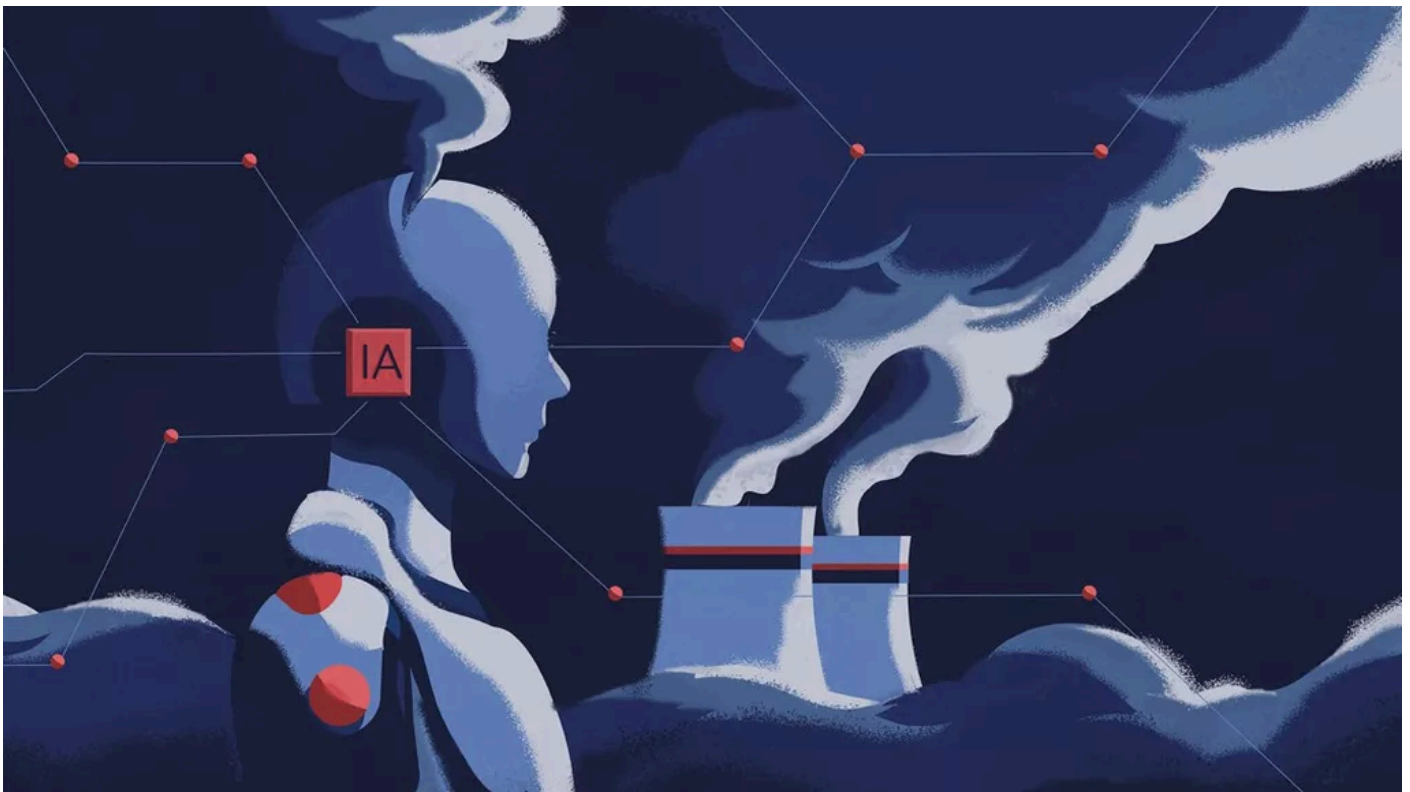


ENQUÊTE

# Comment l'IA cherche à éviter le cauchemar énergétique

Avec ses besoins démesurés en électricité, l'intelligence artificielle devient un enjeu écologique majeur. Rencontre avec des innovateurs qui tentent de lui inventer un avenir plus frugal.



Pour assurer les besoins en énergie de l'IA, Microsoft, Google et Amazon nouent des partenariats avec l'industrie nucléaire. (Illustration Baptiste Stephan pour «Les Echos»)

Par **Gabriel Grésillon**

Publié le 27 nov. 2024 à 07:05 | Mis à jour le 27 nov. 2024 à 13:09



Ils s'appellent HawAI.tech, Illuin, Ampere Computing, Holigrail ou Sharp. Et ils partagent une conviction : quand Goliath apparaît insurmontable, David a tout intérêt à déplacer le terrain sur lequel l'affronter. En l'occurrence, quand les géants américains du numérique affichent une **écrasante avance** dans la course à l'intelligence artificielle (IA), ces start-up

et ces centres de recherche se concentrent sur les moyens de mettre au point des technologies plus agiles et moins consommatrices de ressources.

Une IA plus « frugale » ? L'idée semble aller à contre-courant de l'actuelle course au gigantisme. Le célèbre générateur de texte [Chat GPT 4](#), avec ses 1.000 à 2.000 milliards de paramètres, est dix fois plus gros que son grand frère, GPT 3. Mais les limites de cette fuite en avant commencent à sauter aux yeux : sur une planète confrontée à un [défi énergétique](#) sans précédent, la généralisation de telles usines à gaz apparaît impensable - y compris en termes économiques.

## Voracité énergétique

Une requête sur un « LLM » (large language model) comme Chat GPT consomme environ dix fois plus d'électricité qu'une recherche sur Google. Selon les projections de la banque Wells Fargo pour les Etats-Unis, l'IA consommera [80 fois plus](#) d'électricité en 2030 qu'en 2024. Microsoft, [Google](#) et Amazon nouent donc des partenariats avec [l'industrie nucléaire](#). A Three Miles Island, c'est un réacteur qui va être redémarré pour fournir Microsoft. En attendant, les émissions de gaz à effet de serre des Gafa explosent.

### LIRE AUSSI :

- **ANALYSE - Chat GPT, un succès hallucinant !**
- **Nucléaire : la centrale accidentée de Three Mile Island va revivre grâce à Microsoft**

Pour l'heure, les utilisateurs ne semblent pas obsédés par le problème. C'est ce que reconnaît Frédéric Brajon, cofondateur du cabinet Saegus, spécialisé en intelligence artificielle : « Beaucoup d'entreprises perçoivent les LLM comme des outils extrêmement puissants qu'elles se doivent de déployer pour être dans la course à l'innovation, et la plupart n'ont pas encore de recul sur la surconsommation énergétique que cela engendre. » Mais, ajoute le consultant, « cela va probablement devenir un vrai sujet en 2025 et 2026 ».

Un industriel abonde : « Après une année 2023 où tout le monde a fait joujou avec cette technologie, les directions des services informatiques des sociétés commencent à se demander s'il est pertinent de déployer pour tous les collaborateurs des outils qui coûtent 20 euros par mois... »

## Retour de bâton

D'autant que ce prix ne pourra qu'augmenter. Pour l'instant, les géants du secteur, **croulant sous les financements**, cassent les prix. Jeff Wittich fait partie de ceux qui mettent en garde contre le retour de bâton. Responsable du développement des produits chez la start-up américaine Ampere Computing, il estime que « le coût réel n'est pas payé par le consommateur final ».

Il dresse le parallèle avec l'apparition du Web, « période où les entreprises cherchaient d'abord à recruter des clients ». Quitte à proposer leurs services à perte. On se souvient de ce jour de 2018 où **Google Maps** a subitement multiplié le prix de ses services aux entreprises par plus de dix... Et Jeff Wittich d'ajouter que là où il a fallu « une dizaine d'années pour qu'une normalisation apparaisse avec l'Internet », on peut s'attendre à une évolution « plus rapide cette fois, vu les contraintes énergétiques et les montants gigantesques investis ».

*« Le retour sur investissement des usages bureautiques de l'IA n'est pas facilement démontrable. »*

FRÉDÉRIC BRAJON, Cofondateur du cabinet Saegus

Ce retour sur terre est d'autant plus inéluctable, prévient Frédéric Brajon, qu'un élément réglementaire va changer la donne, en Europe, dès 2025 : la **directive européenne CSRD** va obliger les entreprises à dévoiler des informations extra-financières. « L'IA va probablement devenir un des indicateurs clés qu'elles suivront pour évaluer leurs émissions de tonnes de carbone », estime le consultant. Or, face à ces émissions bien réelles, « le retour sur investissement des usages bureautiques de l'IA n'est, pour l'instant, **pas facilement démontrable** », assure-t-il.

C'est dans cette brèche que cherchent à s'engouffrer de nombreux acteurs, convaincus qu'utiliser un gigantesque LLM pour tout faire revient à mobiliser une armée entière pour des missions que de petits commandos effectueraient au moins aussi bien.

## Une IA « hybride »

La société Illuin, par exemple, a noué un partenariat avec l'école d'ingénieurs [CentraleSupélec](#) pour répondre à la question suivante, formulée par son PDG et cofondateur, Robert Vesoul : « Peut-on déployer des capacités d'IA générative avec des modèles beaucoup plus petits ? » Bilan : en se concentrant exclusivement sur l'anglais et le français, mais aussi en acceptant un résultat « un peu moins bien rédigé qu'un texte littéraire », elle a mis au point un modèle baptisé « Croissant LLM » qui utilise 200 fois moins de paramètres que la dernière version de Chat GPT. « Dans bien des cas d'usage, explique Robert Vesoul, la qualité stylistique est moins importante que l'objectif visé. » Exemple : la synthèse écrite des conversations téléphoniques d'un service clientèle, qui doit surtout faire figurer toutes les idées clés. « Nous avons écarté certaines fonctionnalités en nous recentrant sur les tâches du quotidien et en gardant la vision d'un usage industriel », ajoute Wacim Belblidia, cofondateur et directeur général adjoint de la société.



Le nouveau data center de Meta's Facebook à Eagle Mountain, dans l'Utah. Le complexe couvre une surface égale à vingt terrains de football. (GEORGE FREY/AFP)

Chez Hawaii.tech, une start-up grenobloise, la nécessité de proposer des IA moins gourmandes en énergie découle d'abord de leur usage : déployée dans des systèmes embarqués (drones, robotique, voiture autonome...), l'IA doit fonctionner sans solliciter

un serveur à distance. La solution consiste, pour chaque projet industriel, à comprendre le fonctionnement précis d'un système et à le modéliser de manière probabiliste.

Cette IA dite « bayésienne », du nom du théorème mathématique qui la fonde, est donc différente d'un LLM : ce dernier va suivre un « [apprentissage profond](#) » (deep learning) sans supervision humaine, effectuant des milliards d'opérations itératives. La jeune pousse, elle, va « intégrer de l'expertise métier et modéliser les causalités qui sont en jeu en utilisant les probabilités », explique Raphael Frisch, CEO et cofondateur d'Hawai.tech. Pour effectuer des tâches spécifiques, pas besoin alors d'un « réseau de neurones » : très peu de données d'apprentissage et donc de ressources sont requises dans ces systèmes d'IA « hybride ». Pour créer un système d'assistance à la conduite, par exemple, un réseau de neurones utilise environ 50 fois plus de paramètres qu'un modèle probabiliste.

## Repenser les puces

Mais pour minimiser la consommation d'énergie, Hawai.tech a aussi mis au point un microprocesseur 6 fois plus efficace que celui proposé par [le géant Nvidia](#) pour les technologies embarquées. De fait, c'est aussi dans le « hardware » que sont attendues des avancées en termes de frugalité.

La société Ampere Computing, aux Etats-Unis, est l'une de celles qui se focalisent sur cet enjeu en partant d'un constat : les processeurs qui font tourner l'IA aujourd'hui sont souvent [les fameux GPU](#), pensés à l'origine pour assurer l'interface graphique des ordinateurs. Lorsque la nouvelle génération d'IA est apparue, ses concepteurs ont en effet constaté que son entraînement reposait sur des milliards d'opérations simples effectuées en parallèle, chose que les GPU faisaient à merveille pour assurer, pixel par pixel, l'affichage d'un écran.

### LIRE AUSSI :

- **DECRYPTAGE - L'IA est-elle une bulle spéculative ?**
- **INTERVIEW - IA : « La France est face à une opportunité qui ne se produit qu'une fois par siècle »**

Mais pour Jeff Wittich, le responsable du développement des produits de la société, « ce côté extrêmement répétitif perd une partie de sa pertinence lorsque la phase

d'entraînement du modèle est terminée et qu'il faut répondre aux requêtes des utilisateurs ». Or, on le sait désormais, « l'électricité consommée par une IA générative durant son utilisation pourrait être cinq à dix fois supérieure à celle mobilisée pour son entraînement ».

En s'affranchissant de l'architecture des GPU, Ampere Computing a donc conçu, pour la phase de fonctionnement d'une IA, un microprocesseur qui affiche un rendement deux à cinq fois supérieur aux GPU du leader, Nvidia.

Au final, c'est toute une chaîne de fonctionnement qui va devoir être améliorée. Le but : « faire travailler ensemble ceux qui conçoivent les modèles, ceux qui en déclinent l'exécution informatique et les concepteurs des microprocesseurs optimisés pour cet usage », explique Cédric Auliac, qui dirige le programme IA à la direction de la recherche technologique du CEA (Commissariat à l'énergie atomique).

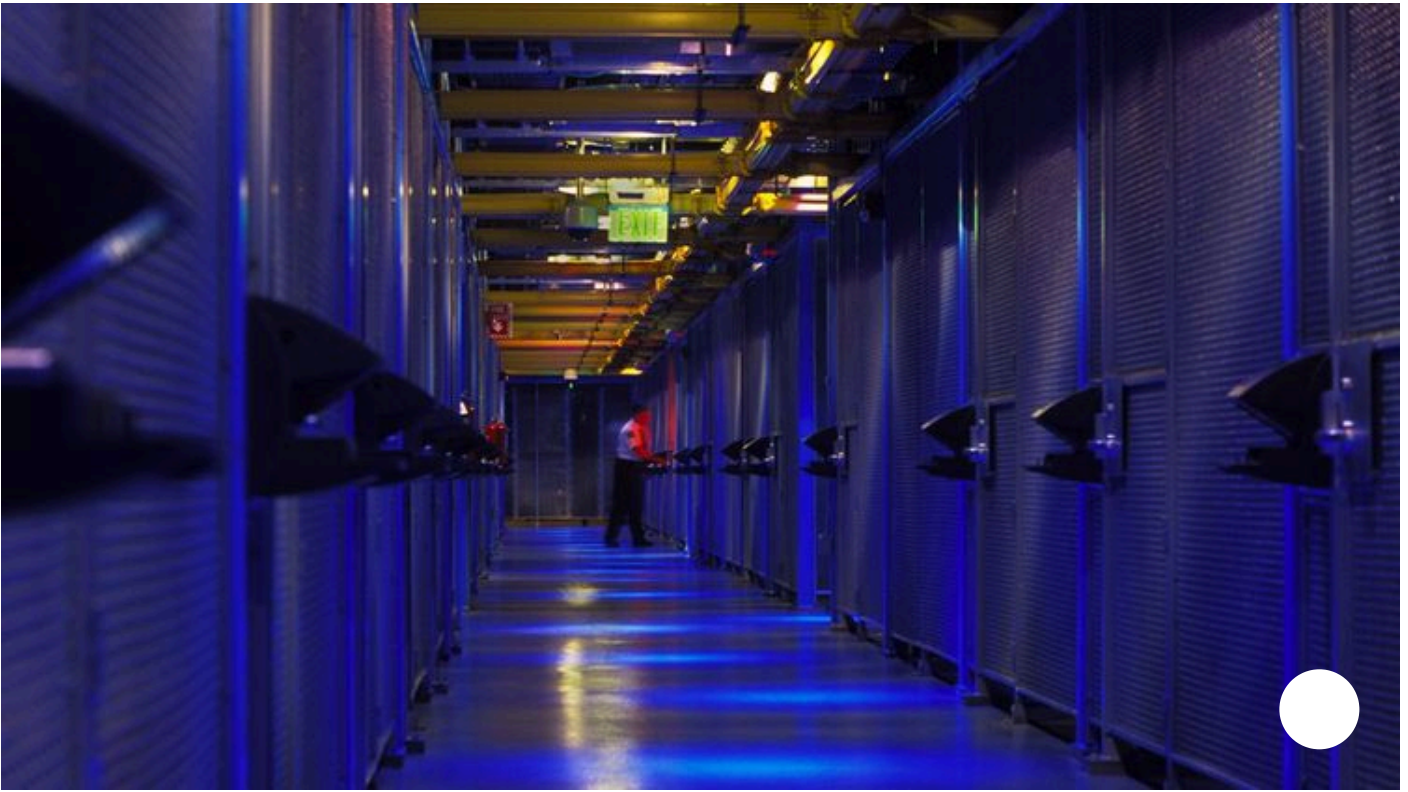
*« Il va au moins falloir multiplier par 1.000 l'efficacité énergétique des IA génératives. »*

François Terrier, Codirecteur du PEPR-IA

C'est aussi le point de vue de François Terrier, l'un des trois directeurs de l'ambitieux projet français baptisé « PEPR-IA » (Programme et équipements prioritaires de recherche pour l'IA). Réunissant l'Inria, le CNRS et le CEA, ce projet doté de 73 millions d'euros par l'Etat français réunit 9 programmes de recherche sur l'IA. « Il va au moins falloir multiplier par 1.000 l'efficacité énergétique des IA génératives », prévient le chercheur.

## **La France bien positionnée**

Sur ce chemin, le programme du PEPR-IA baptisé « Holigrail » vise par exemple à « améliorer le codage des milliards de paramètres d'un réseau de neurones, un peu comme le MP3 a compressé les fichiers musicaux », explique François Terrier. A ce stade, les chercheurs arrivent couramment à diviser par dix la complexité. D'autres recherches sont menées, sur le matériel, pour utiliser des réseaux de neurones génériques, en adaptant uniquement leurs dernières « couches » en fonction de la tâche à accomplir - de quoi multiplier par 1.000 l'efficacité énergétique dans la recherche d'images.



Les requêtes Internet qui utilisent l'IA consomment cinq à dix fois plus d'énergie que sur un moteur de recherche classique. (iStock)

Dans cette quête de frugalité, les modèles « open source », qui permettent de rendre accessible le code d'un programme, ont le vent en poupe. En mutualisant les efforts, ils permettent de capitaliser sur les innovations déjà établies pour les adapter à un contexte précis, sans repartir de zéro. Cette approche peut aussi permettre à une entreprise de faire tourner une IA sur sa propre infrastructure informatique, sans mobiliser des serveurs lointains.

A plus long terme, les réflexions et les recherches ne manquent pas, notamment pour s'inspirer du fonctionnement du cerveau, véritable sommet d'efficacité énergétique - il consomme « en moyenne 20 watts, contre 700 watts pour un GPU », résume Cédric Auliac.

### **La crainte d'un « effet rebond »**

Stratégiquement, ajoute François Terrier, « la France a une carte à jouer, forte notamment de son école de mathématiques qui parvient à travailler avec les informaticiens sur ce sujet, le PEPR-IA en étant une illustration ». Première au monde dans cette démarche, l'Association française de normalisation (Afnor) vient de publier un référentiel permettant d'évaluer la frugalité d'une IA. Pour François Terrier, « la bataille sur le marché de l'IA industrielle n'a pas démarré et nous avons encore

**l'opportunité de prendre de l'avance** sur les enjeux de frugalité et de confiance qui y sont associés ».

Reste une sourde inquiétude : ce que les économistes appellent « l'effet rebond ». Les technologies plus économes en train d'être élaborées, plutôt que de rendre l'IA moins vorace, risquent surtout de banaliser son utilisation jusque dans le moindre téléphone. Comme l'explique un scientifique, « certains chercheurs se refusent à utiliser Chat GPT par conviction écologique, mais redoutent d'être en train de mettre au point les technologies qui contribueront en fait à l'accélérer ».

**ABONNEZ-VOUS À LA NEWSLETTER « LES ECHOS DE L'IA »**

Recevez tous les lundis les dernières nouvelles de l'intelligence artificielle pour vous aider à l'adopter dans votre vie professionnelle > [S'inscrire](#)

**Gabriel Grésillon**