MINALOGIC
Auvergne-Rhône-Alpes

LIVRE BLANC

Intelligence Artificielle et Défi Énergétique

Enjeux et Solutions

ABSTRACT

L'intelligence artificielle est reconnue pour avoir un impact environnemental significatif voire excessif. De nombreux acteurs majeurs investissent massivement dans des capacités de production pour accroître la quantité énergétique nécessaire au fonctionnement de leurs centres de données. Au-delà de ces besoins en électricité, les serveurs soutenant les modèles d'intelligence artificielle sont également très consommateurs en systèmes de refroidissement indispensables au bon fonctionnement des cartes graphiques. Ces systèmes de refroidissement sont particulièrement gourmands en eau et contribuent notamment à réchauffer cette dernière d'une dizaine de degrés. Serveurs, processeurs et autres accélérateurs sont également fortement carbonés tout au long de leur cycle de vie, que ce soit lors de leur production, de leur exploitation ou de leur recyclage.

L'objectif de ce think tank est triple :

- Dresser un état des lieux de l'impact environnemental réel de l'intelligence artificielle, en donnant les bons ordres de grandeurs et en référençant des méthodes existantes qui pourraient servir de base à des calculs plus précis.
- Identifier des solutions existantes, issues de la pratique individuelle de nos entreprises, expertes en IA, afin d'améliorer la situation au plus tôt.
- Formuler des recommandations pour mesurer et pour minimiser collectivement l'impact de l'intelligence artificielle sur l'environnement.

La problématique est adressée à travers deux angles distincts : d'une part la décarbonation de l'intelligence artificielle, et d'autre part l'apport de l'intelligence artificielle pour aider à minimiser l'impact environnemental des activités humaines.

En effet, l'intelligence artificielle est à la fois source et solution du problème.

Ont contribué au think tank :

Adobis Group // AI Tenders // Asygn // ATOS // CEA // Délégation Régionale Académique à la Recherche et à l'Innovation (DRARI) // Direction Régionale de l'Economie, de l'Emploi, du Travail et des Solidarités (DREETS) // ExcellerIA // FlexAI // Neovision // Grenoble Ecole de Management // Grenoble INP // Hawai Technologies // Inceptive // Inddigo // Kairntech // Kaizen Solutions // Leviatan // Minalogic // MountAIn // Naver Labs // Neovision // Orange // Probayes // Région Auvergne-Rhône-Alpes // Schneider Electric // Shynet // SIPearl // SOITEC // Tide Environnement // Université Grenoble Alpes // Université Savoie Mont Blanc // Wikit // Xpdeep

SOMMAIRE

Ce document est interractif !

Préambule

Contexte, objectifs et démarche de l'étude

Contexte : le défi énergétique de l'intelligence artificielle

L'intelligence artificielle (IA) transforme de manière profonde notre société sous bien des aspects. Elle permet notamment de faire face au déluge de données auquel nous sommes quotidiennement confrontés en nous offrant des outils qui nous permettent de traiter ces masses d'informations et d'en extraire l'essentiel. Depuis 2022, l'intelligence artificielle générative révolutionne la manière dont nous produisons l'information dans de nombreux secteurs personnels ou professionnels.

Il convient dès à présent de distinguer l'IA spécialisée de l'IA générative.

Comme son nom l'indique, l'intelligence artificielle spécialisée est destinée à la résolution de tâches très spécifiques. Cette intelligence artificielle spécialisée est en général utilisée dans la prédiction ou l'optimisation de processus industriel. Pour schématiser, son coût énergétique reste très acceptable au regard des bénéfices qu'elle apporte.

L'intelligence artificielle générative de son coté, permet de créer tous types de contenus, qu'ils soient textuels, sonores, ou graphiques. L'IA générative donne l'illusion d'une création comme seuls les humains sont capables de le faire. Cependant, il s'agit bien d'une illusion car les modèles d'IA génératives ont été entraînés sur la base des connaissances humaines et ne sont capables de produire que des contenus plausibles au regard du corpus sur lequel elles ont été entrainées – et uniquement à partir de ceux-ci. L'illusion du raisonnement de l'IA générative provient de l'emploi d'une technique de *« Reinforcement Learning from Human Feedback »*[1]. Cette technique intervient à la suite de la phase de l'apprentissage statistique des grands modèles de langage (LLM – Large Language Model), et permet à l'IA de choisir des actions en fonction des entrées qu'elle reçoit et du contexte qu'elle a déjà produit. C'est une sorte d'incarnation qui donne un *« corps »* à l'IA et qui, grâce à cela, devient drastiquement plus performante[2/3].

Quoi qu'il en soit, contrairement à l'IA spécialisée, l'IA générative est très consommatrice de ressources à la fois dans les phases d'apprentissage de ses modèles, mais également dans les phases d'utilisation des modèles, appelées phases d'inférence.

Et surtout, ce phénomène croît de plus en plus rapidement. Selon le rapport *« International Scientific Report on the Safety of Advanced AI »*[4], le besoin en computation (en FLOPs) est multiplié par 4,1 chaque année depuis 2010.

[1] https://larevueia.fr/quest-ce-que-le-rlhf-rl-from-human-feedback/
[2] https://arxiv.org/pdf/2504.12501
[3] https://arxiv.org/pdf/2204.05862
[4] https://assets.publishing.service.gov.uk/media/6716673b96def6d27a4c9b24/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf

D'après les projections énoncées dans ce rapport, et en supposant que la croissance de l'IA se poursuive au même rythme, le plus gros entraînement de l'IA devrait consommer 90 TWh à la fin de la décennie, soit la moitié de la consommation en énergie des Etats-Unis sur l'année 2022.

Usuellement, nous sommes conditionnés à développer des IA de plus en plus massives, alignant à chaque fois davantage de neurones artificiels, avec des modèles chaque fois plus volumineux. Ces IA consomment toujours plus de hardware et de ressources. Ce conditionnement développe même une transposition du phénomène de Moore à toute forme d'activité informatique. Pour autant, nous savons que la consommation énergétique ne suit pas nécessairement la capacité de calcul d'une IA. Cette quête du *« toujours plus »* est sans doute mue par la quête d'une forme d'IA omnisciente, d'une IA généralisée omnipotente. On sait aussi que cette démarche du *« toujours plus »* est outrageusement facilitée depuis l'origine du développement des réseaux de neurones artificiels. On a cherché à toujours faciliter l'augmentation du nombre de neurones dans un réseau neuronal. Ceci servant la *« scaling hypothesis »*. Cette dernière conjecture dit en somme que plus on aligne de neurones, plus le modèle sait adresser des sujets complexes. Cette quête du graal, logique et compréhensible, répond de façon simple à l'augmentation de la performance d'une IA : *« plus on augmente le nombre de neurones et mieux ça marche »* selon un adage scientifiquement questionnable. Ainsi, la représentation de l'amélioration de la performance conduit à un revers : on oublie que des IA fortement spécialisées sont souvent plus efficaces dans leur domaine que des IA généralistes auxquelles on demande de tout faire, depuis les calculs mathématiques, jusqu'à la rédaction de textes, en passant par le décodage de vidéos. Alors bien sûr, une IA omnisciente et généraliste (IAG) est certainement l'objectif ultime de tout chercheur en IA, mais il est peut-être temps de faire preuve de discernement et de se poser la question suivante : *« ai-je réellement besoin d'une IAG pour traiter de ce sujet particulier ? »*.

En effet, tous les modèles d'intelligence artificielle ont ceci en commun : ils doivent être entraînés avant d'être utilisés. On parle de phase d'apprentissage et de phase d'inférence.

Pour essayer de donner des ordres de grandeur, un modèle typique d'IA spécialisée ne consommera que quelques kilowattheures dans sa phase d'apprentissage et quelques milliwattheures lors de son utilisation. Le bénéfice d'un modèle d'IA spécialisé est très souvent très grand devant son coût énergétique, la plupart du temps négligeable. On peut par exemple penser un système de prédiction de panne qui évitera à une chaîne de production d'être interrompue par la défaillance non anticipée d'une machine, lorsqu'une IA spécialisée pourra prédire la panne et ainsi planifier l'intervention sur la machine incriminée avant que celle-ci ne soit hors service.

À contrario, on estime que la phase d'apprentissage d'une IA générative, comme ChatGPT par exemple, consommerait environ 1,280 MWh, ce qui est l'équivalent de la consommation annuelle de quelques 270 foyers français. Dans le même ordre d'idées, une simple requête à ChatGPT consomme environ 3 Wh, soit 10 fois plus qu'une recherche sur Google[5]. Même si le coût énergétique unitaire peut paraître raisonnable, le facteur aggravant des IA génératives provient de leur extrême popularité. Avec plus d'1,6 milliard d'utilisateurs réguliers, on estime la consommation quotidienne de ChatGPT à environ 560 MWh, soit plus de 200 GWh par an, ce qui représente la consommation de 44 000 foyers français ou encore la consommation électrique annuelle de la République centrafricaine du Bénin ou de Sierra Leone !

[5] https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

La lecture de ces chiffres donne à réfléchir quant à l'utilisation et le bien-fondé de l'IA générative, surtout quand on comprend que les grands modèles de langage ont besoin d'être entraînés régulièrement[6].

L'impact environnemental de l'IA est un défi connu et identifié depuis longtemps. C'était l'un des sujets centraux des échanges qui ont eu lieu lors du Sommet pour l'action sur l'IA qui s'est tenu à Paris en février 2025. La prise de conscience est claire et les gouvernants des différents pays industrialisés sont autour de la table pour discuter de potentielles solutions.

Objectifs du groupe de travail

Mais alors qu'en est-il vraiment ? Est-ce que les chiffres astronomiques indiqués dans le paragraphe précédent sont réellement le reflet de la réalité ?

On peut légitimement douter de ces chiffres, car ils sont basés sur des estimations, au mieux approximatives, voire totalement fausses.

C'est bien tout l'objectif du « think tank » que Minalogic a réuni pour faire le point sur la situation en apportant des repères quantifiés, en identifiant des solutions existantes diminuant l'impact de l'IA sur l'environnement, en identifiant les usages de l'IA dans lesquels son apport est clairement en faveur de la décarbonation, et enfin en émettant des recommandations pour améliorer la situation.

Démarche de l'étude

Une trentaine d'experts, chercheurs en IA, fournisseurs ou utilisateurs de systèmes d'IA, majoritairement issus de la région Auvergne-Rhône-Alpes, ont choisi de se réunir autour du pôle de compétitivité Minalogic, afin de contribuer à ce think tank et de produire ce livre blanc.

Une dizaine de réunions d'échanges passionnants et animés ont permis de produire le présent document. Une première phase d'analyse de la situation actuelle a été suivie d'une phase d'identification des solutions disponibles et des cas d'utilisation pour lesquels l'usage de l'IA est bénéfique à l'environnement, pour terminer par l'expression de recommandations afin de diminuer l'impact environnemental de l'intelligence artificielle.

[6] https://www.inria.fr/fr/llm4code

Synthèse du rapport et des propositions d'actions

Les travaux du think tank ont montré que l'impact environnemental de l'IA est certainement une évidence – du moins en ce qui concerne son volet *« IA générative »*. L'impact énergétique des IA spécialisées est plus modéré, et le retour sur investissement énergétique (coût énergétique vs bénéfice énergétique ou sociétal) est souvent plus évident à démontrer.

Ces travaux ont également mis en évidence que la prise de conscience est réelle, et ce à tous les niveaux de la société : citoyens, décideurs, utilisateurs et spécialistes de l'IA. Cependant, on sait encore mal quantifier cet impact environnemental et il est difficile d'exprimer un retour sur investissement environnemental de l'IA.

Ce sont probablement des pistes de travail pour la poursuite des travaux de ce think tank :

- explorer les méthodes pour mesurer et quantifier le réel coût environnemental de l'IA,
- définir ce qui acceptable et ce qui ne l'est pas, sachant que la notion d'acceptabilité est certainement culturelle et non universelle,
- favoriser la recherche d'IA frugales pour sortir de la démarche du *« toujours plus »* dans le développement des IA.

Le cluster MIAI porte un groupe de travail sur le développement des IA frugales et a financé des chaires concernant l'impact environnemental de l'IA par le passé. Une réunion des deux initiatives va avoir lieu pour adresser cet enjeu majeur de l'IA.

Partie 1. Identifier et comprendre l'impact de l'intelligence artificielle sur l'environnement

1.1. Introduction

1.1.1. La transition « *net zéro* » et la transition énergétique

Afin de limiter le réchauffement de la planète à 1,5 °C (accord de Paris de 2015[7]), il est nécessaire d'atteindre la neutralité carbone avant 2050. Cette transition vers le « *Zéro émission nette* » signifie que « *les émissions de gaz à effet de serre [soient] réduites à un niveau aussi proche que possible de zéro, les émissions restantes présentes dans l'atmosphère étant réabsorbées par les océans et les forêts* ».

Atteindre cet objectif nécessite de stopper l'usage des combustibles fossiles et donc d'opérer une transition énergétique vers un système énergétique décarboné. Cette transition énergétique repose sur trois leviers : la sobriété en priorisant les besoins essentiels, l'efficacité en utilisant moins d'énergie pour un même usage et enfin l'usage des énergies renouvelables et des combustibles nucléaires selon certains scénarios.

1.1.2.Technologies numériques et transition « *net zéro* »

Dans le contexte de la transition énergétique, les technologies numériques ont le potentiel d'améliorer la gestion de l'énergie, l'efficacité énergétique et de promouvoir l'adoption de technologies bas-carbone, y compris les énergies renouvelables. Mais ces effets positifs peuvent être réduits par l'augmentation de la demande de biens et services (un phénomène aussi connu sous le nom d'effet rebond), la consommation des appareils et la production de déchets électroniques. La compréhension des impacts directs et indirects de la digitalisation sur l'utilisation d'énergie reste donc limitée[8]. Les technologies de l'information et de la communication sont responsables de 5 % à 9 % de la consommation globale d'électricité, avec une tendance à la hausse. En particulier, la demande de services numériques a augmenté de 550 % entre 2010 et 2018 et est actuellement estimée à 1 % de la consommation globale d'électricité. Les émissions de carbone attribuées aux technologies numériques représentent environ 2,5 % de l'empreinte carbone de la France, avec une tendance à la hausse, et 3 % au niveau mondial[9].

[7] *https://www.un.org/fr/climatechange/paris-agreement*

[8] *United Nation. (2025). For a livable climate: Net-zero commitments must be backed by credible action. Climate Action. https://www.un.org/en/climatechange/net-zero-coalition*

[9] *Hilty, L.M., Aebischer, B.: ICT Innovations for Sustainability. Advances in Intelligent*

1.1.3. Les types d'impacts du numérique

Les impacts positifs et négatifs des technologies de l'information et de la communication (TIC) peuvent être analysés selon 3 niveaux : la technologie, ses usages et ses effets systémiques.

- **Au niveau technologique**, les impacts environnementaux sont principalement **négatifs**. Les impacts sont liés aux effets directs de la production, de l'utilisation et du traitement des déchets issus des TIC, qui peuvent être analysés via une analyse de cycle de vie (ACV). Cette analyse permet d'évaluer la consommation de matériaux et d'énergie tout au long du cycle de vie des produits et services.
- **Au niveau des usages,** les effets peuvent être positifs ou négatifs. Parmi les effets négatifs, **l'effet d'induction** est le fait que la technologie stimule la consommation d'une autre ressource (ex : une imprimante induit la consommation de papier, d'encre et de cartouches). **L'effet d'obsolescence** est le fait que l'usage d'une nouvelle technologie raccourcit la durée de vie d'une autre technologie existante (ex : une application qui ne fonctionne pas sur d'anciens modèles de smartphone rend ces modèles obsolètes). Parmi les effets potentiellement positifs, **l'effet de substitution** est le fait qu'une TIC remplace l'usage d'une autre ressource (ex : faire une visioconférence au lieu d'un déplacement). **L'effet d'optimisation** désigne le fait qu'une TIC aide à réduire l'usage d'une autre ressource (ex : la domotique peut aider à réduire la consommation de chauffage ou de climatisation).
- **Au niveau systémique**, les effets peuvent également être positifs ou négatifs. Ces effets se manifestent à plus long terme que les effets liés aux usages et incluent les changements structurels et de mode de vie. Parmi les effets négatifs systémiques se trouvent **l'effet rebond** et les **risques émergents**. L'effet rebond se produit lorsque la consommation globale de ressources augmente malgré des gains d'efficacité, les effets d'optimisation sont dépassés par les effets négatifs d'induction (ex : envoyer un email est plus efficace qu'envoyer une lettre, mais on en envoie donc beaucoup plus). Les risques émergents sont des risques nouveaux dont les effets ne sont pas toujours connus, par exemple les vulnérabilités de cybersécurité des infrastructures numériques ou les fake news. Les impacts positifs systémiques impliquent des changements dans les structures socio-économiques.
- **Au niveau économique**, les modèles d'économie de plateforme (ex : covoiturage) permettent par exemple de mutualiser des ressources. Au niveau politique, les TIC peuvent permettre d'implémenter des politiques environnementales par la surveillance plus fine de l'environnement ou des comportements. Au niveau social, une meilleure circulation de l'information via des TIC peut accélérer la diffusion de nouvelles normes sociales plus favorables à la transition écologique.

L'évaluation des impacts directs des technologies par analyse de cycle de vie se développe avec des standards et des méthodes qui deviennent fiables. Cependant, évaluer les impacts plus indirects d'une technologie au niveau des usages ou de ses effets systémiques reste encore aujourd'hui un défi.

1.1.4. Les effets de l'IA en particulier

L'IA générative s'impose aujourd'hui comme un levier majeur de la transformation économique mondiale, portée par une adoption rapide et massive. ChatGPT, par exemple, a atteint 100 millions d'utilisateurs en seulement deux mois[10] et s'est hissé au rang de 5ème site le plus consulté au monde[11]. Cette adoption sans précédent pousse les fournisseurs de modèles à intensifier les usages, certains anticipant jusqu'à 2 heures d'interaction quotidienne par utilisateur avec des outils d'IA générative[12].

Cependant, cette expansion rapide s'accompagne de coûts considérables. Aux États-Unis, le projet Stargate, doté d'un budget de 500 milliards de dollars, illustre l'ampleur des investissements nécessaires pour soutenir l'infrastructure des centres de données indispensables pour l'IA générative[13]. Par ailleurs, OpenAI déclare dépenser quotidiennement 700 000 dollars pour maintenir ses services[14], soulignant les défis financiers liés à ces technologies.

Malgré cet engouement, des disparités subsistent dans la perception de la valeur stratégique de l'IA. Selon l'AI Radar du BCG[15] (Boston Consulting Group), si 75 % des cadres dirigeants considèrent l'IA comme une priorité stratégique majeure, seuls 25 % déclarent en percevoir une valeur significative à ce jour, et 60 % peinent à définir ou suivre des indicateurs financiers clairs. Cet écart met en lumière les difficultés à traduire les ambitions stratégiques en bénéfices économiques mesurables à court terme.

Pourtant, pour rentabiliser ces investissements massifs, les fournisseurs de modèles misent sur deux axes principaux : l'optimisation des coûts opérationnels et la création de nouvelles opportunités de marché. Cette dernière stratégie repose sur l'intensification des usages et la génération de nouveaux besoins, ce qui pourrait accélérer les effets rebonds environnementaux et sociaux. Ces dynamiques soulèvent des questions cruciales[16] sur la durabilité de cette trajectoire, tant sur le plan économique, social, qu'écologique.

Derrière le terme *« intelligence artificielle »* se cachent des réalités technologiques et des usages très variés, avec des impacts énergétiques qui diffèrent dans des proportions considérables. Pour poser les bases d'une réflexion sur la sobriété numérique, il est essentiel de distinguer quatre ordres de grandeur majeurs de consommation énergétique liés à l'IA : l'entraînement de l'IA générative, particulièrement coûteux ; l'entraînement de l'IA non générative, beaucoup plus sobre ; l'inférence de l'IA générative, qui reste significative à grande échelle ; et l'inférence de l'IA non générative, souvent extrêmement faible.

L'IA générative, qui produit du texte, des images ou du code, est aujourd'hui très visible dans le débat public. Mais elle ne représente qu'une partie du champ de l'IA. Son entraînement peut consommer de l'ordre du GWh. 1 GWh : c'est ce qui a été nécessaire à l'entraînement de Llama 2 en 2023, soit l'équivalent de la recharge de 6 000 à 10 000 voitures électriques de type berline[17].

Les modèles non génératifs, souvent regroupés sous les termes d'*« IA traditionnelle »* ou d'*« IA symbolique »*, recouvrent une grande diversité d'usages concrets dans des domaines variés, bien au-delà de l'IA générative. Ces modèles peuvent être entraînés avec des consommations de l'ordre de quelques kilowattheures : c'est ce qu'on observe dans certaines PME qui conçoivent et entraînent localement leurs propres modèles pour des usages ciblés. C'est comparable à l'énergie consommée par un radiateur électrique domestique pendant quelques heures.

L'inférence de l'IA générative, bien qu'elle consomme nettement moins que son entraînement, reste énergivore, en

[10] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[11] https://www.similarweb.com/top-websites/

[12] https://www.technologyreview.com/2025/03/04/1112768/inside-the-wild-west-of-ai-companionship/

[13] https://www.lemonde.fr/en/economy/article/2025/01/22/stargate-trump-s-500-billion-project-to-boost-artificial-intelligence_6737299_19.html

[14] G. Elimian. Chatgpt costs $700,000 to run daily, openai may go bankrupt in 2024, 2023. https://technext24.com/2023/08/14/chatgptcosts-700000-daily-openai/

[15] https://web-assets.bcg.com/0b/f6/c2880f9f4472955538567a5bcb6a/ai-radar-2025-slideshow-jan-2025-r.pdf

[16] https://www.researchgate.net/publication/390920260_Frugal_AI_Introduction_Concepts_Development_and_Open_Questions

[17] https://arxiv.org/pdf/2307.09288

particulier à grande échelle. On estime que l'inférence d'un grand modèle d'IA générative peut consommer jusqu'à environ dix fois plus qu'une recherche Google[18]. Rapportée à des millions d'utilisations quotidiennes, cette consommation devient significative et pèse sur l'empreinte énergétique globale de l'IA générative.

L'inférence de certains modèles d'IA non générative peut se faire avec très peu d'énergie. Dans le cas de l'IA symbolique, on parle de quelques joules seulement – soit l'équivalent de l'énergie nécessaire pour soulever une pomme avec son bras.

Ces écarts illustrent l'importance de ne pas parler de *« l'IA »* comme d'un tout homogène. Par ailleurs, la rapidité avec laquelle les modèles évoluent rend rapidement obsolètes les estimations disponibles : une veille régulière ou des études actualisées sont indispensables pour éclairer les décisions. Une stratégie énergétique ou industrielle pertinente suppose de reconnaître la diversité des approches, des usages, et de leurs impacts respectifs.

1.1.5. Les autres impacts de l'IA sur l'environnement et la société

L'IA a besoin d'une puissance de calcul inconnue jusqu'alors. Sam Altman, le CEO d'OpenAI, a déclaré qu'il lui faudrait plus de 7 000 milliards de dollars d'investissement (soit l'équivalent de tout l'investissement en semi-conducteur) pour mener son projet à bien, et qu'il ne pourrait y arriver sans une rupture dans la production d'énergie.

L'explosion des datacenters et de leurs capacités grandissantes posent de réelles questions par rapport à la production électrique pour les alimenter[19/20/21]. Combiné à l'avènement des voitures électriques, on anticipe aisément que nos sociétés vont devoir faire des choix...

Une autre tendance lourde induite par le besoin énergétique de l'IA est le changement de paradigme que subit le secteur de la production d'énergie. Jusqu'alors, la production d'énergie était une problématique d'Etat, et même si les producteurs d'énergie étaient des entreprises de droit privé, le marché de l'énergie restait souvent régulé par les différents gouvernements.

Tesla a ouvert la brèche en développant l'infrastructure nécessaire à ses superchargeurs. Plus récemment, Google a réhabilité une centrale nucléaire (Five Mile) et Microsoft a annoncé la création de sa propre centrale nucléaire également.

On assiste bien à une privatisation progressive d'un secteur historiquement régalien...

Dans ce contexte de choix politiques et géopolitiques, on pourrait opposer souveraineté et environnement, à l'instar du choix de la Corée du Sud de développer une alternative souveraine à Google Maps. Même si dans ce cas l'enjeu n'est pas l'impact environnemental, mais le contrôle des données GIS et l'indépendance d'un pays par rapport aux Etats-Unis, la Corée du Sud a choisi de maîtriser ses propres données cartographiques, fermant leur accès à Google Maps pour promouvoir une alternative d'un acteur national. Le choix sud-coréen se place dans une volonté d'indépendance technologique vis-à-vis du reste du monde en général et des Etats-Unis en particulier.

En Europe, on pourrait opter pour une utilisation des datacenters existants situés hors du territoire plutôt que d'en construire de nouveaux, malgré leur impact environnement considérable. Il en est de même pour le choix des Grands Modèles de Langage (LLM), dont le développement européen impacte négativement l'environnement, alors que le choix politique pourrait être d'utiliser les LLM existants – américains ou asiatiques.

[18] https://www.cell.com/joule/fulltext/S2542-4351(23)00365-3

[19] https://www.linkedin.com/posts/octave-klaba-3a0b3632_depuis-quelques-jours-on-entend-beaucoup-activity-7294997565928861698-SEJ5?utm_source=share&utm_medium=member_

[20] https://www.linkedin.com/posts/octave-klaba-3a0b3632_milliards-deuros-en-cascade-gigawatts-activity-7295426699184226305-vId6?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAL209UB67wo2gWZ2DFWcw2DWmkaBqTPpos

[21] https://www.linkedin.com/posts/octave-klaba-3a0b3632_pour-mesurer-limpact-co2-de-lai-et-donc-activity-7295714003379474436-LagY?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAL209UB67wo2gWZ2DFWcw2DWmkaBqTPpos

La dépendance à des technologies aussi révolutionnaires que l'intelligence artificielle est certainement une raison suffisante aux yeux de nos gouvernants pour faire le choix d'un difficile équilibre souveraineté/environnement.
L'IA générative est un formidable outil ... à condition de bien savoir l'utiliser !

Plusieurs études récentes ont montré que son utilisation amplifie les différences de niveaux entre les étudiants : les plus brillants en tirent le meilleur, les moins bons l'utilisent trop naïvement et renforcent la médiocrité de leur production[22/23/24]. D'autres études tendent à prouver que l'IA générative tend à faire perdre du temps à ses utilisateurs, tout en leur donnant l'impression d'en gagner[25].

L'IA générative présente de nombreux points communs avec toute avancée technologique qui automatise le travail :

- Risque de perte de compétences humaines : à force de sous-traiter les tâches, l'humain finit par ne plus savoir les réaliser. Ce qui est d'autant plus critique avec l'IA générative est le fait que l'humain doit continuer d'avoir un regard critique sur la production de l'IA, du fait des hallucinations.
- Effet *« boîte noire »*, avec perte de visibilité sur le processus de réflexion, voire de décision. La responsabilité humaine reste prépondérante par rapport aux sorties de modèles d'IA.
- Utilisation de ressources pour nourrir les machines plutôt que les humains (substitution du travail par le capital, industrialisation des métiers intellectuels).
- L'automatisation, bien qu'elle permette de libérer du temps et des ressources et de développer de nouvelles compétences au détriment d'autres compétences, est parfois plus chronophage, avec une valeur ajoutée plus faible pour l'humain.

Pour apporter un peu plus de nuances, ces effets, bien qu'observés, réels et factuels, ne sont pas forcément des fatalités.

[22] https://www.lecho.be/opinions/general/opinion-le-systeme-educatif-tarde-a-integrer-l-intelligence-artificielle-dans-ses-enseignements/10567161.html

[23] https://pedagogie.ac-montpellier.fr/aide-la-differenciation-pedagogique-grace-lia

[24] https://eduscol.education.fr/4188/les-intelligences-artificielles-et-leurs-usages-en-education

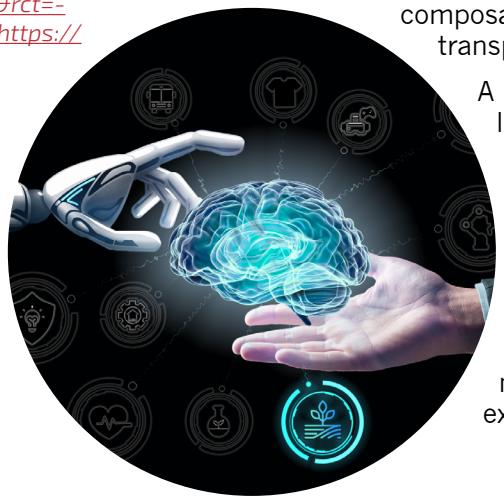
[25] https://intelligence-artificielle.developpez.com/actu/373740/Une-etude-revele-que-les-outils-d-IA-de-codage-ralentissent-les-developpeurs-tout-en-leur-donnant-l-illusion-d-etre-plus-rapides-ils-ont-mis-19-pourcent-plus-de-temps-a-accomplir-les-taches-de-codage

1.2. Mesurer et quantifier : clés d'une objectivation de l'impact de l'IA sur l'environnement

1.2.1. De l'importance de la quantification

De nombreuses études ont souligné l'importance de la localisation du thermostat sur les économies d'énergies réalisées dans l'habitat. Une position centrale et visible du thermostat au sein de la maison responsabilise ses occupants, alors qu'une installation périphérique a tendance à le faire oublier et induit une consommation plus importante.

Pour faire le parallèle avec l'impact environnemental de l'IA, il est critique de mesurer avec certitude l'impact environnemental : énergie, eau, etc. dans toutes les phases de l'IA (apprentissage et inférence) et dans toutes ses composantes (puissance de calcul, stockage et transport des données).



A l'instar du thermostat à la maison, la disponibilité de tableaux de bord de l'IA pourrait sensibiliser les utilisateurs à un usage raisonné de la technologie.

Il existe quelques initiatives émergentes qui commencent à structurer des méthodologies pour mesurer l'impact environnemental des technologies du numérique. On peut par exemple citer le Référentiel Environnemental du Numérique, défini par le Shift Project[26] ou encore le référentiel d'évaluation de la performance environnementale des services numériques (Négaoctet)[27].

Mais ces approches sont généralistes et s'attachent à analyser le cycle de vie des produits et services numériques (ex. : Negawatt), notamment l'impact du matériel et des infrastructures, dont l'IA est très gourmande. Mais il faudrait également inclure dans l'analyse les spécificités liées à l'IA, en distinguant notamment les phases d'apprentissages (très gourmandes en puissance de calcul), mais également les phases d'inférence (consommatrices de moins de ressources par requête unitaire, mais avec une volumétrie conséquente).

Il existe également des travaux normatifs concernant l'impact environnemental de l'IA, notamment à l'AFNOR, et des simulateurs pour estimer l'impact d'un modèle. L'outil Green Algorithms[28] permet par exemple de calculer diverses quantités qui entrent dans l'évaluation de l'impact environnemental d'un modèle d'IA.

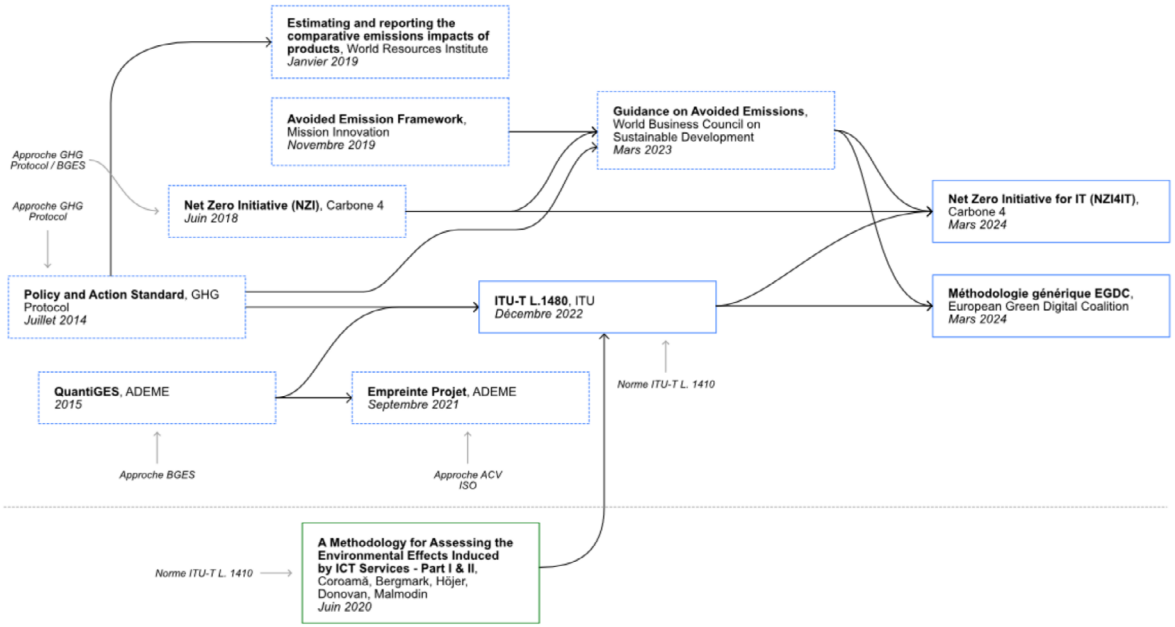
Plusieurs méthodologies existent pour mesurer l'impact du numérique sur l'environnement, et certaines sont spécifiques à l'IA. Les méthodes pour le numérique sont intéressantes, car elles sont éprouvées et peuvent être appliquées à l'IA avec peu de changements.

[26] https://theshiftproject.org/article/pour-une-sobriete-numerique-rapport-shift/

[27] https://www.google.com/url?sa=t&source=web&rct=-j&opi=89978449&url=https://codde.fr/wp-content/uploads/2022/09/APR-PERFEC-TO-2019-Rapport-final-NegaOctet_vfinal_public.pdf

[28] https://www.green-algorithms.org/GAapp-overview/

Le document de référence de l'ADEME[29] propose une synthèse des différentes méthodes existantes pour le numérique ainsi qu'une analyse comparative de méthodes sélectionnées. La figure ci-dessous explicite les liens entre les méthodologies les plus récentes : Empreinte projet de l'ADEME, publiée en 2021, la norme ITU-T L.1480 publiée en 2022, la méthodologie européenne EGDC, nommée *« Net Carbon Impact Assessment »*, publiée en 2024 et la méthodologie de Carbone 4 NZI-IT[30/31/32/33]. Toutes ces méthodologies reposent sur trois piliers : la modélisation des effets directs, des effets indirects et le calcul de l'impact net. Elles se différencient par la manière de les calculer mais aussi par la comparaison ou non avec un scénario de référence (point de comparaison, existant ou fictif, qui représente la situation sans la technologie étudiée) ainsi que l'intégration d'une étude de sensibilité et/ou de gestion de l'incertitude des données.



Cliquez sur l'image pour zoomer

Figure #01 : *Lien entre les principales méthodologies sur les impacts nets de la numérisation*[29]

[29] https://librairie.ademe.fr/ged/9417/ADEME-IT4Green-evaluation-env-numerique_Etape_0.pdf

[30] https://base-empreinte.ademe.fr/empreinte-projet

[31] https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=15030&lang=fr

[32] https://www.greendigitalcoalition.eu/assets/uploads/2024/04/EGDC-Net-Carbon-Impact-Assessment-Methodology-for-ICT-Solutions.pdf

[33] https://www.carbone4.com/publication-nzi-it

[34] https://telechargement.afnor.info/normalisation-livre-blanc-technologies-numeriques-transition-ecologique

Tout récemment, l'Afnor a publié un livre blanc[34] synthétisant une proposition de méthodologie en quatre étapes-clés pour le numérique, représentées sur la figure ci-dessous, qui peut s'appliquer à l'IA : définir le cadre de l'étude, identifier les effets avec un arbre des conséquences, calculer les impacts et analyser les résultats.

Pour le calcul des impacts, la classique Analyse de Cycle de Vie est indispensable. La version spécifique pour le numérique[35] ainsi que la métrique Normalized Carbon Footprint[36] qui mesure l'impact du matériel et de son utilisation peuvent être utilisées.

Cliquez sur l'image pour zoomer

Figure #02 : *Etapes clé de la méthodologie*

Finalement, nous pouvons citer deux initiatives spécifiques à l'IA : l'Afnor spec, référentiel général pour l'IA frugale[37] ainsi que la boussole de l'IA[38] développée par le Shift Project, qui sera présentée dans le rapport final prévu pour la fin de l'année 2025. Ces méthodologies s'orientent sur des analyses fonctionnelles et qualitatives.

Pour clore ce chapitre, voici une anecdote amusante, même si cette analyse ne s'intéresse qu'à la consommation énergétique directe d'un LLM : le cerveau humain consommerait environ trois fois plus qu'une IA générative[39]

1.2.2. Les data centers : partie immergée de l'impact environnemental des IA

Augmentation de la consommation d'énergie

Un domaine clé de préoccupation est la consommation d'énergie. Les modèles d'IA nécessitent une puissance de calcul significative, ce qui se traduit par une forte demande énergétique, représentant actuellement moins de 0,03 % de la consommation énergétique mondiale totale[40]. Bien que ce chiffre puisse sembler faible, la croissance rapide des applications d'IA est préoccupante[41/26]. L'Agence Internationale de l'Énergie (AIE) prévoit une augmentation de la demande énergétique liée à l'IA de plus de 30 % d'ici à 2026[26/42]. Cette augmentation de la consommation d'énergie est principalement due au besoin de centres de données puissants pour héberger et traiter les énormes ensembles de données utilisés pour l'entraînement et l'exécution de ces modèles.

Utilisation de l'eau

De plus, ces centres de données nécessitent des quantités substantielles d'eau pour le refroidissement[43]. Une étude de l'Université de Californie estime que la demande en eau de l'IA pourrait atteindre entre 4,2 et 6,6 milliards

[35] *D. Bol, S. Boyd and D. Dornfeld, « Application-aware LCA of semiconductors: life-cycle energy of microprocessors from high-performance 32nm CPU to ultra-low-power 130nm MCU », in Proc. IEEE ISSST, 2011*

[36] *L. Eeckhout, "FOCAL: A First-Order Carbon Model to Assess Processor Sustainability", ASPLOS 2024.*

[37] *https://www.afnor.org/actualites/referentiel-pour-mesurer-et-reduire-impact-environnemental-de-ia/*

[38] *https://theshiftproject.org/wp-content/uploads/2025/03/Rapport-intermedaire-IA-VF.pdf*

[39] *https://www.linkedin.com/pulse/un-cerveau-humain-consomme-30-w-une-requ %C3 %AAte-sur-llm-des-cirotteau-1ahue/*

[40] *Electricity – Analysis and forecast to 2026 IEA - 2024*

[41] *IA Environnement - CESE 2024*

[42] *AI Index Report 2023 - HAI Stanford University*

[43] *The AI Risk Repository - Whittlestone et al - 2024*

de mètres cubes d'ici à 2027, soit l'équivalent de la moitié de la consommation annuelle du Royaume-Uni[44]. Cela met en évidence l'empreinte hydrique associée au développement et à l'utilisation de l'IA.

Épuisement des ressources

L'IA dépend fortement du matériel contenant des métaux à forts enjeux géo-stratégiques dit *« terres rares »*, ce qui contribue à l'épuisement des ressources. Les appareils terminaux, en particulier les écrans et les téléviseurs, contribuent pour 65 à 90 % à l'impact environnemental des technologies numériques[27], mais les GPU, nécessaires pour entraîner les IA seront également en forte demande.

Impact indirect

Bien que l'impact environnemental de l'IA soit indéniable, l'IA peut aussi apporter des solutions aux problèmes environnementaux[30/45]. Cependant, le fort développement des applications à but lucratif dans des secteurs comme la finance et le marketing soulève des préoccupations quant au fait que ce potentiel pourrait être éclipsé par sa contribution à une consommation accrue[46/47].

Aux États-Unis, la consommation des infrastructures liées aux IA génératives représente actuellement environ 4 % de la demande électrique nationale, un chiffre qui pourrait atteindre entre 4,6 % et 9,1 % d'ici à 2030 selon les projections de l'EPRI (Electric Power Research Institute). L'IA, bien qu'elle ne représente aujourd'hui qu'environ 10 à 20 % de la consommation des centres de données, voit son impact croître rapidement. Une requête effectuée via ChatGPT consomme environ 2,9 Wh, soit dix fois plus qu'une recherche Google classique (0,3 Wh), et d'autres applications, comme la génération d'images et de vidéos, risquent d'amplifier cette tendance[48].

La relation entre complexité des IA et consommation des datacenters est conceptuellement évidente. La quantifier avec précision est une tâche ardue et complexe, car de très nombreux paramètres entrent en ligne de compte :

- Complexité du modèle évidemment, ce qui impacte directement taille de mémoire et puissance de calcul nécessaires.
- Modèle d'hébergement / type de datacenter (on premise, cloud, embarqué...).
- Complexité du design des ressources allouables dans le datacenter par rapport au scaling horizontal (ajout de nœuds), vs le scaling vertical (augmentation des capacités d'un noeud) vs obsolescence rapide des hardwares (et nécessité de maintenance pour faire évoluer le parc).
- La consommation d'un DC générant de la chaleur, il faut bien penser à entrer dans l'équation le coût du refroidissement (consommation énergétique additionnelle + consommation d'eau + réchauffement).
- Granularité des demandes (nombre de demandes utilisant des ressources communes vs nombre de demandes et sollicitations utilisant des ressources énormes, forcément rares, et de façon exceptionnelle).
- La mutualisation des ressources utilisées par l'IA présente une efficacité de consommation énergétique puisqu'un système sert de multiples usages.

[44] *Evaluation de l'impact environnemental du numérique en France - ADEME – Arcep – 2023*

[27] *https://www.google.com/url?sa=t&source=web&rct=-j&opi=89978449&url=https://codde.fr/wp-content/uploads/2022/09/APR-PER-FECTO-2019-Rapport-final-NegaOctet_vfinal_public.pdf*

[30] *https://base-empreinte.ademe.fr/empreinte-projet*

[45] *Les promesses de l'IA grevées par un lourd bilan carbone - https://www.lemonde.fr/*

[46] *ChatGPT consommerait l'équivalent d'une bouteille d'eau par conversation - Novethic - https://www.novethic.fr/*

[47] *L'intelligence artificielle, une "bombe climatique" invisible - Novethic - https://www.novethic.fr/*

[48] *https://www.epri.com/research/products/000000003002028905*

1.3. Exemples d'expérimentations d'IA à impact environnemental positif

1.3.1. L'exemple de Naver Labs Europe

Naver Labs Europe est le plus grand centre de recherche industriel en IA en France, et il se concentre sur le développement de nouveaux modèles pour la robotique. Le centre compte environ 80 chercheurs permanents et il possède ses moyens de calcul, avec plusieurs centaines de GPU de génération récente (typiquement NVDIA A100 ou H100/200) sur le site grenoblois, plus une centaine sur un cloud privé, ainsi qu'une capacité de stockage rapide locale de plusieurs pétaoctets (données précises confidentielles). En 2024, NLE a effectué un bilan carbone portant sur les données de l'année 2023. Le bilan carbone a montré que pour NLE, l'impact de développement de modèles IA – impact du matériel et de la consommation d'énergie (principalement entraînements) – est nettement inférieur à l'impact du transport fait dans le cadre des déplacements professionnels. Il est à noter que les déplacements dans le cadre d'activités de recherche sont nombreux et souvent lointains, en plus des visites effectuées vers la maison mère en Corée.

1.3.2. Les usages dans le domaine de la gestion de l'énergie

« Digital technologies have the potential to reduce energy demand in all end-use sectors through steep improvements in energy efficiency. This includes material input savings and increased coordination as they allow the use of fewer inputs to perform a given task. Smart appliances and energy management, supported by choice architectures, economic incentives and social norms, effectively reduce energy demand and associated GHG emissions by 5 10 % while maintaining equal service levels. »

Au-delà de l'évaluation de l'IA en termes d'énergie et d'empreinte carbone, on peut voir plus largement l'impact des projets IA sur l'environnement.

Outre la décarbonation, Inddigo travaille sur de multiples aspects liés à la transition : préservation et régénération de la biodiversité, résilience climatique des territoires, économie circulaire et déchets...

L'IA, en permettant d'être plus efficaces dans les propositions de solutions, permet d'accélérer la transition: des projets réalisés plus vite, de plus grande qualité, intégrant de plus grandes quantités d'informations...
Les consultants peuvent se concentrer sur l'analyse et

l'expertise, plutôt que sur les tâches de plus faible valeur ajoutée, au bénéfice de l'environnement.

Sur ce même sujet, on peut également citer une étude prospective de Schneider Electric[49] concernant la prédiction et la gestion de l'énergie en utilisant l'IA. Cette étude propose plusieurs scenarii de consommation d'énergie de l'IA et révèle plusieurs trajectoires plus ou moins impactantes à l'horizon 2035. Ces différentes trajectoires sont influencées par de nombreux facteurs : choix stratégiques concernant l'utilisation de l'IA, cadres juridiques et gouvernance associée.

Schneider Electric a également conduit une autre étude montrant le grand bénéfice de l'usage de l'IA dans le contrôle et la gestion des systèmes de chauffage, ventilation et air conditionné (HVAC) des bâtiments d'éducation (lycées, collèges, universités...)[50]. Cette étude démontre l'impact positif de l'IA, et son coût environnement négligeable devant les économies énergétiques réalisées.

Une chaire de l'institut multidisciplinaire sur l'intelligence artificielle MIAI de l'Université Grenoble Alpes, intitulée CIMES, s'intéresse également aux systèmes d'énergie complexe, en rassemblant académiques et industriels du bassin grenoblois, notamment cités dans ce document. En effet, l'intégration des énergies renouvelables et des ressources distribuées, telles que les véhicules électriques, qui accompagnent les changements actuels des systèmes énergétiques, contribue à leur complexité et à leur variabilité accrue. L'IA offre des solutions complémentaires pour optimiser la distribution, le stockage et la consommation d'énergie, améliorant ainsi l'efficacité et la fiabilité de ces systèmes. On parle ici de gestion d'énergie, mais également de conception de nouveaux systèmes résilients. La décentralisation de l'intelligence et l'interconnexion avec les systèmes d'information et de contrôle doivent être développées en gardant l'humain dans la boucle, mais surtout en garantissant des améliorations globales des systèmes énergétiques, basées sur des critères non uniquement techniques et économiques, et en minimisant les effets de rebond.

1.3.3. Optimisations dans le transport routier

En 2020, Transarc, entreprise de transport routier de voyageurs disposant alors d'un parc de 700 véhicules répartis sur 16 dépôts, s'est engagée dans une démarche d'optimisation de ses opérations logistiques afin de réduire l'impact environnemental de ses activités.

L'un des principaux enjeux identifiés concernait les trajets à vide entre les dépôts et les points de prise en charge, appelés *« Haut-Le-Pied »*. Ces déplacements, bien que nécessaires à l'organisation des tournées, représentaient une source importante de consommation énergétique, de coûts opérationnels et d'émissions de gaz à effet de serre.

Dans ce contexte, Transarc a sollicité l'expertise de Neovision pour développer une solution fondée sur l'intelligence artificielle, visant à optimiser l'affectation des conducteurs et des tournées. Une application web a été déployée, permettant de simuler différents scénarios d'organisation et d'en évaluer les impacts sur les trajets à vide.

Les résultats ont été significatifs : 500 000 kilomètres évités sur une année, soit une économie estimée à 500 000 euros, avec un retour sur investissement en moins de trois mois. Cette initiative a également permis de réduire les émissions liées au carburant et d'améliorer les conditions de travail des conducteurs.

[49] https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/artificial-intelligence-electricity-system-dynamics-approach/

[50] https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/ai-powered-hvac-in-educational-buildings/

En 2025, Transarc souhaite aller plus loin en optimisant leur logiciel (Neovision va améliorer/optimiser le code avec des nouveaux modèles publiés plus récemment).

L'utilisation de l'IA pour améliorer la circulation routière et décroître la pollution n'est pas nouvelle.

Il y a plus de 10 ans, le centre de recherche de Xerox à Meylan (aujourd'hui devenu Naver Labs Europe) a travaillé avec les autorités de Los Angeles (Californie, Etats-Unis) pour améliorer les conditions de parking dans le centre-ville.

L'idée sous-jacente est de faire varier dynamiquement le prix du parking en fonction du lieu, de la date et l'heure, et du taux d'occupation. Pour simplifier, le prix des zones sous forte tension est augmenté assez significativement, et le prix des zones sous-utilisées est diminué pour le rendre attractif, et ceci de façon prédictive, car les prix doivent être communiqués en avance.

Un algorithme d'apprentissage a été entraîné à partir des données de parking des 15 années précédentes, en complément des données d'utilisation en temps réel de 6 000 capteurs déployés dans la ville. Le but de cet algorithme est d'établir un prix de parking qui varie pendant la journée pour atteindre un taux d'occupation proche de 80 % sur chaque segment de rue, en indiquant à l'avance quel sera le tarif pour les jours à venir. Ce taux d'occupation de 80 % permet une utilisation optimale des emplacements de parking, tout en garantissant aux nouveaux arrivants de trouver de la place. L'algorithme doit prendre en compte les phénomènes réguliers (tels que les jours de la semaine, l'heure de la journée...) mais aussi exceptionnels (tels que les congés scolaires ou jours de soldes...).

Finalement, les autorités de Los Angeles ont mesuré une baisse du prix moyen du parking, une hausse du revenu total grâce à la meilleure occupation, et une baisse significative du trafic induit par la recherche de places.

L'IA a donc permis de mieux utiliser une infrastructure existante, de réduire la pollution, et de réduire la nécessité de devoir construire des parkings supplémentaires.

1.4. L'émergence d'une approche éthique et raisonnée de l'IA

Certaines entreprises ont, dès leur création, intégré les questionnements sociaux et environnementaux. C'est le cas de Neovision, dont les dirigeants ont toujours incorporé une démarche éthique dans les projets d'IA que la société réalisait, avant de formaliser leur démarche sous la forme d'une charte et d'un comité d'éthique, quelques années après sa création et au fur et à mesure du développement de la société. Cette charte et ce comité éthiques ont pour but d'encadrer les prestations et les modes opératoires afin que tout le business de Neovision reste en accord avec ses valeurs. En particulier, la société informe systématiquement ses prospects et ses clients lorsque des solutions permettent de répondre à leurs besoins, sans passer par de l'IA. Dans cette optique, elle souhaite aujourd'hui aller au-delà de la charte et du comité éthiques portés par ses fondateurs en incluant les questions d'efficacité énergétique des IA qu'elle développe.

Depuis 2024, dans le cadre du projet-pilote Datawise, soutenu par la Région Auvergne-Rhône-Alpes, Neovision expérimente des outils développés en interne, permettant de mesurer précisément la consommation énergétique de ses GPU lors des phases d'entraînement. Cette initiative constitue une première étape vers une meilleure maîtrise de l'empreinte énergétique de ses modèles. Si Datawise sert de terrain d'expérimentation, l'objectif affiché est d'étendre cette démarche à l'ensemble des développements impliquant l'entraînement d'IA, afin d'en faire un standard interne.

Il faut noter que ce type d'approche est intéressant car il mesure concrètement le coût énergétique des phases d'apprentissage, ce qui est un progrès certain par rapport aux pratiques actuelles de l'industrie.

Il ne faut cependant pas oublier le coût des phases d'inférence, car, bien que minimal dans la plupart des cas, l'explosion de l'IA générative a montré qu'un coût unitaire minimal peut devenir très significatif si le nombre de requêtes explose. Ce sont bien les petits ruisseaux qui font les grandes rivières...

De manière similaire, tout projet chez Inddigo est systématiquement évalué avant son lancement en s'appuyant sur une démarche visant à questionner son alignement avec les engagements et les valeurs de l'entreprise.

Les projets sont donc arbitrés avec une *« boussole d'analyse »* commune, et un projet pourra être abandonné en amont si son impact négatif potentiel semble trop important au regard des engagements d'Inddigo.

Si une charte des ESN en IA se dégageait, Inddigo pourrait l'intégrer à sa boussole. En tout état de cause, elle

pourrait refuser d'utiliser une IA si son impact n'est pas mesuré et positif.

Plus spécifiquement concernant l'utilisation de l'IA dans les projets, l'idée est d'évaluer les gains (qualité, temps, plus-value...) et l'impact environnemental (difficile aujourd'hui), et de s'assurer que le bilan est positif.

Chez Soitec, l'usage de l'IA générative est également raisonné et aligné avec les engagements environnementaux de l'entreprise. Chaque cas d'usage doit commencer par le modèle le moins énergivore, suffisant pour tester un prompt ou explorer une idée. Le choix des outils IA se fait en fonction des besoins métiers identifiés, en évitant les surdimensionnements. Un accompagnement par la formation est essentiel pour favoriser une adoption efficace et responsable. L'IA générative est un levier de productivité, mais son usage doit rester sobre, ciblé, et piloté.

Pour encadrer ces usages, une gouvernance dédiée a été mise en place. Elle repose sur une équipe pluridisciplinaire composée de représentants des départements Strategy Office, Juridique, IT, Ressources Humaines, et Sécurité de l'information. Cette diversité permet une approche équilibrée entre innovation, conformité, sécurité et gestion des talents. Cette équipe travaille sous la supervision d'un comité de pilotage, ou Steering Committee, composé de membres du Comité Exécutif. Ce comité se réunit de manière trimestrielle pour valider les orientations stratégiques, évaluer les risques et ajuster la feuille de route des projets IA.

Une charte d'usage de l'IA formalise les bonnes pratiques, les interdictions comme le partage de données sensibles, et les principes éthiques à respecter. Cette charte est automatiquement communiquée à toute personne recevant une licence d'un outil d'IA générative.

Par ailleurs, les formations dispensées incluent systématiquement une sensibilisation à l'impact énergétique des grands modèles de langage (LLM) et recommandent l'usage des modèles les plus sobres et adaptés aux besoins spécifiques.

L'utilisation des outils IA est strictement limitée à ceux validés par les équipes IT et Sécurité Globale, après une revue contractuelle effectuée par le département Juridique. Il est interdit d'utiliser des comptes Soitec pour accéder à des outils non validés. Enfin, la gouvernance encourage une vigilance continue des utilisateurs, qui doivent adopter une posture critique face aux contenus générés, respecter les principes de transparence, rester informés des évolutions technologiques et faire preuve d'éthique en toutes circonstances.

Cette gouvernance garantit un usage sécurisé, sobre et aligné avec les valeurs de Soitec.

Partie 2. Quelques initiatives prometteuses

2.1. Vers une IA frugale ou une frugalité de l'IA

La réponse européenne au besoin de frugalité pour l'intelligence artificielle est de se doter de gigafactories en IA, construites autour d'un principe d'efficience énergétique fort, dans des pays où le mix énergétique est plutôt favorable, car produisant une grande partie de leur énergie sur des ressources renouvelables (solaire, hydroélectrique, nucléaire...), ou faisant appel à des techniques de refroidissement de type *« free-cooling »*. Certes approches ont moins d'impact pour l'environnement, mais sont-elles suffisantes ? Peut-on aller encore plus loin ?

On pourrait par exemple mentionner l'initiative de Flex AI (https://flex-ai.fr/), qui propose une solution de mutualisation de ressources de calcul.

Un système d'IA peu consommateur :

- en ressources (pour l'entrainement et l'exécution)
- et en données (pour l'entrainement).

Quelques bonnes pratiques :

- choisir un service d'entrainement localisé dans un pays où l'énergie est issue d'énergie verte,
- entrainer un modèle à partir d'un modèle de fondation pour éviter d'en recréer un.

De manière globale, il faut systématiquement questionner les outils par rapport à leurs usages. En effet, l'utilisation d'importants modèles d'IA générale soulève de nombreux questionnements, qu'ils soient écologiques, économiques ou sociaux. Leur coût prohibitif nous dirige potentiellement vers des quasi-monopoles. Il est donc raisonnable de se demander s'il n'est pas préférable, pour nos applications du quotidien, d'orienter nos efforts vers le développement d'approches d'IA plus spécialisées et frugales, en charge d'une seule et unique tâche, mais de manière efficiente et à bas coût. Cette relation entre sobriété et frugalité se retrouve aussi sur l'impact à l'usage. Une seule amélioration de la fourniture d'un service entraine souvent au paradoxe de l'effet rebond : plus un service est efficace (et utile), plus on l'utilise et donc plus il a d'impact.

2.2. Approche matérielle

2.2.1. La stratégie d'Asygn

Asygn est une entreprise grenobloise spécialisée dans le développement de petits processeurs neuronaux (NPU) à haute efficacité énergétique. Son objectif est d'imaginer la prochaine génération des dispositifs IoT grâce à sa puce Colibry, qui permet de les rendre à la fois intelligents et autonomes en énergie. Cette puce vise à déployer l'IA traditionnellement effectuée dans les datacenters, au plus proche de la source des données (Edge AI). Les données sont analysées par un réseau de neurones matériel intégré dans la puce qui consomme moins de 10mW. Grâce à cette très faible consommation, il est possible de concevoir des dispositifs IoT complètement autonomes en énergie en récupérant l'énergie disponible sur place, par exemple grâce à des cellules photovoltaïques.

2.2.2. L'ère de l'IA à la périphérie : Comment les substrats FD-SOI transforment notre monde ?

L'intelligence artificielle est communément définie comme la capacité d'une machine à exécuter les fonctions cognitives que nous associons habituellement à l'esprit humain. Elle ne cesse de transformer nos vies quotidiennes à un rythme effréné, avec des applications omniprésentes allant de la création de contenu et des assistants virtuels à la conduite autonome (ADAS/AD), l'Industrie 4.0, les soins de santé, la recherche climatique et la sécurité. L'IA englobe diverses formes, allant de l'apprentissage automatique (Machine Learning) qui permet aux systèmes d'apprendre des données pour prendre des décisions, comme les recommandations personnalisées de plateformes de streaming, au Deep Learning qui utilise des réseaux neuronaux pour tirer des conclusions sans intervention humaine, essentiel pour la reconnaissance faciale ou la traduction. Les avancées récentes incluent les grands modèles de langage (LLM) qui comprennent et génèrent du texte de manière humaine, et l'IA générative qui crée de nouveaux contenus basés sur des données existantes. Pour soutenir cette croissance exponentielle de l'IA, la demande en puissance de calcul est intense, et les *« engineered substrates »* de Soitec sont devenus une composante essentielle de l'industrie des semi-conducteurs.

Traditionnellement, l'entraînement et l'inférence des modèles d'IA se déroulaient principalement dans le cloud, via de grands centres de données dotés d'accélérateurs d'IA. Cependant, cette approche présente des limitations notables. Le Cloud AI souffre de latence, de la nécessité d'une connexion internet fiable, de risques pour la confidentialité des données et d'une forte consommation d'énergie au niveau des centres de données.

C'est pourquoi l'IA à la périphérie (Edge AI), où les puces IA exécutent des modèles optimisés à faible puissance plus près de la source de données, présente des avantages multiples et cruciaux pour l'avenir des technologies intelligentes. En effet, elle offre une latence nulle, ne nécessite aucune connexion pour des opérations essentielles, sécurise la confidentialité des données en les traitant localement, et assure une faible consommation d'énergie. La prolifération des appareils Edge AI est impressionnante, passant de 15 millions en 2020 à 2,5 milliards d'unités prévues pour 2030, soulignant le passage de l'IoT à l'AIoT avec une intelligence embarquée croissante. Le marché des chipsets Edge AI devrait d'ailleurs doubler d'ici à 2027, passant d'environ 15 milliards de dollars à 35 milliards de dollars.

Les substrats FD-SOI (Fully Depleted Silicon-On-Insulator) de Soitec représentent une plateforme idéale pour l'Edge AI et au-delà. Ils offrent des capacités de performance à la demande, sont optimisés pour les appareils alimentés par batterie, intègrent la radiofréquence (RF) et disposent de mémoires non-volatiles (NVM) embarquées. Le FD-SOI est spécifiquement conçu pour assurer une consommation d'énergie active plus faible, ce qui permet un mode *« toujours activé »*, des performances à la demande, des capacités robustes de récupération d'énergie (atteignant même des capacités *« zéro consommation »*), et le traitement le plus économique en termes d'inférences par watt et par dollar. Ces caractéristiques essentielles permettent au FD-SOI de fournir le meilleur rapport performance/puissance pour les applications Edge AI.

L'impact du FD-SOI est particulièrement notable dans le secteur automobile, où il transforme les systèmes avancés d'aide à la conduite (ADAS) et les capacités de conduite autonome (AD). Le contenu en semi-conducteurs par voiture devrait passer d'environ 460 dollars en 2020 à plus de 1 100 dollars en 2030, avec une augmentation significative des radars automobiles. Le FD-SOI permet une réduction d'environ 30 % des émissions de gaz à effet de serre, une réduction de 50 % de la taille des puces (die size) et une augmentation de 50 % de la portée de détection des radars. Ces avancées soutiennent des applications d'IA cruciales telles que la reconnaissance d'objets et de piétons, la détection des panneaux de signalisation, l'assistance au franchissement de ligne, et l'amélioration de l'expérience en véhicule grâce à la surveillance du conducteur, le suivi comportemental et la reconnaissance vocale. De plus, le FD-SOI est idéalement positionné pour les applications radar de nouvelle génération, offrant une performance supérieure en matière de coût système, de performance radar et d'efficacité énergétique par rapport aux alternatives.

Dans le contexte de l'Industrie 4.0, l'Edge AI alimentée par FD-SOI peut améliorer de manière significative les opérations de fabrication. Elle permet une amélioration du rendement de 30 %, une réduction des déchets de 15 % et des coûts d'exploitation inférieurs de 10 %. En accélérant les tests de conception via la prédiction des performances, en identifiant rapidement les causes profondes des problèmes et en réduisant le temps d'ingénierie, l'IA et le FD-SOI transforment la fabrication en des opérations intelligentes, sécurisées et adaptables.

Au-delà de l'automobile et de l'industrie, le FD-SOI est également l'architecture de choix pour une multitude d'appareils intelligents (Smart Devices), y compris les capteurs environnementaux intelligents, les dispositifs médicaux IoT, les appareils portables (wearables) et la maison intelligente. En outre, les substrats de Soitec jouent un rôle clé dans la connectivité à haute vitesse pour les centres de données et l'interconnexion GPU, essentiels pour l'IA/ML en nuage, permettant des débits

de données plus rapides et une consommation d'énergie réduite.

Soitec continue d'étendre son empreinte mondiale pour répondre à la demande croissante, avec une capacité maximale de production de wafers pouvant atteindre 2,7 millions par an grâce à ses sites de Bernin 2 en France et Pasir Ris à Singapour. Grâce à sa feuille de route technologique innovante et ses investissements en recherche et développement, le FD-SOI est positionné pour être le *« terreau »* des innovations futures, de l'automobile à l'Edge AI, en améliorant de manière significative la connectivité, l'efficacité énergétique et la performance globale.

2.2.3. L'infrastructure d'inférence frugale AI-over-space de MountAIn

MountAIn développe une infrastructure révolutionnaire d'intelligence artificielle par satellite (AI-over-space) qui repense fondamentalement l'architecture traditionnelle de l'IA en déplaçant l'inférence au plus près de la source des données. Cette approche d'Edge AI représente une rupture technologique majeure dans la réduction de l'impact environnemental des systèmes de vision intelligente.

Le capteur de vision intelligent IBEX : la pointe de l'iceberg

Au cœur de cette infrastructure se trouve IBEX, un capteur de vision intelligent qui incarne parfaitement les principes de l'IA frugale. Ce *« gardien numérique »* autonome effectue l'inférence d'IA pour la vision par ordinateur directement sur le dispositif, sans nécessiter de connexion permanente au cloud ou d'alimentation électrique traditionnelle.

IBEX intègre dans un format compact :

- un capteur d'imagerie haute résolution,
- un processeur d'IA embarqué pour l'inférence en temps réel,
- un panneau solaire pour l'autonomie énergétique complète,
- un module de connectivité satellitaire,
- une passerelle LoRa® pour l'intégration multi-capteurs.

L'innovation majeure réside dans sa capacité à analyser les flux vidéo à 10-20 images par seconde avec la même précision que dans le cloud et à ne transmettre que les événements pertinents sous forme de messages compacts de 50 octets, optimisant ainsi drastiquement l'utilisation de la bande passante et réduisant la consommation énergétique.

Architecture réseau spécialisée pour l'inférence IA

L'infrastructure de MountAIn ne se limite pas au dispositif IBEX mais s'étend à un écosystème complet de protocoles réseau spécialement conçus pour l'inférence IA. Cette architecture hybride edge-to-cloud comprend :

1. Protocoles de transmission optimisés : Les protocoles développés par MountAIn sont spécifiquement adaptés au transfert d'images et de données d'inférence sur des connexions à bande passante limitée (LoRa, NB-IoT), permettant une transmission efficace même dans les environnements les plus contraints.
2. Backend MLOps avancé : L'infrastructure backend constitue un véritable système d'exploitation hybride qui orchestre intelligemment les tâches entre l'edge et

le cloud. Ce système permet :

- la mise à jour over-the-air (FUOTA) des modèles d'IA sans intervention physique,
- la gestion distribuée des algorithmes d'inférence
- et l'optimisation dynamique des ressources selon les besoins opérationnels.

3. Modèle d'IA adaptatif à deux couches : L'architecture logicielle d'IBEX repose sur un modèle d'IA modulaire permettant une personnalisation fine selon les cas d'usage :
   - Détection et suivi : identification d'objets, personnes ou animaux avec attribution d'identifiants anonymisés et application de règles de géofencing
   - Analyse de mouvement : suivi des trajectoires, analyse comportementale et prédiction basée sur l'évolution multi-images
   - Impact environnemental révolutionnaire

L'approche de MountAIn génère une réduction de l'impact environnemental de 99 % par rapport aux solutions on-premise traditionnelles (telles que celles utilisées durant les Jeux Olympiques de Paris) tout en exécutant des modèles de détection d'objets similaires sur des images de taille comparable et en préservant nativement les données personnelles.

Cette performance exceptionnelle s'explique par plusieurs facteurs :

1. Consommation énergétique ultra-faible : avec moins de 50mW de consommation totale, IBEX consomme significativement moins que les capteurs d'imagerie standards, permettant un fonctionnement 24/7 sur simple panneau solaire.
2. Élimination des infrastructures lourdes : en supprimant le besoin de centres de données dédiés pour le traitement d'images, l'architecture de MountAIn évite la construction et l'exploitation d'infrastructures énergivores.
3. Optimisation du transport de données : seules les images d'intérêt sont transmises vers le cloud, réduisant drastiquement les besoins en bande passante et l'énergie associée au transport des données.

L'infrastructure AI-over-space de MountAIn représente ainsi un paradigme nouveau pour l'IA environnementalement responsable, prouvant qu'il est possible de concilier intelligence artificielle avancée et sobriété énergétique extrême grâce à une architecture repensée de bout en bout.

2.3. Approche logicielle

L'un des moyens d'atteindre une certaine frugalité en IA est de réduire la complexité de ses modèles. Il est en effet fréquent de voir des modèles d'IA, notamment les grands modèles de langage (LLM), comporter plusieurs centaines de milliards de paramètres, qui sont autant de consommation de ressources (eau, électricité, puissance de calcul) lors de la phase d'apprentissage, mais également lors de la phase d'inférence à cause de la taille des modèles en résultant.

2.3.1. Réduction des modèles par le moteur de deep learning : la startup Xpdeep

La startup deeptech Xpdeep, basée à Grenoble, née d'un transfert technologique de l'Université Grenoble Alpes, soutenue par le programme Confiance AI, le GICAT et l'EIT Digital, élue Meilleur projet où investir en 2025 selon le magazine Challenges, a développé la première plateforme de deep learning conçue pour la transparence, la gouvernance et l'impact opérationnel.

L'IA basée sur le deep learning est puissante, mais souvent incompréhensible, d'où son surnom de boite noire. Pour de nombreuses organisations, cela pose des obstacles majeurs :

- pas de confiance dans les décisions du modèle,
- impossible à auditer ou à certifier (réglementations, conformité),
- inadaptée aux contraintes métier ou aux objectifs réels
- et difficile à déployer dans des environnements critiques ou régulés.

La réponse de la société Xpdeep est de proposer une nouvelle génération de modèles IA, une technologie unique qui permet de créer des modèles de deep learning auto-explicables, c'est-à-dire :

- compréhensibles par les humains dès leur conception (pas de boîte noire),
- audités, contrôlés et certifiables,
- alignés avec les contraintes opérationnelles ou réglementaires
- et capables de suggérer ce qu'il faut changer pour obtenir un meilleur résultat — une IA qui ne prédit pas uniquement, mais qui agit, une IA explicable.

Xpdeep n'ajoute pas de l'explication après coup. L'approche rend l'IA explicable par construction. C'est ce qui permet aux modèles d'être :

- certifiables,
- gouvernables
- et actionnables par les équipes opérationnelles.

Quelques exemples d'application qui montre tout le potentiel de la technologie et en quoi elle permet également de réduire l'impact environnemental de l'AI :

- un modèle de scoring bancaire peut indiquer quelles variables financières ajuster pour passer un client de *« risqué »* à *« solvable »*,
- un modèle médical peut suggérer les changements physiologiques minimaux pour réduire le risque de récidive d'un patient,
- un modèle industriel peut recommander des paramètres d'usage pour éviter une panne avant la maintenance programmée,
- l'application de Xpdeep au cas de maintenance prédictive sur le jeu de données *« Failure Detection of SCANIA Component X »*) a permis :
  - une réduction du nombre de variables d'entrée de 87, passant de 105 à 19 variables effectivement utilisées — soit 82 % de réduction – ce qui permet également de réduire le nombre de capteurs physiques à l'origine de ces variables,
  - une optimisation de l'efficacité de calcul : inférence divisée par deux (FLOPs),
  - un modèle allégé : nombre de paramètres réduit d'un tiers.

Xpdeep offre en plus une actionnabilité à ses utilisateurs, qui peuvent à présent déclencher les ajustements nécessaires pour non seulement prédire les pannes, mais surtout les retarder. Des coûts environnementaux sont ainsi également évités.

2.3.2. L'approche de Wikit de l'IA générative spécialisée

Wikit a choisi de développer ses propres modèles d'IA générative en les spécialisant à l'usage auquel ils sont destinés, au lieu d'utiliser les traditionnels grands modèles de langage disponibles sur le marché (Mistral AI, Llama, OpenAI, deepseek, etc.).

Le gain environnemental est certain, car les modèles spécialisés de Wikit ont des complexités sans commune mesure avec leurs grands frères LLM, et le coût environnemental de leur phase d'apprentissage est quasiment négligeable par rapport à leurs aînés.

La phase d'inférence est également très faible, de même que la volumétrie, car l'usage est très spécialisé du fait de leur spécialisation.

Bien entendu, l'étendue de leur domaine d'application est très restreinte, du fait de la spécialisation voulue dès leur conception.

Cet exemple illustre bien le paradigme de l'impact environnemental de l'IA : il faut choisir entre une IA capable de tout faire avec un coût élevé, ou un coût plus raisonné, mais avec un domaine d'application plus étroit. .

2.3.3. La proposition d'ExcellerIA

Chez ExcellerIA, des IA, et plus spécifiquement des réseaux de neurones artificiels, sont entrainés en interne via un framework propriétaire d'entraînement de modèles. Cet outil permet à la société de spécifier un espace de paramètres de modèle à explorer afin de trouver le *« petit »* modèle d'IA répondant justement au problème donné. La Neural Architecture Search (NAS) est le champ des techniques permettant la recherche automatisée d'un modèle d'IA optimal, répondant à un problème donné[51]. L'équipe TAU de l'INRIA (spécialisée en apprentissage automatique et optimisation) travaille actuellement sur la Neural Architecture Growth, qui permet de partir d'un neurone et identifier puis intégrer successivement les composants les plus pertinents afin d'obtenir un modèle complet répondant au problème donné.

Cette procédure d'entraînement réalisée chez ExcellerIA n'est pas l'unique façon de produire des modèles plus petits[52].

La *« distillation »* consiste à entraîner un petit modèle (student) à partir d'un plus gros (teacher), en apprenant à reproduire ses sorties. Cette technique a une double utilité. Elle permet d'entraîner un petit modèle qui n'aurait pas pu l'être directement à partir du jeu de données d'entraînement. Elle permet ainsi la production de modèles spécialisés à partir d'un modèle généraliste.

Lorsque le modèle d'IA a terminé son entraînement, il existe plusieurs techniques permettant de réduire sa taille finale et donc sa consommation de ressources une fois déployé :

1. Le *« pruning »* est la première technique et repose sur le principe de suppression des neurones et des connexions les moins efficaces du modèle[53/54]. L'application de cette technique a un impact variable sur la taille du modèle, réduisant de 30 % la taille de GPT3. Cependant l'impact peut être plus important selon la précision du modèle à laquelle on est prêt à renoncer.
2. La *« quantization »* est une autre technique qui consiste à réduire la précision des poids du modèle, par exemple passer de nombres à virgule flottante encodés sur 32 bits à 16, 8 ou 4 bits. Par exemple cette technique permet d'atteindre une réduction de la taille d'un LLM de 68 %[55]. Bien que cette technique soit efficace une fois le modèle entraîné, l'appliquer préalablement à l'entraînement augmente drastiquement l'instabilité du modèle et donc les risques d'échouer. Une pratique courante consiste donc à utiliser des représentations numériques spécifiquement conçues pour les réseaux de neurones artificiels. C'est le cas des Brain Floating Point (BFloat) qui sont encodés sur 16 bits et qui permettent de réduire la taille du modèle, dès la phase d'entraînement, sans perdre en stabilité.

Enfin, les travaux sur les architectures de réseaux de neurones artificiels portent sur la réalisation de composants qui implémentent la fonction recherchée. Cela permet de sélectionner et de mettre en valeur, pour des neurones, les propriétés importantes des données pour réaliser la tâche. En un sens, cette technique soulage les neurones d'apprendre à réaliser cette transformation par eux-mêmes.

Par exemple, l'architecture de réseaux de neurones la plus simple consiste à disposer par couche des neurones connectés entre eux. Pourtant ces neurones performent mal dans les tâches de classification d'images. Ainsi, les Convolutional Neural Networks (CNN)[56] reprennent ces mêmes couches de neurones auxquelles on vient interca-

[51] https://arxiv.org/pdf/1905.01392
[52] https://arxiv.org/pdf/1503.02531
[53] https://arxiv.org/pdf/2011.00241
[54] https://www.researchgate.net/publication/221618539_Optimal_Brain_Damage
[55] https://arxiv.org/html/2411.06084v1#S6

ler différentes fonctions mathématiques, forçant les neurones à apprendre les motifs souhaités.

Toutes ces techniques permettent de réduire la consommation des modèles de réseaux de neurones artificiels. Cependant, les espoirs de réduire drastiquement cette consommation est limitée par le fonctionnement même de ces types de réseaux. Deux mécanismes sont à l'origine de cette limite, le premier est que le réseau est mobilisé dans sa totalité à chaque requête, or très peu de neurones ont vraiment un impact dans la prédiction pour une requête donnée. Le second mécanisme limitant est la rétro-propagation, ce processus est ce qui permet au réseau d'apprendre en corrigeant les poids du modèle vers la solution optimale, mais c'est un calcul très coûteux, qui s'applique une fois de plus sur tous les neurones lors de l'entraînement.

On comprend que les méthodologies classiques atteignent leurs limites du fait du mécanisme fondamental des neurones artificiels. On entend par là que pour réaliser une étape d'entraînement, c'est-à-dire un cycle comprenant une prédiction du réseau suivi par la mesure de l'erreur entre la prédiction et l'attendu puis de la mise à jour des poids du modèle, il faut nécessairement réaliser les calculs sur chaque neurone composant le modèle.

Selon ce fonctionnement, seul une faible partie de la consommation en ressources de l'IA, lors des phases d'entraînement et d'inférence, sert effectivement à produire et maintenir les fonctions intelligentes recherchées alors que l'entièreté du réseau de neurones est mobilisée.

Face à ces limitations fondamentales, de nouvelles approches neuromorphiques émergent comme alternative prometteuse. Les réseaux de neurones spikants (SNN)[57] sont la 3ème génération de réseaux de neurones artificiels, et permettent de propager de l'information via des signaux discrets aux autres neurones. Ce mécanisme répond fondamentalement à la limite précédemment évoquée en permettant aux neurones de répondre uniquement sur réception d'un signal. Ces neurones permettent aussi d'utiliser le matériel de computation spécifique comme les FPGAs (Field-Programmable Gate Arrays, circuits intégrés reconfigurables) ou des processeurs neuromorphiques, qui rendent leur consommation significativement inférieure aux neurones artificiels classiques. Certaines études montrent une réduction de 69 % dans la consommation en énergie en utilisant les SNNs[58]. Il faut tout de même souligner que les SNNs sont aujourd'hui très précoces et se confrontent à l'état de l'art des Neurosciences Computationnelles. Les SNNs fonctionnels implémentent une forme de rétro-propagation comme les neurones artificiels classiques, alors que le fonctionnement de l'apprentissage et de l'adaptation des neurones biologiques reste encore à élucider.

En synthèse, l'approche logicielle pour l'IA frugale combine plusieurs techniques complémentaires : la recherche automatisée d'architectures (NAS), les techniques de compression post-entraînement (distillation, pruning, quantization), et les architectures émergentes comme les réseaux de neurones spikants. Ces approches permettent des réductions significatives de la consommation énergétique, allant de 30 % à 69 % selon les techniques employées.

[56] https://www.cs.princeton.edu/courses/archive/spr08/cos598B/Readings/Fukushima1980.pdf

[57] https://link.springer.com/article/10.1007/s11063-021-10562-2

[58] https://arxiv.org/html/2409.08290v1

2.4. Approche méthodologique

En tant que société qui développe des SIAs, Neovision applique également sa charte éthique à ses pratiques commerciales : si des solutions qui ne sont pas de l'IA sont plus adaptées pour résoudre les besoins des clients, Neovision l'en informe et ne pousse pas ses solutions.

Inddigo ne se lance pas non plus tous azimuts sur l'IA, mais de manière raisonnée. L'IA est vue comme un outil auquel les consultants peuvent avoir recours, mais avec parcimonie et à certaines conditions. De manière similaire aux projets numériques chez Inddigo, une méthodologie a été mise en place, gérée par une équipe dédiée et formée :

- recensement des cas d'usage,
- évaluation des cas pertinents
- et accompagnement des utilisateurs sur les choix techniques.

Plutôt qu'une IA généraliste, Inddigo privilégie une IA très spécialisée, restreinte, laissant la main au consultant.

Par exemple, il est envisageable de proposer un outil d'assistance à la rédaction, mais il sera bien spécifique à un usage. Le traitement se fera paragraphe par paragraphe, tant pour les informations d'entrée, que pour le paramétrage, et la génération. On évite ainsi l'utilisation d'une IA généraliste qui serait plus énergivore.

Entraîner un modèle de réseaux de neurones artificiels peut reposer sur beaucoup de spéculations et dans ce cas elle provoque nécessairement un gaspillage important de ressources. Dans le cas d'une analyse prédictive, il est assez spontané d'itérer le travail de développement du modèle en commençant par un modèle Long Short Term Memory (LSTM) pour voir ce que ça donne, puis d'ajuster manuellement les hyperparamètres du modèle dans l'espoir d'identifier une solution suffisante. La problématique majeure de cette approche est que la quantité d'information récoltée au fur et à mesure des essais est faible et donc une quantité importante d'entraînements de modèles se trouvent inutiles, démultipliant le coût du développement de ce modèle. Un document produit par l'équipe Brain chez Google expose ce problème ainsi qu'une méthodologie pour développer un modèle d'IA répondant à un problème spécifique[59].

Ainsi une méthode pertinente de développement d'un modèle d'IA peut se décliner selon ces étapes :

1. Analyser le problème et le formaliser à travers deux aspects : l'aspect mathématiques et l'aspect métier. Il est important de définir une notion d'utilité quantifiable via plusieurs mesures. Ceci permet de s'assurer

[59] https://github.com/google-research/tuning_playbook?tab=readme-ov-file#a-scientific-approach-to-improving-model-performance

que le développement de l'IA avance dans la bonne direction. Il est aussi important d'envisager les valeurs de performance suffisantes pour ces mesures, afin d'arrêter les efforts de développement lorsque l'IA répond aux attentes du métier. Ne pas le faire engendre le risque de poursuivre une course à la performance qui est coûteuse puisqu'il est nécessaire de fournir beaucoup d'effort pour améliorer les performances d'un modèle alors qu'il est déjà très optimisé.

2. Identifier les contraintes techniques qui guideront la stratégie de développement du modèle d'IA. Ces contraintes auront un impact sur la taille de l'espace de recherche des hyperparamètres, l'identification des hyperparamètres qui seront intéressants à rechercher et les architectures et tailles envisageables des modèles.

3. Identifier un ensemble d'architectures de modèles qui devraient avoir la capacité de résoudre le problème.

4. Appliquer une recherche systématique des hyperparamètres lors de l'entraînement des modèles en respectant une hiérarchie dans les types de paramètres à explorer. La principale problématique, ici, est que des hyperparamètres optimaux trouvés via un entraînement court ne sont pas forcément les hyperparamètres qui seront optimaux quand on entraînera le modèle plus longtemps. L'idée est d'obtenir rapidement les modèles les plus pertinents à partir d'entraînements courts où l'on aura recherché des paramètres structuraux du modèle (comme le nombre de couche, etc.). Ensuite on réalise une seconde phase d'entraînement en faisant varier les paramètres qui sont sensibles à la durée de l'entraînement.

5. Optimiser les modèles sélectionnés avec les techniques de réduction de taille de modèle.

Une méthode de recherche systématique du modèle optimal permet d'optimiser l'usage des ressources nécessaire au développement de l'IA, tout en progressant assurément vers un modèle répondant de mieux en mieux à la problématique.

Il est important de connaître les dynamiques des réseaux de neurones afin de ne pas se tromper dans l'interprétation des résultats obtenus au fil du développement du modèle d'IA.

Pour une combinaison entre la taille de notre jeu de données et la taille de notre réseau de neurones artificiels, il existe un phénomène de double descente qui caractérise l'existence d'un régime intermédiaire lors duquel augmenter la taille du modèle ou du jeu de données dégrade les performances du modèle[60].

En réalité, il existe trois régimes de traitement de l'information par le modèle selon le rapport entre la taille du modèle et la taille du jeu de données. Un premier régime est celui de la mémorisation ; ici, le modèle a tendance à *« overfitter »*, plus précisément, il va apprendre à reconnaître chaque point de donnée. Ce régime se caractérise par un modèle qui performe uniquement sur la donnée d'entraînement. Le second régime est celui de la généralisation ; ici, le modèle apprend à reconnaître des caractéristiques à partir de la donnée. C'est dans ce régime que le modèle performe sur la donnée d'entraînement, mais aussi et surtout sur la donnée de test. Le dernier régime est une hybridation des deux autres, où mémorisation et généralisation ont lieu en même temps.

[60] *https://transformer-circuits.pub/2023/toy-double-descent/index.html*

Lorsque le jeu de données est grand et que le modèle est petit, ce dernier ne peut pas mémoriser la donnée et par conséquent il généralise. En augmentant la taille du modèle, celui-ci aura tendance à mémoriser la donnée. Le phénomène appelé *« grokking »* correspond à une transition abrupte du modèle de la mémorisation à la généralisation, après un très long entraînement[61]. En continuant d'augmenter la taille du modèle, ou la taille du jeu de données, de façon très importante, le modèle généralise à nouveau.

Cette dynamique est illustrée dans les figures ci-dessous (Figure #03 et Figure #04), tirées des travaux de Henighan et al. d'Anthropic[62], qui démontrent empiriquement les phénomènes de superposition et de double descente.

Avoir conscience des dynamiques des réseaux de neurones artificiels nous permet de progresser sans s'étonner des performances du modèle d'IA à travers son développement. Les travaux d'explicabilité des modèles d'IA ont beaucoup progressé au cours des dernières années et les prendre en compte sont nécessaires pour produire de l'IA frugale.

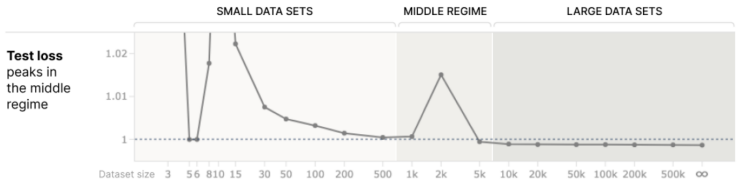
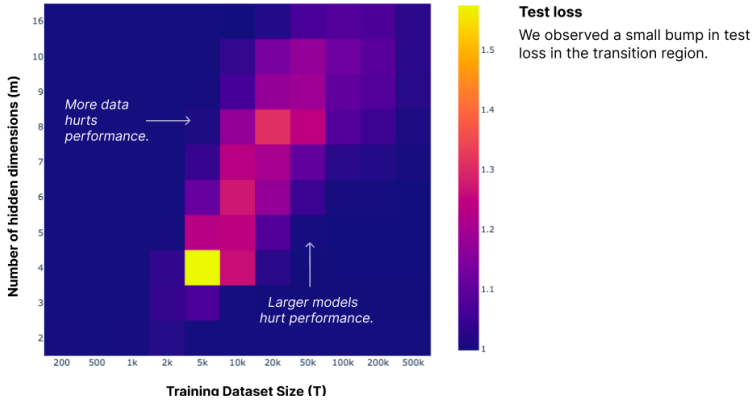


Figure #03:
Performance du modèle d'IA en fonction de la taille du jeu de données



Cliquez sur les images pour zoomer

Figure #04:
Performance du modèle selon les variations de la taille du modèle et de la taille du jeu de données

[61] *https://pair.withgoogle.com/explorables/grokking/*

[62] *Superposition, Memorization, and Double Descent, Heninghan et al, Anthropic, 2023 (https://transformer-circuits.pub/2023/toy-double-descent/index.html)*

2.5. Approche stratégique

Cette section affirme que le caractère frugal de l'IA dépend non seulement de la manière dont elle est conçue et entrainée (voir sections précédentes), mais aussi de l'usage qui en est fait. En outre, en s'appuyant sur Bohnsack et al. (202), nous proposons qu'il soit important de prendre en compte à la fois les résultats escomptés et les conséquences de second ordre sur la société et en déduisons trois conditions à remplir pour que l'usage de l'IA permette plus de frugalité énergétique.

L'IA rend de nouvelles pratiques possibles. On parle ainsi *« d'affordances digitales »*. Dans l'industrie, l'IA permet par exemple de créer des modèles prédictifs afin d'améliorer le contrôle de la qualité et faciliter la maintenance prédictive en prédisant les performances. Les IA de reconnaissance visuelle peuvent aider à améliorer les processus industriels et éviter des dysfonctionnements (ex. : identifier des gravats de taille anormale sur une chaîne de broyage). L'IA a également de nombreuses applications qui n'ont pas d'impact positif sur l'environnement. Par exemple, l'IA permet la reconnaissance faciale, outil notamment expérimenté par des écoles pour vérifier que les repas ont bien été payés par les familles.

Pour mémoire, dans une perspective de transition énergétique, l'ordre des priorités est le suivant :

1. sobriété dans les usages,
2. efficacité,
3. usage des énergies renouvelable/décarbonées.

Ce que ces exemples révèlent, c'est que l'IA est un outil qui n'amènera de la frugalité énergétique que si c'est son objectif (condition 1).

L'IA permet également d'apporter des informations supplémentaires pour aider les humains et les machines à prendre des décisions. L'IA peut par exemple être intégrée dans les outils de management des entreprises (ex. : rapports dynamiques Power Bi). L'IA permet aussi de modéliser différents scénarios de conception en amont afin de guider la prise de décision. L'entreprise Oris utilise par exemple l'IA pour créer plusieurs scénarios de conception d'infrastructures routières ou ferroviaires permettant aux donneurs d'ordres de comparer différentes options, leurs coûts, leurs impacts carbones et la disponibilité en matériaux locaux. Dans tous ces cas, le potentiel de réduction des émissions carbone dépendra in fine de l'intégration ou non des critères environnementaux aux prises de décision par les donneurs d'ordre (condition 2).

Enfin, l'IA peut faciliter l'innovation. Prenons l'exemple des cimentiers. Ces derniers utilisent l'IA dans des systèmes experts afin d'optimiser des processus de production du ciment, afin de les rendre moins énergivores. Il se

servent aussi de l'IA pour créer des outils de reconnaissance visuelle afin d'identifier la nature des matériaux dans une benne de déchets et ainsi faciliter leur recyclage ou leur réemploi.

Pour illustrer ce propos, on peut par exemple citer les initiatives suivantes :

- TOMRA[63]
- RECYCLEYE[64]
- 3Recycleo[65]

Ces innovations ont un impact direct positif sur les émissions de CO_2. En revanche, si on considère les conséquences de second ordre, le premier exemple, optimisant des systèmes de production non durable présente le risque de verrouiller l'engagement dans ces systèmes. Le second a un potentiel de réduction des émissions plus élevé car permettant une restructuration plus en profondeur du secteur (condition 3).

[63] https://www.tomra.com/fr-fr/waste-metal-recycling/media-center/news/2025/a-short-history-of-ai

[64] https://recycleye.com/fr/lia-et-la-reconnaissance-des-dechets-pourquoi-cela-fonctionne-si-bien-2/

[65] https://www.3recycleo.fr/2022/02/21/la-technologie-de-lia-pour-optimiser-la-gestion-des-dechets/

Partie 3. Recommandations et propositions du groupe de travail

Les parties précédentes ont mis en évidence le problème – sans toutefois identifier d'outils ou de méthodes pour réellement mesurer et quantifier le problème, prérequis à toute amélioration de la situation – et ont également dévoilé quelques pistes de solutions émergentes.

Cette partie s'attache à prendre un peu plus de recul pour imaginer des pistes de réflexion et des recommandations pour aller plus loin.

L'intention est également de faire émerger des idées pour nourrir le think tank, si ce dernier devait se poursuivre.

3.1. Définir des indicateurs pour éco-concevoir les IA de demain

Afin d'en limiter les effets sur l'environnement, il est critique et prioritaire de standardiser l'évaluation d'impact de l'IA – et plus généralement des services numériques (par exemple le poids carbone de l'IA) – et de rendre cette évaluation intelligible et audible par les clients/usagers, afin qu'ils puissent faire des choix technologiques éclairés. La définition de ces indicateurs demandera certainement une analyse fine de tous les éléments qui composent une IA, de la puissance de calcul dans les datacenters à la mise à disposition des données permettant le calcul d'un ROI environnemental.

Cette réflexion devrait également permettre de penser l'impact de l'IA sur les infrastructures, actuelles ou à venir, par exemple en anticipant l'augmentation du trafic lié aux AI crawlers (AI Crawlers and Scrapers Are Contributing to an 86 % Increase in General Invalid Traffic[66]).

Une fois l'impact de l'IA mesurable, on pourra alors imaginer d'objectiver son développement, et mettre en place une démarche d'écoconception de l'IA, à l'instar de ce qui se fait pour de nombreux produits industriels et autres biens de consommation. .

[66] https://doubleverify.com/ai-crawlers-and-scrapers-are-contributing-to-an-increase-in-general-invalid-traffic/

3.2. Structurer les données de l'IA

L'IA reste dépendante des données disponibles pour son entraînement et n'est qu'un outil au sein de systèmes d'informations complexes qui transcendent parfois les frontières des organisations. Pour que l'IA serve à la gestion des problèmes environnementaux, il faut articuler les systèmes d'information de manière cohérente. Cela requiert tout d'abord d'identifier le type d'information (précision, qualité, format) dont ont besoin les acteurs d'une filière. Il faut ensuite rendre obligatoire la génération de ces données de manière standardisée. Enfin, il faut établir des protocoles d'échange de données et assurer l'interopérabilité entre différents systèmes. Cette dernière étape est d'autant plus importante que la transition écologique se joue non pas au sein des organisations, mais requiert une réorganisation des chaînes de valeur et repose donc sur des flux d'informations inter-organisationnels.

Dans ce contexte, l'IA n'est qu'un outil possible parmi d'autres pour traiter les données. Nos observations montrent que les IA utilisées en pratique dans le secteur du bâtiment sont plutôt des IA spécialisées (moins gourmandes en énergie que des IA généralistes ou génératives). Il conviendrait donc d'encourager le développement de ce type d'IA spécifiquement dédiées à des usages utiles pour la transition énergétique. Par ailleurs, d'autres programmes peuvent s'avérer plus adaptés aux usages que l'IA (ex. : les architectes utilisent des logiciels et conception paramétrique, plutôt que de l'IA générative pour dessiner des bâtiments). Il importe donc de codévelopper les solutions numériques avec les usagers afin de choisir les outils les plus adaptés aux usages.

3.3. Définir et utiliser le « ROI environnemental »

Le monde commence à prendre conscience du coût environnemental de la technologie, et notamment de celui de l'intelligence artificielle.

L'une des grandes questions que l'on se pose lorsque l'on a recours à l'IA est de mesurer son retour sur investissement (ROI). Non seulement par rapport à son coût financier, mais également son coût énergétique.

Il est des situations où le coût importe moins : c'est par exemple le cas pour des applications de santé, où l'utilisation de l'IA améliore considérablement l'état du patient ou lui sauve la vie. Pour rester dans le domaine de la santé, doit-on adopter le même raisonnement quand le bénéfice de la technologie s'adresse plus au soignant qu'au patient ? La question mérite d'être posée, même s'il semble évident que dégager du temps à un spécialiste lui permettra de traiter plus de patients, et donc de sauver plus de vies...

Mais qu'en est-il pour d'autres utilisations de l'IA ? Un système de recommandation d'achat en ligne par exemple, mérite-t-il son coût énergétique ?

Il n'est pas simple de mesurer le ROI de l'utilisation de l'IA, si ce n'est en adoptant le raisonnement simpliste, mais efficace, de comparer la dépense au gain : seules les IA permettant de gagner plus d'énergie qu'elles n'en consomment auraient alors droit de cité... probablement irréaliste. Il y a là une vraie question de société qui va bien au-delà de ce livre blanc...

3.4. Questionner le recours systématique à l'IA

L'intelligence artificielle est tellement prégnante aujourd'hui qu'on ne peut imaginer un produit innovant qui n'y fasse pas appel. C'est devenu un argument marketing tellement indispensable, que les fabricants l'utilisent pour tout et (parfois) n'importe quoi.

Cela conduit parfois à utiliser des modèles d'IA pour des applications où un simple algorithme de recherche opérationnelle, ou une base de règles (GOFAI) pourrait amplement atteindre l'objectif, avec un coût environnemental beaucoup plus réduit.

Toutes les Entreprises de Services Numériques qui ont contribué à ce livre blanc ont adopté une méthodologie qui questionne l'utilisation de l'IA avant d'engager tout nouveau projet de développement.

La plupart des acteurs utilisateurs de l'IA optent pour un usage raisonné de la technologie, et ont plus ou moins formalisé leur approche.

Le think tank pourrait proposer de construire une méthodologie commune, qui pourrait prendre la forme d'une charte signée par tous les acteurs de l'IA, ayant pour objectif de promouvoir un usage à bon escient de l'IA, respectueux de l'environnement, avec un retour sur investissement environnemental positif.

Cette charte pourrait définir des usages prioritaires (sobriété) et encourager la co-conception de solutions IA autour de ces usages, afin de limiter les effets rebonds et ainsi bénéficier au mieux des améliorations d'efficacité permis par l'IA.

3.5. Passer des usages à la maîtrise et au contrôle de la technologie

L'Europe accuse un léger retard technologique en ce qui concerne l'IA générative, qui, à l'exception de Mistral, est dominée par des acteurs américains ou asiatiques. C'est nettement moins vrai pour les IA spécialisées, et les différents instituts de recherche en IA européens sont au niveau de l'état de l'art international.

Quoiqu'il en soit, il est important que l'Europe ne rate pas la diffusion de l'IA et son adoption dans les entreprises et par les particuliers.

Des programmes commencent à voir le jour en ce sens, mais au-delà des usages, la clé reste la maîtrise de la technologie, sa compréhension et son contrôle.

Audrey Tang (Ministre des affaires numériques de Taïwan) a prononcé une allocution inspirante lors du sommet pour l'action sur l'IA de Paris en février 2024 : pour favoriser l'adoption de l'IA par la société, Taïwan a mis en place des cours d'IA dès le plus jeune âge (10-12 ans).

Il ne s'agit pas là d'apprendre aux enfants à concevoir et développer un modèle d'IA et de les transformer en data scientists, mais simplement de leur donner les bases de la compréhension de la technologie pour qu'ils en cernent les limites, les utilisent au mieux en évitant les hallucinations par exemple, et in fine gagnent confiance dans l'IA. L'exemple taïwanais pourrait inspirer nos politiques en matière d'éducation...

L'IA générative est un outil très récent, mais dont l'utilisation a été massivement adoptée par le grand public, dans des situations d'une grande variété. Son utilisation nécessite une compréhension suffisante de ses avantages, mais aussi, et surtout, de ses limitations. On parle d'éducation au numérique de manière plus générale. Que ce soit dans la vie professionnelle, mais aussi dans sa professionnalisation, cet outil est à double tranchant. En effet, les enseignants constatent, notamment à l'université, que l'IA générative semble renforcer l'écart entre les apprenants. Les plus brillants en tirent le meilleur (ils sont plus rapides à produire des résultats pertinents) tout en préservant leur capacité à acquérir et solidifier des compétences, là où les moins bons l'utilisent sans recul. Ce faisant, ils délèguent leur apprentissage à un outil qui n'en a pas besoin et sortent eux-mêmes de leur formation sans réel acquis durable.

3.6. Réguler l'IA : une approche encore timide du législateur vis à vis de l'impact environnemental.

Selon le rapport de l'Agence Internationale de l'Energie (AIE) publié le 10 avril 2025[67], la consommation d'électricité des centres de données devrait *« plus que doubler »* d'ici à 2030.

Dans leur position paper publié en février 2025 et intitulé *« Les principaux défis à relever pour favoriser la performance environnementale de l'IA »*[68], l'Institut national de recherches en sciences et technologies du numérique et le Ministère de la transition écologique précisent que *« les technologies numériques représentent jusqu'à 12 % de la consommation mondiale d'électricité »*, et alertent en même temps sur le fait que le *« rythme du développement de l'intelligence artificielle dépasse largement celui de la capacité de production d'électricité à partir de sources d'énergies renouvelables »*.

Il est ainsi manifeste que l'un des enjeux cruciaux de l'intelligence artificielle générative, particulièrement avide en énergie, est environnemental.

Or, les débats autour de l'intelligence artificielle laissent souvent au second plan le sujet de l'impact environnemental, pour se concentrer sur la souveraineté technologique, l'éthique ou encore la productivité.

Dans ce contexte, la fameuse loi d'Amara trouve tout son sens. Formulée dans les années 70 par Roy Amara, prospectiviste au Standford Research Institute, cette loi stipule que *« nous avons tendance à surestimer les effets d'une technologie à court terme et à les sous-estimer à long terme »*.

Dès lors, il est intéressant d'examiner comment les régulateurs nationaux et européens ont pris la mesure de l'enjeu de l'empreinte environnementale de l'intelligence artificielle.

En outre, il convient de se demander si l'arsenal juridique existant en France et en Europe garantit un développement de l'intelligence artificielle qui ne se fasse pas au détriment de la transition écologique et soit compatible avec le principe de sobriété énergétique.

Les pouvoirs publics se sont emparés depuis de nombreuses années de la question des conséquences environnementales du développement de l'intelligence artificielle.

À ce titre, dès 2018, le rapport du député Cédric Villani *« Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne »*, ne manquait pas de mettre en exergue les importantes limites climatiques et de disponibilité des ressources du développement de l'intelligence artificielle.

Dans le prolongement de la vision apportée par la mission Villani, la stratégie nationale pour l'intelligence artificielle a posé les jalons d'une structuration de l'éco-

[67] https://doubleverify.com/ai-crawlers-and-scrapers-are-contributing-to-an-increase-in-general-invalid-traffic/

[68] https://www.inria.fr/sites/default/files/2025-03/PositionPaper_IA_environnement_Inria_FR.pdf

système d'intelligence artificielle visant notamment à promouvoir le développement de l'intelligence artificielle frugale, c'est-à-dire de solutions d'intelligence artificielle dont les besoins en ressources, tant matérielles qu'énergétiques, sont minimisées grâce à une réflexion tenue en amont sur les usages et la mise en place de bonnes pratiques pour leurs développements.

C'est dans ce cadre que le Ministère de la transition écologique et de la cohésion des territoires a défini sa feuille de route *« Intelligence artificielle et transition écologique »* et que le Conseil Économique Social et Environnemental a rendu en séance plénière le 24 septembre 2024, à l'unanimité et sans abstention, son avis *« Impacts de l'intelligence artificielle : risques et opportunités pour l'environnement »*, dans lequel sont notamment formulées des propositions sur la réduction de l'impact de l'implantation des data centers.

S'agissant des instruments juridiquement contraignants, la Commission européenne a rendu publique le 21 avril 2021 une proposition de Règlement établissant des règles harmonisées concernant l'IA[69], visant à encadrer le développement et l'utilisation de l'intelligence artificielle, tout en renforçant la compétitivité de l'Union européenne en la matière.

À la suite d'intenses discussions, le Règlement, directement applicable dans tous les États membres de l'Union, a finalement été adopté le 13 juin 2024[70], avec une entrée en vigueur progressive à compter du 1er août 2024 et jusqu'au 1er août 2030 concernant certains systèmes d'intelligence artificielle.

En ce qui concerne l'intégration des considérations d'ordre environnemental, l'article 3.5 de l'exposé des motifs de la Proposition de règlement mettait, en son temps, l'accent sur le nécessaire respect, dans le cadre de l'utilisation de l'intelligence artificielle, des droits fondamentaux, tels que consacrés dans la Charte des droits fondamentaux de l'Union Européenne, ainsi que le renforcement du droit à un niveau élevé de protection de l'environnement et l'amélioration de la qualité de l'environnement.

Dans le texte voté par le Parlement européen en mars 2024, peu de dispositions ont, in fine, trait aux enjeux environnementaux soulevés par les usages de l'intelligence artificielle.

À ce titre, si l'on analyse plus finement les 180 considérants inclus dans le texte, on constate que seuls 6 d'entre eux renvoient aux préoccupations environnementales de l'intelligence artificielle. Ceux-ci s'articulent autour des 5 axes ci-après :

1. le soutien et la promotion de solutions d'intelligence artificielle permettant l'obtention de résultats bénéfiques pour la société et l'environnement (considérants n°4 et n°142) ;
2. la promotion, le développement et l'utilisation de l'intelligence artificielle en assurant un niveau élevé de protection de l'environnement (considérant n°8) ;
3. l'intégration, dans la définition des sept principes éthiques non contraignants pour l'intelligence artificielle, du bien-être sociétal et environnemental (considérant n°27) ;
4. l'encouragement de certains acteurs de la chaîne de valeur de l'intelligence artificielle à appliquer, sur une base volontaire, des exigences supplémentaires liées notamment à la viabilité environnementale (considérant n°165) ;

[69] *Doc. COM (2021) 206 final, 2021/0106 COD*

[70] *Règlement 2024/1689 du Parlement européen et du Conseil du 13 mai 2024*

5. l'inclusion des parties prenantes environnementales dans le cadre du processus de normalisation (considérant n°121).

En second lieu, force est de constater que sur les 113 articles du Règlement, seuls 3 font état de considérations d'ordre environnemental, à savoir :

- L'alinéa 2 de l'article 40, aux termes duquel les demandes de normalisation doivent notamment porter sur *« des éléments livrables concernant les processus d'établissement de rapports et de documentation visant à améliorer les performances des systèmes d'IA en matière de ressources, (...) »*. Est ainsi attendue une norme sur la consommation d'énergie et d'autres ressources sur le cycle de vie de l'intelligence artificielle ;
- L'article 95 – 2.b, qui invite les acteurs concernés de la chaîne de valeur de l'intelligence artificielle, dans le cadre des codes de conduite d'application volontaire, à *« évaluer et minimiser l'impact des systèmes d'IA sur la durabilité environnementale (...) »* ;
- L'article 112-7, aux termes duquel la Commission se doit d'inclure la durabilité environnementale dans le cadre de l'évaluation de l'impact et de l'efficacité des codes de conduite volontaires.

Au regard des éléments précédemment exposés, l'intégration des enjeux environnementaux liés à l'intelligence artificielle dans le Règlement européen est encore perfectible.

Par ailleurs, il est important de souligner que l'arsenal juridique interne existant permet d'appréhender, de manière somme toute très incomplète, certains aspects des enjeux environnementaux de l'intelligence artificielle et notamment la régulation de l'implantation des data centers qui, du fait de la complexité de leur fonctionnement et de leur impact non négligeable sur l'environnement, est soumise au régime des installations classées pour la protection de l'environnement (ICPE) et aux dispositions du droit de l'urbanisme. De surcroît, aux termes des dispositions de l'article 28 de la loi n°2021-1485 du 15 novembre 2021 visant à réduire l'empreinte environnementale du numérique en France, l'étude des possibilités de récupération de la chaleur fatale des centres de données pourrait, sous réserve de l'adoption du décret annoncé, devenir une obligation pour les opérateurs.

Enfin, au titre des mesures non-contraignantes, l'initiative actuellement portée par le Ministère de la transition écologique et de la cohésion des territoires en partenariat avec l'association française de normalisation (AFNOR), aux fins d'élaboration d'un *« AFNOR-Spec IA frugale »* dont l'intention est de proposer à tous les acteurs un cadre méthodologique opérationnel permettant de mesurer utilement les coûts/bénéfices des modèles et systèmes d'intelligence artificielle au vu, notamment, de leurs impacts environnementaux, est à souligner.

En conclusion, le Règlement européen, dans sa version actuelle, peine encore à apporter des réponses concrètes et efficaces aux défis environnementaux de l'intelligence artificielle.

Dans ce contexte, il convient d'espérer que les actes d'exécution, les normes harmonisées à paraître ainsi que les éventuelles évolutions du texte permettront d'apporter des améliorations en ce sens.

Conclusion

Sans se prétendre exhaustif ou refléter les recherches les plus avancées sur le sujet, ce livre blanc fait l'inventaire de nombreuses solutions au niveau ''hardware'' et ''software'' qui peuvent permettre de réduire l'impact environnemental de l'IA. Souvent issues de la pratique et d'un souci d'optimisation normal des entreprises et laboratoires de l'IA et de la microélectronique, ces initiatives individuelles et spontanées méritent d'être partagées, structurées et diffusées dans notre écosystème IA.

C'est tout l'objet de ce travail collectif qui montre également que des techniques de calcul d'impact environnemental issues du numérique et déjà utilisées, pourraient être adaptées à l'IA. Ce travail préliminaire est jugé indispensable pour mesurer avec justesse l'impact environnemental, avant de se lancer dans l'éco-conception de solutions à base d'IA. Il pourrait faire l'objet d'une suite aux travaux du think tank.

Par ailleurs, ce livre blanc démontre que les IA spécialisées, sont bien plus frugales que les IA génératives. C'est en fait une bonne nouvelle, puisque les IA génératives sont maitrisées par très peu de sociétés dans le monde et que celles-ci (hors Mistral AI) sont principalement situées hors Europe. Il reste donc un champ important à exploiter par notre écosystème sur les IA spécialisées.

Enfin, comme pour la plupart des technologies ''disruptives'', l'effet rebond nous menace si nous ne réfléchissons pas aux usages, en nous focalisant principalement sur les problèmes à forts enjeux sociétaux, comme l'IA pour la santé ou pour l'optimisation des activités industrielles. Il faut également décider quelle part de notre consommation énergétique totale nous souhaitons réserver au numérique et à l'IA, dans l'avenir. Une véritable éthique ''environnementale'' doit être définie et intégrée à la formation des professionnels et des utilisateurs – même les plus jeunes – de ces technologies.

Annexe

Liste des experts ayant contribué au groupe de travail et au présent document

Philippe WIECZOREK
Directeur des projets innovants & expert Intelligence Artificielle
Minalogic (coordonnateur du think tank et du livre blanc)

A

Fabrice ARROYO
Program Director of the Advanced Master (MS) in Energy Marketing & Management
Grenoble Ecole de Management

Astrid ASTIER
Déléguée Régionale Académique Adjointe de la Recherche et de l'Innovation en Auvergne-Rhône-Alpes
Ministère chargé de l'Enseignement supérieur et de la Recherche

B

Marie-Astrid BARRAL
Cheffe de projet partenariats et innovation
MIAI Cluster IA

Anne-Thérèse BOURREAU
Data scientist
Inddigo

Christophe BOUVARD
Cofounder & CTO/CPO
Wikit

C

Nathalie CHARBONNIAUD
Chercheuse en IA de confiance
Orange

Stanislas CHESNAIS
CEO and co-founder
Xpdeep

Renaud COLLIN
Responsable des Produits Data-Numériques
Inddigo

D

Antoine DE DECKER
Directeur associé
Inceptive

Vincent DEBUSSCHERE
Maître de Conférences
Grenoble INP UGA

Eric DOMINGUEZ
Public Affairs and Partnership Program Manager
Soitec

E

David EXCOFFIER
Responsable du programme de Recherche Passive Infrastructures & Augmented Interventions
Orange Innovation

F

Philippe FLATRESSE
Product Manager
Soitec

Raphael FRISCH
CEO & Cofounder
HawAI.tech

G

Débora GALLÉE
Marketing Communication Manager
Neo

Michel GASTALDO
Directeur
Naver Labs Europe

Thomas GILLOT
Business Developer
Asygn

Christophe GIRARD
Strategic Office Technology Director
Soitec

H

Vincent HUARD
Fondateur et PDG
MountAIn SAS

I

Céline ISSARD-GUILLOT
Responsable filière numérique, IA, référente French Tech & solutions Industrie du Futur
DREETS Auvergne-Rhône-Alpes

J

Ilhem JOULALI
Avocate Associée / Droit commercial Droit du numérique & IA
Cabinet JOULALI

L

Hubert LÈBRE
CEO
ExcellerIA

M

Laurent MALNOË
Chargé de mission Innovation technologique Service Recherche, Technologie, Innovation Direction de l'Enseignement supérieur, de la recherche et de l'innovation
Conseil régional Auvergne-Rhône-Alpes

Rodolphe MARBOT
Machine Learning Engineer et Data Scientist
ExcellerIA

Sandra MATHIEU
DAF
ADOBIS Group

Sébastien MONNET
Professeur
Université Savoie Mont Blanc

N

Lucas NACSA
Président & co-fondateur
Neovision

P

Rémi PACCOU
Director of Sustainability Research
Schneider Electric

Adrien PARÉ
Responsable Innovation Inno'Lab Energie
Atos

Matthieu PETIT GUILLAUME
Chief Innovation & Research Officer
Leviatan - Laboratoire de recherche appliquée en IA

Sophie PHOMSOUVANDARA
Chargée des Relations Partenariales
ExcellerIA

R

Adélie RANVILLE
Chercheuse post-doctorante
Grenoble Ecole de Management

Marina REYBOZ
Directrice de recherche
CEA

Grégory ROJAS
Chef de projet nouvelles pédagogies
Université Grenoble Alpes

S

Laurent SAROUL
Directeur scientifique
Probayes

Lisa SCANU
Responsable d'équipe
Probayes

Pierre SEDILLE
Ingénieur Qualité Environnementale du Bâtiment
Inddigo

T

Mathis TAILLAND
Ingénieur de recherche
Inria

Madigan TRAFFEY
Consultant IA
Inceptive

V

Brice VARINI
Expert technique
Atos

Anne-Lorène VERNAY
Professeure Associée, cheffe de l'équipe de recherche en management de l'énergie et de l'environnement et co-directrice du Msc Energy Business and Climate Strategy
Grenoble Ecole de Management

Adrien VIALLETELLE
AI Business Unit Director
Asygn

Références

[1] https://larevueia.fr/quest-ce-que-le-rlhf-rl-from-human-feedback/

[2] https://arxiv.org/pdf/2504.12501

[3] https://arxiv.org/pdf/2204.05862

[4] https://assets.publishing.service.gov.uk/media/6716673b-96def6d27a4c9b24/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf

[5] https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

[6] https://www.inria.fr/fr/llm4code

[7] https://www.un.org/fr/climatechange/___paris-agreement

[8] United Nation. (2025). For a livable climate: Net-zero commitments must be backed by credible action. Climate Action. https://www.un.org/en/climatechange/net-zero-coalition

[9] Hilty, L.M., Aebischer, B.: ICT Innovations for Sustainability. Advances in Intelligent

[10] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[11] https://www.similarweb.com/top-websites/

[12] https://www.technologyreview.com/2025/03/04/1112768/inside-the-wild-west-of-ai-companionship/

[13] https://www.lemonde.fr/en/economy/article/2025/01/22/stargate-trump-s-500-billion-project-to-boost-artificial-intelligence_6737299_19.html

[14] G. Elimian. Chatgpt costs $700,000 to run daily, openai may go bankrupt in 2024, 2023. https://technext24.com/2023/08/14/chatgptcosts-700000-daily-openai/

[15] https://web-assets.bcg.com/0b/f6/c2880f9f4472955538567a5bcb6a/ai-radar-2025-slideshow-jan-2025-r.pdf

[16] https://www.researchgate.net/publication/390920260_Frugal_AI_Introduction_Concepts_Development_and_Open_Questions

[17] https://arxiv.org/pdf/2307.09288

[18] https://www.cell.com/joule/fulltext/S2542-4351(23)00365-3

[19] https://www.linkedin.com/posts/octave-klaba-3a0b3632_depuis-quelques-jours-on-entend-beaucoup-activity-7294997565928861698-SEJ5?utm_source=share&utm_medium=member_

[20] https://www.linkedin.com/posts/octave-klaba-3a0b3632_milliards-deuros-en-cascade-giga-watts-activity-7295426699184226305-vld6?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAL209UB67wo2gWZ2DFWcw2DWmkaBqTPpos

[21] https://www.linkedin.com/posts/octave-klaba-3a0b3632_pour-mesurer-limpact-co2-de-lai-et-donc-activity-7295714003379474436-LagY?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAL209UB67wo2gWZ2DFWcw2DWmkaBqTPpos

[22] https://www.lecho.be/opinions/general/opinion-le-systeme-educatif-tarde-a-integrer-l-intelligence-artificielle-dans-ses-enseignements/10567161.html

[23] https://pedagogie.ac-montpellier.fr/aide-la-differenciation-pedagogique-grace-lia

[24] https://eduscol.education.fr/4188/les-intelligences-artificielles-et-leurs-usages-en-education

[25] https://intelligence-artificielle.developpez.com/actu/373740/Une-etude-revele-que-les-outils-d-IA-de-codage-ralentissent-les-developpeurs-tout-en-leur-donnant-l-illusion-d-etre-plus-rapides-ils-ont-mis-19-pourcent-plus-de-temps-a-accomplir-les-taches-de-codage

[26] https://theshiftproject.org/article/pour-une-sobriete-numerique-rapport-shift/

[27] https://www.google.com/url?sa=t&source=web&rct=-j&opi=89978449&url=https://codde.fr/wp-content/uploads/2022/09/APR-PERFECTO-2019-Rapport-final-Nega-Octet_vfinal_public.pdf

[28] https://www.green-algorithms.org/GAapp-overview/

[29] https://librairie.ademe.fr/ged/9417/ADEME-IT4Green-evaluation-env-numerique_Etape_0.pdf

[30] https://base-empreinte.ademe.fr/empreinte-projet

[31] https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=15030&lang=fr

[32] https://www.greendigitalcoalition.eu/assets/uploads/2024/04/EGDC-Net-Carbon-Impact-Assessment-Methodology-for-ICT-Solutions.pdf

[33] https://www.carbone4.com/publication-nzi-it

[34] https://telechargement.afnor.info/normalisation-livre-blanc-technologies-numeriques-transition-ecologique

[35] D. Bol, S. Boyd and D. Dornfeld, « Application-aware LCA of semiconductors: life-cycle energy of microprocessors from high-performance 32nm CPU to ultra-low-power 130nm MCU », in Proc. IEEE ISSST, 2011

[36] L. Eeckhout, "FOCAL: A First-Order Carbon Model to Assess Processor Sustainability", ASPLOS 2024.

[37] https://www.afnor.org/actualites/referentiel-pour-mesurer-et-reduire-impact-environnemental-de-ia/

[38] https://theshiftproject.org/wp-content/uploads/2025/03/Rapport-intermedaire-IA-VF.pdf

[39] https://www.linkedin.com/pulse/un-cerveau-humain-consomme-30-w-une-requ %C3 %AAte-sur-llm-des-cirot-teau-1ahue/

[40] Electricity – Analysis and forecast to 2026 IEA - 2024

[41] IA Environnement - CESE 2024

[42] AI Index Report 2023 - HAI Stanford University

[43] The AI Risk Repository - Whittlestone et al - 2024

[44] Evaluation de l'impact environnemental du numérique en France - ADEME – Arcep – 2023

[45] Les promesses de l'IA grevées par un lourd bilan carbone - https://www.lemonde.fr/

[46] ChatGPT consommerait l'équivalent d'une bouteille d'eau par conversation - Novethic - https://www.novethic.fr/

[47] L'intelligence artificielle, une "bombe climatique" invisible - Novethic - https://www.novethic.fr/

[48] https://www.epri.com/research/products/000000003002028905

[49] https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/artificial-intelligence-electricity-system-dynamics-approach/

[50] https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/ai-powered-hvac-in-educational-buildings/

[51] https://arxiv.org/pdf/1905.01392

[52] https://arxiv.org/pdf/1503.02531

[53] https://arxiv.org/pdf/2011.00241

[54] https://www.researchgate.net/publication/221618539_Optimal_Brain_Damage

[55] https://arxiv.org/html/2411.06084v1#S6

[56] https://www.cs.princeton.edu/courses/archive/spr08/cos598B/Readings/Fukushima1980.pdf

[57] https://link.springer.com/article/10.1007/s11063-021-10562-2

[58] https://arxiv.org/html/2409.08290v1

[59] https://github.com/google-research/tuning_playbook?tab=readme-ov-file#a-scientific-approach-to-improving-model-performance

[60] https://transformer-circuits.pub/2023/toy-double-descent/index.html

[61] https://pair.withgoogle.com/explorables/grokking/

[62] Superposition, Memorization, and Double Descent, Heninghan et al, Anthropic, 2023 (https://transformer-circuits.pub/2023/toy-double-descent/index.html)

[63] https://www.tomra.com/fr-fr/waste-metal-recycling/media-center/news/2025/a-short-history-of-ai

[64] https://recycleye.com/fr/lia-et-la-reconnaissance-des-dechets-pourquoi-cela-fonctionne-si-bien-2/

[65] https://www.3recycleo.fr/2022/02/21/la-technologie-de-lia-pour-optimiser-la-gestion-des-dechets/

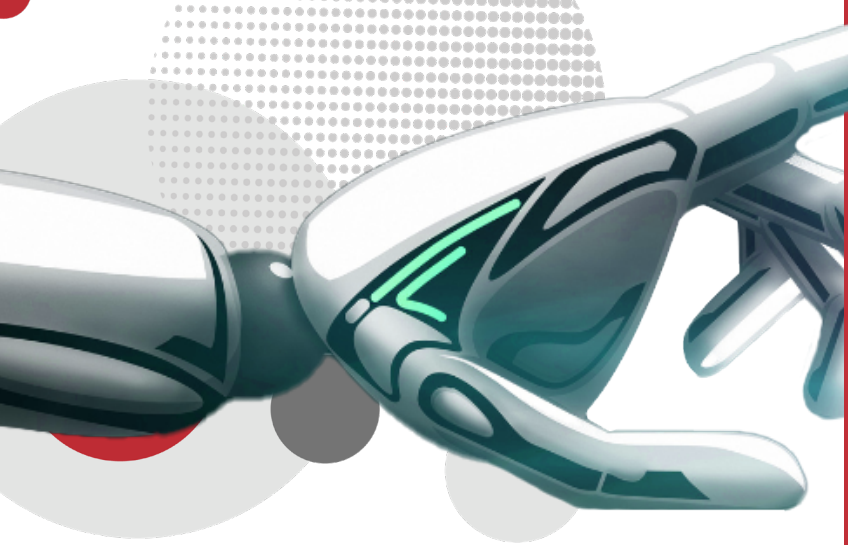
[66] https://doubleverify.com/ai-crawlers-and-scrapers-are-contributing-to-an-increase-in-general-invalid-traffic/

[67] https://doubleverify.com/ai-crawlers-and-scrapers-are-contributing-to-an-increase-in-general-invalid-traffic/

[68] https://www.inria.fr/sites/default/files/2025-03/PositionPaper_IA_environnement_Inria_FR.pdf

[69] Doc. COM (2021) 206 final, 2021/0106 COD

[70] Règlement 2024/1689 du Parlement européen et du Conseil du 13 mai 2024



MINALOGIC
Auvergne-Rhône-Alpes

Nos partenaires publics

les pôles de compétitivité
RÉPUBLIQUE FRANÇAISE
AFPC
Commission européenne
MINISTÈRE DE L'ÉCONOMIE, DES FINANCES ET DE LA SOUVERAINETÉ INDUSTRIELLE ET NUMÉRIQUE
MINISTÈRE DES ARMÉES
La Région Auvergne-Rhône-Alpes
GRENOBLE ALPES MÉTROPOLE
MÉTROPOLE GRAND LYON
SAINT-ÉTIENNE la métropole
Grand Annecy
Communauté du Pays Voironnais
valence romans AGGLO
Le GRÉSIVAUDAN communauté de communes

Nos partenaires privés

abgi part of visiativ absiskey
CIC
Entreprise DU FUTUR
SWARM Innovation & transformation center
visiativ