LLM).

Les usages potentiels de l'IA dans la santé ou l'éducation sont bien connus, mais ils ne s'arrêteront pas à ces secteurs et toucheront toutes les échelles de nos activités, des tâches les plus englobantes aux plus précises. Dans cette économie «LLM-isée», la majorité des tâches cognitives routinières – analyse de documents, transformation de données, traduction, rédaction, programmation, recherche d'information, prise de micro-décisions – seront augmentées ou automatisées par IA. Dans l'industrie manufacturière, des modèles spécialisés optimiseront en temps réel les chaînes de production, détecteront les anomalies et prévoiront les maintenances. Les PME s'appuieront sur des assistants IA pour automatiser leur comptabilité, leur relation client ou encore leurs processus RH. Dans la construction, des modèles de langage analyseront en continu les données des capteurs de l'Internet des objets sur les chantiers pour anticiper les risques structurels et séquencer les travaux. Les architectes génèreront et testeront des milliers de variations de leurs plans en fonction des contraintes techniques, environnementales et réglementaires. Des modèles spécialisés en droit analyseront la jurisprudence en temps réel, prépareront des contrats sur-mesure et détecteront les incohérences réglementaires. Ces transformations toucheront jusqu'aux artisans : plombiers et électriciens utiliseront des assistants pour diagnostiquer les pannes, suggérer des réparations ou générer le modèle d'une pièce manquante avant une impression 3D. Les LLM et autres modèles fondamentaux sont en passe de devenir une infrastructure critique, aussi essentielle que l'électricité ou le réseau Internet, intégrée dans la plupart des processus productifs et interactions économiques. Cette «LLMisation» de l'économie fera de l'accès à la puissance de calcul un facteur de production aussi crucial que l'était l'accès au charbon durant la révolution industrielle.

La dépendance technologique européenne aux infrastructures américaines pose de ce point de vue un risque systémique.

Une rupture d'approvisionnement en capacités de calcul – qu'elle soit due à des tensions géopolitiques, des sanctions économiques ou des choix stratégiques des fournisseurs – paralyserait des pans entiers de l'économie européenne. Les secteurs critiques comme la santé, l'énergie ou la défense perdraient leur capacité à exploiter l'IA, c'est-à-dire, dans dix ans, à fonctionner correctement. Cette vulnérabilité n'est pas théorique : les restrictions américaines sur l'export de puces vers la Chine et d'autres parties du monde – y compris certaines parties de l'Europe – montrent la réalité de ce levier géopolitique.

Ce nouveau risque survient alors que l'Europe subit déjà un décrochage économique par rapport aux États-Unis. Depuis 2000, le revenu disponible réel par habitant a progressé deux fois plus vite aux États-Unis que dans l'Union. L'écart de PIB par habitant en parité de pouvoir d'achat s'explique à 70 % par une productivité

plus faible en Europe. Sans maîtrise des infrastructures d'IA – de l'entraînement à l'inférence – ce fossé risque de devenir un gouffre. Les gains de productivité massifs permis par l'IA profiteront d'abord aux économies disposant des capacités de calcul nécessaires – et laisseront de côté ceux que l'on appelle déjà les «*GPU-poors*».

Cette situation exige un changement radical de paradigme en Europe.

D'abord sur l'allocation des ressources : plutôt que de continuer le saupoudrage actuel de subventions et d'investissements – qui divise ses ressources – l'Europe doit accepter des arbitrages difficiles et concentrer ses investissements. Les dépenses à réaliser sont gigantesques et elles impliqueront nécessairement de renoncer à d'autres projets et d'autres champs d'action. C'est un pari que nous ne pouvons pas nous permettre de ne pas faire. Notons qu'il ne concerne pas que les gouvernements : les fortunes privées françaises et européennes ont les moyens de positionner l'Union sur la carte mondiale de l'IA, à condition de le faire, là encore, de manière concentrée.

Sur l'approche des données, l'Europe doit dépasser une vision centrée uniquement sur la protection des données personnelles pour embrasser l'enjeu de la collecte et de l'accessibilité des données d'entraînement. Les administrations publiques, qui disposent de trésors de données dans la santé, l'éducation ou l'énergie, doivent montrer l'exemple en facilitant leur exploitation pour l'IA : ce n'est pas incompatible avec la confidentialité des données proprement personnelles et cela permettrait de débloquer de véritables avancées.

En visant une couverture des besoins en inférence essentiels de notre future économie, nous proposons un objectif clair et lisible. Le calcul et les hypothèses peuvent faire l'objet de discussions. Mais c'est, à notre sens, le critère sur lequel il faut se concentrer : il ne s'agit pas de faire de la recherche pour la recherche ni de décider à l'avance des choix technologiques qui doivent être laissés aux entreprises. Il s'agit de prévoir, par tous les moyens, notre future indépendance computationnelle.

Il y a deux cents ans, la révolution industrielle a mis un siècle à reconfigurer les équilibres anthropologiques de l'Europe et du monde. Ce que nous voyons des possibilités offertes par l'IA promet des bouleversements d'une ampleur comparable mais sur un horizon temporel beaucoup plus court. Une Europe dépendante d'infrastructures étrangères perdrait toute capacité à façonner son propre destin économique et social. Automatiser toutes les tâches cognitives qui peuvent l'être est un enjeu trop important pour que nous le laissions à d'autres qu'à nous-mêmes. En 2030, la liberté aura un prix concret : celui des processeurs.

SCALABILITÉ DES MODÈLES D'IA : IMPLICATIONS MATÉRIELLES ET ÉNERGÉTIQUES

Entraînement et inférence

Le développement des modèles d'IA se divise en deux grandes phases : l'entraînement, qui consiste en une phase d'apprentissage à partir de données, et l'inférence, qui correspond à leur utilisation effective pour générer des réponses et accomplir des tâches en temps réel.

Pour les modèles d'IA les plus avancés, l'entraînement repose sur des clusters massifs de processeurs graphiques (GPU) interconnectés dans des centres de données spécialisés. Lors du développement d'un modèle, il est essentiel de réaliser des entraînements exploratoires à petite et moyenne échelle afin de tester et valider des choix architecturaux, des optimisations d'entraînement ou des stratégies d'allocation de données avant de lancer un entraînement final à grande échelle. Ces phases intermédiaires doivent être intégrées dans le coût total du modèle lorsque l'on veut estimer la capacité de calcul nécessaire pour opérer à la frontière technologique. L'inférence, quant à elle, peut être exécutée sur des clusters GPU moins puissants ou des dispositifs *«edge»* – processus d'exécution des modèles d'IA directement sur les appareils locaux (tels que les smartphones, les appareils IoT ou dans les voitures) – selon les besoins d'usage. Contrairement aux centres de données dédiés à l'entraînement, ceux spécialisés dans l'inférence sont localisés proche des lieux d'utilisation afin de réduire la latence entre les utilisateurs et les serveurs.

Lois d'échelles

Les lois d'échelle empiriques de l'IA indiquent que toutes choses égales par ailleurs – qualités des données d'entraînement notamment – pour un entraînement optimal, la puissance de calcul doit être répartie à parts égales entre l'augmentation de la taille du modèle et l'augmentation de la quantité de données. Ainsi, alors que les budgets pour entraîner les modèles ne cessent de croître, la taille des jeux de données ainsi que celle des modèles augmentent en proportion. Globalement, le coût d'entraînement des modèles les plus avancés a augmenté d'un facteur 2 ou 3 depuis les huit dernières

années[1] atteignant des dizaines à centaines de millions de dollars. Ainsi, le modèle GPT4 d'OpenAI entraîné en 2022 (environ 2e25 FLOPs) utilisait alors un cluster de 20K GPUs A100 pour une consommation énergétique de 15-20 MW. Le modèle Llama 3 de Meta entraîné début 2024 (3.8e25 FLOPs) utilisait 16K GPUs H100 d'un cluster de 24K GPUs alors que Llama 4 devrait utiliser plus de 100K GPUs H100[2].

De même, l'émergence des *reasoning models* (*DeepSeek-R1, O1-mini, O3-mini* etc) ou modèles de raisonnement a montré qu'il était possible de passer à l'échelle selon une seconde dimension : à l'inférence, la performance est également fonction croissante de la quantité de calcul allouée pour que le modèle déroule et teste plusieurs raisonnements.

Efficacité algorithmique et matérielle (hardware)

Sous l'effet de la recherche et de l'innovation technologique, on observe une double dynamique. L'efficacité algorithmique – architectures de modèles, méthodes d'optimisation et d'entraînement – progresse pour réduire la quantité de calcul nécessaire pour atteindre une performance donnée. Les performances des GPUs, à prix donné, progressent – entre 2006 et 2021, au rythme d'un doublement tous les deux ans[3]. En conséquence, des estimations[4] établissent que le niveau de calcul nécessaire – et donc le coût – pour atteindre un niveau de performance donné diminue de moitié environ tous les 8 mois. D'autres observations avancent qu'à performance donnée, le coût d'un modèle sera divisé par 4 chaque année grâce aux avancées technologiques. Autrement dit, si entraîner un modèle coûte 100 millions de dollars aujourd'hui, ce coût tombera à 25 millions un an plus tard, puis à 6 millions dans deux ans etc.

L'exemple récent le plus marquant de gain d'efficacité algorithmique est celui de DeepSeek.

L'entreprise chinoise a trouvé des gains d'efficacité notamment en innovant à la fois sur l'architecture du modèle (facteur de sparsité inédit, MLA, GRPO, etc.) et en réécrivant en langage assembleur (PTX) les communications entre les GPUs et les noeuds de leurs clusters afin de palier aux limitations de vitesse d'interconnection des GPUs H800. Ces innovations ont permis à DeepSeek une meilleure utilisation de ses ressources à la fois à l'entraînement et à l'inférence. Cependant, cela ne signe pas non plus la fin des lois d'échelle en IA.

Les entreprises développant l'IA à la frontière technologique ne vont pas se mettre à dépenser moins pour l'entraînement de leurs modèles. À capacité de modèle donnée, les gains d'efficacité algorithmique et matérielle impliquent une réduction du coût d'entraînement pour développer le modèle ainsi qu'une réduction des coûts d'inférence – diminution d'un facteur 10 chaque année depuis 3 ans[5]. Mais lorsqu'ils opèrent à la frontière technologique, les laboratoires d'IA trouvent de nouvelles dimensions à faire passer à l'échelle[6] (*pretraining*, RL, temps de calcul à l'inférence, etc.) nécessitant une capacité de calcul inédite. Celle ci se reflète dans le CapEX des grands fournisseurs de *cloud* américains : en 2024, les *hyperscalers* ont dépensé un total de plus de 100 milliards de dollars pour les infrastructures de l'IA[7]. On estime le prix d'un gigawatt d'un centre de données équipés des dernières puces NVIDIA GB300 à 40 ou 50 milliards de dollars. À titre d'indication, le coût en puissance de calcul de DeepSeek est estimé à 100 millions par an[8], certainement autour de 500 millions depuis le début de fonctionnement de l'entreprise.

COMPARAISON INTERNATIONALE ET ÉTAT DES LIEUX DES STRATÉGIES SUR LES INFRASTRUCTURES DE L'IA

Une infrastructure européenne de l'IA et des financements sous-dimensionnés

1 — Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej et David Owen. 'The rising costs of training frontier AI models'. ArXiv [cs.CY], 2024. arXiv. https://arxiv.org/abs/2405.21015.

2 — Jowi Morales, «Meta is using more than 100,000 Nvidia H100 AI GPUs to train Llama-4 — Mark Zuckerberg says that Llama 4 is being trained on a cluster "bigger than anything that I've seen"», Tom's Hardware, 31 octobre 2024.

3 — Konstantin Pilz, Lennart Heim, Nicholas Brown, «Increased Compute Efficiency and the Diffusion of AI Capabilities», 13 février 2024.

4 — Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson et Jaime Sevilla. 'Algorithmic progress in language models'. ArXiv [cs.CL], 2024. arXiv. https://arxiv.org/abs/2403.05812.

5 — Guido Appenzeller, «Welcome to LLMflation — LLM inference cost is going down fast», Andreessen Horowitz, 12 novembre 2024.

6 — Nikhil Sardana, Jacob Portes, Sasha Doubov, Jonathan Frankle, «Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws», 31 décembre 2023.

7 — Jérôme Marin, «En 2024, Microsoft, Amazon, Google et Meta ont dépensé 100 milliards de dollars dans leurs infrastructures d'IA», L'Usine Digitale, 19 décembre 2024.

8 — Nathan Lambert, «DeepSeek V3 and the actual cost of training frontier AI models», Interconnects, 6 janvier 2025.

L'Europe représente aujourd'hui seulement 4-5 % de la puissance de calcul mondiale déployée pour l'IA[9].

Les entreprises européennes du *cloud* détiennent une part de marché inférieure à 5 %[10]. Si l'on considère les principales startups mondiales dans le domaine de l'IA, 61 % des financements mondiaux vont aux entreprises américaines, 17 % aux entreprises chinoises et seulement 6 % aux entreprises de l'Union[11]. Concernant les centres de données, au total l'Europe héberge 18 % de la capacité mondiale des centres de données, dont moins de 5 % appartiennent à des entreprises européennes, contre 37 % pour les États-Unis[12], une économie comparable. Les tarifs industriels européens (0,18 USD/kWh en moyenne) sont jusqu'à trois fois supérieurs à ceux des États-Unis, rendant les infrastructures d'IA plus coûteuses : certaines estimations calculent un coût d'installation de *data centers* de 1,5 à 2 fois plus élevé en Europe qu'aux États-Unis[13][14]. L'entreprise française Mistral AI a ainsi alerté en juin 2024 sur le manque de capacité de calcul pour l'entraînement des modèles d'IA sur le sol européen[15].

Stratégie européenne pour les infrastructures en IA

La Commission européenne a annoncé un plan d'*AI Factories* articulé autour des projets de supercalculateurs des États membres principalement dédiés à la recherche publique.

Cet investissement d'1,5 milliards d'euros s'inscrit dans le programme Digital Europe qui finance l'IA – infrastructures de données, d'évaluation et diffusion de l'IA dans l'économie – à hauteur de 2,1 milliards d'euros pour la période 2021-2027 et 2,2 milliards pour la mise à jour ou la construction de supercalculateurs[16].

Stratégies internationales pour les infrastructures en IA

Les États renforcent leurs efforts pour capter les investissements privés afin de financer les infrastructures stratégiques de l'IA – un levier clef des dynamiques géopolitiques et économiques.

Ces initiatives s'inscrivent dans une compétition accrue pour le leadership technologique.

Aux États-Unis, même si un certain scepticisme entoure les 500 milliards d'investissements et l'exécution du projet Stargate, il n'annule pas la nécessité fondamentale pour l'Europe de développer à l'échelle ses infrastructures de l'IA – l'éclatement de la bulle Internet n'a pas freiné l'émergence des acteurs du *cloud*.

Au Royaume-Uni, le plan d'action britannique pour les opportunités en IA introduit quant à lui des «*AI growth zones*», qui accélèrent les autorisations de construction de centres de données et facilitent l'accès au réseau énergétique.

La banque de Chine a annoncé un plan de financement de 1 trillion de yuans (140 milliards de dollars) pour soutenir les entreprises d'IA engagées dans la recherche fondamentale et l'industrialisation de l'IA[17]. Un plan pour la construction de huit centres de calcul et dix *data centers* nationaux a également été approuvé[18].

Demande et production mondiale

TSMC prévoit que la demande de serveurs dédiés à l'IA augmentera de 50 % par an au cours des cinq prochaines années[19]. Du côté de la production, des estimations prévoient une croissance annuelle de 35 %-60 % du volume

9 — ylan Patel, Daniel Nishball et Jeremie Eliahou Ontiveros, «AI Datacenter Energy Dilemma – Race for AI Datacenter Space», *SemiAnalysis*, 13 mars 2024.

10 — Alexander Sukharevsky, Eric Hazan, Sven Smit, Marc-Antoine de la Chevasnerie, Marc de Jong, Solveigh Hieronimus, Jan Mischke et Guillaume Dagorret, «Time to place our bets: Europe's AI opportunity», McKinsey, 1er octobre 2024.

11 — Mario Draghi, «The Future of European Competitiveness», Commission européenne, septembre 2024.

12 — Alexander Sukharevsky, Eric Hazan, Sven Smit, Marc-Antoine de la Chevasnerie, Marc de Jong, Solveigh Hieronimus, Jan Mischke et Guillaume Dagorret, «Time to place our bets: Europe's AI opportunity», McKinsey, 1er octobre 2024.

13 — Dylan Patel, Daniel Nishball et Jeremie Eliahou Ontiveros, «AI Datacenter Energy Dilemma – Race for AI Datacenter Space», *SemiAnalysis*, 13 mars 2024.

14 — Alexander Sukharevsky, Eric Hazan, Sven Smit, Marc-Antoine de la Chevasnerie, Marc de Jong, Solveigh Hieronimus, Jan Mischke et Guillaume Dagorret, «Time to place our bets: Europe's AI opportunity», McKinsey, 1er octobre 2024.

15 — Cynthia Kroet, «Mistral AI warns of lack of data centres and training capacity in Europe», *Euronews*, 14 juin 2024.

16 — Digital Europe Programme (DIGITAL) | EU Funding & Tenders Portal, Commission européenne.

17 — Sharveya Parasnis, «Bank of China Announces Investments Worth 1 Trillion Yuan to Develop AI Industry», Medianama, 28 janvier 2025.

18 — «China approves mega project for greater computing power, digital future», République populaire de Chine, 18 février 2022.

19 — «Q1 2024 Taiwan Semiconductor Manufacturing Co Ltd Earnings Call», 18 avril 2024.

de GPUs disponible[20].

LA TAILLE ET LE COÛT DES INFRASTRUCTURES DE L'IA EN EUROPE ET EN FRANCE

En Europe

La demande mondiale en puissance critique pour l'informatique des *data centers* passera de 49 gigawatts (GW) (dont 5 GW pour l'IA) en 2023 à 130 GW d'ici 2030[21], dont environ 40 GW seront consommés par l'IA.

La situation aux États-Unis est la suivante : en 2023, la puissance totale des *data centers* américains est estimée à 23 GW, représentant environ 5 % de la capacité électrique totale du pays – dont 3,3 GW spécifiquement alloués à l'IA. En se basant sur la demande en cartes graphiques spécialisées, les projections pour 2028 indiquent que la puissance totale des *data centers* aux États-Unis atteindra 83 GW, dont 56 GW dédiés à l'IA. Ceci confirmerait la part des États-Unis à environ 70 % de la puissance mondiale dédiée à l'IA – dont 80% détenus par les *hyperscalers* américains.

L'Europe représente aujourd'hui 4 à 5 % de la capacité de calcul pour l'IA déployée dans le monde, soit 0.25 GW[22]. Si elle se fixait l'objectif de représenter 16 % – en proportion de son poids dans l'économie mondiale – de la puissance de calcul globale en IA à horizon 2030, elle devrait porter à 20 GW sa puissance énergétique dédiée à l'IA. On retrouve un ordre de grandeur similaire à celui-ci, si l'Europe se fixe pour objectif de se mettre au niveau des États-Unis dans la part de puissance électrique allouée à l'IA dans les *data centers* en 2030 (installation de 17 GW[23]).

En France

Pour la France, un objectif minimal serait de sécuriser sur son territoire une capacité de calcul dédiée à l'IA équivalente à 10 % de celle des États-Unis, reflétant ainsi son poids relatif dans le PIB américain – soit environ 5-6 GW à horizon 2028.

Selon la répartition de cette puissance entre inférence et entraînement de modèle à la frontière technologique – de l'ordre de 10^25 FLOPS aujourd'hui et 10^26 FLOPS d'ici 2027[24] –, cela permettrait à la France de supporter sur son sol la puissance de calcul pour 3 à 5 acteurs de niveau mondial. En 2024, 40% des revenus de Nvidia pour ses *data centers* étaient liés à l'inférence. Google indique qu'entre 2019 et 2021, l'inférence représentait environ 60 % du calcul total utilisé en IA dans l'entreprise[25].

Coût

Comme nous l'avons rappelé, le coût d'installation de 1 GW de la prochaine génération de GPU Nvidia GB300 est estimé à 40 ou 50 milliards d'euros[26]. Pour la génération actuelle Hopper (H100), 1 GW d'installation pourrait représenter un coût de 15 à 23 milliards en fonction d'un facteur de dépréciation du matériel et d'ajustement du marché.

En d'autres termes, cela chiffrerait l'objectif français à 250-300 milliards d'euros d'investissements et l'objectif européen à 600-850 milliards.

COMMENT FINANCER LES INFRASTRUCTURES DE L'IA

Pour financer les colossaux investissements requis, la mobilisation de fonds à la fois privés et publics est nécessaire.

La capacité d'emprunt collective des États membres de l'Union, aujourd'hui sous-utilisée, pourrait être mise à disposition, en particulier par la réorientation des fonds non utilisés ou pas encore engagés du plan *NextGenerationEU*, et par l'émission nouvelle d'obligations pan-européennes dédiées au financement de l'infrastructure en clusters et en production d'électricité.

Du côté des financements privés, la modulation des primes de risque régulatoires associées au secteur de l'IA et au financement des infrastructures pour les

20 — Jaime Sevilla et al. (2024), «Can AI Scaling Continue Through 2030?», Epoch AI.

21 — Tim Fist et Arnab Datta, «How to Build the Future of AI in the United States», IFP, 23 octobre 2024.

22 — Cela représente environ 3% du total des 10 GW de puissance électrique installée dans les data centers européens.

23 — La demande en puissance électrique totale installée dans les data centers en Europe est estimée à 35 GW d'ici 2030.

24 — Pilz, Konstantin F., Yusuf Mahmood et Lennart Heim, AI's Power Requirements Under Exponential Growth: Extrapolating AI Data Center Power Demand and Assessing Its Potential Impact on U.S. Competitiveness. Santa Monica, CA: RAND Corporation, 2025.

25 — Tim Fist et Arnab Datta, «How to Build the Future of AI in the United States», IFP, 23 octobre 2024.

26 — «NVIDIA GB300 «Blackwell Ultra» Will Feature 288 GB HBM3E Memory, 1400 W TDP», 23 décembre 2024.

assureurs (directive Solvency 2) et pour les banques (accords Bâle III) est une piste potentielle pour mobiliser les capitaux issus en particulier de l'épargne-retraite et de l'assurance-vie. Plus généralement, la possibilité de lever des «fonds IA» dédiés, accessibles aux épargnants européens sans limite de montant et éligibles aux produits d'investissement défiscalisés (à l'exemple du PEA en France) pourrait être étudiée.

Du point de vue des entreprises, la généralisation et l'harmonisation par le haut à l'échelle de l'Union des crédits d'impôts innovation et recherche permettrait d'assurer la continuité fiscale du traitement des investissements en clusters entre pays de la zone. Le traitement fiscal des dépenses relevant de l'acquisition, de la maintenance et du stockage des données peut également être ajusté pour permettre la dépréciation accélérée de ces investissements, au vu de la nécessité d'un rapide rattrapage des pays les plus avancés en la matière.

LA NÉCESSITÉ D'UNE SIMPLIFICATION RÉGLEMENTAIRE

La hausse des moyens de financement est une condition nécessaire mais insuffisante de la puissance de calcul. Le développement des infrastructures sur le sol européen et français est freiné par le cumul de procédures administratives, de recours judiciaires et de délais de raccordement électrique des centres de données.

Au total, l'installation de *data centers* sur le territoire français dure au moins cinq ans.

Pour réduire ces délais, dans la continuité des recommandations de la Commission de l'IA, le projet de loi de simplification de la vie économique prévoit que les *data centers* de dimension industrielle soient qualifiés de projet d'intérêt national majeur, permettant d'accélérer certaines procédures et d'être exempté de l'application de certaines réglementations, comme le Zéro artificialisation nette. Déposé au Parlement en avril 2024, ce texte n'a pas encore été voté par les deux assemblées. Cette démarche trouve un écho outre-Manche, dans le cadre des «*AI Growth zones*» britanniques, qui vise à réduire les obstacles réglementaires à la construction des *data centers*.

Au-delà de l'achèvement impératif de ce processus de simplification normatif et procédural, il conviendra de faciliter le raccordement par RTE – et par ses équivalents nationaux européens – des infrastructures de données au réseau électrique. En France, la puissance électrique nucléaire, décarbonée, ne pourra devenir un atout pour l'essor de l'IA que si son accès est priorisé et le prix de son raccordement maîtrisé.