

Intelligence artificielle : le vrai coût environnemental de la course à l'IA

« L'IA et la réussite de la transition énergétique vont de pair. Nous n'assisterons pas à une transition à grande échelle vers une énergie sans carbone sans les avancées significatives que l'IA promet d'apporter. » Cette citation de Mélanie Nakagawa, la responsable développement durable de Microsoft en février 2025, est à l'image de la pensée magique que l'on rencontre de plus en plus fréquemment dans les médias et la société. Celle d'une IA mystique et providentielle qui serait la solution à tous nos problèmes, et qui justifierait son développement effréné depuis l'arrivée de ChatGPT en 2022.

Alors qu'Emmanuel Macron a annoncé en février 2025 des investissements privés de 109 milliards d'euros pour développer l'IA et que plus de 35 nouveaux projets de centres de données ont été annoncés en France, nous vous proposons une analyse approfondie pour déconstruire ces discours et découvrir la face cachée de l'intelligence artificielle. Comme le résume la chercheuse en IA Kate Crawford, elle repose sur une triple extraction : extraction de ressources naturelles, extraction de données, exploitation humaine. Dans cet article nous nous concentrerons sur les impacts environnementaux directs de l'IA (consommation énergétique, émissions carbone, impacts locaux des *data centers*), mais insistons sur la nécessité d'appréhender ces technologies dans leur ensemble pour comprendre les bouleversements sociétaux que son adoption massive génère (enjeux de pouvoir, surveillance de masse, militarisation, inégalités sociales, etc.)

Nous commencerons par définir ce qu'est l'intelligence artificielle, son évolution récente, les acteurs qui la sous-tendent et détaillerons le fonctionnement particulier des IA génératives. Puis dans un second temps nous décrirons l'infrastructure matérielle qui la soutient. Loin de flotter au-dessus de toute réalité physique, comme le champ lexical du *cloud* et de la « dématérialisation » aime à nous le faire croire, ces technologies ont des conséquences matérielles très tangibles.

Sommaire

- Qu'est ce que l'intelligence artificielle ?
- L'ère de l'IA générative depuis la sortie de ChatGPT en 2022
- L'IA, bien plus qu'une technologie
- Qui est derrière l'IA ?
- Le voyage d'une requête ChatGPT
- L'inférence pour l'usage quotidien
- Le problème de la transparence
- Quel est l'impact d'une requête ChatGPT ?
- L'optimisation de la consommation électrique de l'IA
- Les impacts directs de l'intelligence artificielle
- La consommation électrique des data centers explose à cause de l'IA
- Émissions de CO₂, la roue libre
- Google, Microsoft, Meta. Quid des émissions des géants du numérique ?
- La consommation d'eau
- Ressources abiotiques pour la fabrication des équipements et bâti : dépendance aux métaux et à l'industrie chimique
- Conclusion
- Infographie

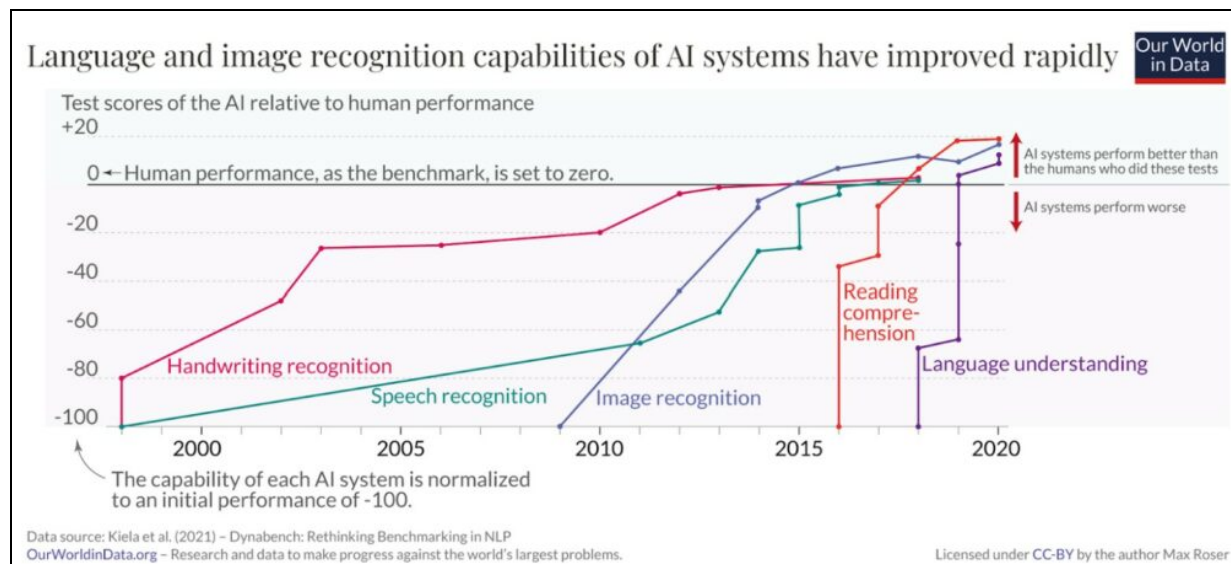
Qu'est ce que l'intelligence artificielle ?

Un peu d'histoire et de définitions

Commençons par définir les termes. La définition technique de l'intelligence artificielle est la capacité d'un logiciel, via des algorithmes, à effectuer et émuler des tâches typiquement associées à l'intelligence humaine, comme l'apprentissage, le raisonnement, la résolution de problème, la perception ou la prise de décision. Il est important de comprendre qu'il n'existe pas une intelligence artificielle mais *des* intelligences artificielles, et qu'elles sont présentes dans notre quotidien depuis bien plus longtemps qu'on ne le pense.

La conférence de Dartmouth en 1956 lance l'intelligence artificielle comme un domaine de recherche à part entière. Le terme *Intelligence Artificielle* est même choisi par rapport à d'autres expressions comme *Automata Studies* pour des raisons marketing plus à même d'attirer des financements.

Depuis, l'histoire de l'IA a connu des rebondissements et des désillusions face aux potentiels de commercialisation de cette technologie. Depuis les années 2010, le domaine connaît une nouvelle vague d'intérêt scientifique et industriel pour au moins quatre raisons : une grande quantité de données disponibles pour nourrir les algorithmes suite à l'avènement des réseaux sociaux et de la numérisation, plus de puissance de calcul, et des investissements sans précédent basés sur les promesses répétées d'une technologie révolutionnaire auto-alimentée par les géants de la tech et les États.

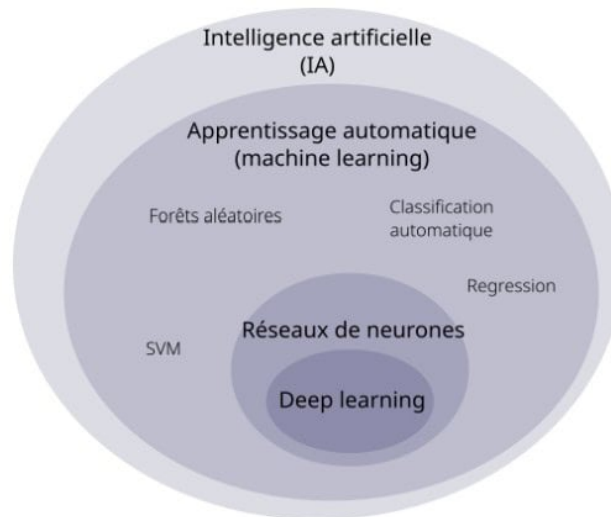


Par exemple comme le montre ce graphique retraçant la performance des systèmes d'IA comparée à celles d'humains sur des tâches précises, les systèmes d'IA sont progressivement devenus performants à la reconnaissance manuscrite (comme par exemple retranscrire des montants sur un chèque), puis sur la reconnaissance vocale, puis sur la reconnaissance d'image (par exemple pour reconnaître un passant sur une caméra, un visage pour déverrouiller son téléphone, ou un monument dans une photo de réseaux sociaux), puis sur l'analyse de texte (pour analyser les avis des utilisateurs d'un produit ou améliorer la recherche sur Google). L'amélioration des performances sur ces tâches spécifiques s'est accélérée depuis les deux dernières décennies.

L'ère de l'IA générative depuis la sortie de ChatGPT en 2022

L'IA traditionnelle

Si l'Intelligence Artificielle est un domaine de recherche et d'ingénierie vaste et qui existe depuis bientôt huitante ans, depuis 2022 nous sommes entrés dans une nouvelle ère : celle de l'explosion de l'IA générative – symbolisée par l'ouverture au grand public de ChatGPT en novembre 2022. Pour appréhender les impacts environnementaux de l'IA, il est important de comprendre la différence entre les familles de l'IA et ce que cette vague récente de l'IA générative a de particulier.



Source : Wikipedia, « Apprentissage automatique »

Le Machine Learning

Depuis les années 2000, quand on parle d'IA c'est principalement pour parler d'apprentissage machine (*Machine Learning*), c'est-à-dire un programme informatique qui apprend à partir d'exemples plutôt que d'être explicitement programmé avec des règles. Par exemple, dans le cadre d'une traduction, plutôt que d'écrire toutes les règles possibles pour traduire un texte de français à anglais (conjugaison, grammaire, dictionnaires de mots, ...) l'algorithme va « lire » les mêmes textes en français et en anglais et « découvrir » les règles complexes pour passer d'une langue à une autre sans forcément les expliciter. Il est important de préciser ici qu'il n'y a rien d'intelligent ou d'autonome : l'explication est simplifiée pour vulgariser le fonctionnement (guillemets sur « lire » et « découvrir »), mais c'est en réalité un calcul de probabilités qui est réalisé.

C'est grâce à ce nouveau paradigme informatique que les logiciels de traduction ne traduisent plus mot à mot.

Deep Learning et apparition des LLM (Large Language Models)

L'apprentissage machine est également une grande famille d'algorithmes avec des dizaines de techniques mathématiques possibles. L'une d'entre elles est l'apprentissage profond (*deep learning*) qui consiste à utiliser une modélisation mathématique appelée « réseaux de neurones artificiels », très vaguement inspirée du fonctionnement de notre cerveau.

En 2017 les équipes de recherche de Google publient un papier de recherche « Attention is all you need », qui propose une nouvelle technique mathématique pour l'apprentissage profond : le *Transformers* (le fameux T de ChatGPT). Cette technique va accélérer la construction d'algorithmes d'IA généralistes, c'est à dire non dédiés à une seule tâche, grâce à ce qu'on appelle l'auto-apprentissage ou le pré-entraînement. Ces algorithmes sont souvent appelés *Large Language Models* ou *LLMs* (grands modèles de langage). Prenons un exemple : trier automatiquement un email dans les spams et traduire un texte de français à anglais, deux tâches qui n'ont a priori rien à voir. Cependant les deux font appel à un problème sous-jacent identique : comprendre le langage et analyser un texte. L'auto-apprentissage va donc consister à montrer une quantité massive de textes pour modéliser statistiquement le fonctionnement du langage (vocabulaire, syntaxe, grammaire, expressions). En 2019, un laboratoire à but « non lucratif » encore inconnu appelé OpenAI sort alors l'algorithme GPT-2 (pour *General Pretrained Transformers*). GPT-3 sort en 2020 et montre des résultats qui impressionnent les chercheurs et le grand public ; le modèle est 100 fois plus gros que la version précédente.

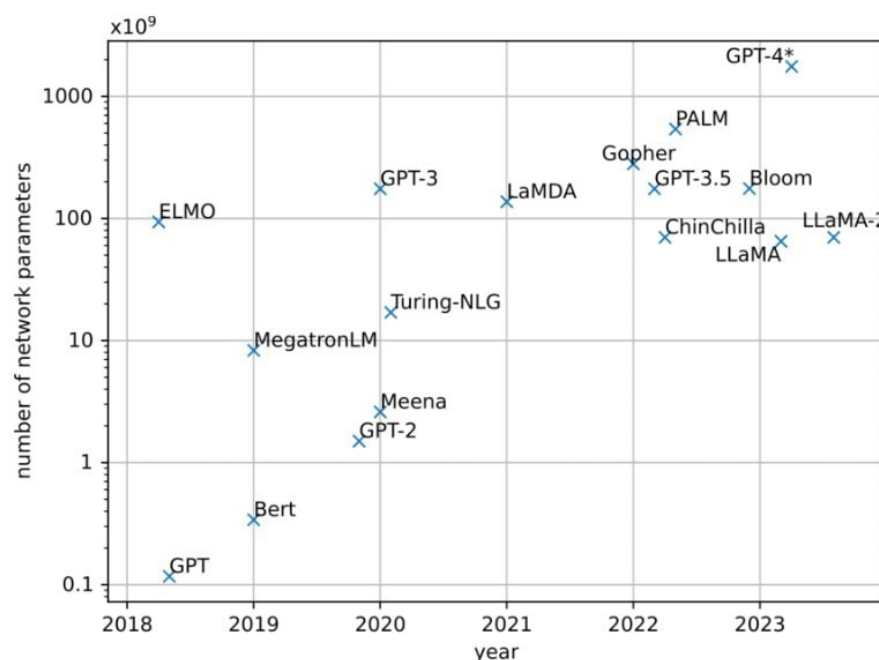
En novembre 2022, OpenAI lance ChatGPT et déclenche la nouvelle vague d'IA générative grand public. Si générer du contenu synthétique (texte, image, son...) n'est donc pas nouveau, ce qui va être déterminant avec ChatGPT, c'est l'interface conversationnelle (avec les *prompts*) et un modèle spécialisé pour la conversation grâce à une technique appelée RLHF (*Reinforcement Learning from Human Feedback*), où des annotateurs humains classent et améliorent les réponses du modèle, principalement des employés Kenyans expérimentant des conditions de travail et de rémunération très précaire.

L'ampleur de cette adoption est vertigineuse. ChatGPT a atteint 400 millions d'utilisateurs actifs hebdomadaires en février 2023, après avoir conquis son premier million d'utilisateurs en seulement cinq jours – une vitesse d'adoption 12 à 60 fois plus rapide que les applications de réseaux sociaux populaires. Aux États-Unis et au Royaume-Uni, 40% des ménages utilisent désormais ces outils, tandis que dans des pays comme le Brésil, l'Inde, l'Indonésie, le Kenya et le Pakistan, plus de la moitié des utilisateurs d'internet emploient l'IA générative au moins une fois par semaine. L'adoption en entreprise suit le même rythme effréné : parmi les grandes firmes des pays de l'OCDE, le taux d'adoption est passé de 15% en 2020 à près de 40% en 2024.

Triple Effet Kiss Cool

L'explosion des impacts environnementaux de l'IA provient donc de trois effets qui se renforcent : Premièrement, une croissance exponentielle de l'IA générative liée à l'usage quotidien de millions d'utilisateurs. Cette adoption est accélérée par la facilité d'utilisation (une interface conversationnelle pas chère facile à manipuler avec les instructions ou *prompts*) et par une excitation économique des entreprises et États qui poussent à l'adoption de l'IA dans un nombre toujours croissant d'applications. La plupart du temps sans demander aux utilisateurs leur autorisation. Une étude du projet de recherche Limites numériques montre que "jamais une fonctionnalité n'aura autant été poussée en si peu de temps, spatialement, graphiquement, interactivement et de manière autant répétée dans nos sites web, services et logiciels."

Deuxièmement, avec une complexification des modèles d'IA qui sont plus versatiles, plus gros, demandent plus de données et plus de puissance de calcul. Avec ces outils dit généralistes (General Purpose AI), on peut autant générer un rap dans le style d'Eminem sur la compote de pommes, que générer des vidéos de Will Smith qui mange des spaghetti, que de faire des résumés de réunions à partir d'enregistrements vocaux. La taille des LLMs a été multipliée par 10000 entre 2018 et 2023.



La taille des Large Language Models a été multipliée par 10 000 en cinq ans.

Troisièmement, une accélération de la construction des centres de données dans le monde entier pour faire face à l'augmentation de la demande en services d'IA générative, provoquant des impacts locaux sur le réseau électrique, sur l'eau, ou la pollution.

L'IA, bien plus qu'une technologie

Enfin, dans cette tentative de définition de ce qu'est l'intelligence artificielle, il est essentiel de ne pas se cantonner à une simple définition technique. L'IA est « un champ de recherche pluridisciplinaire, un enjeu géopolitique, un domaine industriel, un discours publicitaire et parfois presque religieux, un outil informatique complexe mais également un phénomène social planétaire » comme l'indique Thibault Prévost dans son livre *Les prophètes de l'IA*.

La mention du terme intelligence donne l'impression d'un système infaillible, en maîtrise, et participe d'une mythification de la technologie. Or, un système d'IA fait des calculs probabilistes, des erreurs, comporte de nombreux biais. Le terme artificiel quant à lui tente de faire oublier les données sous-jacentes créées par des humains, souvent extraites sans leur consentement et le travail d'annotation manuel indispensable à leur fonctionnement.

L'intelligence artificielle n'est donc jamais neutre : à partir des données sur lesquelles elle a été entraînée, de sa modélisation, de la manière dont elle s'insère dans la société, elle incarne et reproduit une vision du monde. Son développement et ses impacts sont intrinsèquement liés à l'industrie qui l'a fait sortir du laboratoire. « Pour comprendre en quoi l'IA est fondamentalement politique, nous devons aller au-delà des réseaux de neurones et de la reconnaissance statistique des formes et nous demander *ce* qui est optimisé, pour *qui*, et *qui* décide » écrit Kate Crawford. Commençons alors avec ceux qui la façonnent.

Qui est derrière l'IA ?

Ainsi pour appréhender les impacts environnementaux de l'IA, il faut comprendre son industrie et saisir les rapports de force économiques et politiques qui la sous-tendent. Les acteurs de l'IA s'articulent autour de cinq maillons essentiels : les données, les modèles, la puissance de calcul (*compute*), l'infrastructure, et les applications et services.

Les données sont la matière première de l'IA, si une grande quantité de données brutes existe en licence libre sur internet, les géants du numérique (Meta, Google, Amazon, Microsoft) dominent la collecte massive de données via leurs plateformes sociales, moteurs de recherches, services cloud et terminaux numériques. S'ajoutent des courtiers en données personnelles (Experian, Equifax, ...) ou de marché (Nielsen, Acxiom, ...), ainsi que de nombreux acteurs spécialisés pour nettoyer, étiqueter, et annoter la donnée pour l'IA (Amazon Mechanical Turk, Scale AI, ...) qui emploient une armée invisible de "travailleurs du clic" précaires.

Les modèles d'IA sont développés autant par des universités et des entreprises privées. En revanche, les modèles les plus avancés d'IA génératives sont développés majoritairement (90% en 2024) par des géants privés Meta (Llama), Google (Gemini) ou Alibaba (Qwen), ou des entreprises spécialisées en IA comme OpenAI (GPT), Anthropic (Claude), xAI d'Elon Musk (Grok) ou Mistral AI. En effet, entraîner un modèle peut coûter en 2025 plusieurs dizaines à centaines de millions d'euros : ce qui est complètement inaccessible pour les universités ou d'autres entreprises et institutions. En plus, les laboratoires publics de recherche en IA n'ont accès qu'à quelques centaines de puces de calculs GPUs, quand les laboratoires privés en ont des centaines de milliers.

La puissance de calcul (*compute*) nécessite aussi une chaîne de valeur avec en premier lieu les entreprises du *cloud* : le marché est dominé par Amazon (AWS), Microsoft (Azure), Google (GCP), IBM, Oracle ou Alibaba. Au niveau du matériel comme la conception des puces de calcul GPUs, NVIDIA écrase la concurrence avec +80% des parts de marché, suivi d'AMD et Intel. La production de ces puces est-elle réalisée en grande majorité par TSMC (*Taiwan Semiconductor Manufacturing Company*) qui produit +60% des semi-conducteurs mondiaux, tandis qu'ASML (Pays-Bas) détient un quasi-monopole complet sur les machines de lithographie EUV indispensables à leur fabrication.

L'infrastructure physique du calcul s'articule autour des centres de données. Ils sont détenus directement par les géants de la tech (Meta, Microsoft, Amazon, Google, xAI, Alibaba) ou loués

auprès de spécialistes comme Equinix ou Digital Realty. Les réseaux et fournisseurs d'énergie se réorganisent autour de l'IA : nucléaire (Constellation Energy, EDF, ...), turbines à gaz (Shell, ExxonMobil, ...), systèmes de refroidissement et de gestion énergétique (Schneider Electric, Vertiv, ...), transport et distribution électrique. Toute cette infrastructure nécessite des investissements monumentaux de plusieurs dizaines à centaines de milliards de dollars par des fonds d'investissement, des fonds souverains ou des banques.

Les applications et services constituent le dernier maillon qui rend l'IA accessible aux utilisateurs finaux (particuliers et entreprises). Au niveau grand public, les interfaces conversationnelles comme ChatGPT, Gemini, Claude ou Perplexity dominent, tandis que les APIs de ces mêmes entreprises permettent aux développeurs d'intégrer l'IA dans leurs propres applications. Des milliers de *startups* spécialisées développent des applications sectorielles. Les entreprises de services numériques (ESN) comme Accenture, Capgemini ou CGI intègrent l'IA dans les systèmes d'information des grandes entreprises, et les cabinets de conseil comme McKinsey, BCG ou Deloitte vendent des « stratégies de transformation par l'IA ».

La dimension géopolitique de l'IA

La domination des intérêts privés sur l'IA est donc partout : de la recherche et développement de modèles, des infrastructures logicielles et matérielles aux terminaux et plateformes numériques. Cette asymétrie se traduit par une influence politique considérable : le secteur de la tech est celui qui dépense le plus en lobbying à Bruxelles avec 110 millions d'euros par an – Google et Meta en tête. Cette concentration du pouvoir économique et politique autour de l'IA explique pourquoi les enjeux environnementaux ne peuvent être dissociés de ces questions de pouvoir et de démocratie. Les décisions qui vont déterminer l'ampleur de l'impact climatique de l'IA – déploiement et adoption forcée des usages, système énergétique, allocation des ressources, expansion des *data centers* – sont prises par une poignée d'acteurs privés en fonction de leurs intérêts financiers sans consultation démocratique.

Cette situation arrange d'ailleurs parfaitement les États qui, percevant l'IA comme un enjeu de souveraineté technologique dans la course à l'innovation mondiale, déroulent le tapis rouge à ces entreprises par le biais d'investissements publics massifs, d'exemptions fiscales et de réglementations complaisantes (comme l'article 15 de la loi PLS votée en France en Juin 2025 pour encourager l'expansion des *data centers* en France). Comprendre cette réalité industrielle et géopolitique est indispensable pour appréhender les enjeux de la transition écologique à l'ère de l'IA.

L'IA est un gouffre financier en quête de rentabilité

Cependant au sein du grand banquet de l'IA, ce n'est pas si évident de manger à sa faim. Les investissements sont tellement monumentaux qu'il est difficile pour tous ces acteurs d'atteindre la rentabilité et de trouver un retour sur investissement. OpenAI perd de l'argent tous les mois (au moins 1 milliard de \$ en 2024) et à chaque requête sur ChatGPT. Microsoft admet que l'IA générative ne génère pas suffisamment de valeur économique. Au global, malgré les centaines de millions d'utilisateurs, il n'y a pas assez de revenus et d'utilisations pour justifier les centaines de milliards investis dans l'infrastructure matérielle, les modèles d'IA, et les salaires de stars de football des chercheurs en IA.

Pour combler cette bulle économique, les entreprises d'IA ont plusieurs idées et stratégies. Déjà, elles sont obligées de réduire les coûts autant que possible et de faire du forcing pour encourager l'adoption par tous les moyens. Elles testent cherchent à intégrer de la publicité dans les applications d'IA générative (comme Snapchat ou Perplexity). après les posts sponsorisés sur les réseaux sociaux, préparez-vous pour les réponses sponsorisées sur ChatGPT. Sam Altman, PDG d'OpenAI se dit réticent comme le disaient les fondateurs de Google ou Facebook, mais s'y prépare quand même. Enfin, elles font surtout ce qui marche le mieux : vendre un futur radieux, où les promesses d'une IA générale qui nous sauverait de la dette publique, du changement climatique, du cancer et des inégalités, permettent de continuer à attirer des financements. Mais derrière l'intérêt est

cyniquement bien économique, la définition d'une IA générale entre Microsoft et OpenAI est d'ailleurs celle qui permettrait de générer 100 milliards de profits.

Pourquoi c'est important ? Parce que pour le moment en 2025, tout va très vite. Pour anticiper les impacts environnementaux, il faut anticiper l'évolution des usages. Énormément d'argent est investi pour provoquer une accélération la plus rapide possible mais le déploiement de l'IA pourrait aussi connaître de nombreuses transformations et revers, si la rentabilité ne progresse pas.

Le voyage d'une requête ChatGPT

Pour comprendre simplement l'impact environnemental d'une IA générative, essayons de décomposer ce qu'il se passe quand on pose une question à ChatGPT.

ChatGPT fonctionne grâce à un LLM (*Large Language Model*) : ces algorithmes récents dans la grande famille des algorithmes d'IA qui ont appris à prédire le mot suivant après avoir appris sur plusieurs milliards d'exemples. Comme leur nom l'indique, ces algorithmes sont "grands", et nécessitent une puissance de calcul conséquente largement inatteignable pour les puces de calcul que l'on retrouve dans nos outils du quotidien (téléphone, ordinateur, etc.). Il faut donc aller chercher cette puissance ailleurs, et c'est dans les "*data centers*" (centre de données) qu'on la trouve.

Les centres de données sont de grands bâtiments qui vont stocker des milliers de serveurs qui auront pour simplifier deux fonctions principales :

- Le stockage : pour stocker, gérer et récupérer des données – comme des photos, des vidéos, des archives.
- Le calcul ou *compute*: pour traiter et analyser des données à grande échelle. C'est dans ces infrastructures spécifiques qu'ont lieu l'entraînement et les calculs des modèles d'IA, grâce à des puces spécialisées comme les ASIC, FPGA ou encore GPU (Graphical Processing Unit ou cartes graphiques) que l'on retrouve également dans les ordinateurs pour les jeux vidéo. Ces puces peuvent réaliser des millions d'opérations en parallèle, et ce faisant vont consommer de l'électricité et chauffer.
-

Mais avant de pouvoir poser une question à ChatGPT il a fallu fabriquer le modèle. Dans le monde de l'IA, on appelle ça l'entraînement.

L'entraînement d'un modèle d'IA

L'entraînement consiste à montrer à un modèle des millions d'exemples pour qu'il modélise la donnée pour reproduire des modifs et faire ensuite des prédictions justes sur de nouvelles données qu'il n'a jamais vues. L'entraînement est un processus intensif et chronophage qui s'effectue sur des puces spécialisées comme les GPU. Un seul GPU haut de gamme (par exemple la puce H100 de Nvidia) a une puissance d'environ 700 watts, soit autant qu'un micro-ondes. Les modèles d'IA de pointe nécessitent des clusters (ensembles) de milliers de GPU fonctionnant en parallèle dans un même data center pour des raisons d'efficacité de communication entre les processeurs. Ils vont concentrer une puissance électrique gigantesque sur une période relativement courte.

La consommation électrique et les émissions lors de la phase d'entraînement va varier drastiquement selon la taille du modèle, la durée d'entraînement (le nombre d'heures de GPUs mobilisés), et l'intensité carbone de l'électricité utilisée à l'emplacement du *data center* (au moins 10 fois plus élevée aux Etats-Unis qu'en France). Voici quelques exemples :

Modèle	Paramètres	Heures GPU	Consommation	Émissions CO ₂
BLOOM (2022)	176B	~1M heures	433 MWh	30 tCO ₂ eq
Llama 3.1 8B (2024)	8B	1,46M heures	~400 MWh	420 tCO ₂ eq
Llama 3.1 70B (2024)	70B	7M heures	~1,9 GWh	2 040 tCO ₂
Llama 3.1 405B (2024)	405B	30,8M heures	~8,6 GWh	8930 tCO ₂ eq
GPT-3 (2020)*	175B		1,3 GWh	552 tCO ₂ eq
GPT-4 (2023)*	~280B		~42 GWh	

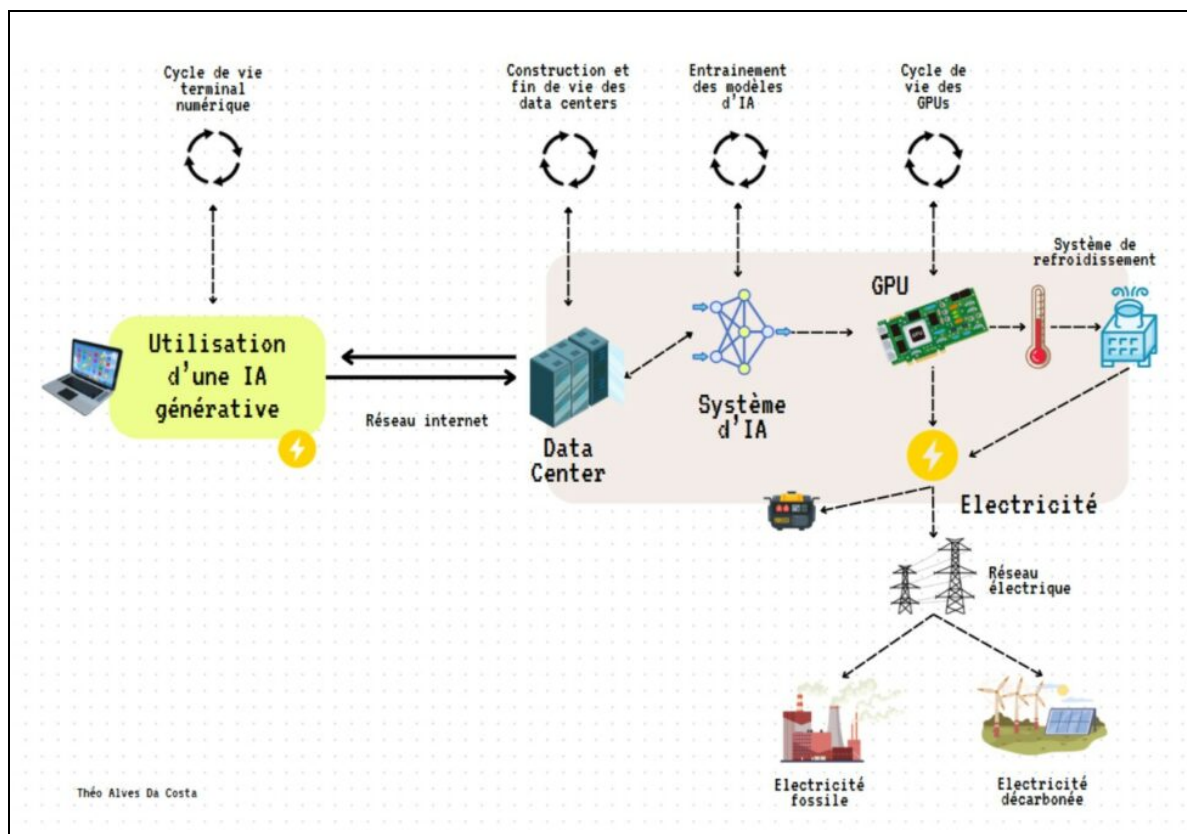
Pour BLOOM, la consommation électrique est connue, pour les modèles Llama de Meta la consommation électrique est estimée en fonction des heures de GPUs et des émissions de CO₂ déclarées. Pour les modèles GPT, nous avons encore moins de données précises déclarées – cependant des études externes ont fait des estimations (IEA, EpochAI)

Pour se rendre compte de ce que cela représente, *quand* Meta utilise 39 millions d’heures de GPU au total pour les différentes versions de Llama3.1, c’est l’équivalent de 4500 années de calcul en continu. Cela a nécessité plus de 16 000 puces de calcul H100, et 11 GWh de consommation énergétique. Pour GPT4, cela a nécessité 3 mois de calcul sur 25 000 GPUs.

Selon une étude d’Epoch AI (2024), la situation s’accélère : la puissance requise pour entraîner les modèles d’IA de continue de doubler chaque année malgré les optimisations matérielles et logicielles.

D’un point de vue CO₂ si vous connaissez vos ordres de grandeur, les émissions associées à l’entraînement d’un modèle peuvent paraître au final plutôt petites relativement à d’autres industries. Déjà, les hypothèses déclarées par les concepteurs sont souvent incomplètes : Carbone4 réhaussait par exemple les émissions de Llama de 9000 à 19000 tCO₂eq en prenant en compte un périmètre complet (sans compter encore l’usage). Et pour connaître l’impact total de l’entraînement, il faudrait pouvoir additionner tous les modèles entraînés. Mais surtout, il manque un élément majeur de l’équation : l’usage.

Dans l’IA classique, c’était traditionnellement la phase d’entraînement qui concentrait le plus d’émissions de CO₂ car il y avait très peu d’utilisateurs et des algorithmes plus simples. Mais avec l’IA générative, l’équation s’est totalement renversée car les usages se sont massifiés. Sasha Luccioni, scientifique et experte des questions environnementales de l’IA chez Hugging Face, a établi qu’autour de 100 millions d’utilisations pour ce type de modèle, la phase d’inférence devient plus gourmande en énergie que l’entraînement initial. Ce seuil critique est atteint en seulement quelques heures sur ChatGPT, qui compte désormais au moins 1 milliard de requêtes par jour. Pour les *data centers*, autour de 80% de la consommation électrique est réservée à l’inférence. En résumé, cela signifie qu’aujourd’hui pour les modèles d’IA générative populaires, l’impact environnemental de l’usage quotidien dépasse très largement celui de leur création.



L'inférence pour l'usage quotidien

Une fois le modèle entraîné, on peut l'utiliser pour poser une question :

- La question est envoyée depuis votre ordinateur ou téléphone au service de ChatGPT. Elle va rejoindre un centre de données plus ou moins loin de chez vous. Par exemple, si le calcul était réalisé dans des nombreux centres de données en Virginie ou en Oregon aux Etats-Unis, elle traverserait un câble sous-marin sous l'Océan Atlantique.
- La question est traitée par le système d'IA de ChatGPT dans les serveurs du data center – qui réalise également les calculs des millions de questions reçues en simultanément par les utilisateurs du monde entier, parfois en mutualisant les calculs pour gagner en efficacité.
- Pour donner une réponse, ChatGPT réalise un calcul. Ce calcul s'effectue dans les serveurs (et en particulier les cartes graphiques GPU) des centres de données qui vont nécessiter de l'électricité. Cette électricité peut être fournie par le réseau électrique local (à l'endroit du data center) ou par des générateurs supplémentaires de secours, non raccordés au réseau. L'intensité carbone de l'électricité va varier grandement en fonction des modes de production d'électricité utilisés au moment et à l'endroit de son utilisation (charbon, gaz, renouvelable ou nucléaire).
- La réalisation de nombreux calculs fait chauffer les serveurs. Pour les garder à une température optimale, le centre de données possède un système de refroidissement, qui lui aussi va nécessiter de l'électricité et de l'eau.
- La réponse calculée est ensuite renvoyée à l'utilisateur et traverse les câbles en sens inverse vers votre terminal utilisateur.

Chaque élément de cette chaîne a lui-même un cycle de vie qui apporte son lot d'impacts environnementaux matériels – à la fabrication ou à la fin de vie – que l'on décortique en détails plus bas.

Le problème de la transparence

Peu de données fiables sur l'impact de l'IA

« Une requête sur ChatGPT c'est équivalent à 10 recherches sur Google ! » Depuis 2023, ce chiffre est martelé dans les médias et symboliserait la consommation excessive d'énergie de l'IA. 75% des articles parlant de l'impact environnemental de l'IA mentionnent cette statistique. En réalité, nous n'en savons pas grand-chose. L'IA est en 2025 une industrie très opaque : les géants de l'industrie ne communiquent quasiment aucune donnée ou information fiable et vérifiable.

Si on déroule le fil de cette affirmation reprise dans tous les médias, une publication scientifique récente a retracé sa généalogie : elle provient d'une remarque faite à la volée par John Hennessy, le président d'Alphabet (maison mère de Google), lors d'une interview avec Reuters en 2023. Il avait déclaré qu'« avoir un échange avec une IA connue sous le nom de grand modèle de langage coûte probablement 10 fois plus qu'une recherche standard par mots-clés ».

Cette remarque a ensuite été utilisée comme base pour estimer « environ 3 Wh par interaction LLM », le chiffre de la recherche Google étant tiré d'un *blog post* de Google de 2009 qui indiquait 0,3 Wh d'énergie par recherche. Cette affirmation est problématique à plusieurs niveaux. D'abord, John Hennessy n'a aucun lien avec OpenAI ou Microsoft (qui fournit l'infrastructure de calcul pour ChatGPT), donc son commentaire est basé sur des informations de seconde main. Ensuite, le chiffre de Google date d'il y a plus de seize ans, une époque où la recherche web utilisait des techniques de recherche par mots-clés sans modèle d'IA plus énergivores.

Quel est l'impact d'une requête ChatGPT ?

Et on connaît encore moins l'impact d'une requête ChatGPT qui peut être significativement différente en fonction des paramètres suivants :

- Quelle version de ChatGPT est utilisée – GPT5, GPT4o, GPT3.5, GPT4o-mini, o3-mini, o4, GPT4o-20250325 ? Combien de paramètres y a-t-il dans le modèle utilisé ?
- Quelle quantité d'information est ingérée ? C'est-à-dire est-ce que vous posez juste une question d'une centaine de mots ? Ou est-ce que vous copiez-collez 4 pages ou ajoutez un PDF à votre question ?
- Quelle quantité d'information est générée ? Une réponse en une phrase, ou un rapport d'une vingtaine de pages ?
- Quelles fonctionnalités sont utilisées ? De la génération de texte, ou bien également de la génération d'image ou la capacité d'aller chercher sur internet ?

La réponse à ces questions va déterminer le temps nécessaire pour réaliser le calcul ainsi que le nombre de puces mobilisées, et donc la consommation électrique totale pour chacun de ces usages. Et pour connaître l'impact environnemental total, il faudrait également connaître le nombre d'utilisateurs, le modèle de carte graphique utilisée, les optimisations serveurs réalisées, le type de data center et son système de refroidissement, sa localisation ou encore l'intensité carbone du réseau électrique local.

Or, nous n'en savons rien. Il n'existe que très peu de données sur l'impact environnemental de l'IA, comme le rappelle très justement une nouvelle étude du MIT. Les fournisseurs d'IA orchestrent cette opacité et ne dévoilent quasiment aucune information pour des raisons de compétitivité (la consommation électrique pouvant montrer des choix économiques et stratégiques ou des secrets industriels), de sécurité (certaines normes de sécurité numérique imposent de ne pas dévoiler l'emplacement des data centers), réputation (pour masquer des impacts environnementaux galopants).

Cette opacité n'est pas anecdotique. Selon une analyse récente de Sasha Luccioni et al, 84% de l'utilisation des modèles de langage se fait via des modèles sans aucune transparence environnementale. En analysant les données de mai 2025 d'OpenRouter, les chercheurs ont montré que parmi les 20 modèles les plus utilisés, un seul (Meta Llama 3.3 70B) a directement publié des données environnementales. En termes de volume, seulement 2% de l'utilisation concerne des modèles transparents sur leur impact environnemental, 14% ne publient rien mais sont suffisamment ouverts pour faire au moins des études énergétiques tierces, et une écrasante majorité

de 84% des usages passent par des modèles sans aucune information environnementale. Cette situation s'est d'ailleurs dégradée depuis 2022 avec la compétition grandissante provoquée par ChatGPT.

Peu de données... et du greenwashing

A noter que pendant l'été 2025 ont été publiés par Mistral et Google deux études pour documenter pour la première fois quelques impacts environnementaux d'une requête "médiane" faite sur leurs IA grand public. Cependant ces études sont intentionnellement sélectives et ne donnent pas assez de détails pour permettre de comparer et comprendre les impacts tout en détournant l'attention des impacts globaux cumulés en mettant en valeur les gains d'efficacité et en sur-responsabilisant les individus.

En résumé en 2025, même s'il est possible d'obtenir des approximations, nous ne connaissons pas la consommation électrique d'une requête sur ChatGPT, de la génération d'une image studio Ghibli ou d'un starter pack. Comme le résumait Sasha Luccioni pour Wired : « Cela me sidère qu'on puisse acheter une voiture et connaître sa consommation au 100 kilomètres, mais qu'on utilise tous ces outils d'IA tous les jours sans avoir la moindre mesure d'efficacité, aucun facteur d'émission, rien. »

L'hypocrisie de Sam Altman

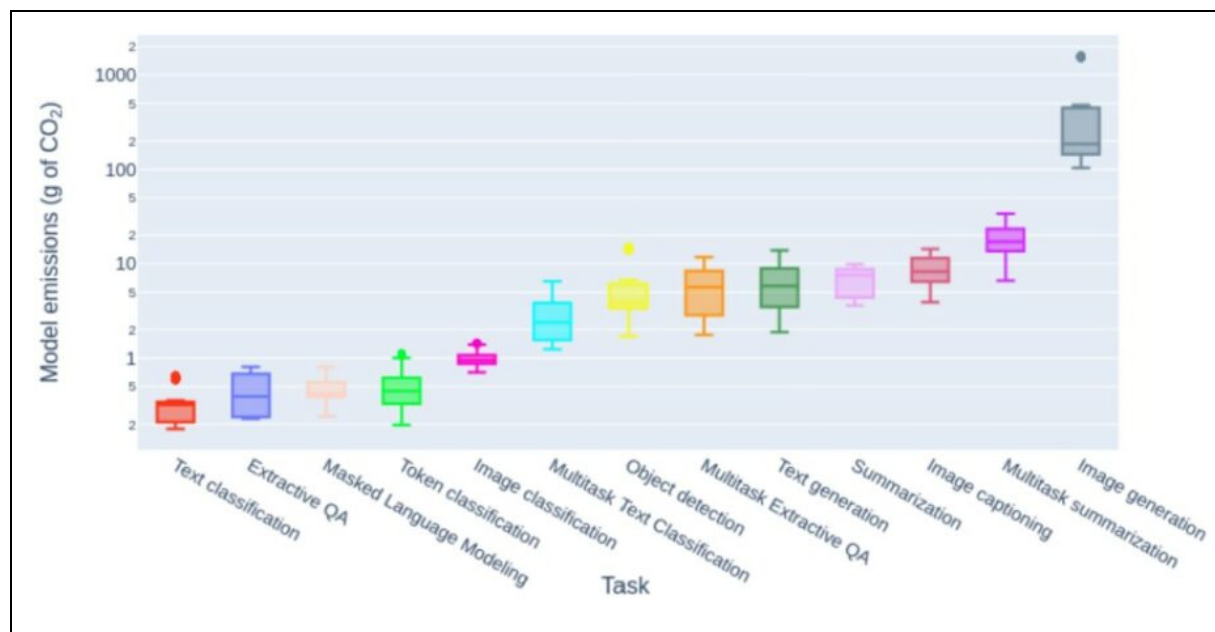
Sam Altman, PDG d'OpenAI – l'entreprise qui commercialise ChatGPT – déclarait en juin 2025 dans son dernier billet de blog "qu'une requête moyenne sur ChatGPT consomme 0,34 Wh d'énergie, soit "environ ce qu'un four utiliserait en un peu plus d'une seconde, ou ce qu'une ampoule à haute efficacité utiliserait en quelques minutes" dans le but de balayer la critique sur la consommation énergétique de l'IA.

Aucune précision, définition du périmètre ou de la méthodologie, ni audit externe ne permet de se fier à ce chiffre le rendant inutilisable mais particulièrement efficace dans la fabrique du doute. Paradoxalement au même moment, il déclarait également au cours d'une conférence « qu'une partie significative de l'électricité mondiale devrait être utilisée pour l'IA ». Résumons : ça ne consomme rien, mais en même temps on a besoin de toute l'électricité du monde.

Que sait-on alors sur la consommation d'une requête ?

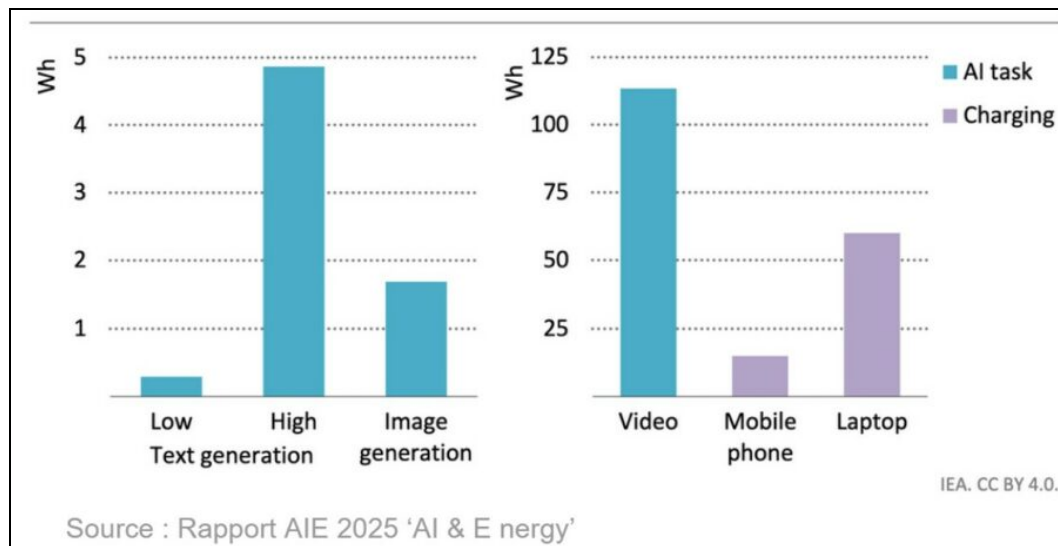
S'il est impossible de connaître la consommation de ChatGPT pour le moment, nous avons tout de même quelques études scientifiques plus large sur l'impact environnemental de l'IA générative. En particulier, il est possible de mesurer la consommation électrique de modèles d'IA « ouverts » (*open weights*), c'est-à-dire des modèles que l'on peut lancer sur ses propres serveurs et dont on peut donc mesurer la consommation énergétique (avec des outils comme CodeCarbon). Cela permet d'étudier différents usages, et d'approximer les ordres de grandeur.

C'est ce qu'a fait la chercheuse Sasha Luccioni en 2023, dans l'étude scientifique de référence "Power Hungry Processing" qui chiffre pour la première fois la consommation électrique d'un grand nombre d'algorithmes ouverts. Elle découvre que classifier un texte (par exemple en 'spam') est une tâche relativement anodine. Dans le contexte précis de l'étude réalisée en 2023, générer du texte est en moyenne 25 fois plus énergivore que classifier. Générer une image est en moyenne 60 fois plus énergivore que générer du texte, soit de l'ordre de 1 à 3 Wh.



Ce graphique montre les émissions de CO₂ (en abscisse) en fonction des différentes tâches réalisées à partir du papier de recherche 'Power hungry processing' de Sasha Luccioni. Attention : l'échelle d'impact est logarithmique. Depuis 2023 ces chiffres ont évolué.

Quant à la génération de vidéo, elle explose tous les compteurs : d'après le dernier rapport Energy & AI de l'AIE, une vidéo de 6 secondes à 8 images par seconde nécessite environ 115 Wh, soit l'équivalent de la recharge de deux ordinateurs portables.



Cette hiérarchisation des consommations énergétiques selon les usages est importante pour comprendre l'évolution de l'impact environnemental de l'IA et permettre des projections plus précises de la demande énergétique globale en fonction de l'adoption de chaque type d'usage.

L'optimisation de la consommation électrique de l'IA

Les techniques d'optimisation

Pour pouvoir vraiment comprendre l'évolution des usages et de la consommation électrique, il faut aussi prendre en compte les optimisations qui sont effectuées pour rendre les systèmes d'IA plus économes en énergie. Face à l'explosion de la consommation énergétique de l'IA et donc des coûts financiers opérationnels, l'industrie de l'IA multiplie les innovations pour optimiser l'efficacité de ses modèles. Ces techniques peuvent être logicielles, liées au déploiement technologique, ou matérielles.

La première famille d'optimisation concerne des techniques logicielles d'optimisation pour baisser la consommation énergétique à l'usage qui peuvent être mises en place par les chercheurs et concepteurs des systèmes d'IA. Par exemple, la *quantization* permet de stocker les informations d'un modèle d'IA avec moins de précision. L'élagage (*pruning*) revient à supprimer des parties inutiles d'un modèle de réseaux de neurones. La distillation de connaissance permet de créer des versions allégées des plus gros modèles en transférant une partie de la connaissance à un plus petit modèle moins énergivore. L'architecture dite de *MoE (Mixture Of Experts)* va permettre d'éviter de mobiliser toute la connaissance du modèle à chaque fois. Toutes ces méthodes permettent de réduire la consommation électrique par requête (à l'inférence) tout en conservant quasiment la même qualité de réponse.

La deuxième famille des techniques d'optimisation est liée à ce qu'on appelle le déploiement des systèmes d'IA, c'est-à-dire rendre le modèle accessible et utilisable par des millions d'utilisateurs, sans bug et sans latence. Il est possible par exemple de traiter plusieurs demandes en même temps (*batch inference*) pour exploiter au maximum la puissance des ordinateurs pour baisser le coût énergétique par requête. Il existe également des logiciels d'optimisation automatique (compilateurs spécialisés) qui réorganisent intelligemment les calculs pour éviter les tâches redondantes.

Enfin la troisième famille d'optimisation s'applique au niveau du matériel et des infrastructures de calcul elles-mêmes. Les cartes graphiques (GPU) spécialisées pour l'IA sont plus rapides que les processeurs classiques tout en consommant moins d'énergie par calcul effectué. Les puces ultra-spécialisées pour l'IA (comme les TPUs de Google ou les LPUs de Groq) peuvent fournir plus de performance par watt consommé. Les *data centers* ont également gagné en efficacité grâce à l'optimisation de leur PUE (*Power Usage Effectiveness*) avec différentes innovations allant de l'amélioration des systèmes de refroidissement à de l'optimisation énergétique en temps réel.

Une autre idée qui émerge est l'utilisation de modèles légers (*Small Language Models* ou *SLMs*). En théorie certains de ces modèles « de poche », avec quelques millions à quelques milliards de paramètres contre des centaines de milliards pour les mastodontes, peuvent fonctionner directement sur des smartphones ou des ordinateurs portables, évitant les calculs très énergivores dans les *data centers*. Cependant, rien ne garantit que ces petits modèles vont complètement remplacer les gros : ils risquent plutôt de s'empiler aux usages existants, créant de nouveaux besoins là où il n'y en avait pas, comme des assistants IA dans chaque objet connecté ou application mobile, tout en accélérant l'implantation de puces de calculs GPUs dans nos téléphones et ordinateurs et donc l'obsolescence des terminaux numériques.

Malgré toutes ces avancées techniques indéniables, l'équation globale reste déséquilibrée. La puissance de calcul nécessaire pour entraîner des modèles d'IA double tous les 6 mois depuis 2012 et la consommation électrique totale continuent d'augmenter.

Les effets rebonds de l'IA

Ce phénomène s'illustre donc parfaitement par ce que les économistes appellent les « effets rebonds », récemment analysés en détail dans le contexte de l'IA par Luccioni *et al.* (2025).

Les effets rebond directs : quand l'amélioration de l'efficacité d'un produit améliore son accessibilité, cela peut entraîner une augmentation de sa consommation. Par exemple pour l'IA, le coût en euros par mot généré est divisé par 10 tous les ans pour les mêmes performances. Malgré les améliorations constantes d'efficacité des puces et des algorithmes, NVIDIA a expédié 3,7 millions de GPU en 2024, soit plus d'un million d'unités supplémentaires par rapport à 2023. Parallèlement,

même si l'efficacité énergétique des centres de données s'est améliorée (même si le PUE peine à baisser en dessous de 1,1), leur consommation électrique totale explose avec plus de centres, des centres plus gros et qui consomment plus d'énergie. Ces effets rebonds directs sont souhaités par l'industrie pour rentabiliser leurs investissements en augmentant les usages.

Les multiples effets rebonds indirects de l'IA (utilisations dans d'autres secteurs carbonés, nouveaux produits manufacturés...) feront l'objet d'un second article à paraître plus tard.

Le cas DeepSeek

Le cas emblématique d'effet rebond est celui du modèle d'IA chinois DeepSeek-R1. Lancé le 20 janvier 2025 par l'entreprise chinoise DeepSeek, ce modèle d'IA open source a fait sensation et fait chuter les actions de Nvidia de 17% en une seule journée. Contrairement aux annonces médiatiques qui présentent ce modèle de raisonnement chinois comme une révolution énergétique, les premiers tests indépendants sont en fait bien plus nuancés. Bien que DeepSeek ait effectivement réduit ses coûts d'entraînement grâce à des techniques d'optimisation comme le *mixture of experts* (MoE), cette efficacité théorique se heurte à un effet rebond majeur lors de l'utilisation quotidienne.

Un des autres problèmes réside dans une architecture du modèle d'IA popularisée par DeepSeek (en particulier son modèle R1), basée sur le « *chain-of-thought* » (chaîne de pensée), une technique qui permet au modèle de décomposer les problèmes complexes en étapes de raisonnement intermédiaires avant de donner sa réponse finale. Les tests menés par le MIT Technology Review montrent que cette approche est au final plus énergivore à l'usage : DeepSeek tend à générer des réponses beaucoup plus longues que les modèles traditionnels : en moyenne sur 40 requêtes testées, le modèle a consommé 87% d'énergie supplémentaire simplement parce qu'il produisait des textes de raisonnement étendus.

Ces résultats font écho aux recherches académiques récentes sur l'impact énergétique des modèles de raisonnement. La publication *The Energy Cost of Reasoning* démontre que ces modèles de raisonnement peuvent multiplier la consommation énergétique par un facteur conséquent – jusqu'à 97 fois plus dans certains cas – car les modèles génèrent en moyenne 4,4 fois plus de mots que leurs équivalents traditionnels, atteignant parfois 46 fois plus dans les cas extrêmes. Comme l'alerte la chercheuse Sasha Luccioni: « Si nous adoptons largement ce paradigme, la consommation énergétique d'inférence exploserait. »

Les impacts directs de l'intelligence artificielle

Maintenant que nous avons exploré ce qu'est l'IA, qui est derrière son développement, les particularités de l'IA générative avec ce qu'on sait et ce qu'on ne sait pas, la différence entre les usages, et les optimisations énergétiques possibles, nous sommes prêts pour analyser l'ensemble des impacts environnementaux liés à l'IA au niveau global et à l'échelle locale des centres de données.

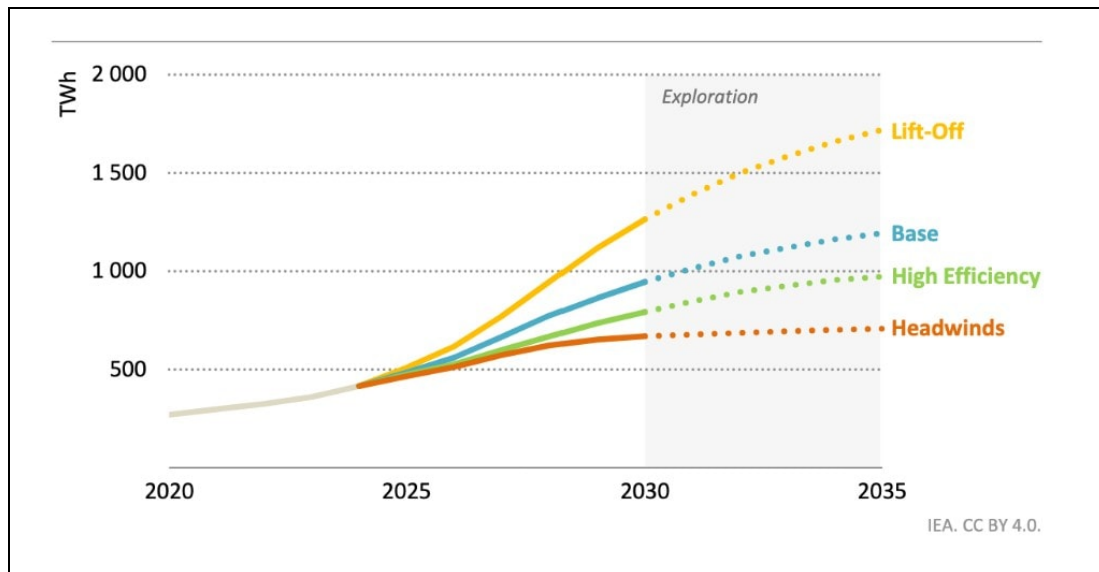
Soif d'électricité

« De nombreuses personnes prévoient que la demande d'électricité pour [l'IA] passe de 3 % à 99 % de la production totale... soit 29 gigawatts supplémentaires d'ici à 2027 et 67 gigawatts de plus d'ici à 2030. » Cette phrase a été prononcée par Eric Schmidt, ancien PDG de Google devant le Congrès, à l'image des délires de grandeurs des géants de la tech. Qu'en est-il actuellement ?

Nous l'avons vu, les centres de données consomment énormément d'électricité, à la fois pour réaliser les calculs lors des phases d'entraînement et d'inférence, mais également pour refroidir les serveurs.

La consommation électrique des *data centers* explose à cause de l'IA

D'après le dernier rapport de l'Agence Internationale de l'Énergie (AIE), les centres de données représentent 1,5% (415 TWh) de la demande d'électricité mondiale en 2024. L'IA représenterait autour de 10% de cette demande (le reste servant au stockage des données, au streaming, aux cryptomonnaies...).



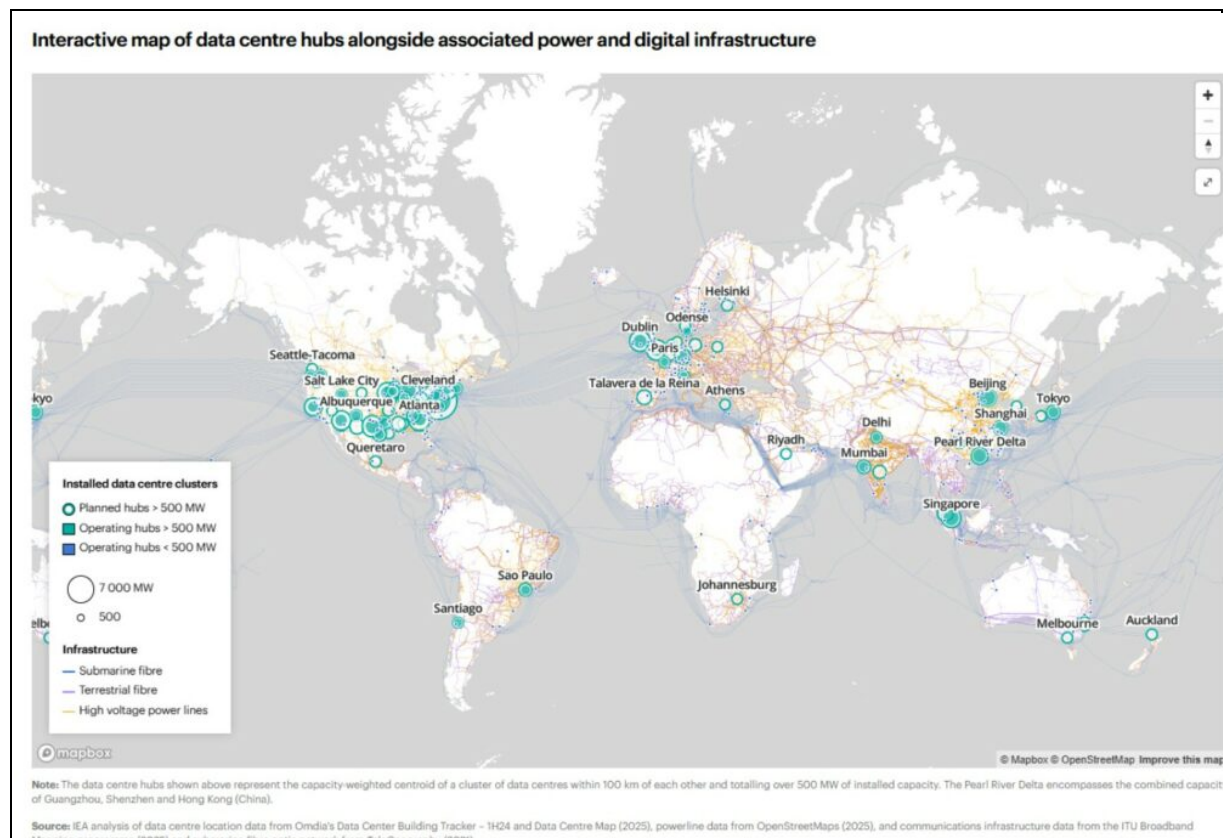
Les différents scénarios de consommation d'électricité des centres de données de l'AIE : dans le cas de base la consommation atteint 3% (1000 TWh) de la demande mondiale en 2030 – dans le cas d'adoption massive 'Lift off' elle atteint 4,5% (1600 TWh)

Mais plus encore que les chiffres absolus, c'est la dynamique qui est vertigineuse : dans son scénario de base, l'AIE prévoit que les centres de données atteignent 3% (975 TWh) de la demande d'électricité mondiale d'ici à 2030. Cela représente une augmentation de 15% en moyenne par an, 4 fois plus rapide que l'augmentation de la consommation d'électricité de tous les autres secteurs.

Aux États-Unis, en 2030, la demande d'électricité des centres de données pourrait représenter plus que la demande en électricité réunit des secteurs de l'aluminium, de l'acier, du ciment, de la chimie et d'autres biens industriels intensifs en électricité. Cette augmentation est largement tirée par le développement de l'intelligence artificielle générative et des terminaux associés, qui selon les scénarios devrait au moins représenter plus de 50% de la consommation électrique des data centers, ce qui représente une accélération pour l'IA de au moins entre x5 et x10 par rapport à 2022.

Les disparités géographiques de la consommation d'électricité

Cette moyenne de consommation d'électricité mondiale cache en réalité d'énormes disparités, car les centres de données sont très concentrés géographiquement, ce qui augmente les conflits d'usage locaux autour des ressources.



Les centres de données sont répartis dans des zones géographiques très spécifiques, créant des tensions locales. Source : Rapport AIE Avril 2025 – Energy and IA

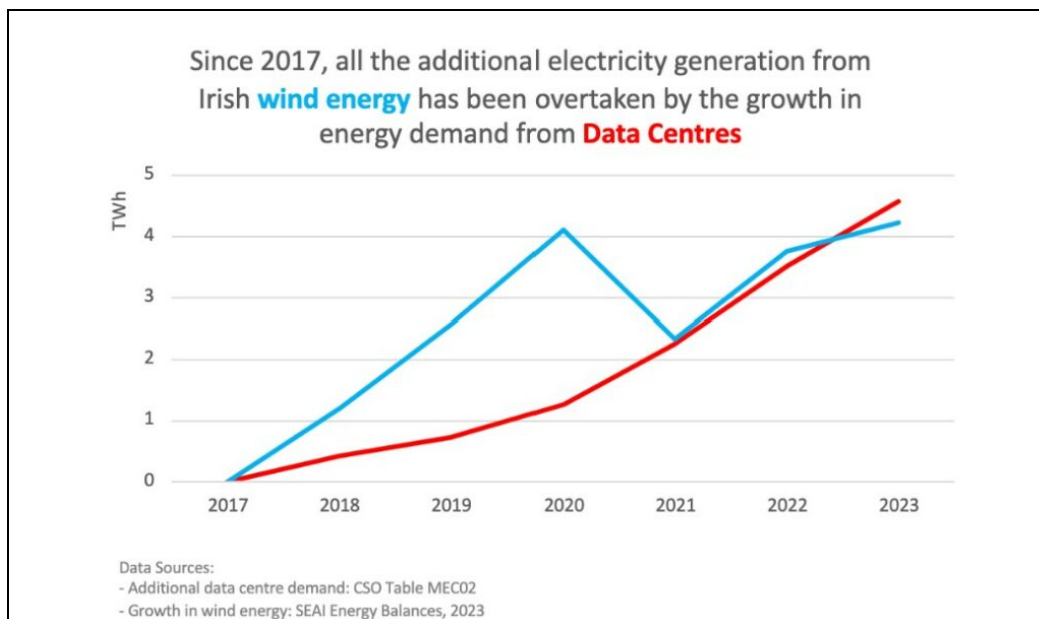
On comptait environ 8000 centres de données dans le monde en 2024. Des centaines d'autres ont été annoncés depuis. La carte ci-dessus met en évidence les grands clusters mondiaux : Les États Unis, l'Asie pacifique avec la Chine en tête, et l'Europe en troisième position. La région la plus concentrée au monde est la Virginie du Nord dans une zone surnommée la « *Data Center Alley* », grâce à une réglementation très favorable à l'implantation, une zone historiquement centrale dans le développement d'internet et une électricité à bas coût.

En Malaisie, dépendante à plus d'un tiers du charbon pour sa production d'électricité, la ville de Johor Bahu, un ancien village de pêcheurs, se transforme en hub de centre de données, et se fait surnommer la « Virginie d'Asie ». Microsoft et Amazon investissent massivement dans la région. En Europe, c'est en Irlande que se trouvent en 2024 le plus de *data centers*, et ça n'est pas un hasard : l'Irlande est un véritable paradis fiscal européen et a mis en place une politique agressive d'accueil des centres de données.

Les conflits d'usages liés à l'électricité

Cette concentration génère des conflits d'usage, même en France où on observe des conflits d'usages à Marseille, où la demande en électricité des centres de données entre en concurrence et pénalise l'électrification des transports publics et des ferrys, qui contribuent à la pollution locale. En Irlande, les centres de données consomment 21% de l'électricité du pays et les habitants doivent vivre avec le spectre de potentiels *black out* à mesure que les centres se développent.

En Irlande, la demande en électricité des centres de données a dépassé la croissance de capacité de l'éolien, les empêchant de substituer d'autres usages fossiles.



En bleu les TWh produit par l'énergie solaire – en rouge la demande en électricité des centres de données – leur croissance se suivent, annihilant toute possibilité de substitution.

Les impacts locaux liés à l'électricité

Par ailleurs, les centres de données ne sont pas flexibles : ils ne supportent pas les coupures de courant qui entraînent un arrêt total des serveurs. Par sécurité, ils sont donc systématiquement équipés de batteries et de groupes électrogènes (dont la très grande majorité fonctionnent au fioul) pour prendre le relais en cas de coupure de courant. Ces derniers doivent être testés dans tous les cas au moins quelques fois par mois pour vérifier leur fonctionnement, ce qui alourdit les émissions de CO₂ et les pollutions atmosphériques.

A titre d'exemple, le centre de données « PAR7 » de la Courneuve, propriété d'Interxion France, dispose de 8 groupes électrogènes, chaque moteur fait plus de 6 tonnes et équivaut à celui d'un « paquebot traversant la mer de Chine » d'après une enquête d'Usbek et Rica en 2017. 280 000 litres de fioul sont stockés, et préchauffés en permanence pour pouvoir démarrer à tout moment. En Irlande, les générateurs de secours du *data center* de EdgeConneX ont émis 130 000 tCO₂eq supplémentaires depuis 2017.

Enfin les chiffres mentionnés plus haut ne n'intègrent pas l'électricité nécessaire au cycle de vie d'un centre de données comme l'empreinte environnementale nécessaire à la fabrication des équipements IT ou celles de la fabrication d'un centre.

Les impacts sur le réseau de distribution et de transport électrique

L'essor de l'IA met sous tension l'ensemble du système de distribution et de transport électrique. Ce n'est pas seulement le nombre de data centers qui augmente, mais également leur puissance. Il y a quelques années les centres demandaient une puissance de quelques dizaines de MW. En 2025, avec l'explosion de l'IA, ils représentent plusieurs centaines de MW. Les plus grands data centers annoncés auront une puissance de 5GW – soit l'équivalent de la production d'une grosse centrale nucléaire. Cette concentration de demande très rapide dans des zones géographiques précises crée des points de tension sur le réseau électrique actuel.

Les goulots d'étranglement au niveau des connexions au réseau constituent aujourd'hui une limite physique au développement de l'IA. Le dernier rapport de l'AIE Energy & AI résume par exemple (Table 2.4) que les temps d'attente pour connecter de nouveaux centres de données s'étendent désormais à 7-10 ans dans certaines juridictions américaines, avec des cas extrêmes comme les Pays-Bas où les délais peuvent atteindre 10 ans.

Ces tensions se traduisent par des impacts techniques immédiats et des risques de pénurie à moyen terme. Les charges de travail d'IA présentent des défis particuliers pour les opérateurs de réseau en raison de leurs caractéristiques distinctes : l'entraînement d'IA génère une consommation électrique soutenue et élevée avec des pics et des chutes périodiques, tandis que l'inférence peut provoquer des fluctuations rapides de la demande selon ce que font les utilisateurs.

Sur le plan opérationnel, les centres de données provoquent déjà des dysfonctionnements du réseau électrique (Box 2.10, AIE) : surchauffe des équipements électriques par surcharge, fluctuations dangereuses de la tension qui peuvent endommager les appareils, perturbations de la qualité du courant électrique, et instabilités du réseau provoquant des variations de fréquence. Lorsqu'une perturbation du réseau pousse un centre de données à basculer sur son alimentation de secours, cela retire brutalement une charge massive du réseau, pouvant provoquer des changements de tension ou de fréquence et potentiellement déclencher des coupures en cascade.

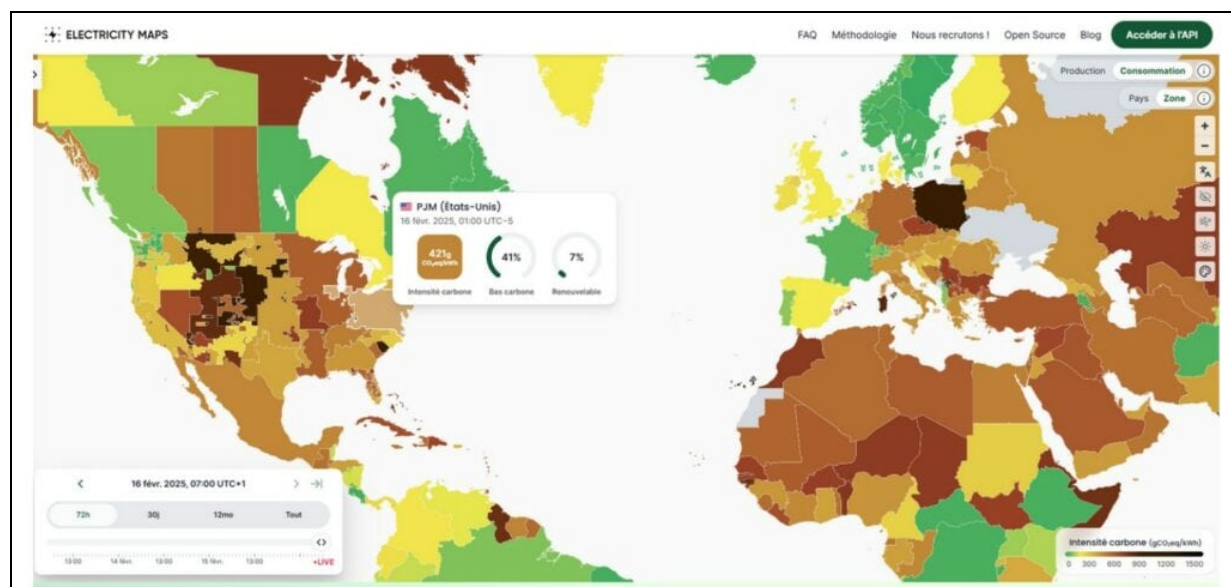
Cette situation contraint les opérateurs électriques et les États à engager des investissements massifs de modernisation du réseau électrique, avec leurs propres impacts environnementaux. Par exemple, cela implique la construction de nouvelles lignes à haute tension, l'installation de transformateurs plus puissants – autant d'infrastructures nécessitant du béton, de l'acier et des métaux rares, générant des émissions carbone supplémentaires.

Ces investissements et ces coûts opérationnels pour le réseau se traduisent même parfois par une autre injustice sociale : une augmentation de la facture d'électricité pour les communautés présentes sur ces réseaux électriques. Dans l'est des États-Unis, la facture mensuelle a pu augmenter pendant l'été 2025 de 10 à 27\$ par mois supplémentaires face à l'augmentation de la demande en IA.

Émissions de CO₂, la roue libre

Les émissions de CO₂ liées au mix carbone de l'électricité

En ce qui concerne la demande en électricité lors de l'usage du centre de données, les émissions de CO₂ sont fortement dépendantes de l'intensité carbone de l'électricité consommée.



Visualisation de l'intensité carbone de l'électricité dans le monde le 22 mai 2025 à 14h03 sur Electricity Map

L'augmentation brutale de la demande d'électricité avec le déploiement très rapide de l'IA a des conséquences inquiétantes pour le climat car elle est assurée par un raccordement à des moyens de production d'énergie rapide à mettre en place et largement carbonés.

Ainsi, de nombreuses centrales à charbon dans le monde voient leur date de fermeture reportée, mettant en danger nos objectifs climatiques. C'est le cas par exemple de trois centrales d'une

capacité de 8,2 GW (dans le Mississippi et en Géorgie) pour répondre à la demande des centres de données ou encore d'une centrale à charbon dans le Nebraska, pour répondre aux besoins de Meta et Google.

Par ailleurs, la création de nouveaux projets de centrales à gaz se voient justifiés par la nécessité de fournir de l'énergie aux centres de données. La base de données Open Source "Global Energy Monitor" a recensé l'installation de 44 GW de nouveaux projets de gaz dans le monde uniquement pour répondre à la demande des *data centers*, dont 37 GW aux US. À titre d'exemple, ExxonMobil a annoncé en décembre 2024 l'ouverture d'une centrale à gaz de 1,5 GW exclusivement dédiée aux centres de données.

La question du « *behind-the-meter* », de l'électricité non raccordée au réseau

Les géants de la tech, qui se heurtent aux limites de la disponibilité de l'énergie et au temps réglementaire de raccordement aux réseaux, sont en train de développer des solutions alternatives telles que l'installation directe d'infrastructures de génération électrique adjacentes aux centres de données, contournant ainsi les contraintes du réseau public et ses mécanismes de gouvernance collective, mais en ayant recours à des énergies carbonées.

Bien qu'ils misent beaucoup sur la géothermie ou le nucléaire (Amazon investit dans des *start up* de SMR, des petits réacteurs nucléaires modulaires, et le fondateur d'OpenAI mise sur la fusion nucléaire), ces installations ne seront pas prêtes avant plusieurs années, et à court-terme les choix se portent vers des turbines à gaz.

Zoom sur xAI à Memphis

C'est précisément la stratégie qu'a adoptée Elon Musk avec xAI à Memphis pour son centre de données *Colossus* – présenté fièrement comme « construit en 19 jours ». Pour satisfaire ses besoins énergétiques immédiats, xAI a déployé 15 turbines à gaz vétustes fonctionnant déjà sans notification publique ni surveillance réglementaire depuis l'été 2024. Selon un permis d'exploitation récemment déposé auprès du département de santé du comté de Shelby, xAI prévoit de faire fonctionner ces turbines sans interruption de juin 2025 à juin 2030. D'après *The Commercial Appeal*, ces équipements émettent chacun 11,51 tonnes de polluants atmosphériques dangereux par an, dépassant ainsi le plafond annuel de 10 tonnes fixé par l'EPA pour une source unique.

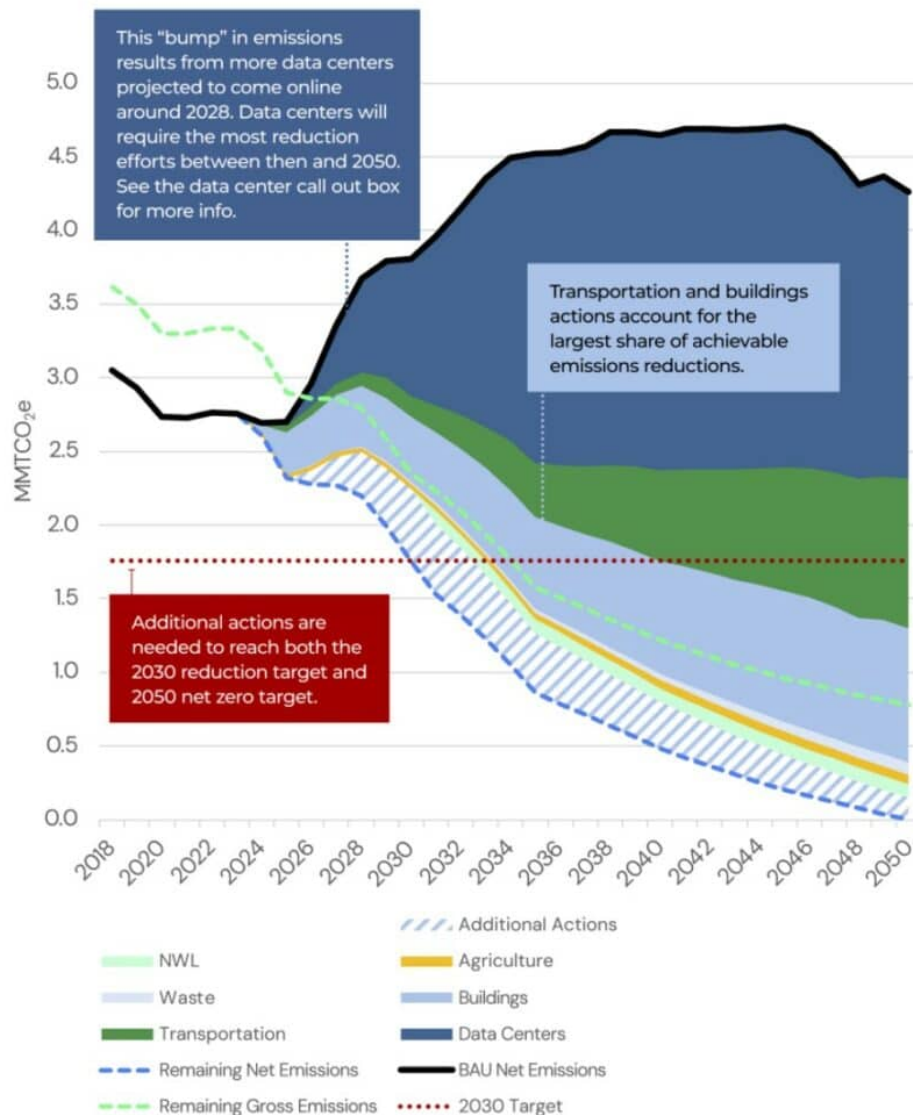
Cette infrastructure s'inscrit dans un contexte géographique inégalitaire. Elle est implantée à proximité de Boxtown, quartier historiquement noir, déjà accablé par une pollution industrielle chronique. Le comté de Shelby est actuellement noté F pour ses niveaux de smog par l'*American Lung Association*, avec un taux de cancer local quatre fois supérieur à la moyenne nationale et une espérance de vie réduite de dix ans par rapport au reste de Memphis.

Le fait que les GAFAM commencent à générer eux même leur propre énergie participe d'une opacification croissante du secteur sur leurs impacts réels et est une très mauvaise nouvelle pour la démocratie.

L'AIE estime que les émissions dues à la consommation d'électricité des centres de données représentent 180 millions de tonnes (Mt) aujourd'hui, et passeront à 300 Mt dans le scénario de base d'ici à 2035, et jusqu'à 500 Mt dans le scénario d'adoption massive de l'IA, connaissant une des croissances sectorielles les plus importantes. (Précisons encore une fois que ces émissions sont partielles : elles ne tiennent compte que de la consommation d'électricité et pas de tous les autres composants de la chaîne de valeur mentionnés plus haut.)

Au niveau local, voir un data center arriver dans son territoire peut signifier ne plus pouvoir respecter ses objectifs climatiques quand un plan était peut-être déjà en marche. Dans le comté Frederick dans le Maryland, un nouveau data center de 2GW alimenté par des turbines à gaz rend finalement impossible la trajectoire de réduction d'émissions de GES travaillée par le gouvernement local.

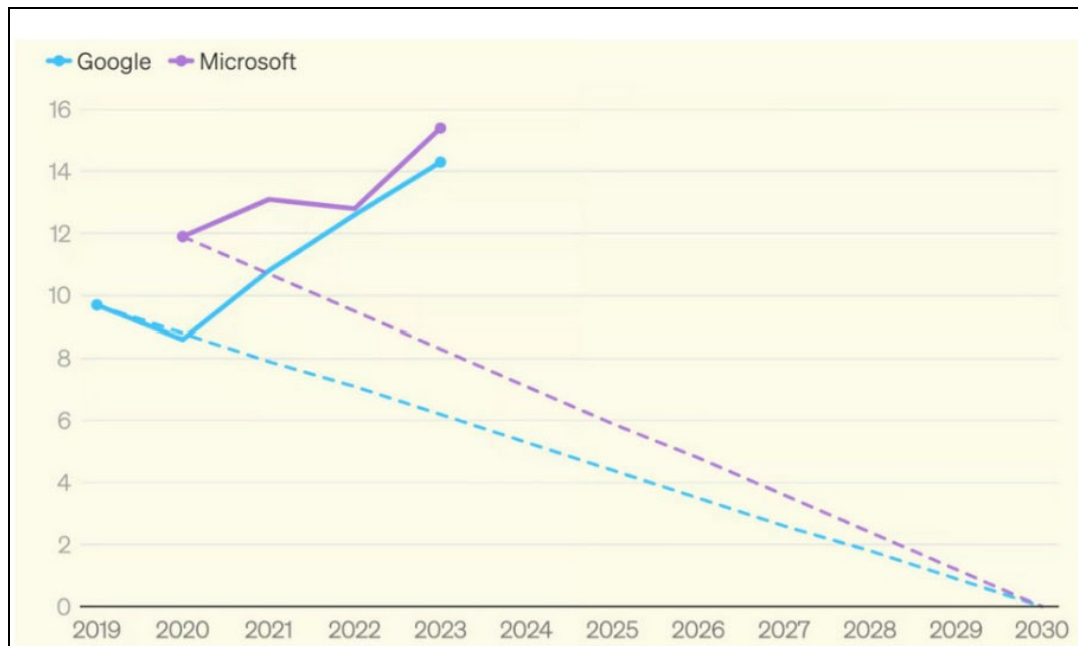
Figure 14. Modeled Emission Reductions by Mitigation Strategy



Frederick's 2025 Climate and Energy Action Plan.

Google, Microsoft, Meta.. Quid des émissions des géants du numérique ?

Ce qu'on constate aujourd'hui, c'est une explosion des émissions des géants de la tech à des années lumière des engagements de neutralité : Google a multiplié par 2 ses émissions depuis 2019 ; 2023 Microsoft les a augmentées de 30%, et se situe donc 50% au-dessus de son objectif affiché.



Source : IFP – Empreinte carbone de Google et de Microsoft de 2019 à 2023
et projections de leur trajectoire bas-carbone jusqu'en 2030 (scope 1, 2 et 3 en millions de tonnes de CO₂).

Mais ces chiffres ne représentent qu'une partie des émissions, car les géants jouent avec les règles de la comptabilité carbone actuelle pour effacer une partie de leurs émissions liées à la consommation d'énergie (scope 2). Tout se joue dans la définition de la méthode de calcul de la consommation d'énergie, entre une approche *location based* et une approche *market based* :

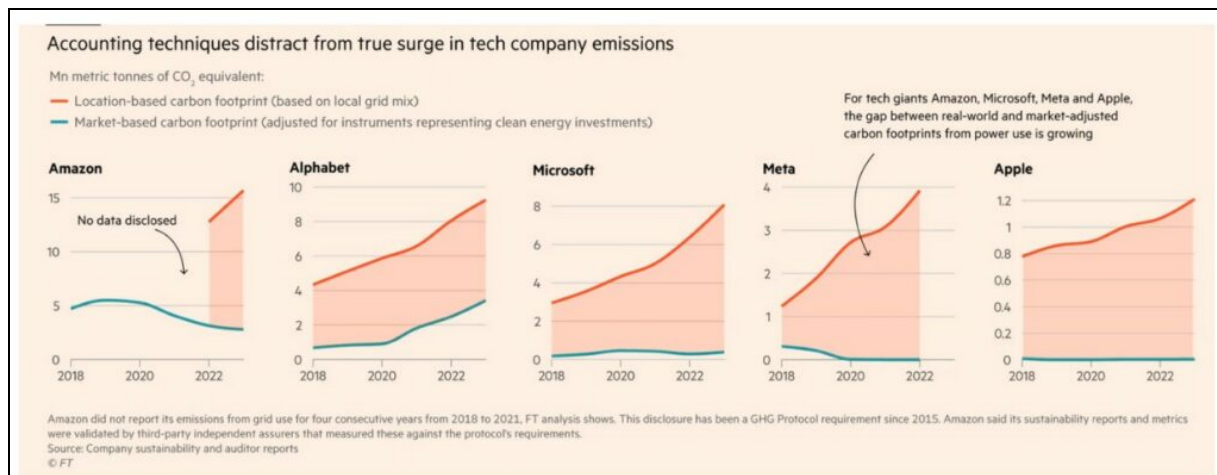
- *Location based* : cette méthode calcule les émissions de gaz à effet de serre en fonction du mix énergétique réel du réseau auquel l'entreprise est connectée. Elle reflète la réalité physique de l'électricité réellement consommée.
- *Market based* : cette méthode calcule les émissions de gaz à effet de serre en se basant sur la nature du contrat de fourniture d'électricité. Elle permet de compter à zéro l'électricité achetée issue de sources renouvelables, indépendamment de la réalité physique de la consommation. Les mécanismes de marché utilisés sont :
 - les certificats d'énergie renouvelable (RECs) : ce sont des options d'achat d'électricité bas carbone qui peuvent être utilisées jusqu'à un an après la date d'achat, n'importe où dans le monde. Elles peuvent donc être très éloignées de la consommation réelle de l'entreprise. Par exemple, des *data centers* qui tournent la nuit avec un mix fort en charbon et en gaz comme en Virginie peuvent être « effacés » par l'achat d'un certificat attaché à de l'énergie renouvelable produite deux mois plus tard au cours de la journée dans le Nevada.
 -

Les contrats d'achat d'énergie verte (PPA), parmi lesquels on distingue :

– les PPA physiques, qui impliquent la livraison réelle d'électricité depuis le site de production renouvelable jusqu'au site de consommation de l'acheteur. Les lieux de production et d'achat sont souvent géographiquement plus proches. Les PPA physiques sont donc considérés comme des instruments à haute intégrité environnementale, de meilleure qualité que les RECs.

– les PPA virtuels (vPPAs), qui sont des contrats financiers, sans livraison physique d'électricité. Comme pour les RECs, il n'existe pas nécessairement de synchronisation spatiale et temporelle entre la production et la consommation.

La comptabilité *Market based* a été développée initialement pour inciter les acteurs à investir dans de nouvelles capacités et décarboner le mix électrique. Malheureusement, le recours aux instruments de marché (RECs, PPA virtuels, etc.) n'a pas nécessairement d'impact direct sur la décarbonation réelle du mix électrique. Il permet dans le même temps à de grandes entreprises d'afficher un « zéro » comptable dans les émissions liées à leur consommation d'énergie, et de faire ainsi disparaître la réalité physique de leur consommation.



Ce graphique, issu d'une enquête du Financial Times représente l'écart entre les émissions liées à l'énergie (scope 2) reportées par les géants de la tech en 'market based' – en bleu – et les émissions réelles en 'location based' – en rouge.

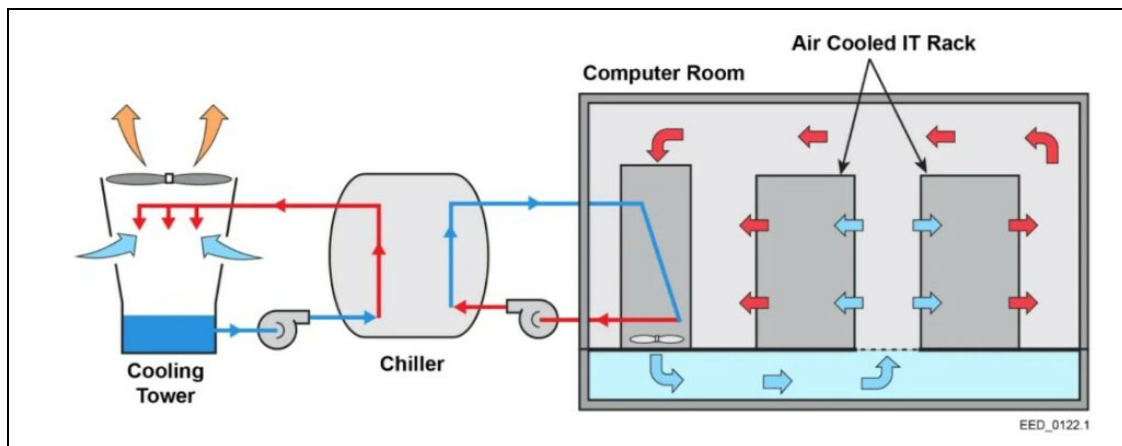
Si on prend l'exemple de Microsoft qui déclare 16 millions de tonnes de CO₂ en 2023 (1^{er} graphique de la section) en *market-based*, il faut rajouter 8 millions de tonnes de CO₂ supplémentaires liées à sa consommation d'électricité (2e graphique) pour obtenir ses émissions réelles en *location-based*, soit 50% de plus que ce qui est déclaré.

La consommation d'eau

L'utilisation de l'eau des data centers

Les centres de données sont des grands consommateurs d'eau. Bien que cette consommation soit encore une fois difficile à estimer, on peut la découper en plusieurs niveaux :

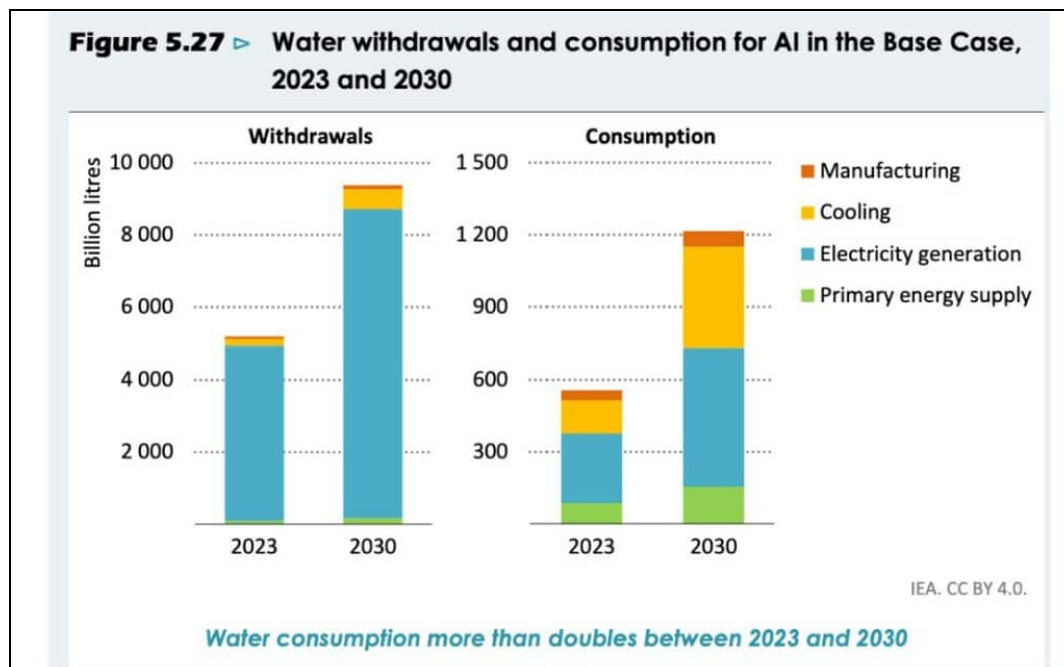
- consommation directe d'eau pour refroidir les serveurs. Le refroidissement des serveurs se fait principalement soit par air (systèmes de climatisation ou de récupération d'air extérieur), soit par liquide (pulvérisation d'air, liquide diffusé directement sur les composants). La méthode utilisée va grandement influencer la consommation d'eau. Moins la technique de refroidissement requiert de l'eau, plus elle consomme de l'énergie et inversement. Les nouveaux grands clusters Nvidia (racks > 40kW) utilisent du *direct liquid cooling*, une technique de refroidissement où le liquide est amené directement au contact des composants à refroidir. Cela ne consomme pratiquement pas d'eau mais soulève de nombreuses questions quant aux impacts et au recyclage de cette eau glycolée.



U.S. Department of Energy : représentation simplifiée d'un système de refroidissement qui utilise l'évaporation. Il y a deux mécanismes qui vont déterminer la consommation d'eau : la génération du froid et le refroidissement des serveurs.

- consommation de manière indirecte via la fabrication d'électricité qui alimente les centres de données. En ce qui concerne l'électricité, la quantité d'eau prélevée et consommée est fortement dépendante du mode de production (dans l'ordre de consommation) : l'hydroélectricité, le nucléaire, le charbon, et le gaz prélèvent et consomment de grandes quantités d'eau pour faire tourner les turbines ou refroidir les systèmes, là où le solaire ou l'éolien ne nécessitent pratiquement aucun prélèvement.

Ce graphique représente l'eau prélevée et l'eau consommée par les centres de données en fonction des différentes étapes du cycle de vie projetée jusqu'en 2030 par l'AIE.



Manufacturing : fait référence à la fabrication des composants IT Cooling : au refroidissement des serveurs. Electricity Generation : à l'eau nécessaire à la production de l'électricité consommée par le centre de données. Primary Energy Supply : à l'eau nécessaire pour l'énergie primaire utilisée dans le centre (hors électricité). Sources: IEA modelling and analysis based on Harris *et al.* (2019), Hamed *et al.* (2022), IEA (2016), Lei *et al.* (2025), Lei and Masanet (2022), and Shehabi *et al.* (2024)

L'Agence Internationale de l'Energie table sur un doublement de la consommation d'eau des centres de données d'ici à 2030 dans son scénario de base, atteignant 1200 milliards de litres en 2030. On observe déjà une demande galopante au cours de ces dernières années, surtout aux Etats-Unis. Entre 2019 et 2023, la consommation d'eau des *data centers* de Virginie a augmenté de plus de 60% d'après les chiffres des fournisseurs d'eau obtenus par le *Financial Times*.

Les conflits d'usages liés à la consommation d'eau des *data centers*

Les impacts de la consommation d'eau vont beaucoup dépendre de la concentration des centres et de la situation hydrologique de la zone : plus la zone est en stress hydrique, plus un prélèvement même faible vient perturber l'environnement. Car même quand l'eau est prélevée sans être consommée, elle n'est pas reversée au même endroit et génère donc des conflits d'usage. Aux USA, cela va jusqu'à empêcher des gens d'être logés : en juin 2023, les autorités du comté de Maricopa en Arizona, qui abrite deux centres de données de Microsoft, ont annulé les permis de construire de nouvelles habitations en raison du manque d'eau souterraine et d'une « sécheresse extrême » dans la région, d'après une enquête du Guardian. Depuis 2020, elle entre aussi en conflit avec l'agriculture comme à Taiwan, où les agriculteurs se voient contraints de flécher l'eau disponible vers l'entreprise de fabrication de puces TSMC.

Une grande partie des projets en cours de développement se situe dans des zones déjà arides : dans le rapport RSE de Microsoft datant de 2023, on découvre que 42% de son eau provient de « zones en stress hydrique » tout en ayant augmenté de 34% sa consommation d'eau par rapport à l'année précédente. Quant à Google, 15% de sa consommation d'eau provient de zones en tension d'après son rapport de 2024. Dans l'Aragon en Espagne, une région déjà aride où Amazon prévoit d'ouvrir trois nouveaux centres de données, l'entreprise vient de demander au gouvernement régional la permission d'augmenter sa consommation d'eau sur ses centres existants de 48%.

Ces tensions sont aggravées par le fait que les centres de données sont aujourd'hui majoritairement raccordés à l'eau potable. À titre d'exemple, Google utilise de l'eau potable pour plus de ⅓ de sa consommation.

Bien que certains critères expliquent ces choix de localisation (évacuation de la chaleur plus facile dans un climat sec, plus grand ensoleillement) le manque de prise en compte de la raréfaction de la ressource et de son impact sur l'environnement est préoccupant.

« *Water positive* », le nouveau greenwashing à la mode

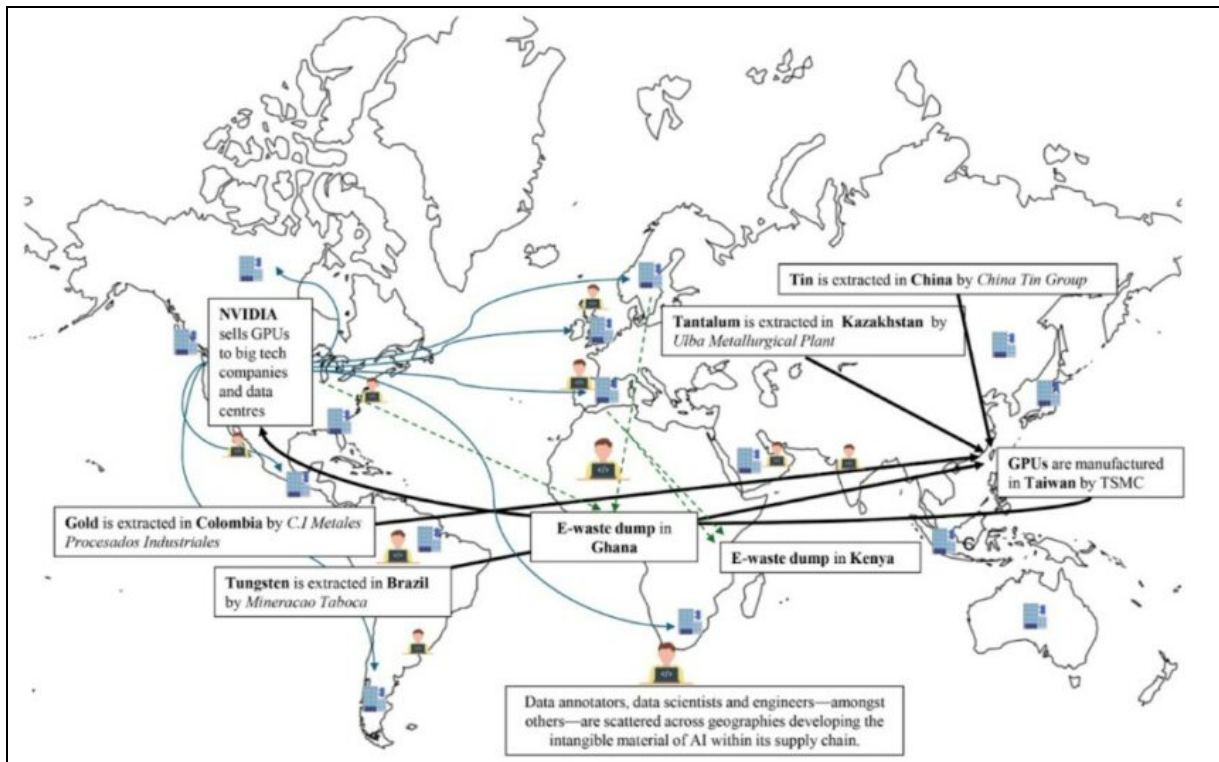
En parallèle, les GAFAM continuent à prendre de nouveaux engagements, comme celui de devenir « *Water positive* » (Amazon, Google, Microsoft à l'horizon 2030). La notion de neutralité à l'échelle d'une entreprise était déjà problématique avec le carbone, la neutralité n'étant scientifiquement valable qu'à l'échelle d'un pays ou du monde. Lorsque qu'il s'agit de l'eau, cette allégation de « neutralité », voire de « positivité » est encore plus trompeuse car les impacts sont plus localisés et complexes que pour les émissions de CO₂ : « compenser » une consommation d'eau ne peut en aucun cas annuler les impacts sur les populations et les écosystèmes à l'endroit où elle est consommée. Par ailleurs, évaporer de l'eau ou la prélever à un endroit en la restituant à un autre perturbe les équilibres et le cycle de l'eau. Ces engagements ne tiennent par ailleurs pas compte de l'eau consommée ou prélevée pour générer l'électricité consommée par les *data centers* et sont par ailleurs très difficilement vérifiables. Comble de l'ironie : le Guardian rapporte dans sa grande enquête sur l'eau qu'Amazon prévoit d'aider les agriculteurs à utiliser plus efficacement l'eau... grâce à l'IA.

Ressources abiotiques pour la fabrication des équipements et bâti : dépendance aux métaux et à l'industrie chimique

Métaux, chimie et déchets électroniques

L'infrastructure de l'IA dépend d'un grand nombre de matériaux, de produits de base et de ressources qui sont enchevêtrés dans des chaînes d'approvisionnement mondiales.

Pas d'IA sans serveurs, sans puces, et en particulier sans GPU. Or ces puces nécessitent elles-mêmes de l'électricité, de l'eau, des ressources abiotiques pour être fabriquées.



Chaîne de valeur de l'IA – Ana Valvidia

Ce graphique illustratif, extrait d'un papier de la chercheuse Ana Valvidia, met en lumière la chaîne de valeur internationale de l'IA. Elle commence avec de l'extraction minière : étain en Chine, tantale au Kazakhstan, or provenant de Colombie, Tungsten du Brésil, etc. Tous ces métaux sont réceptionnés par l'entreprise Taiwan Semiconductor Manufacturing Company (TSMC) pour fabriquer des GPU, qui sont conçus par l'entreprise américaine NVIDIA, en quasi-monopole avec plus de 85% des parts de marché. Ces GPU sont ensuite envoyés dans des centres de données à travers le monde. Ils ont en moyenne une durée de vie de 5 ans.

La diversité et la quantité des métaux mobilisés par le numérique a beaucoup augmenté depuis les années 80 comme le retrace le graphique ci-dessous. Depuis la création des semi-conducteurs, les puces évoluent en étant de plus petites et performantes. Les besoins de l'IA ont fait naître de nouveaux défis : un besoin de métaux de plus en plus pur, qui entraîne une augmentation de la consommation d'énergie, de l'empreinte écologique et une dépendance accrue à l'industrie chimique, comme le montre le chercheur Gauthier Roussilhe, spécialiste des impacts du numérique. Ces impacts sont mal connus et difficilement quantifiables, mais ils augmentent indéniablement en valeur absolue, malgré les gains d'efficacité et génèrent des pollutions locales fortes dans les pays où l'extraction est implantée. Par ailleurs, ils perpétuent des logiques d'exploitation néocoloniales des ressources et du travail des pays du Sud global.

Liste des éléments mobilisés par l'industrie électronique des années 80 aux années 2010.
Grabel and Allendy, TSMC. Reproduction et adaptation par Gauthier Roussilhe.

Pollutions et toxicité

Déchets électroniques

Les composants ultrasophistiqués utilisés dans les *data centers* contiennent une quantité exceptionnelle de matériaux toxiques (plomb, cadmium, mercure) dans un espace réduit. Seulement 22% des déchets électroniques sont actuellement recyclés correctement. Le reste finit souvent dans des décharges informelles principalement dans les pays du Sud global, où des milliers de travailleurs démantèlent à mains nues les GPU obsolètes de l'Occident, s'exposant à des contaminations jusqu'à 220 fois supérieures aux seuils européens. Une étude publiée dans *Nature Computational Science* estime que l'IA générative pourrait générer jusqu'à 5 millions de tonnes de déchets électroniques supplémentaires d'ici à 2030, s'ajoutant aux 60 millions de tonnes produites aujourd'hui. Cette accélération s'explique par trois facteurs : les besoins énergétiques exponentiels, l'évolution rapide du matériel spécialisé qui impose des mises à niveau constantes, et les guerres commerciales des semi-conducteurs qui obligent certains acteurs à déployer deux fois plus de puces moins performantes.

Bruit

Le bruit constitue la nuisance la plus immédiate : à l'intérieur d'un grand centre, les serveurs atteignent 96 dB(A), soit le niveau d'une tronçonneuse, tandis qu'à l'extérieur le bourdonnement constant des systèmes de refroidissement porte à des centaines de mètres.

Pollution atmosphérique

Notamment dû aux générateurs diesel de secours qui émettent jusqu'à 600 fois plus d'oxydes d'azote que les centrales à gaz modernes. Selon une étude d'UC Riverside et Caltech, même utilisés à seulement 10% de leur capacité en Virginie, ils causent déjà 14 000 cas d'asthme supplémentaires par an. Les projections pour 2030 sont vertigineuses : l'expansion de l'IA pourrait causer 600 000 cas d'asthme et 1300 morts prématurées par an aux États-Unis, avec des coûts sanitaires de 20 milliards de dollars – soit plus que l'impact des 35 millions de véhicules californiens.

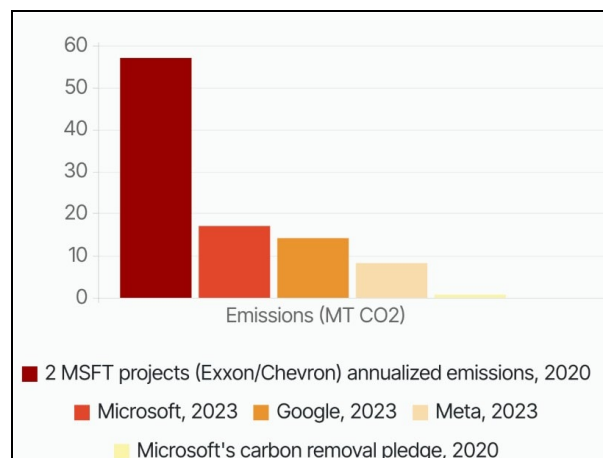
Pollution locales

Une autre pollution concerne certains systèmes de refroidissement qui peuvent utiliser des fluides avec des PFAS (« polluants éternels »), des molécules chimiques quasi indestructibles dans l'environnement. Ces pollutions locales sont un schéma d'injustice environnementale supplémentaire puisque les *data centers* sont le plus souvent localisés près de communautés déjà vulnérables.

Conclusion

L'intelligence artificielle est matérielle. Pour comprendre ses impacts il faut remonter sa chaîne de valeur qui mobilise des métaux, de l'eau, de l'énergie, des terres. Mais plus qu'une image statique, il faut apprécier sa tendance : les géants de la tech nous font basculer vers une accélération incontrôlée de l'IA, provoquant une course à l'expansion des centres de données dans le monde entier, la forçant insidieusement dans un nombre croissant de nos outils du quotidien, transformant nos interactions et notre perception du monde. La croissance actuelle semble incompatible avec nos objectifs climatiques. Par ailleurs, elle renforce des mécanismes de domination : les entreprises américaines et chinoises exploitent plus de 90 % des centres de données que d'autres entreprises et institutions utilisent pour leurs travaux d'IA et seuls 32 pays dans le monde, situés pour la plupart dans l'hémisphère Nord, disposent de centres de données spécialisés dans l'intelligence artificielle.

Une question fondamentale reste à être traitée. Au-delà des impacts direct répertoriés plus haut, il nous faut nous interroger sur la finalité des usages que vient servir l'IA, ce qu'elle permet et accélère. Explorer comment elle s'inscrit dans notre économie actuelle et comment elle renforce sa nature ultra carbonée. Et cela ne va pas dans la bonne direction : les géants de la tech ont noué des relations très particulières avec le secteur pétro-gazier. Amazon avec BP ; Google avec TotalEnergies et Gazprom ; ou encore récemment Mistral avec TotalEnergies. Il est encore une fois très difficile d'estimer les émissions supplémentaires générées par ces partenariats, mais les quelques chiffres disponibles sont déjà édifiants. À titre d'exemple, les émissions générées par l'usage de l'IA de Microsoft sur seulement deux projets avec Exxon et Chevron représentent 300 % des émissions reportées par Microsoft en 2023, d'après la lanceuse d'alerte Holly Alpine. Ces émissions facilitées par l'IA ne sont pas cantonnées à l'Oil and Gas : ciblage publicitaire, fast-fashion, exploitation minière, etc. : autant d'activités et d'usages qui viennent accélérer notre système productif.



En rouge foncé les émissions des deux projets, comparées à celles de Microsoft en 2023. Pour découvrir toute la méthodologie détaillée, rendez-vous sur le site d'Holly et Will Alpine et leur campagne Enabled Emissions.

Par ailleurs, les IA de recommandations jouent également un rôle fondamental sur le climat : en façonnant nos opinions et en promouvant certains contenus plus que d'autres, elles contribuent à invisibiliser des sujets tout en propageant de la désinformation sur d'autres. Le changement climatique est un des sujets les plus exposés à la désinformation en ligne. Cela contribue à ralentir

l'adhésion aux mesures de transition, accentue la polarisation de nos sociétés et met profondément en danger nos démocraties.

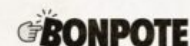
La réalité est très loin des discours technosolutionnistes des barons de la tech qui prétendent que toute lutte contre le changement climatique sera impossible sans IA pour décarboner notre économie. Ce faisant, ils ne mentionnent jamais qu'une majorité de leurs revenus proviennent de la publicité et de l'industrie fossile. Les algorithmes d'IA utiles pour le climat, utilisant principalement des techniques d'apprentissage machine spécialisées et relativement sobres énergétiquement, servent d'alibi pour justifier une grande fuite en avant.

Leurs promesses, systématiquement formulées au conditionnel, sont un leurre. Il n'existe à ce jour aucune étude globale qui permette de corroborer ces affirmations. Les impacts positifs du développement de l'IA restent profondément hypothétiques et surestimés, sans prise en compte des contextes dans lesquels les technologies sont développées et des bouleversements sociaux qu'elles génèrent. Les effets négatifs sont quant à eux largement sous-évalués et masqués.

Ces promesses nous détournent d'un questionnement sur la sobriété de nos usages.

Enfin, ces discours s'inscrivent dans un contexte politique nouveau, où les géants de la tech embrassent les idées d'extrême droite de Donald Trump. Ce faisant, ils mettent leur technologie au service de son projet politique : obscurantisme scientifique, déni total du réchauffement climatique, État autoritaire qui conjugue son pouvoir sur Big Data. Nous aborderons toutes ces questions dans un prochain article.

L'IMPACT DE L'IA SUR L'ENVIRONNEMENT



ÉLECTRICITÉ

Alimenter les calculs des IA et systèmes de refroidissement

PROJECTION 2030 (DATA CENTER)

975–1 600 TWh

soit 3–4,5 % de la consommation mondiale

IRLANDE **Impact local**

21 % de l'électricité du pays absorbée par les data centers, au détriment des habitants

MARSEILLE **Impact local**

La demande en électricité des centres de données **pénalise l'électrification des transports publics** et des ferrys,

La création de nouveaux projets de centrales à gaz se voient **justifiés par la nécessité de fournir de l'énergie** aux centres de données.

EAU

Refroidir les serveurs et produire l'électricité

PROJECTION 2030 (DATA CENTER)

+100% (x2)

soit 1200 Mds L/an. (l'IA représentant ~50% de la consommation des data centers)

ARIZONA **Impact local**

Avec l'arrivée des centres de donnée Microsoft, les permis de construire sont annulés en raison du **manque d'eau souterraine** et d'une **"sécheresse extrême"**

TAÏWAN **Impact local**

L'eau disponible est utilisée par TSMC (Fabrication de puces) **au détriment des agriculteurs**

Google utilise **⅓ d'eau potable** pour ses data centers.

MÉTAL

Fabriquer les serveurs et les puces (GPU)

PROJECTION 2030

Extraction massive

et pollutions locales fortes dans les pays où l'extraction est implantée

DÉPENDANCE **Impact global**

Par ex : Étain (Chine), Tantale (Kazakhstan), Or (Colombie), Tungstène (Brésil) ...

EXPLOITATION SOCIALE **Impact global**

Perpétuent des logiques **d'exploitation néo-coloniales** des ressources et du travail des pays du Sud global.

La durée de vie moyenne d'un GPU n'est que d'environ **5 ans**

LES MOTEURS DE L'ACCÉLÉRATION :



CROISSANCE EXPONENTIELLE DES USAGES
une adoption ultra rapide de l'IA générative **souvent forcée** par les entreprises de la tech.



COMPLEXIFICATION DES MODÈLES
Des IA de plus en plus grosses (**×10 000 en 5 ans**), qui génèrent texte, image, son, vidéo...



EXPLOSION DES DATA CENTERS
Des centres toujours plus grands, **jusqu'à la taille énergétique d'une centrale nucléaire.**

Source : "Intelligence artificielle - le vrai coût environnemental de la course à l'IA", Bonpote.com

Lou Welgryn et Théo Alves Da Costa, « Intelligence artificielle : le vrai coût environnemental de la course à l'IA », BonPote.com, 2 septembre 2025.