Comment adopter une démarche green AI en entreprise ?

La boîte à outils d'Impact AI

Impact AI est le Think & Do Tank de référence pour l'intelligence artificielle éthique en France. Notre vision est de travailler ensemble avec l'écosystème numérique (entreprises, startups, institutions, organismes de recherche ou de formation, acteurs de la société civile…) pour créer une approche de l'IA collective qui répond aux besoins et aux attentes des citoyens. Notre Taskforce dédiée à l'environnement s'attache à étudier, approfondir et donner des clés de compréhension et d'applications concrètes autour des deux concepts clés à la croisée de l'IA et de l'environnement : Green AI & AI for Green.
Nous vous souhaitons une bonne lecture de notre papier dédié au Green AI, qui vous aidera à en comprendre les grands enjeux et démêler le vrai du faux, mais également à initier opérationnellement une démarche Green AI ambitieuse et à l'état de l'art grâce au partage de bonnes pratiques, outils et méthodologies[1].

Contributeurs principaux : Axionable, DC Brain, Intel, Microsoft, Orange – membres de la taskforce Environnement d'Impact AI

SOMMAIRE

[1] *(note de bas de page)* : Le sujet du Green AI évolue constamment et de nouvelles méthodes pourraient voir le jour à très court terme après la sortie de ce papier ; nous invitons ainsi nos lecteurs à continuer de s'informer régulièrement sur les pratiques les plus récentes.

0/ Introduction

L'été 2022 en a été le triste témoin : la multiplication des événements météorologiques extrêmes (dômes de chaleur, inondations, incendies, sécheresse, ouragans...) soulignée par les différents rapports du GIEC, nous confrontent chaque jour au dérèglement climatique provoqué par les activités humaines qui émettent des gaz à effet de serre dans l'atmosphère. L'augmentation de la température sur la Terre dépend, en partie, de la quantité totale de carbone présente dans l'atmosphère, et non pas de la vitesse à laquelle nous émettons. Pour limiter la hausse de la température à 1.5 degré en 2050, nous devons en priorité cesser d'ajouter du carbone dans l'atmosphère, puis atteindre une neutralité sur les émissions résiduelles. Ce qui signifie que pour chaque gramme de carbone que nous émettons, nous devons retirer un gramme : la masse globale de carbone dans l'atmosphère restant ainsi fixe.

Dans ce contexte de changement climatique, le numérique apparaît pour l'environnement comme un « *Pharmakon* » qui en Grec désigne à la fois le remède et le poison.
D'un côté, le numérique représente actuellement selon l'ADEME 3,5 % des émissions de gaz à effet de serre (GES); et la forte augmentation des usages laisse présager un doublement d'ici 2025 si rien n'est fait pour en limiter l'impact.
De l'autre, le numérique est indispensable à la transformation des organisations et peut jouer un rôle positif dans la diminution des émissions GES engendrées par les autres secteurs (énergie, transport, chauffage, industrie…).
Dans l'objectif de réconcilier la transition écologique et la transition numérique, il faut mettre la durabilité au cœur de la technologie (Green IT) et utiliser la technologie au service de la durabilité (IT for Green). Le cas particulier de l'usage de l'Intelligence Artificielle (IA) est un parfait exemple de cette injonction contradictoire.
L'intelligence artificielle présente des cas d'usages ayant des impacts positifs sur l'environnement (AI for Green). Leur impact peut être mesuré par leur contribution aux 17 objectifs de développement durable des Nations Unies, englobant les résultats sociétaux, économiques et environnementaux : l'IA pourrait contribuer à l'atteinte de 134 cibles mais aurait un effet négatif en inhibant 59 cibles (étude de 2020 publiée dans Nature Communications). Le programme AI for Earth ou l'initiative Climate Change AI sont des exemples d'initiatives faisant émerger des cas d'usage de l'IA au service d'un développement durable.
De l'autre côté, l'IA a un fort impact carbone en constante augmentation et contribue donc au changement climatique. En effet, la disponibilité des données et des capacités de calcul a entraîné une course à la performance (Red AI) et une forte augmentation des coûts de calcul. On constate que la relation entre la performance et la complexité du modèle (mesurée en nombre de paramètres ou en temps d'inférence) est au mieux logarithmique : pour un gain linéaire en performance, la complexité du modèle croît exponentiellement. Par exemple, un entraînement du modèle émet 50% de son CO2 uniquement pour atteindre une diminution finale de 0,3 du taux d'erreur de reconnaissance des mots (« The Energy and Carbon Footprint of Training End-to-End Speech Recognizers ») ou encore GPT-3, un modèle de langage puissant (175 milliards de paramètres) et récent d'OpenAI, aurait consommé suffisamment d'énergie à l'entraînement pour laisser une empreinte carbone équivalente à la conduite d'une voiture pour un aller-retour de la Terre à la Lune.
La prise de conscience et l'urgence du changement climatique a permis la structuration du mouvement "Green AI", initié par des chercheurs en traitement du langage naturel, proposant un compromis entre la précision du modèle et son coût carbone. Certaines conférences (NeurIPS 2019, EMNLP 2020, SustaiNLP2020) exigent maintenant les coûts de calcul nécessaires à la génération des résultats proposés dans toutes les soumissions.
La citation de Peter Drucker "You can't manage what you can't measure", s'applique bien sûr à une IA durable pour laquelle il est impératif de pouvoir 1/ comprendre et appréhender le sujet dans son ensemble, 2/ mesurer les émissions de gaz à effet de serre et autres impacts environnementaux puis 3/ piloter et réduire ces impacts.
Le collectif Impact AI vous propose ci-dessous une boîte à outils autour des trois volets 1/ comprendre 2/ mesurer 3/ réduire, dans le but de vous aider à structurer et initier votre démarche Green AI et accélérer sa mise en œuvre opérationnelle.
L'exemplarité d'une démarche Green AI est clé : des méthodologies robustes et outils existent, mais il est nécessaire de bien les comprendre et les utiliser afin d'adopter une approche

globale, ambitieuse et cohérente. A travers ce papier, nous souhaitons ainsi contribuer à éveiller les consciences et faire monter la maturité collective autour de la compréhension, de la mesure et de la réduction des impacts environnementaux de l'IA.

Gwendal Bihan, CEO d'Axionable, Vice-Président d'Impact AI et leader de la taskforce environnement

1/ Comprendre

La notion de "Green AI" est généralement associée à l'empreinte carbone liée à la consommation d'énergie de l'IA et notamment aux phases d'entraînement des réseaux de neurones, parfois très complexes tels que GPT3. Or, l'impact environnemental de l'IA ne se réduit pas uniquement à cela : la notion de Green AI doit être prise en compte de manière plus globale lorsque l'on souhaite adopter une approche holistique.

La nécessité d'inscrire la démarche Green AI dans une démarche globale Green IT

L'impact spécifique de l'IA est encore difficile à estimer ou isoler par rapport à l'ensemble de l'empreinte carbone du numérique, du fait de l'apparition relativement récente du sujet (voir les premiers textes fondateurs en introduction et bibliographie de ce papier) et l'absence de normes et référentiels établis autour du Green AI. Les bornes de l'IA sont ainsi à définir au sein d'une organisation qui souhaite mener une démarche de Green AI : quelle part de l'impact peut être attribuée spécifiquement à l'IA par rapport à l'impact global du numérique au sein de mon organisation ? La limite n'est pas simple à établir, il est ainsi important de ne pas occulter une partie des émissions de gaz à effet de serre liée à l'IA et au contraire d'adopter une vision holistique et complète des impacts carbone sur toute la chaîne de valeur. Il apparaît ainsi nécessaire d'inscrire la démarche de Green AI dans une démarche plus globale autour du green IT.

Aussi, il est important de conserver à l'esprit les ordres de grandeur de l'impact carbone liés au numérique : une récente étude de l'Arcep a montré que parmi les équipements numériques, les terminaux représentent la majorité de l'empreinte carbone (79%) suivi par les centres de données (16%) et les réseaux (5%). Également, la majorité de l'empreinte carbone des équipements est émise lors de la fabrication des équipements (78%) par rapport à leur utilisation (21%) (source). Ainsi la majeure partie des émissions de GES liées au numérique se situent au niveau de la fabrication, l'acheminement et la fin de vie des terminaux et équipements. Ces ordres de grandeur permettent de souligner la nécessité d'avoir une démarche globale et cohérente entre Green AI et Green IT.

La comptabilisation des émissions de gaz à effet de serre sur l'ensemble des scopes 1, 2 et 3

L'empreinte carbone de l'IA est l'indicateur communément utilisé pour initier et piloter une démarche Green AI et mettre en place et mesurer des actions d'amélioration. Lorsque l'on

parle d'empreinte carbone, il est absolument nécessaire de prendre en compte l'ensemble des 3 scopes d'émissions de GES :

- les émissions de scope 1 recouvrent les émissions directes liées aux consommations de gaz, fioul ou encore des fuites de fluides frigorigènes, présents dans les circuits de refroidissement et climatisation des data centers notamment ;
- les émissions de scope 2 recouvrent les émissions indirectes liées à l'énergie, c'est-à-dire la production et consommation d'électricité et de vapeur (chaud / froid) ;
- les émissions de scope 3 recouvrent l'ensemble des autres émissions indirectes, dont les principaux postes se concentrent en général autour de :
- la fabrication, l'acheminement et la fin de vie des équipements informatiques liés à l'entraînement et la mise en production de l'IA et des équipements edge sur lesquels l'IA est déployée ;
- les achats des services et prestations techniques et informatiques dédiées aux projets d'IA (licence logicielle, infogérance, etc.)
- l'utilisation des produits / services visés par le projet d'IA.

L'analyse de l'empreinte carbone à toutes les étapes du cycle de vie de l'IA

Lorsque l'on parle de Green AI, on pense souvent en premier lieu à l'empreinte carbone liée à la phase d'entraînement de l'IA ; des études permettent d'ailleurs d'obtenir des premières évaluations de cette empreinte carbone, on estime par exemple à 85 tonnes de CO2e l'entraînement de GPT-3 (source).

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Cependant, la mesure de l'empreinte carbone dans le cadre d'une démarche Green AI globale ne doit pas se limiter à la phase d'entraînement, mais doit bien porter sur l'ensemble des étapes du cycle de vie de l'IA, depuis la conception jusqu'à l'inférence. La tendance est à une augmentation forte de l'empreinte carbone de l'IA. En effet, les capacités de calcul et la quantité de données disponibles augmentant, on assiste à une course à la performance se traduisant par une augmentation de la taille et de la complexité des modèles. Or, un modèle complexe nécessite plus de ressources à son entraînement mais aussi en production. Cela a des conséquences à toutes les étapes du cycle de vie de l'IA :

- L'augmentation de la quantité des données utilisées dans les systèmes IA a des impacts sur les infrastructures nécessaires : le stockage des données et le pipeline d'ingestion représentent une part importante de la consommation énergétique.
- L'augmentation de la taille des modèles permet de meilleures performances mais nécessite des ressources pour une mise à l'échelle d'un modèle l'IA qui dépassent clairement ceux du hardware existant.
- La croissance des modèles entraîne une croissance des ressources requises pour l'entraînement (x2.9) et l'inférence (x2.5).

Cela illustre la façon dont les émissions carbone se produisent tout au long du cycle de vie d'un projet d'IA : développement, déploiement, utilisation… Il faut donc considérer l'empreinte des données existantes et exploitées, des algorithmes et des systèmes physiques, depuis la fabrication des matériaux jusqu'à l'utilisation opérationnelle de tous les composants de l'IA pour avoir une vision complète de l'empreinte carbone d'un système d'IA.

Au-delà du carbone, la prise en compte des autres impacts environnementaux

Il est enfin important de souligner que l'empreinte carbone n'est pas le seul impact environnemental généré par l'IA et plus généralement par le numérique. L'étude publiée fin 2021 par NégaOctet et GreenIT (source) sur l'évaluation du cycle de de vie des technologies numériques en Europe a par exemple intégré les impacts liés à l'épuisement des ressources abiotiques naturelles (minéraux, métaux), la consommation de ressources fossiles, les effets sur le changement climatique, la consommation d'eau douce et l'écotoxicité, les émissions de particules, les radiations ionisantes impactant la santé humaine, ou encore la production de déchets électriques et électroniques. L'empreinte carbone est ainsi un indicateur mesurable et largement utilisé pour quantifier l'impact environnemental d'une activité, il permet de suivre les évolutions et les impacts des actions. C'est cet indicateur que nous avons choisi de retenir pour la suite de l'article autour de la mesure et l'identification des actions de réduction.

Cinq points clés à retenir pour comprendre la notion Green AI et adopter une approche holistique :

- il est nécessaire de prendre en compte l'impact de l'ensemble des 3 scopes d'émissions de GES et ne pas se limiter à l'impact de la consommation d'électricité ;

- il est nécessaire de prendre en compte l'ensemble du cycle de vie de l'IA, depuis l'idéation et la conception jusqu'à l'inférence, en passant par l'entraînement et la mise en production des modèles ;
- il est nécessaire de considérer l'impact de l'ensemble des infrastructures et services associés au projet d'IA : hébergement, réseaux, équipements, applications et logiciels, terminaux et edge ;
- il est fortement recommandé d'inscrire la démarche Green AI au sein d'une démarche plus globale Green IT, pour assurer une cohérence aux bornes de l'IA, qui sont encore difficiles à définir précisément ;
- bien qu'il s'agisse aujourd'hui de l'indicateur clé du Green AI, le carbone n'est pas le seul impact environnemental de l'IA et d'autres impacts peuvent ainsi être considérés (eau, ressources abiotiques, changement climatique, déchets électriques et électroniques,...)

Sources / Pour aller plus loin :

- https://www.researchgate.net/publication/354088344_Understanding_and_Co-designing_the_Data_Ingestion_Pipeline_for_Industry-Scale_RecSys_Training
  https://cs.stanford.edu/~matei/papers/2020/iclr_svp.pdf
- https://arxiv.org/pdf/2101.11714.pdf
- https://www.impact-ai.fr/app/uploads/2022/03/IMPACT-AI-FICHES-PRATIQUES-NUMERIQUES.pdf?utm_source=mailchimp&utm_campaign=0300fd4ce0f0&utm_medium=page
  https://www.arcep.fr/uploads/tx_gspublication/etude-numerique-environnement-ademe-arcep-note-synthese_janv2022.pdf
- https://ecoresponsable.numerique.gouv.fr/publications/guide-pratique-achats-numeriques-responsables/
- https://www.greenit.fr/le-numerique-en-europe-une-approche-des-impacts-environnementaux-par-lanalyse-du-cycle-de-vie/

2/ Mesurer :

L'exercice de mesure de l'empreinte carbone de l'IA n'est pas nécessairement un exercice simple :
- il n'existe pas de référentiel ou méthodologie spécifique et reconnue pour mesurer précisément l'empreinte carbone des activités numériques, y compris sur le sous-périmètre de l'IA ;
- les référentiels carbone "traditionnels" sont peu adaptés à l'évaluation carbone des activités numériques ;
- l'ensemble des données d'activité permettant de calculer les émissions de GES ne sont pas toujours disponibles, notamment lorsque celles-ci dépendent de prestataires et partenaires externes ;
- des solutions de mesure commencent à émerger sur le marché mais n'ont pas toutes les mêmes objectifs, la même méthodologie ou encore le même périmètre couvert ;

- des écarts méthodologiques sont observés entre les différents acteurs et référentiels

Nous vous proposons ainsi ci-dessous un premier panorama de méthodologies et ressources disponibles afin de vous aider à travailler sur la mesure de l'empreinte carbone de l'IA.

Mesure des émissions carbone liées à la consommation d'énergie des modèles IA

L'énergie de fonctionnement est l'énergie utilisée lors du fonctionnement de l'IA, que ce soit pour l'entraînement des modèles ou pour les modèles en production. On peut la mesurer à trois niveaux : mesure a priori, mesure a posteriori et mesure à la volée.

Lors de la mesure de l'empreinte carbone associée à la consommation d'énergie des modèles IA, il est important de prendre en compte l'énergie consommée durant l'entraînement du modèle mais aussi l'inférence une fois le modèle en production. En effet, cette énergie n'est pas négligeable et pourrait représenter 80 à 90% de l'empreinte carbone d'un réseau de neurone (source). Nous avons distingué trois grandes méthodes permettant de mesurer la consommation d'énergie des modèles IA : mesure a priori, mesure a posteriori et mesure à la volée.

Mesure a priori

La mesure a priori se base sur l'estimation du nombre d'opérations à réaliser par l'ordinateur lors de l'exécution du code. Ce nombre d'opérations se calcule en général en FLOPS (floating point operations per second), ce qui correspond au nombre d'opérations primaires. Il est ensuite converti en énergie en fonction des caractéristiques de l'équipement utilisé pour l'exécution puis en équivalent carbone.

Plusieurs librairies sont disponibles en fonction des types de modèles. Par exemple, Keras Flops permet d'estimer le nombre de Flops sur un modèle Tensor Flow et Keras ou Torchstat qui fait l'équivalent pour un modèle Pytorch.

Pour généraliser cette approche et permettre d'estimer le coût d'un projet complet, on peut utiliser l'approximation suivante (source) :

$$Cost \propto E \,.\, D \,.\, H$$

Avec

- **Exemple** : coût de calcul pour un exemple
- **Dataset** : nombre d'exemples dans le dataset d'entraînement
- **Hyperparamètres** : nombre de combinaison d'hyper paramètres testées (nombre d'expérimentations)

Cette formule n'est qu'une approximation et ne prend pas en compte certains facteurs comme le nombre d'epochs (spécificité des réseaux de neurones).

Mesure a posteriori

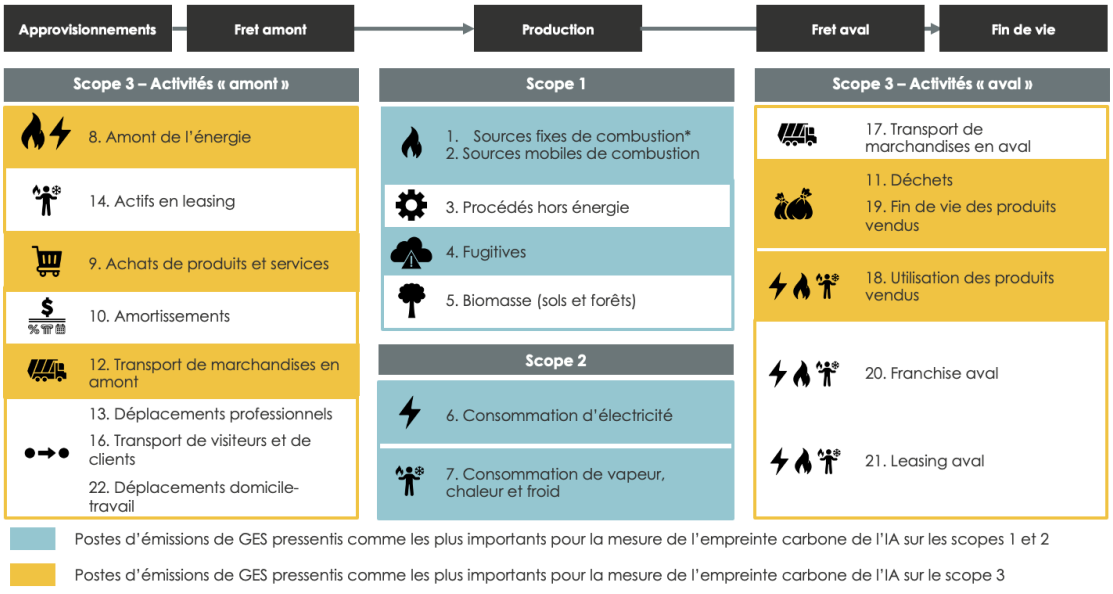
La mesure a posteriori est une estimation de l'empreinte carbone à partir du temps de calcul et de l'infrastructure utilisée. L'outil ML CO2 Impact développé par le Mila (Quebec AI Institute) permet d'estimer l'empreinte carbone d'un entraînement à partir du temps d'entraînement, d'un type d'hardware, d'un cloud provider et de son pays. L'outil fournit aussi la quantité de carbone compensé par le cloud provider ainsi que des conseils d'optimisation de l'infrastructure pour réduire la consommation.

Mesure à la volée

La mesure à la volée consiste à mesurer directement la consommation électrique du processeur pendant l'exécution du code. Pour cela, on peut utiliser la librairie CodeCarbon qui permet, en ajoutant quelques lignes de code, de mesurer la consommation du processeur pendant l'exécution du code.
Une mesure plus générale est effectuée par les cloud providers et peut être suivie sur des dashboards avec un détail plus ou moins important en fonction des cloud. Ainsi par exemple, les ressources collectées par Azure Monitor à partir d'un espace de travail Azure Machine Learning permettent de collecter l'indicateur Énergie par intervalle d'une minute en joules sur un nœud GPU.

Mesure des autres émissions carbone

Comme déjà évoqué, pour mesurer l'impact carbone de l'IA il est important de pouvoir tenir compte du cycle de vie complet qui va donc couvrir l'ensemble des 3 scopes, avec un poids très important des émissions de GES situées au niveau du scope 3, tant sur les activités amont que les activités aval. Nous avons ainsi repris ci-dessous les postes d'émissions de GES proposés par l'Ademe sur les 3 scopes et avons identifié les postes pressentis comme les plus impactants dans le cadre de la mesure de l'empreinte carbone de l'IA :



Une des principales difficultés dans l'évaluation du scope 3 se situe au niveau des équipements et infrastructures utilisées par les projets IA, qui nécessitent la prise en compte des émissions de GES sur l'ensemble du cycle de ces équipements, notamment :

- L'extraction ou le recyclage des matières premières pour fabriquer de nouveaux composants (CPU, GPU, lecteur de disque,…)
- Assemblage des composants dans les serveurs et équipements
- Transport vers le datacenter avec différents modes (avion, bateau, train...)
- Utilisation de ces matériels pour fournir les services
- Décommissionnement (réutilisation ou recyclage…)

La mesure de ces émissions de GES sont fortement dépendantes des données mises à disposition par les fournisseurs et prestataires externes, notamment à travers les analyses de cycle de vie des équipements, y compris lorsque l'on se situe sur des infrastructures on-premise. Or ces analyses ne sont pas systématiquement réalisées et / ou communiquées, d'autant plus lorsque l'on s'intéresse à des infrastructures externalisées ou Cloud.
A noter que certains fournisseurs de Cloud comme Microsoft ou Google sont capables de fournir à leurs clients un tableau de bord d'émission carbone par géographie et spécifique à leur utilisation des services Cloud. Pour le Cloud Microsoft Azure par exemple, la méthode de calcul des émissions de GES sur les scopes 1, 2 et 3 est décomposée en sept étapes et basée sur 2 sous-méthodes principales, l'une pour calculer les émissions carbone au niveau d'une région géographique, l'autre pour les répartir au niveau individuel de chaque client à partir d'une unité d'utilisation. Ces données sont synthétisées dans un tableau de bord permettant de créer différentes vues agrégées des émissions spécifiques et pertinentes à l'usage du client, notamment à l'échelle des services, régions, datacenters et limites de temps spécifiques (voir en Annexe la synthèse de la méthodologie de calcul du Cloud Microsoft Azure, validée par l'Université Stanford en 2018 et expliquée dans un livre blanc paru en 2021).

Limites des mesures

La précision de la mesure de l'empreinte carbone diffère selon le périmètre et les scopes d'émissions de GES couverts, les étapes de cycle de vie de l'IA retenues ou encore la méthode de calcul choisie. Aussi, les informations à disposition ne permettent pas toujours le choix de la méthode la plus précise, cela est très dépendant de la disponibilité, de la qualité et de la fiabilité des données d'activités. Du fait de l'absence de méthodologies reconnues et partagées par tous, on observe également un manque d'homogénéité et de comparabilité entre les données utilisées par les entreprises ou celles fournies par les prestataires et partenaires, à l'instar des méthodologies proposées par les hyperscalers. Cette forte dépendance à des parties prenantes externes pour le calcul de l'empreinte carbone globale de l'IA complexifie et allonge le travail, notamment sur la mesure du scope 3 qui constitue la très grande majorité des émissions de GES.
Ainsi, la mesure de l'empreinte carbone d'un modèle IA constitue un point de départ et de comparaison ; il s'agit d'un exercice évolutif qui a tendance à s'améliorer d'année en année avec la précision des méthodologies et la publication de nouveaux référentiels et papiers / études de référence. Il faut garder à l'esprit que les mesures d'empreinte carbone doivent être interprétées davantage comme des ordres de grandeur que des chiffres parfaitement

exacts et précis, et constituent ainsi une base de référence pour comparer et suivre l'évolution dans le temps.

3/ Réduire

Nous proposons ci-dessous des pistes d'actions concrètes de réduction de l'empreinte carbone de l'IA, tout au long des phases du cycle de vie de l'IA, afin d'initier et opérationnaliser une démarche Green AI.

Les étapes du cycle de vie de l'IA (source : Impact AI)

étape #0 - Introduction

- Une fois que la mesure de l'impact carbone a été réalisée, on dispose d'un point de départ pour la réduire et pouvoir suivre les progrès. **Publier la mesure** permet de sensibiliser la communauté aux impacts environnementaux mais aussi de donner un point de comparaison.
- Dans les bonnes pratiques du cycle de vie d'un produit IA, on retrouve souvent la documentation. Appliquée au Green AI, **documenter la démarche et la justification des choix aux regards des contraintes environnementales** permet de suivre la démarche, d'argumenter ses choix et de diffuser les bonnes pratiques.
- Nous proposons ici un ensemble de mesures permettant la réduction de l'impact carbone des IA tout au long du cycle de vie. Chacun a des avantages et des inconvénients et un impact différent sur la réduction de l'empreinte carbone. Ce sont des pistes de solutions à prendre en compte dans la construction des produits d'IA.

étape #1 - Idéation

- **Sensibiliser et former les salarié.e.s à l'impact carbone de l'IA** : la sensibilisation permet de comprendre les conséquences et de justifier la nécessité de prendre des actions.
- **Intégrer la mesure et la réduction de l'empreinte carbone dans les principes éthiques de l'entreprise**: la réduction de l'empreinte carbone est un processus plus global et s'inscrit dans la stratégie de l'entreprise. Elle sera alors déclinée pour le développement du produit d'IA.

étape #2 - Qualification

- Evaluer la création de valeur du produit d'IA par rapport aux impacts environnementaux :
- S'assurer que l'IA est la bonne solution pour le cas d'usage et qu'une autre solution moins émettrice en carbone et ne nécessitant pas l'usage de l'IA n'existe pas
- Mesurer les impacts environnementaux négatifs : impact carbone du développement et de l'usage de la solution, impact carbone de la solution actuelle

- Évaluer la pertinence du produit : contribution aux objectifs de durabilité, réponse aux besoins utilisateurs, réduction de l'impact environnemental d'un processus
- Fixer des contraintes pour limiter l'impact carbone de l'IA :
- Dimensionner les besoins et les usages en fonction des besoins utilisateurs et respecter les besoins minimums
- Fixer un budget environnemental : budget carbone pour l'entrainement, temps maximal d'inférence, impact carbone de la solution en production
- Définir une performance acceptable minimale au delà de laquelle on arrêtera les optimisations

étape #3 - Développement

- Faire des compromis entre exactitude et impact environnemental, considérer l'efficacité du modèle à la place de l'accuracy classique :

    - Identifier les modèles les plus frugaux permettant de répondre au besoin : faire une analyse d'impact des différentes solutions possibles, limiter l'usage du Deep Learning, favoriser des modèles moins gourmands mais tout aussi performants, favoriser des modèles efficaces en termes de mémoire (exemple). Si le scaling est nécessaire, un petit scaling du modèle et des données sera moins gourmand que le scaling important de l'un ou l'autre.
    - Optimiser efficacement les hyperparamètres du modèles pour limiter le nombre d'entraînements : bayesian search ou random search (pas grid search), réduire l'espace de recherche des hyperparamètres (nombre d'hyperparamètre et valeurs possibles), arrêter l'optimisation lorsque la performance minimale est atteinte, Neural architecture search (NAS) et hyperparameter optimization (HPO)
    - Commencer les expérimentations avec un algorithme simple et augmenter en complexité tout en évaluant le gain de performance par rapport aux ressources nécessaires
    - Optimiser la performance (Throughput (débit = quantité de donnée traitée/seconde en général) et Latency (temps nécessaire à traiter 1 donnée) pour limiter la consommation et permettre un déploiement dans un environnement contraint

- Optimiser l'environnement de développement :

    - Choisir l'infrastructure optimale (on-premise vs cloud vs edge) en adoptant une approche globale pour tenir compte notamment du PUE, de l'alimentation en énergie renouvelables mais aussi de l'effet rebond lié à l'usage du cloud et la redondance utilisée pour assurer la robustesse des services cloud
    - Optimiser les ressources : CPU, GPU, serverless. En effet, au même titre que le software peut être pensé de façon frugale, il faut aussi réfléchir au hardware frugal et donc à l'infrastructure existante pour la partie training. Il faut aussi regarder si on est dans un cas de Machine Learning ou de deep learning. En effet, les processeurs (CPU) sont massivement déjà déployés dans les infrastructure IT (data center, cloud, workstations ..) et excellent dans un large éventail de workload de machine learning, ce qui est le besoin de la plupart des entreprises (versus deep learning). Les CPUs

excellent également dans les tâches à faible latence et permettent de gérer un grand ensemble de données difficiles à diviser ou à sous-échantillonner pour rentrer dans la mémoire des accélérateurs. Là où les workloads l'exigent, il existe une gamme variée d'accélérateurs spécifiques par domaine ou des GPUs.

- Réutiliser ce qui existe déjà :
  - Réutilisation des données pour éviter le stockage multiple et la multiplication des traitements
  - Réutilisation des algorithmes et librairies et optimisation de ceux-ci : les développeurs doivent aussi penser aux cas de la réutilisation du code sur divers type de Hardware à l'edge (sur divers types de machines ou d'environnements contraints comme dans des trains, des voitures, des bornes, des satellites...). Dans ce cas, le code pourrait être amené à être porté et redéployé sur des Hardware différents (par exemple de GPU à CPU à XPU) et il faut penser à utiliser des compilateurs/outil et librairies polyvalents (par exemple : Intel One API), afin d'éviter de devoir re-écrire tout ou partie du code.
  - Réutilisation de modèles déjà entrainés : fine-tuning, Transfer Learning, Incremental Training

- Optimiser le stockage et l'utilisation des données :
  - Définir et mettre en place une politique d'archivage, d'expiration et de suppression des données
  - Limiter et optimiser les traitements de données, enregistrer les données traitées pour les réutiliser sans refaire les traitements
  - Adapter le stockage et le format (format de fichier, compression) des données en fonction de l'usage
  - Échantillonner les données pour réduire leur quantité sans compromettre la performance, la périssabilité des données est un des axes à étudier (source)
  - Penser l'infrastructure IT cible où l'inférence aura lieu (du hardware à l'edge) et à ses contraintes techniques (chaleur, alimentation, encombrement...) :
  - Prendre en compte l'infrastructure IT cible en amont d'un projet ou d'un pilote ; souvent oubliée des développeurs, cela peut empêcher le passage à l'échelle à cause de la consommation énergétique de la solution complète. Si l'infrastructure cible est par exemple un petit boitier dans un train disposant de quelques watts, il faut penser à un Hardware cible peu gourmand en consommation électrique, de type xPU ou CPU.

étape #4 - Mise en production et MCO

- Définir le SLA (Service Level Agreement) en prenant en compte les objectifs de durabilité, choisir les bonnes ressources en fonction des contraintes utilisateurs (temps réel, latence, disponibilité…)

- Optimiser les modèles pour l'inférence (exemple d'outils : Treelite, Hugging Face Infinity, SageMaker Neo…)
- Monitorer le modèle en production et réentrainer uniquement lorsque nécessaire (passage d'un seuil de performance, model ou data drift, nouvelles données…)
- Intégrer le suivi de l'impact environnemental en tant que métrique dans le dashboard de monitoring du modèle en production
- Archiver ou supprimer les fichiers non nécessaires : versions non utilisées du modèle, historique des logs

étape #5 - Usage

- Sensibiliser les utilisateurs aux impacts environnementaux de l'usage du produit
- Veiller à l'utilité, l'utilisabilité et la bonne utilisation des fonctionnalités du produit IA
- Evaluer les opportunités liées à l'émergence de nouvelles technologies
- Interroger la pertinence de mettre fin au produit IA

Sources / pour aller plus loin

- https://aws.amazon.com/fr/blogs/architecture/optimize-ai-ml-workloads-for-sustainability-part-1-identify-business-goals-validate-ml-use-and-process-data/
- https://aws.amazon.com/fr/blogs/architecture/optimize-ai-ml-workloads-for-sustainability-part-2-model-development/
- https://aws.amazon.com/fr/blogs/architecture/optimize-ai-ml-workloads-for-sustainability-part-3-deployment-and-monitoring/
- AFNOR SPEC 2201
- https://www.statworx.com/en/content-hub/blog/how-to-reduce-the-ai-carbon-footprint-as-a-data-scientist/#:~:text=According%20to%20a%20recent%20estimation,2.4%20Gts%20of%20CO2e.
- https://www.techrxiv.org/articles/preprint/The_Carbon_Footprint_of_Machine_Learning_Training_Will_Plateau_Then_Shrink/19139645
- https://www.carbone4.com/analyse-empreinte-carbone-du-cloud
- Discussions Impact AI

Zoom sur la communication autour de la démarche Green AI : quelles précautions à prendre ?

- **Green AI ne signifie pas IA neutre en carbone :** le concept de Green AI ne doit pas être confondu avec le concept d'IA neutre en carbone. La notion de neutralité carbone est à utiliser avec grande précaution, comme le rappelle ici l'Ademe : "L'objectif de neutralité carbone, défini comme l'équilibre arithmétique entre les émissions et séquestrations anthropiques de GES, n'a réellement de sens qu'à l'échelle de la planète. Au travers de l'Accord de Paris, les États s'approprient l'objectif pour permettre une coordination internationale de l'action. La définition de neutralité carbone, telle que décrite ci-avant, ne doit pas s'appliquer à une autre échelle : territoire infranational, organisation (entreprises, associations, collectivités, etc.), produit ou service". Ainsi, le collectif Impact AI recommande fortement de ne pas utiliser la notion d'"IA neutre en carbone" ou de "neutralité carbone de l'IA" dans le cadre des démarches Green AI. Les éventuelles démarches de compensation mise en œuvre peuvent être présentées indépendamment et séparément de la mesure de l'empreinte carbone totale de l'IA, et ne peuvent en aucun cas être "déduites" du total des émissions calculées dans le cadre de la mesure d'empreinte carbone globale de l'IA.
- **Transparence autour des objectifs d'une démarche green AI pour éviter le greenwashing :** il est important de communiquer de manière transparente sur les objectifs et sous-jacents d'une démarche green AI pour éviter les écueils du greenwashing, en rappelant qu'une telle démarche contribue certes à l'efficacité environnementale mais également à l'efficacité opérationnelle, et doit permettre de répondre à des exigences de performance globale, à la fois technique, pratique, économique et environnementale.

Annexe - Microsoft Azure : synthèse de la méthodologie de calcul des émissions de GES

Étape 1 : Calculer les émissions pour les composants et le matériel

- Les émissions en amont incluent les étapes du cycle de vie pour la fabrication au niveau des composants, l'entreposage et le transport vers le dock de déchargement du centre de données Microsoft
- Les émissions d'élimination du matériel en aval inclut les facteurs d'élimination, de traitement et de transport pour chaque composant matériel.

Les émissions de GES du matériel sont calculées en utilisant les facteurs d'émission pour les différents composants (lecteur de disque, FPGA, server blades, les racks, unités d'alimentation,...).

Étape 2 : Calculer les émissions du centre de données pour un mois donné

Il faut croiser les émissions du cycle de vie au niveau matériel de l'étape 1 avec les bases de données d'approvisionnement en matériel pour le matériel de centre de données en tenant compte d'une durée de vie moyenne (par exemple six ans), si la durée de vie réelle de l'équipement s'étend au-delà de cela, ses émissions scope 3 pour la durée prolongée seront nulles. Si sa durée de vie réelle est plus courte, les émissions continueront d'être comptabilisées pour l'estimation sur six ans afin d'assurer une comptabilisation complète.

Étape 3 : Calculer les émissions de la région du centre de données

Les émissions des centres de données individuels sont agrégées au niveau régional. (par exemple une région composée de 3 zones de disponibilité va agréger les émissions de ces 3 datacenters.

Étape 4 : Calculer les émissions liées à l'utilisation spécifique à l'échelle d'un client

Calcul de l'utilisation totale des services cloud par les clients sur la base d'une mesure de mesure de coût normalisée associée aux services IaaS/PaaS/SaaS. L'utilisation totale des clients inclut à la fois l'utilisation directe des ressources par le client et une quantité proportionnelle de capacité de serveur de surcharge dédiée à la fourniture de services cloud.

Étape 5 : Calculer les facteurs d'émission spécifiques à la région

Nous divisons les émissions totales de la région par l'utilisation totale des clients dans cette région. Le résultat est un facteur d'émission spécifique à la région par unité d'utilisation client pour une période donnée.

Étape 6 : Calculer les émissions totales propres au client

Pour quantifier les émissions spécifiques au client, nous multiplions l'utilisation individuelle et mesurée des services d'un client par les facteurs d'émissions spécifiques à la région calculée à l'étape 5.

Étape 7 : Combiner les données et les résumer

Dans cette étape, les clients cloud peuvent utiliser le tableau de bord d’émissions, créer différentes vues agrégées des émissions pertinentes pour des services, des régions, des datacenters et des limites de temps spécifiques.