IA frugale : l'inatteignable Graal pour les DSI ?

PAR FABRICE DEBLOCK

Soumis à des pressions contraires, les DSI ont-ils les moyens de déployer une véritable démarche d'IA frugale au cœur de leurs projets ? La réponse avec trois experts aux points de vue complémentaires.

Les algorithmes d'intelligence artificielle (IA), et plus particulièrement ceux liés à l'IA générative, sont très consommateurs en énergie. Dans un rapport paru en 2024 et intitulé « *Electricity 2024 : Analysis and Forecast to 2026* », l'Agence internationale de l'énergie *(International Energy Agency)* anticipe un doublement de la consommation électrique des centres de données, de la cryptomonnaie et de l'IA d'ici 2026. Toujours selon l'Agence internationale de l'énergie, leur consommation électrique annuelle se serait élevée à 460 térawattheures en 2022, soit près de 2 % de la demande mondiale. Et les projections ne sont pas réjouissantes : d'ici deux ans, elle devrait dépasser les 1 000 térawattheures.

> “
>
> La sensibilisation à un usage modéré des outils d'IA générative est un levier clé pour le DSI [...]
>
> Pierre Delort, président de l'ANDSI

Et, dans les faits, ces prévisions semblent déjà s'être réalisées. Microsoft vient en effet d'annoncer avoir signé un accord avec Constellation, la société qui exploite l'un des deux réacteurs de la centrale nucléaire de Three Mile Island, en Pennsylvanie, arrêtée depuis 2019. L'autre réacteur est tristement connu pour avoir partiellement fusionné en 1979, provoquant l'un des accidents

En 2026, la consommation électrique annuelle des centres de données, des cryptomonnaies et de l'IA devrait dépasser les 1000 térawattheures.

nucléaires les plus graves aux États-Unis. L'objectif de Microsoft est de générer une production de 835 mégawatts d'électricité à partir de 2028, pour au moins deux décennies, afin d'alimenter une partie de ses centres de données.

Dès lors, la question se pose de savoir si, au sein des entreprises, les DSI ont la volonté, mais aussi la capacité, de se saisir du sujet en activant un certain nombre de leviers permettant de tendre vers une IA frugale. En matière d'IA, la frugalité se définit comme toute démarche visant à s'interroger sur les usages et les besoins, ainsi que les niveaux de résultat attendus. L'objectif d'une telle politique est de minimiser la consommation de ressources, tout en délivrant un résultat qui fasse consensus.

LA SENSIBILISATION DES UTILISATEURS ET LE *RAG* : DEUX PREMIÈRES PISTES D'ACTION

Pour Pierre Delort, président de l'ANDSI (Association nationale des directeurs de systèmes d'information), enseignant à l'École nationale supérieure des mines de Paris et auteur de *Le Big Data* (collection « Que sais-je ? »), le sujet ne fait pas forcément partie des

© Freepik

thématiques prioritaires des DSI. « L'IA frugale correspond à une réalité, mais pour le moment celle-ci provient essentiellement de démarches volontaristes. Ce n'est pas le grand sujet du moment pour les CIO. Certes, les DSI ont des projets liés au *green IT*, mais je ne suis pas certain que la majorité d'entre eux pousse la réflexion jusqu'à l'IA », déclare-t-il.

Un avis que partage Rémi Sabonnadiere, CEO et cofondateur d'Effixis (du groupe Artefact), société de conseil dans le domaine de l'intelligence artificielle. « Pour le moment, dans les entreprises, l'impact écologique de Copilot ou de n'importe quelle autre IA n'est pas du tout pris en considération. Et, même nous, en tant que spécialistes de l'IA, nous avons du mal à estimer combien d'eau consomme, par exemple, une conversation avec ChatGPT. Ce sujet est très neuf. Je dirais donc que c'est dans la tête de tout le monde, mais que ce n'est pas vraiment une priorité, il faut être transparent sur ce point », a déclaré Rémi Sabonnadiere lors d'une conférence organisée par l'Asage.

Quant à Pierre Delort, il relativise cependant ses propos en rappelant que, en matière d'intelligence artificielle, le paiement à l'utilisation se développe de plus en plus, avec une facturation proportionnelle au nombre de requêtes, de mots ou de *tokens*. C'est notamment vrai chez OpenAI, ce qui incite les entreprises à une certaine modération dans leur consommation. « Le DSI peut communiquer sur cet aspect-là. C'est une incitation pour les utilisateurs à ne pas demander dans la même journée la synthèse de multiples rapports annuels d'entreprises, par exemple. La sensibilisation à un usage modéré des outils d'IA générative est un levier clé pour le DSI, même si, in fine, les collaborateurs restent dans la plupart des cas maîtres de leurs usages », complète-t-il.

Autre piste pour les DSI, selon Pierre Delort : opter pour la *RAG (retrieval-augmented generation*, soit la « génération augmentée de récupération » en français). « En adjoignant au *LLM* utilisé des modules permettant d'ajouter des connaissances propres à l'activité de l'entreprise, et en rafraîchissant souvent ces informations, le DSI a la possibilité d'éviter de trop fréquemment réentraîner le modèle et, surtout, de le réentraîner avec des données très éloignées des métiers de son organi-

> “
>
> **Pour le moment, dans les entreprises, l'impact écologique de Copilot ou de n'importe quelle autre IA n'est pas du tout pris en considération.**
>
> Rémi Sabonnadiere, CEO et cofondateur d'Effixis


© Freepik

sation, ce qui a des effets bénéfiques sur la consommation énergétique », commente-t-il.

RAISONNER « EN BORDURE DU RÉSEAU » ET APPROFONDIR LA MESURE

Du côté des laboratoires de recherche, des pistes émergent également. Denis Trystram est professeur à l'Institut polytechnique de Grenoble (Grenoble INP-UGA) et membre du Laboratoire d'informatique de Grenoble (LIG). Il est également titulaire depuis 2019 d'une chaire à l'institut Miai (Multidisciplinary Institute in Artificial Intelligence) Grenoble Alpes, intitulée *« Edge Intelligence »* : « Cette chaire s'intéresse de près à l'IA frugale, c'est-à-dire à l'IA qui fonctionne avec un minimum de ressources et en bordure du réseau ».

Pour Denis Trystram, étant donné que la plupart des données sont produites localement, on doit s'interroger sur l'opportunité de les envoyer dans des centres de données distants pour y être analysées. « Je m'intéresse à ce qu'il est possible de faire en bordure du réseau. Nvidia a, par exemple, récemment sorti des composants (des cartes) qui ne consomment que quelques dizaines de watts, avec pas mal de mémoire. On peut donc déjà lancer certains traitements sans forcément faire traverser tout le réseau aux données. De nombreux industriels, comme Orange, Atos/Eviden, Berger-Levrault ou Qarnot, nous accompagnent pour creuser ces questions de frugalité », note-t-il.

Denis Trystram oriente également une partie des travaux menés dans sa chaire vers la mesure de la consommation énergétique. « Nous avons contribué à systématiser la

comparaison de tous les algorithmes comme Carbon Tracker ou PowerAPI pour savoir s'ils sont efficaces et précis, et s'il est possible de les faire évoluer facilement. Nous avons également contribué à l'Afnor Spec 2314, un référentiel sur l'IA frugale rassemblant les méthodes et les bonnes pratiques utiles à l'évaluation et à la réduction de l'impact environnemental d'un système d'IA. Cela donne des outils supplémentaires aux DSI qui s'intéressent à cette thématique pour évaluer l'impact des applications d'IA », précise le chercheur.

IA FRUGALE : UNE DÉMARCHE TOUT SAUF TECHNIQUE

Mais quand Denis Trystram cherche à résumer les principaux points caractérisant une IA frugale, il met en avant le sujet suivant : la problématique dépasse largement la technique. « On ne peut pas dissocier un service de l'analyse du besoin et de l'utilité de ce service. Et il faut même aller plus loin encore et essayer de comprendre, en amont, les éventuels effets rebonds et les effets indirects. Si l'on n'interroge pas les effets multidisciplinaires liés en amont à l'IA, je pense que l'on passe à côté des enjeux essentiels. » Le chercheur reconnaît cependant que cette démarche n'est pas facilement applicable par un DSI soumis à la pression de sa direction, des utilisateurs et des clients de l'entreprise.

« Faute d'aller aussi loin dans la réflexion, les DSI peuvent a minima gérer leurs projets d'IA proprement, dans une démarche de *green AI*. Ils peuvent commencer par choisir des *data centers* dont le PUE [*power usage effectiveness*, en français l'"indicateur d'efficacité énergétique", NDLR] est faible. Et, au lieu de faire tourner à l'aveugle un gros réseau de neurones comme GPT-4, ils peuvent ne garder – une fois que le réseau de neurones a été entraîné – que les poids essentiels et les coder sur 4 bits, ce qui est largement suffisant, même s'il y a une petite incertitude. Cela permet de faire des mathématiques avec beaucoup moins de calculs. Plutôt qu'une démarche frugale, on est dans ce cas de figure dans une démarche de sobriété », analyse-t-il.

> “
>
> **[…] les DSI peuvent a minima gérer leurs projets d'IA proprement, dans une démarche de *green AI*.**
>
> Denis Trystram, professeur à Grenoble INP-UGA et membre du LIG

Par ailleurs, les DSI peuvent décider d'instaurer au sein de leurs projets un certain nombre de petits gestes. « Sur la question du renouvellement du matériel, un petit geste consiste, par exemple, à allonger la durée de vie des équipements. Cela a un impact positif immédiat sur les émissions de CO2 liées à leur fabrication. Autre exemple, quand vous déployez un service, vous pouvez insérer un petit outil de monitoring peu intrusif qui permet de suivre la charge d'utilisation au sein d'un *data center* », conclut Denis Trystram. Autant de petits pas qui contribuent, sur le long terme, à faire avancer les projets dans le bon sens. ■