Partie 4

Enjeux environnementaux

Introduction

Dans ce chapitre, nous allons traiter des enjeux environnementaux liés à l'IA générative. Pour rappel, il est nécessaire de considérablement diminuer les émissions de gaz à effet de serre pour maintenir l'augmentation des températures moyennes à + 1,5 °C par rapport à l'ère préindustrielle. Chaque secteur d'activité doit effectuer sa transition pour atteindre une réduction globale des émissions de gaz à effet de serre. Cela comprend donc le secteur de l'Intelligence Artificielle, ou plus largement, le secteur du numérique ICT (*Information and Communication Technology*).

En 2020, on estimait que les émissions de gaz à effet de serre du secteur du numérique étaient responsables de 2,1 à 3,9 % des émissions mondiales. Ce secteur est toujours en forte croissance et l'IA encore plus particulièrement. En 2025, ces émissions auront doublé et constitueront 4 à 8 % des émissions mondiales (Freitag, et al. [146]). Par ailleurs, l'impact du numérique est préoccupant, non seulement du point de vue des émissions de gaz à effet de serre, mais aussi en termes de consommation énergétique, d'utilisation de l'eau, des ressources minérale et métallique, et de réduction de la biodiversité.

Le domaine de l'IA doit également faire sa part dans cette transition environnementale et réduire son impact. Malheureusement, les récents travaux et projets en IA générative ne semblent pas tous prendre cette direction. C'est pourquoi dans ce chapitre, nous traiterons de l'impact environnemental entraîné par cette nouvelle technologie. Comment évaluer cet impact, et sur quoi se baser ? Quelles pourraient être les éventuelles pistes de réduction de ces impacts, aussi bien au niveau technique que pratique ? Quels seraient les cas d'usages positifs (*for good*) qui pourraient encourager l'utilisation d'une telle technologie afin de construire un avenir plus responsable et durable ?

Matérialité du numérique et de l'IA

Les services numériques, dont l'IA générative fait partie, reposent systématiquement sur des équipements informatiques, et ces derniers consomment de l'énergie pour leur fonctionnement. L'impact environnemental de l'IA générative est souvent associé à sa consommation énergétique. Il ne faut pas oublier pour autant que la production d'équipements informatiques a des impacts sur l'environnement au moins aussi importants que la consommation énergétique.

Consommation d'électricité

Les programmes offrant des services d'IA générative sont, dans la grande majorité des cas, hébergés sur des serveurs installés dans des *data centers* localisés partout dans le monde. Pour fonctionner, les serveurs consomment de l'électricité, qui, n'étant pas une énergie primaire, provient nécessairement d'une centrale de production d'électricité.

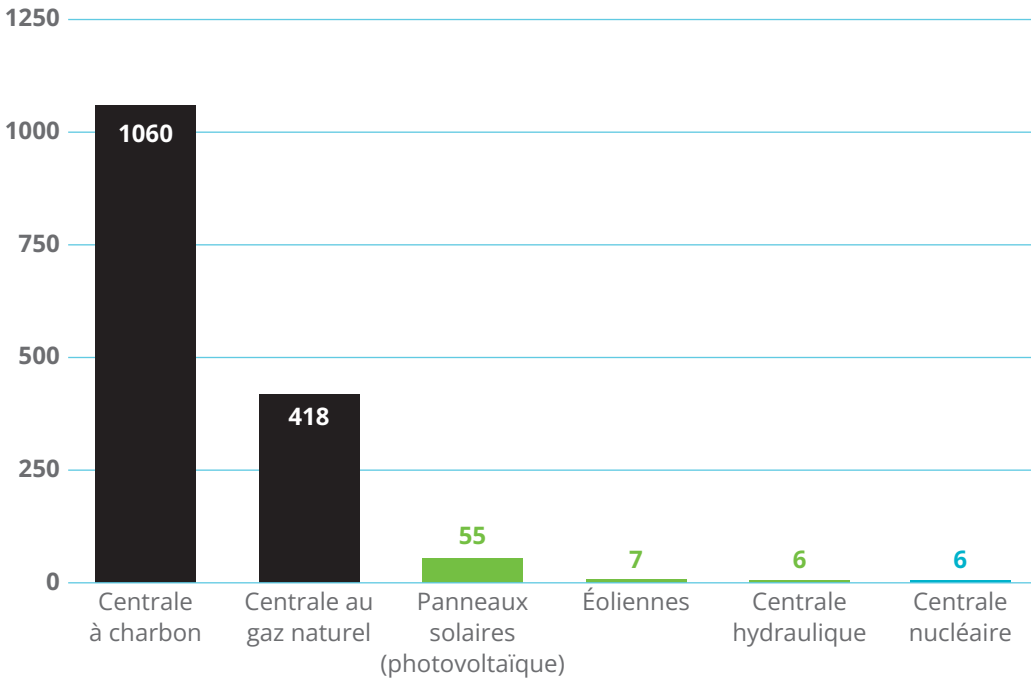
La production d'électricité a un impact sur l'environnement qui dépend du moyen de production. On peut comparer les impacts sur l'environnement provoqués par différents modes de production en considérant les facteurs d'impact associés. Le plus

[146] https://doi.org/10.1016/j.patter.2021.100340

connu et le plus utilisé dans ce domaine est l'intensité carbone, qui mesure la quantité de gaz à effet de serre rejetés dans l'atmosphère. Elle s'exprime en « équivalent CO_2 » (CO_2eq), c'est-à-dire la quantité de gaz à effet de serre émis rapportée à la participation au réchauffement climatique globale du CO_2. En effet, les gaz à effet de serre englobent non seulement le CO_2, mais aussi d'autres gaz tels que le méthane ou le protoxyde d'azote. Bien que l'intensité carbone soit le facteur d'impact le plus regardé, il en existe bien d'autres permettant de quantifier d'autres types d'impacts, tels que la consommation d'eau, l'extraction de ressources minérales, ou bien l'impact sur la biodiversité.

Toutes les centrales ne sont pas égales du point de vue de leur impact sur l'environnement. En effet, les centrales à combustible fossile (charbon, gaz naturel, fioul) émettent beaucoup plus de gaz à effet de serre que les centrales à énergie renouvelable (solaire, éolien, hydraulique) ou encore les centrales nucléaires.

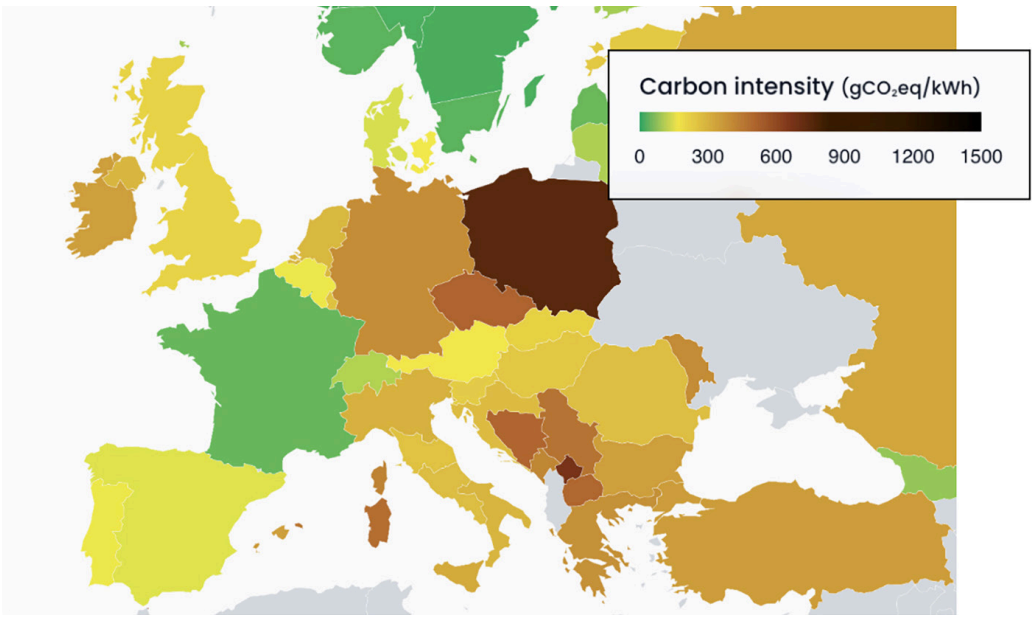
Émissions de gaz à effet de serre par type de centrale (gCO_2 / kWh)



Base Empreinte ADEME [147].

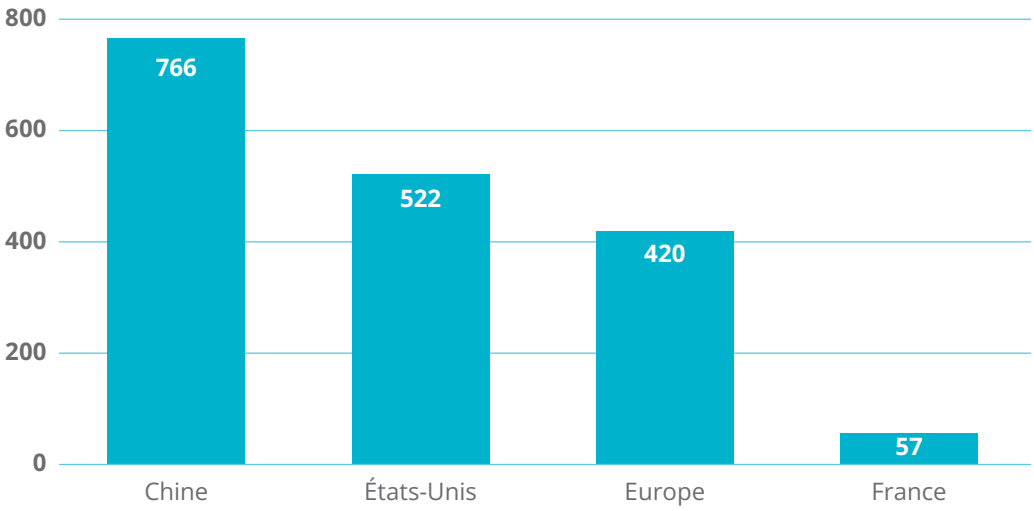
Au sein d'un même pays, plusieurs types de centrales électriques sont installés afin de répondre à la demande d'électricité. En particulier, les centrales à combustible fossile peuvent être utilisées pour absorber les creux de production des énergies renouvelables (lorsqu'il n'y pas de vent ou de soleil). À l'échelle d'un pays, on parle de mix électrique pour quantifier la part de production de chaque type de centrale. En France, où il y a une majorité de centrales nucléaires, l'intensité carbone du mix électrique reste faible comparé à d'autres pays.

[147] https://base-empreinte.ademe.fr/donnees/jeu-donnees



ElectricityMaps.org [148].

Intensité carbone du mix électrique par pays ($gCO_2eq./$ kWh)



Base Empreinte ADEME [147].

L'impact environnemental de l'IA générative est dépendant des impacts du mode de production de l'électricité. À titre d'exemple, un même modèle d'IA peut émettre presque 10 fois plus de gaz à effet de serre lorsqu'il s'exécute aux États-Unis plutôt qu'en France. Ce facteur multiplicatif est d'autant plus important que l'intensité en calcul de l'IA générative peut être très élevée en phase d'entraînement.

Cycle de vie des équipements numériques

La consommation électrique des serveurs accueillant des modèles d'IA générative n'est pas le seul facteur qui entre en compte dans l'évaluation des impacts sur l'environnement. Le cycle de vie des équipements numériques comprend en amont l'extraction des matières premières, la fabrication et la distribution des serveurs et de ses composants électroniques, auxquels il faut ajouter la fin de vie. Toutes ces étapes du cycle de vie représentent une part importante des impacts qu'il faut considérer.

[148] https://app.electricitymaps.com/map

Les impacts liés au cycle de vie complet des équipements informatiques sont plus difficiles à estimer, et ce pour plusieurs raisons que nous aborderons dans les parties suivantes. Néanmoins, la première raison reste le manque de données d'impact de la part des constructeurs et fournisseurs.

Pour estimer les impacts sur l'environnement d'un serveur, on peut estimer les impacts de chacun de ses composants séparément, que l'on peut ensuite agréger. Pour l'IA générative, on se base très souvent sur des composants de type GPU (*Graphics Processing Unit*), en plus des composants d'un serveur classique : CPU, RAM, disques, carte mère, alimentation, etc. Dans un certain nombre de ces composants, on trouve une ou plusieurs puces de silicium dont la fabrication à un impact important. L'impact environnemental de la fabrication d'un serveur servant à l'entraînement de modèle d'IA se situe autour de 3 700 kg CO_2eq [149], pour une durée de vie d'environ 6 ans.

6 ans de **serveur** = **3 ×** Paris ↔ New York ***ou*** **25 ×** Paris ↔ Toulouse *Impactco2.fr* [150].

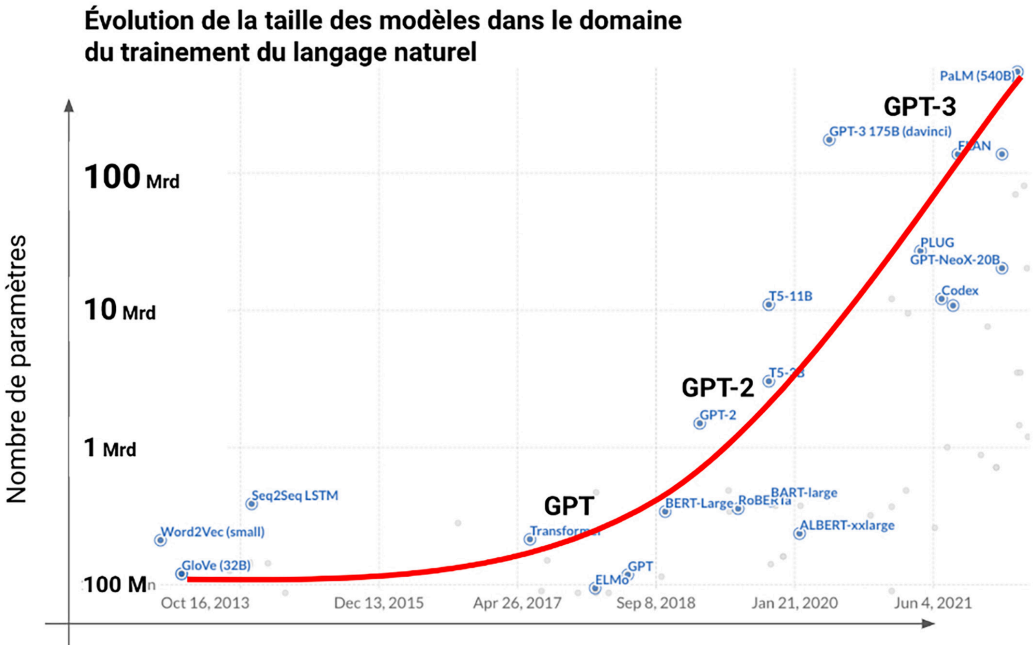
Aujourd'hui, les estimations des impacts environnementaux des serveurs ou même plus largement des équipements numériques prennent généralement uniquement en compte le critère des émissions carbone. Or, il a été démontré que les impacts sont en réalité multicritères [151]. Évaluer les émissions de gaz à effet de serre n'est pas suffisant pour estimer l'impact réel de la production des équipements numériques. Il faut également étudier les critères de raréfaction des ressources naturelles, de la consommation et pollution de l'eau, la consommation d'énergie, de la diminution de la biodiversité, pollution de l'air, etc.

Impacts directs de l'IA générative

Puisque l'IA reste un programme informatique, ses impacts directs sur l'environnement proviennent notamment de l'extraction des ressources, de la fabrication, du transport et de la fin de vie des équipements numériques et de la consommation d'électricité de ces derniers. Dans cette partie, nous traiterons des impacts lors des phases d'entraînement et d'inférence (utilisation) du modèle en nous basant sur la littérature scientifique. Nous aborderons également plus en détail le cas des modèles de langage ou *large language model* (LLM) qui sont utilisés dans les services comme ChatGPT. Le même type de développement peut être conduit pour des modèles génératifs d'image, de son ou de vidéo.

Les modèles d'IA générative de texte connus aujourd'hui sont fondés sur une architecture *transformer* inventée en 2017 par une équipe de recherche chez Google Brain. OpenAI a publié GPT en 2018. Depuis cette invention, le nombre et la taille des modèles de langage n'a cessé de croître, entraînant ainsi l'augmentation de la quantité d'énergie et d'équipement informatique nécessaires pour entraîner et utiliser ces modèles. Le premier modèle GPT contenait 120 millions de paramètres, contre 175 milliards pour GPT-3, sorti en 2020. Aujourd'hui, l'architecture exacte de GPT-4 est inconnue, mais il a été estimé qu'il pouvait être encore 10 fois supérieur à son prédécesseur.

[149] https://arxiv.org/pdf/2211.02001.pdf
[150] https://impactco2.fr/
[151] https://arxiv.org/pdf/2011.02839.pdf



Sevilla, et al. (2022) [152].

Par la suite, nous nous intéresserons à deux LLM comparables : GPT-3 et BLOOM. GPT-3 est le dernier modèle fourni par OpenAI dont on connaît l'architecture et la méthode de création, ce qui en permet l'estimation des impacts. BLOOM est un modèle de langage très semblable à GPT-3, *open-source* pour lequel une analyse d'impact détaillée a été conduite [153], que nous utiliserons en référence. Nous étudierons séparément la phase d'entraînement et d'inférence (ou utilisation) de ces modèles d'IA générative.

Entraînement des modèles

Lorsque l'on parle de l'impact environnemental des modèles d'IA, il est souvent mentionné que c'est lors de la phase d'entraînement que ces derniers consomment le plus de ressources et d'énergie. L'entraînement des modèles d'IA générative de texte comme GPT-3 nécessite par exemple de rassembler de nombreuses ressources numériques (serveurs, GPUs) et de données. Cette phase peut durer plusieurs jours, semaines ou mois. Le principe est d'utiliser des bases de données de texte massives provenant d'internet. Une tâche d'entraînement classique pour un modèle de langage consiste à prédire le mot suivant dans une phrase ou un paragraphe. Dans cette partie, nous n'entrerons pas plus dans les détails de l'entraînement, mais nous concentrerons plutôt sur les ressources employées pour réaliser cet exercice.

Nous comparerons les impacts sur l'environnement des deux modèles GPT-3 et BLOOM, comparaison rendue possible par la proximité des modèles et par notre connaissance suffisante des détails techniques.

	GPT-3 (OpenAI)	BLOOM (BigScience)
Nombre de paramètres	175 milliards	176 milliards
Architecture	Transformer (decoder-based)	Transformer (decoder-based)

[152] https://ourworldindata.org/grapher/artificial-intelligence-parameter-count
[153] https://arxiv.org/pdf/2211.02001.pdf

Dans l'étude de Patterson et al. [154], l'impact environnemental de GPT-3 est estimé à 552 tonnes CO_2eq. Ce chiffre a été calculé sur la base de la consommation électrique des serveurs utilisés pour entraîner le modèle GPT-3. Cette phase a mobilisé 10 000 GPUs [155] (NVIDIA V100) pendant presque 15 jours, un chiffre rarement égalé. L'estimation inclut aussi le PUE (*Power Usage Effectiveness*) sur la période d'entraînement. Le PUE est un ratio toujours supérieur à 1 qui sert à tenir compte de la consommation des équipements non informatiques dans un *data center* (climatisation, éclairage, espaces bureaux, etc.).

Entraînement de GPT-3

Énergie consommée (15 jours)
1 287 MWh
→ ~270 ménages (1 an) en France [156]

Impact carbone (15 jours)
552 tonnes CO_2eq
→ ~200 allers-retours Paris ↔ New York [157]

Dans le cadre de l'évaluation d'impact environnemental de BLOOM mené dans l'étude de Luccioni et al. [158], la même approche d'estimation a été conduite et nous donne des résultats assez différents. D'abord l'énergie consommée est presque 3 fois inférieure, cela peut s'expliquer par les serveurs et GPUs utilisés qui peuvent être plus efficaces énergétiquement parlant, mais également par les techniques d'entraînement du modèle. Cependant, l'écart le plus important porte sur l'impact carbone, environ 18 fois inférieur. Cette réduction s'explique, tout d'abord, par le fait que la consommation énergétique est plus faible dans le cas de BLOOM. Différents facteurs entrent en compte dans la consommation d'énergie, comme le profil de consommation des serveurs utilisés, la durée de l'entraînement ou encore l'architecture du modèle. Par ailleurs, l'entraînement du modèle a eu lieu en France dont l'intensité carbone de la production d'électricité est environ 7,4 fois inférieur (57 g CO_2eq / kWh contre 427 g CO_2eq / kWh pour GPT-3).

	Entraînement GPT-3	Entraînement BLOOM
Énergie consommée	1 287 MWh	433 MWh
Intensité carbone de la production d'électricité	423 g CO_2eq / kWh	57 g CO_2eq / kWh
Impact carbone	552 tonnes CO_2eq	30 tonnes CO_2eq

Il est cependant capital de noter que ces chiffres sont des estimations et que le périmètre de l'analyse est restreint. Dans l'étude de Luccioni et al. [159], une estimation plus poussée a été menée, afin d'étendre le périmètre de calcul. Trois améliorations sont envisageables. Tout d'abord, quantifier la vraie consommation résiduelle

[154] https://arxiv.org/pdf/2104.10350.pdf
[155] https://news.microsoft.com/source/features/innovation/openai-azure-supercomputer/
[156] https://particuliers.engie.fr/electricite/conseils-electricite/conseils-tarifs-electricite/consommation-moyenne-electricite-personne.html
[157] https://impactco2.fr/
[158] https://arxiv.org/pdf/2211.02001.pdf
[159] https://arxiv.org/pdf/2211.02001.pdf

d'électricité du *data center*. Ensuite, introduire les impacts liés à la fabrication des serveurs de calcul. Enfin, tenir compte de la phase de développement du modèle.

L'un des problèmes de l'approche précédente est de ne pas tenir compte du fonctionnement d'un *data center* et de sa consommation électrique réelle. Le premier résultat inclut certes la consommation des serveurs ainsi que des externalités comme la climatisation, ou l'éclairage du bâtiment grâce au PUE. Cependant, cette approche ignore complètement la consommation des autres équipements présents dans le data center et qui sont essentiels à son fonctionnement. Cela inclut, par exemple, les *switchs* et *routers* pour les communications réseaux ou les serveurs de stockage de données. Si l'on tient compte de ces appareils supplémentaires, la consommation nécessaire à l'entraînement du modèle BLOOM passe de 433 MWh à **690 MWh, soit une augmentation de presque 60 %**. On peut recalculer l'impact carbone sans tenir compte du PUE (inclus dans la consommation résiduelle) pour arriver à un impact de 39,3 tonnes CO_2eq.

Pour calculer un impact environnemental complet selon les méthodologies d'Analyse de Cycle de Vie (ACV), il est nécessaire d'intégrer l'impact de la fabrication des serveurs utilisés. Comme développé dans la partie précédente, la fabrication des équipements numériques a un impact environnemental fort. Il est par conséquent important de l'intégrer dans les calculs d'impacts. L'ajout des impacts à la fabrication se fait souvent en considérant le ratio du temps d'utilisation des ressources informatiques sur leur durée de vie. Dans le cadre de l'étude des impacts du modèle BLOOM, il a été considéré une durée de vie de 6 ans et l'impact de la fabrication d'un serveur de 2 500 kg CO_2eq et d'un GPU de 150 kg CO_2eq. Ces chiffres d'impacts sont très probablement des sous-estimations, mais restent cohérents en termes d'ordre de grandeur. Intégré à notre calcul précédent, l'impact de l'entraînement du modèle BLOOM passe maintenant de 39,3 à 50,5 tonnes CO_2eq.

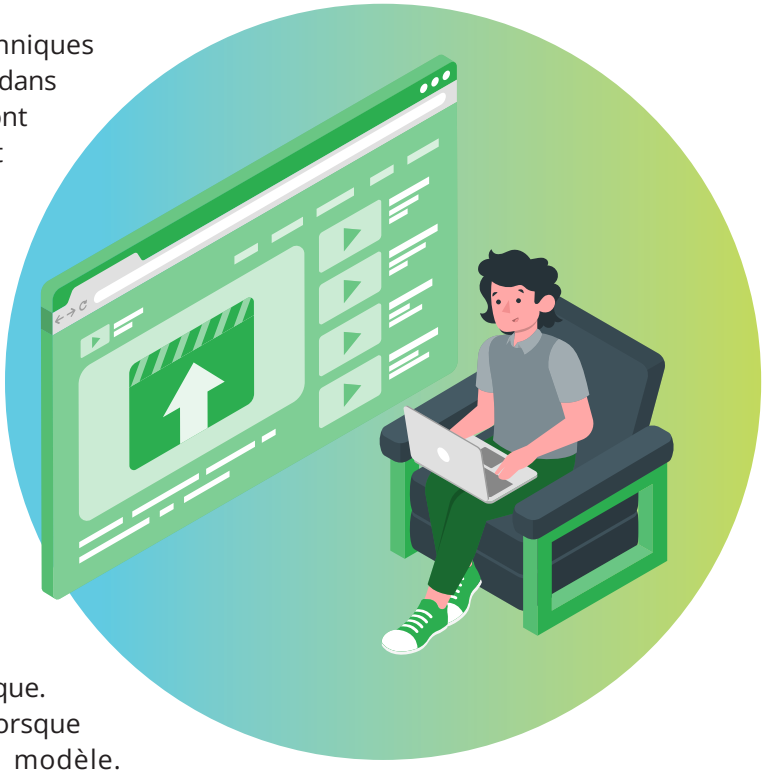
Le développement de modèle d'IA générative n'est pas un processus linéaire, en réalité un certain nombre de tests, évaluations et corrections ont lieu lors de l'entraînement. Dans le cadre de BLOOM, il a été estimé que cette phase de développement ajoute 73,32 tonnes CO_2eq supplémentaires. La phase de développement a plus d'impact que l'entraînement complet du modèle de bout en bout. Il est donc primordial de quantifier également cette phase, qui est très souvent ignorée dans les études. L'impact total du modèle calculé est maintenant de 123,82 tonnes CO_2eq, soit 4 fois plus que la première estimation plus naïve.

Bien que méthodologiquement cette approche ne soit pas applicable aux estimations d'impact réalisées pour GPT-3, l'impact carbone augmenterait à ~ 2200 tonnes CO_2eq. Cela représente cette fois l'impact de 1 600 allers-retours Paris – New York, soit plus de 4 vols par jour sur une année.

Phase d'inférence

Après la phase d'entraînement, on parle de phase d'inférence lorsque le modèle est rendu disponible aux utilisateurs finaux. Au cours de cette phase, les paramètres du modèle sont figés et on calcule uniquement le résultat d'une requête. Le mode d'inférence consomme moins de ressources que le mode d'entraînement et il est

possible d'avoir recours à des optimisations techniques pour réduire le temps de calcul que nous traiterons dans une partie dédiée. Ces opérations d'optimisation sont couramment mises en place, car elles permettent en général de faire des économies en réduisant les ressources nécessaires, et donc les impacts.



La méthodologie d'estimation des impacts d'un modèle en phase d'inférence reste proche de celle liée à l'estimation des impacts de la phase d'entraînement. Il est nécessaire de prendre en compte la consommation électrique des équipements informatiques. À laquelle on ajoute aussi les impacts de la fabrication des équipements en fonction du temps d'utilisation et de leur durée de vie.

La comparaison des impacts liés aux phases d'entraînement et d'inférence sont encore plus difficilement trouvables dans la littérature scientifique. C'est un manquement qui peut s'avérer critique lorsque l'on regarde l'intégralité du cycle de vie d'un modèle. **Cependant dans le cas de l'IA générative, il semblerait que la part des émissions de GES liées à la phase d'inférence peut dépasser largement celle liée à l'entraînement.** Dans ce cas, l'utilisation du modèle prend le dessus sur son entraînement en termes d'impact, notamment poussé par le nombre d'utilisateurs. Autrement dit, la consommation unitaire est plus faible mais le volume est beaucoup plus élevé.

Dans l'étude de Luccioni et al. [160] du modèle BLOOM, une expérimentation en phase d'inférence a été conduite sur 18 jours. Durant cette période, l'API exposant le modèle traitait en moyenne 558 requêtes par heure. Le modèle était déployé sur un serveur Google Cloud Platform. Il a été mesuré que le serveur a consommé 914 kWh d'électricité, soit 50,8 kWh par jour. En comparaison, lors de l'entraînement de BLOOM, les serveurs ont consommé 3 671 kWh par jour, soit 72 fois plus. Il est important de noter que dans ce cadre, la différence dans le nombre de requêtes pèse dans la balance, avec 558 pour BLOOM contre plusieurs millions de requêtes par heure pour ChatGPT.

Une autre étude publiée par Meta (Wu et al. 2022) [161] démontre un cas d'usage un peu plus réaliste du déploiement d'un modèle de langage en phase d'inférence. Il s'agit d'un modèle de langage à base d'architecture *Transformer* utilisé pour faire de la traduction de texte. Les modèles sont utilisés pour produire plusieurs milliards de milliards de prédictions par jour. Pendant deux ans, les impacts de la phase d'entraînement et d'inférence du modèle de langage ont été quantifiés. Sur cette période, l'impact de l'entraînement correspond à 35 % de l'impact global, les 65 % restant correspondant à la phase d'inférence. Utilisés à grande échelle, les modèles peuvent être source d'émissions de GES importantes durant leur phase d'utilisation.

Les calculs d'impacts en phase d'inférence, bien qu'ils soient similaires aux calculs faits lors de la phase d'entraînement, sont très souvent absents des études. Cette absence peut s'expliquer par les mêmes freins rencontrés pour la phase d'entraînement,

[160] https://arxiv.org/pdf/2211.02001.pdf
[161] https://arxiv.org/pdf/2111.00364.pdf

notamment le manque de méthodologie standard. Néanmoins, on peut y ajouter le manque de données relatives au déploiement dans des services cloud dont leurs fonctionnements restent opaques et rendent les estimations difficiles.

Étant donné que les impacts lors de l'inférence ont lieu après l'entraînement du modèle et sur de plus longues périodes, il se pose la question des usages. Pour estimer les impacts d'inférence, il faut anticiper les usages liés aux modèles ce qui peut s'avérer être difficile. C'est encore plus le cas lorsque les modèles sont open-source, il devient impossible d'estimer l'impact total en inférence. Cependant, il est important de noter qu'il est tout-à-fait envisageable de fournir (comme dans le cadre de l'analyse de BLOOM) des données sur l'utilisation du modèle sur une période de test. Il serait intéressant de calculer des facteurs d'émissions standards afin de pouvoir comparer les modèles.

Afin de nous donner une idée plus claire de ce que peut représenter l'impact d'utilisation de ChatGPT avec le modèle GPT-3, nous allons extrapoler certaines données des études précédentes. Une approche possible est d'estimer l'énergie consommée pour répondre à une requête utilisateur, en partant des données de l'étude du modèle BLOOM [162]. L'énergie consommée pour une requête serait donc d'environ 3,96 Wh. Si on estime le nombre moyen de requêtes par jour sur ChatGPT, on peut donc calculer l'énergie totale consommée. En se basant sur les hypothèses développées par Patel et Ahmad [163], il y a 13 millions d'utilisateurs actifs par jour faisant chacun 15 requêtes. L'énergie consommée chaque mois par les serveurs est d'environ 23 GWh. En considérant un *PUE* classique d'un *data center hyperscaler* d'environ 1,10, l'énergie consommée par le data center passe à 25,3 GWh. On peut calculer l'impact GES avec l'intensité carbone du mix électrique moyen mondiale de 0,39 $kgCO_2eq./kWh$ [164], ce qui résulte en 9 867 tCO_2eq uniquement liée à la consommation d'énergie directe. Ces estimations de consommation d'énergie et d'impact GES ne nous donnent qu'un ordre de grandeur. Dans d'autres publications de Kasper Groes Albin Ludvigsen [165], il est tenté également selon différents scénarios d'estimer la consommation énergétique de ChatGPT entre 1 GWh et 23 GWh par mois.

À l'estimation précédente, on peut ajouter une estimation de l'impact de la fabrication du matériel. Dans la modélisation de Patel et Ahmad [166], il est estimé que OpenAI utilise environ 3 617 serveurs HGX A100 et un total de 28 936 GPUs. L'impact GES d'un serveur de calcul haute performance peut être estimé à 3 700 kg $CO_2eq.$ auquel sont ajoutés 1 200 kg $CO_2eq.$ pour les 8 GPUs par serveur. Si on considère une durée de vie de 6 ans, l'impact de la fabrication des serveurs est ramené à un mois d'utilisation et donne 246 t $CO_2eq.$ Ce chiffre reste une très large sous-estimation de l'impact total des équipements numériques présents dans un *data center*. Si l'on combine l'impact de la consommation d'électricité et l'impact de la fabrication du matériel, on a donc un impact mensuel d'environ 10 113 t CO_2eq par mois. Si l'on se projette sur **une année complète, cela représente l'émission d'environ 122 351 t CO_2eq**. Bien que cette modélisation soit très incomplète, on se rend compte que les impacts à l'utilisation peuvent dépasser largement les impacts de l'entraînement du modèle, qui était autour de 552 t $CO_2eq.$ Le déploiement de services similaires à ChatGPT dans d'autres domaines comme la recherche sur internet pourrait d'autant plus décupler les impacts de ces modèles, en

[162] https://arxiv.org/pdf/2211.02001.pdf
[163] https://www.semianalysis.com/p/the-inference-cost-of-search-disruption
[164] https://base-empreinte.ademe.fr/donnees/jeu-donnees
[165] https://towardsdatascience.com/chatgpts-electricity-consumption-7873483feac4#:~:text=ChatGPT's%20electricity%20consumption%20per%20query%20is,as%20BLOOM's%2C%20i.e.%200.00396%20KWh
[166] https://www.semianalysis.com/p/the-inference-cost-of-search-disruption

touchant cette fois non pas une dizaine de millions de personnes, mais des milliards.

Notons également, que nous n'avons évoqué que les impacts liés à GPT-3. Or, selon les dernières rumeurs, GPT-4 serait composé de 8 sous-modèles, chacun plus gros que GPT-3 pour un nombre total de paramètres d'environ 1 760 milliards. L'énergie consommée et les impacts à l'usage d'un tel modèle sont probablement entre 10 et 100 fois plus importants.

Entraînement GPT-3 pour ChatGPT

552 tonnes CO_2eq

Utilisation de ChatGPT (GPT-3) en janvier 2023

10 113

tonnes CO_2eq

Notons que nous avons abordé de façon assez large la question des impacts directs des modèles d'IA générative sous le prisme des phases d'entraînement et d'inférence de ces modèles. Nous avons omis un certain nombre de facteurs comme le *fine-tuning* des modèles, l'impact de la collecte, du stockage des données, la fréquence de mise à jour, etc. Les impacts directs sont par conséquent sous-estimés. Même lorsqu'on se concentre sur une phase spécifique, il n'est pas évident (et rarement établi) de prendre en compte l'ensemble du périmètre d'émissions liées à l'utilisation des équipements numériques. Enfin, nous traiterons plus en détail dans les parties suivantes les facteurs de difficultés à produire une évaluation des impacts fiable, la question de la réduction des impacts et enfin les impacts indirects et les effets rebond découlant des usages des modèles d'IA générative.

Les recommandations de Data For Good pour les **utilisateurs**

S'interroger sur ses pratiques et ses besoins, limiter l'usage des modèles génératifs au nécessaire. Il faut envisager et préférer des solutions techniques moins gourmandes quand cela est possible (recours à des templates, des moteurs de recherches classiques).

Les recommandations de Data For Good pour les **data scientists**

Ne pas nécessairement utiliser de l'IA générative pour tous les cas d'usage. Par exemple, les expressions régulières sont parfois plus pertinentes pour extraire de l'information. Le *fine-tuning* d'un modèle de classification est beaucoup plus pertinent, frugal et facile à entraîner avec des techniques de *few shot learning* qu'avec de l'IA générative [167].

Ne pas toujours utiliser le modèle le plus puissant (par exemple GPT-4) par défaut, voire mettre en place des systèmes de routage automatique qui choisiront le modèle le plus pertinent et frugal en fonction du cas d'usage.

Difficulté pour évaluer les impacts

Si des approximations et des tentatives d'appréciation des impacts écologiques peuvent être réalisées, il est cependant difficile d'estimer un impact complet et réel de l'IA générative.

Une première difficulté est inhérente à toute technologie informatique et n'est pas forcément propre à l'IA générative : la chaîne de valeur, de production et de développement d'un algorithme intègre de nombreux paramètres et acteurs à prendre en compte (utilisation des serveurs, effets indirects, etc.).

L'IA générative implique tout de même quelques spécificités qui rendent plus complexe la mesure des impacts de celle-ci sur l'environnement. D'une part, la taille des modèles et des jeux de données utilisés est sans précédent. La durée d'entraînement nécessaire pour produire des modèles est variable d'un modèle à un autre, peut se révéler fortement coûteuse et est souvent inconnue du grand public. L'utilisation massive de GPU interroge également quant aux émissions de ces technologies. Le tout dans un environnement opaque, où chaque acteur individuellement ne fait pas preuve de la transparence nécessaire à la compréhension des impacts environnementaux de ses travaux.

Manque de transparence des acteurs de l'IA

Pour comprendre les impacts de l'IA générative, nous devons prendre en compte chaque acteur de la chaîne de production.

Tout d'abord, en ce qui concerne la collecte, si certains jeux de données sont mis à disposition du public dans un effort de transparence et d'ouverture pour la Recherche, la plupart des acteurs manquent de transparence quant aux jeux de données utilisés pour l'entraînement des modèles. De la méthode de *scrapping* (collecte de données sur le web) à l'hébergement des données, les informations disponibles sont rares, ou présentées de manière partielle. De plus, le travail de labellisation et de collecte des données est souvent délégué à des acteurs opérant depuis des pays en voie de développement (comme longuement démontré par Antonio Casilli dans son ouvrage « *En attendant les robots. Enquête sur le travail du clic* », ou plus récemment dans un article « *Enquête : derrière l'IA, les travailleurs précaires des pays du Sud* »[168] écrit en collaboration avec Clément Le Ludec, sociologue du numérique et Maxime Cornet, doctorant en sociologie de l'IA).

Vient ensuite l'entraînement des algorithmes. Lorsque l'IA générative en était à ses balbutiements, de nombreux articles de recherche ont été publiés et ont permis de comprendre l'architecture des modèles utilisés, avec parfois des informations sur les temps d'entraînement et le hardware utilisé. Mais au fur et à mesure que l'IA générative s'est démocratisée, le nombre d'informations communiquées sur l'architecture des modèles, les méthodes et les durées d'entraînement a drastiquement diminué pour ce qui concerne les acteurs privés. Certaines informations fuitent plus ou moins volontairement (par exemple, GPT-4 posséderait 170 000 milliards de paramètres, contre 175 milliards de paramètres de GPT-3.5), il semble que nous sommes davantage dans une course « au plus gros modèle » et à la performance

[167] Par exemple en utilisant https://github.com/huggingface/setfit.
[168] https://theconversation.com/enquete-derriere-lia-les-travailleurs-precaires-des-pays-du-sud-201503

que dans une volonté de faire progresser l'état de l'art. Avec de plus en plus de cas d'utilisation des modèles et de leur commercialisation, cette transparence pourrait encore s'amenuiser.

Du côté des *cloud providers*, un effort global engagé depuis quelques années a été constaté autour de la consommation de leurs services, avec l'utilisation de nouvelles méthodes de refroidissement, et la volonté d'aller vers la neutralité carbone, ainsi que le développement d'équipement numérique spécifique. Là encore, pour quiconque chercherait à mesurer l'impact de l'IA générative sur les machines virtuelles et services managés, l'information fournie reste relativement faible à un niveau plus granulaire.

Les producteurs de hardware ne sont pas en reste : alors que la demande en GPU (et TPU) explose, provoquant des pénuries tant pour les data scientists que pour les joueurs de jeux vidéos ou les mineurs de crypto-monnaies, aucun acteur majeur (tel que Nvidia ou Google) ne propose de manière publique une analyse du cycle de vie de ses outils. Certaines innovations, telles que les nouvelles puces d'accélération de réseaux de neurones, doivent permettre de réaliser des optimisations importantes, mais il reste difficile à ce stade d'avoir des idées précises de mesures.

Enfin, du côté des opérateurs de modèles, la transparence là encore n'est pas de mise. D'une part, les business modèles diffèrent : entre des modèles utilisables en SaaS sur abonnement (comme par exemple ChatGPT+) ou des partenariats entre entreprises privées pour enrichir leurs offres de services. Par exemple, l'intégration à des outils existants (comme le partenariat Bing et OpenAI) ou la possibilité d'embarquer des modèles dans des infrastructures propriétaires (par exemple Azure). Cette grande variété dans les modes opératoires ne permet pas d'avoir une étude de mesure d'impacts précis.

Pour les modèles SaaS par exemple, il n'est pas possible de connaître le nombre ni la taille des requêtes reçues, le nombre d'instances en parallèle ou la redondance des services. De même, il n'y a pas de possibilité de connaître les potentielles optimisations générées par la mise en place de mémoire cache.

Des méthodologies à développer

Certains articles de recherche ont tenté d'apporter des réponses à l'impact de l'IA générative sur l'environnement. Ces études sont utiles pour commencer à comprendre ces impacts et surtout pouvoir agir.

Malheureusement elles délaissent souvent de nombreux critères de mesure, en se focalisant principalement sur l'entraînement des modèles ou leur mise en production. Ce n'est pas dénué d'intérêt : on note une forte corrélation entre la taille du modèle et son empreinte carbone. Il est complexe aujourd'hui d'appliquer une méthodologie claire et complète qui nous permettrait d'avoir une idée précise des impacts de l'IA générative avec le nombre d'informations disponibles actuellement.

La plupart des études ne présentent par exemple pas la phase du développement du modèle ou encore sa mise en production. Il est de même que rarement abordé, l'impact de tout ce qui a trait à la collecte, au stockage, au transfert des réseaux de données.

Les études tendent également à se concentrer sur la consommation électrique. Or un modèle n'est pas qu'un « produit » numérique dénué de toute matérialité physique. À la consommation électrique, nous devons aussi ajouter l'extraction des ressources, la fabrication, le transport, les usages et la fin de vie de tous les équipements numériques.

Certains groupes de travail comme Boavizta [169] tendent à développer des méthodologies et outils d'évaluation d'impact, fondés sur des méthodes de construction de l'impact en « *bottom-up* » en se basant sur l'analyse de cycle de vie des équipements. Nous avons besoin de définir des standards, des normes, des méthodologies d'évaluation afin de pouvoir vraiment approcher les impacts environnementaux de l'IA générative.

Des évaluations d'impacts encore trop focalisées sur les gaz à effet de serre

Enfin, mais ce n'est sans doute pas propre à l'IA générative, la majorité des tentatives d'évaluation des impacts environnementaux restent centrées autour du seul critère des émissions carbone. L'extraction des métaux, le rapport à la biodiversité, la consommation d'eau ne sont que rarement quantifiés.

Par exemple, le *water cooling* est aujourd'hui souvent utilisé du côté des *cloud providers* pour réduire la climatisation et la consommation électrique. Mais celui-ci n'est pas neutre dans notre consommation d'eau [170] et son gaspillage, ou encore dans l'impact sur la biodiversité.

De même, remplacer un GPU par une nouvelle génération peut avoir des impacts positifs sur la consommation en électricité en réduisant et optimisant les temps de calcul. Mais la fabrication d'un nouveau GPU nécessite des matières comme le silicium avec un fort impact sur l'extraction des ressources. Il faut trouver le bon équilibre entre les différents critères et ne pas se concentrer sur les GES.

En cela, avec une approche basée sur un seul critère, on pourrait assister à un **transfert de pollution** : baisser les émissions de gaz à effet de serre n'est pas une solution si elle devient dommageable à l'environnement dans son ensemble, ou si cette solution se contente de transférer les responsabilités vers les autres.

Il est important de prendre en compte différents critères dans notre rapport à la technologie : déplétion de ressource métallique et des minerais, consommation d'eau, biodiversité, sources d'énergie primaire en amont de la production d'électricité, conservation des données, typologie de stockage, etc. Évaluer l'empreinte carbone des modèles d'IA générative est important, mais cela ne doit pas être le seul critère d'évaluation.

Les impacts indirects et effets rebonds non évalués

Les effets indirects de l'IA générative concernent les impacts environnementaux résultant d'une utilisation spécifique de cette technologie qui créent de nouveaux usages ayant un impact négatif ou induisent des effets rebond.

Premièrement, **l'IA générative permet de nombreux nouveaux usages qui ont un impact environnemental négatif** :

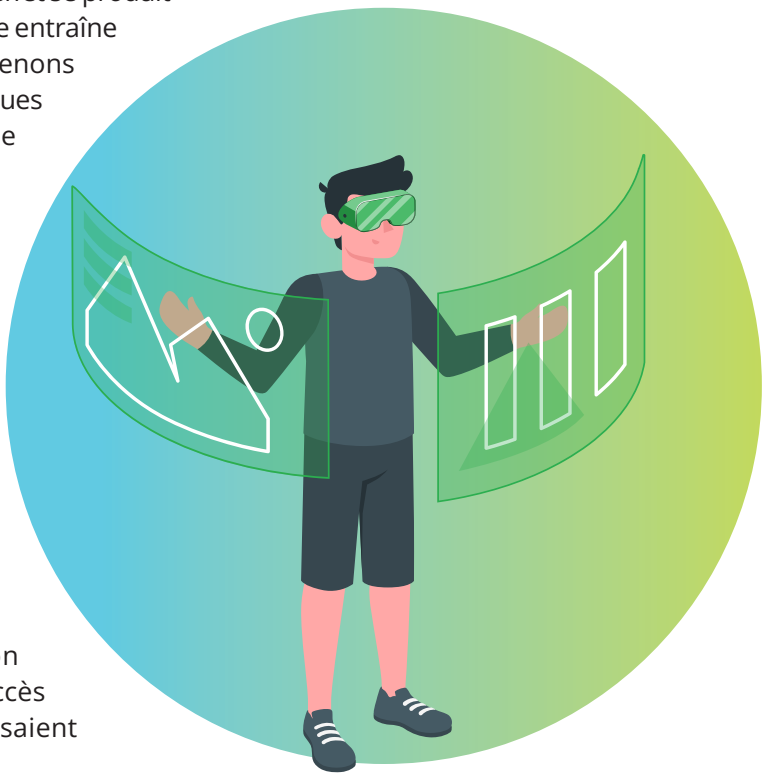
- L'IA générative permet de générer du nouveau contenu (images, textes, audio, ...) qui n'existait pas avant et nécessite d'être calculé, stocké, et montré aux utilisateurs. Par exemple, si auparavant des campagnes marketing montraient une dizaine de contenus différents en fonction des utilisateurs, demain l'IA générative permet la

[169] https://www.boavizta.org/
[170] https://www.nature.com/articles/s41545-021-00101-w

génération de milliers de variations personnalisées, allant jusqu'à avoir un contenu unique pour chaque utilisateur et un calcul en temps réel.

- Par ailleurs, l'IA générative permettant de personnaliser et d'adapter un message de communication pour la vente d'un produit, cela contribue à l'accélération de la vente de produits et services souvent carbonés ou la rétention à un service, de la manière que la personnalisation a permis l'émergence de Netflix et de vendre des millions de produits sur Amazon. Par exemple, si l'IA générative contribue à vendre 100 télévisions supplémentaires par jour sur Amazon, cela équivaut en ordre de grandeur à 40 tCO_2eq (pour des émissions de 400 $kgCO_2eq$ moyenne pour un téléviseur de 40 pouces sur sa durée de vie). Soit autour de 25 fois plus sur un an que les émissions liées à l'entraînement de GPT-3. Et nous pouvons imaginer cela sur l'ensemble de l'économie – des véhicules thermiques aux billets d'avion.
- L'IA générative permet également d'extraire de l'information, et de diffuser et démocratiser une expertise sur une base de connaissance choisie. Par exemple, cela peut permettre de débloquer des scientifiques travaillant sur un traitement médical, cela peut aussi débloquer une entreprise pétro-gazière dans la recherche de lieux et moyens d'extraction.
- L'IA générative complexifie des usages numériques qui étaient plus sobres au préalable. Par exemple : un moteur de recherche génératif au lieu d'un moteur d'indexation plus standard, ou un personnage dans un jeu vidéo qui avait une instruction unique à répondre au joueur aura demain un moteur d'IA générative embarquée pour rendre l'expérience plus immersive. Nous pourrons également citer la réalité augmentée et virtuelle – elles-mêmes des usages numériques gourmands en ressource et en énergie – accélérées par les potentiels de l'IA générative.

Deuxièmement, l'IA générative induit des effets rebond, un aspect crucial à considérer dans ses impacts environnementaux. Cet effet se produit lorsque l'amélioration de l'efficacité d'une technologie entraîne une augmentation de sa consommation totale. Prenons l'exemple du transport routier [171] : les progrès techniques tels que l'amélioration du rendement des moteurs, le développement de moteurs hybrides et électriques, ainsi que la modernisation des infrastructures routières ont contribué à réduire les émissions de CO_2 par véhicule. Cependant, ces améliorations rendent les véhicules plus économiques et pratiques, ce qui incite les consommateurs à les utiliser davantage. Par conséquent, le nombre de véhicules en circulation a augmenté, aboutissant à une augmentation globale des émissions en Europe de 25 % entre 1990 et 2005, malgré les avancées technologiques visant à réduire les émissions individuelles. L'IA générative présente ainsi de nombreux effets rebond, par exemple :



- L'IA générative par sa simplicité d'utilisation permet à des millions de citoyens d' avoir accès à l'IA pour des tâches du quotidien qu'ils faisaient

[171] https://fr.wikipedia.org/wiki/Effet_rebond_(%C3%A9conomie)#Transport_routier

autrement avant et sans avoir recours à des *data scientists* ou des compétences particulières. Par exemple pour écrire des mails, résumer ou créer un texte. C'est l'effet rebond le plus majeur de l'IA générative, en diminuant les barrières à l'entrée et avec l'appropriation culturelle permise par ChatGPT, le nombre d'utilisateurs moyens de l'IA est décuplé.

- Nous pouvons également citer la visio-conférence souvent exhibée comme un impact positif du numérique pour éviter les transports. Microsoft Teams ou Google Meet proposent maintenant une nouvelle fonctionnalité consistant à enregistrer l'audio et la vidéo d'une visio-conférence d'une réunion pour résumer la prise de notes et les tâches. Ce nouvel usage déjà énergivore crée un effet rebond supplémentaire en incitant à enregistrer et stocker les réunions en vidéo, ce qui n'arrivait aujourd'hui que rarement.

En résumé, l'IA générative est un formidable catalyseur de productivité et de performance, ce qui engendre un effet rebond sociétal. Est-ce que cela permet de travailler plus, accélérant un système socio-économique encore trop émissif et destructeur du vivant ? Ou est-ce que cela permet de libérer du temps, un temps libre qui peut lui aussi être très carboné ?

Il est crucial de prendre en compte et de quantifier ces effets pour réduire les impacts environnementaux de l'IA générative. Pour y parvenir, il est possible de restreindre les usages des technologies aux applications les plus pertinentes et soutenables. Comprendre et limiter ces effets rebond représente un défi majeur. Cependant, il est primordial d'en être conscient afin d'améliorer la conception et la mise en œuvre des projets d'IA générative, en optimisant leur rentabilité environnementale et en minimisant leurs conséquences négatives sur l'environnement.

La quantification de ces effets indirects peut être complexe en raison de la diversité des cas d'usage, mais reste réalisable pour un projet donné. L'évaluation de la rentabilité environnementale d'un projet, en parallèle de sa rentabilité économique, permet d'évaluer les impacts directs et indirects, et de juger si la valeur ajoutée vaut les émissions engendrées.

Il faut introduire des méthodologies d'évaluation environnementale permettant une meilleure prise de conscience des émissions indirectes de GES associées à un projet d'IA générative. Ceci est particulièrement important pour les projets d'optimisation et de réduction des impacts dans d'autres secteurs. Tous les projets entraînent ces impacts, mais dans certains cas, comme les projets sociaux ou environnementaux bénéfiques, les bénéfices peuvent justifier ces conséquences.

La difficulté n'exclut pas les efforts d'évaluation et de réduction

Ainsi, il n'est à ce stade pas possible d'avoir une compréhension globale ni une évaluation complète des impacts de l'IA générative. Un certain nombre de facteurs entrent en compte : manque de transparence des acteurs, difficulté d'avoir une vision complète et une méthodologie générale, difficulté de gérer des impacts multiples.

Pourtant, il faut garder une approche pragmatique sur le sujet : chaque estimation, même partielle, permet ensuite de travailler sur des solutions actionnables, localisées, à impact.

Il est important pour une entreprise de pouvoir communiquer auprès de ses clients, ses investisseurs, ses collaborateurs, son écosystème, les efforts réalisés. Mais il n'est pas nécessaire pour cela d'avoir nécessairement des chiffres précis. L'incapacité à mesurer précisément ne doit pas servir d'excuse à l'inaction.

Sobriété numérique, diminution de la taille des modèles par distillation, usage de hardware ou de cloud providers plus sobres, optimisations, etc. Individuellement, ces actions peuvent paraître limitées et avec des impacts relativement faibles. Cependant, mises bout-à-bout, non seulement ces solutions peuvent avoir un impact, mais elles vont permettre également à des acteurs de définir des normes, des standards, des réglementations qui seront plus générales et qui pourront être mises en place. Il est important de garder en tête la règle des 5 R [172] de la démarche zéro-déchet dès la phase conception : Refuser, Réduire, Réutiliser, Recycler et Rendre.

Nous aborderons dans la partie suivante quelques pistes pour avoir de l'impact rapidement.

Les recommandations de Data For Good pour les **data dcientists**

Évaluer l'impact écologique des modèles. Cela doit inclure une évaluation de l'impact direct de l'entraînement (les *data scientists* pourront utiliser des outils comme Code Carbon) mais ne pas s'y limiter. En particulier l'impact de l'inférence doit être mesuré. De même, d'autres impacts indirects doivent être mesurés (recours à des terres rares pour le matériel technique, impact publicitaire / politique indirect, *greenwashing*).

Une fois évalués, les impacts écologiques, directs ou indirects, doivent être largement documentés, comme les autres limites des modèles génératifs. Les utilisateurs doivent être informés de l'impact écologique pour éventuellement préférer une solution moins gourmande.

Proposer des solutions plus efficaces et moins impactantes pour l'environnement. Étant souvent à l'origine du choix technique retenu pour l'utilisateur, les *data scientists* doivent prendre en compte la contrainte environnementale dans leur choix de solution technique.

Pistes de réduction des impacts

Première piste, la sobriété

Pour réduire l'empreinte sur l'environnement des ces technologies d'IA, la sobriété numérique reste la meilleure solution. Elle se matérialise par la prise de conscience des impacts environnementaux, puis par les actions de mitigation que peuvent prendre les utilisateurs, les entreprises et leurs employés.

Traitons des actions qu'un *data scientist* peut aujourd'hui mettre en place pour atteindre la sobriété dans son entreprise. Il faut dans un premier temps convaincre ses collègues et sa hiérarchie de l'importance de répondre aux problèmes liés à la crise climatique, pour appliquer dans un second temps des mesures de réduction des

[172] https://fr.wikipedia.org/wiki/R%C3%A8gle_des_5_R

impacts. D'un point de vue plus technique, il est possible de changer son processus de développement pour intégrer les enjeux environnementaux. Dans l'ordre, il faut se concentrer au minimum sur les points suivants :

1. Questionner l'utilité du projet,

2. Estimer les impacts du projet,

3. Évaluer la finalité d'un projet,

4. Restreindre les cas d'usage aux finalités souhaitées.

En tant que fournisseur de service, la sobriété peut passer par plusieurs vecteurs. Le premier est l'*open-source*, qui permet d'éviter de multiplier les services, les logiciels, les données, ainsi que le développement de nouveaux modèles. Un grand nombre d'entreprises ont déjà montré qu'il était possible d'offrir des services dont le cœur de développement est ouvert. Cela permet même de bénéficier d'un plus large soutien de la communauté qui peut se former autour d'un projet libre. Un second aspect, aligné avec le précédent, est la transparence : transmettre des informations claires sur les impacts environnementaux et mettre à disposition leurs méthodologies de calcul. La communication de chiffres sans les détails sous-jacents s'apparente le plus souvent à du *greenwashing*. Enfin, un dernier axe majeur est le fait de sélectionner explicitement les clients avec lesquels on souhaite travailler. Si le service vendu peut potentiellement être utilisé pour des cas d'usages *for bad*, c'est au fournisseur de service de refuser la collaboration.

Pour un utilisateur final, la sobriété correspond à limiter les usages de ces technologies. Il faut remettre en question ses pratiques, limiter le nombre de services que l'on utilise et y recourir avec modération. Lorsque ces informations sont disponibles, il est possible de se renseigner sur les impacts de l'utilisation d'un service. Autrement, il faut être pragmatique et accepter que toutes les technologies que nous consommons reposent sur des équipements informatiques bien réels : il est bien connu que toutes ne sont pas frugales, l'IA générative en fait partie. En somme, l'IA générative la plus sobre reste celle que l'on ne crée pas.

Opportunités de réduction liées à la technique

Nous venons de le voir, cette réflexion autour de la sobriété commence avant même la conception d'un modèle d'IA, en en étudiant ses usages. En effet, un modèle qui n'aura pas d'utilisation concrète, autre que des activités de recherche, ne verra pas l'impact de son utilisation compensé par la réduction d'autres activités qu'il pourra remplacer. Le travail du *data scientist* consistera d'abord à valider l'intérêt du futur modèle auprès de ses potentiels utilisateurs, qu'ils soient internes à une entreprise ou auprès du grand public, puis définir son champ d'application et être transparent sur les choix techniques.

Une recherche utilisateur, opérée pendant ou en amont du développement d'un modèle, permet de valider l'appétence et d'optimiser la collaboration humain-IA. Cela réduit les efforts nécessaires à l'aboutissement d'un modèle utilisable et utilisé, tout en favorisant le développement des usages d'un tel outil. Besmira Nushi [173] propose même de travailler au partage d'un modèle mental entre les collaborateurs humains et leurs outils basés sur l'IA, pour aller au-delà de la notion de performance des algorithmes. Cependant, cet effort ne peut être fait que dans des cadres structurés, comme celui de l'entreprise, où nous pouvons trouver une proximité relative entre les utilisateurs

[173] http://erichorvitz.com/gbansal-hcomp19.pdf

et les concepteurs, ce qui est le cas lors d'une collaboration B2B, par exemple.

Heureusement, la prolifération des modèles d'IA générative, soutenue par une publication en o*pen-source* d'une partie de ses acteurs, offre des options moins contraignantes pour tester la pertinence d'un concept technique ou d'une idée de modèle innovant.

En effet, de nombreux modèles, souvent généralistes, mais parfois très spécialisés, peuvent être testés simplement, via des APIs. C'est le cas, par exemple, pour la génération de modèles GPT-1 et GPT-2 ou via des démonstrateurs ouverts sur le site de Hugging Face.

Une fois l'appétence validée, le champ d'application défini et la transparence sur les choix techniques confirmée, il est temps de s'intéresser à la phase de développement. Pour ce faire, nous nous concentrerons sur deux leviers : la **méthodologie d'entraînement et de validation des modèles**, ainsi que **l'efficacité et l'impact des infrastructures** supportant ce développement.

Un troisième levier peut être évoqué, celui de l'implémentation et de méthodologie d'entraînement des réseaux de neurones, notamment au sujet du mécanisme de back-propagation. Une des limites techniques de ce mécanisme est la manipulation des gradients d'erreurs au long du réseau de neurones, qui a pour conséquence de ralentir la convergence et l'entraînement du réseau *in fine*. Si certaines alternatives au mécanisme de back-propagation ont été proposées par DeepMind [174], ou le MILA [175], aucune n'a réussi – à date – à faire consensus dans la communauté du ML et c'est pour cette raison que nous ne détaillerons pas plus ce levier.

Comment optimiser ces modèles d'IA génératifs ?

1. Méthodologie d'entraînement et de validation des modèles

Lors de cette phase du cycle de vie de modèles d'IA, qu'ils soient génératifs ou non, l'impact environnemental peut être décomposé comme un certain nombre d'autres produits du numérique. En premier lieu, l'utilisation des serveurs pendant cette phase, correspondant au scope 2 (du point de vue de l'hébergeur) de la norme Bilan Carbone proposée par l'Ademe [176]. Cet impact peut être approximé par le taux d'utilisation d'un certain matériel, pondéré par l'équivalent carbone généré par la consommation électrique du matériel utilisé. **Nous chercherons donc dans cette phase d'entraînement à minimiser l'utilisation en énergie du matériel.**

Cette première phase peut donc être abordée avec l'angle de l'usage du modèle génératif et peut bénéficier des avancées déjà permises par les modèles publiés en *open-source* par la communauté. Deux approches – potentiellement complémentaires – sont envisageables :

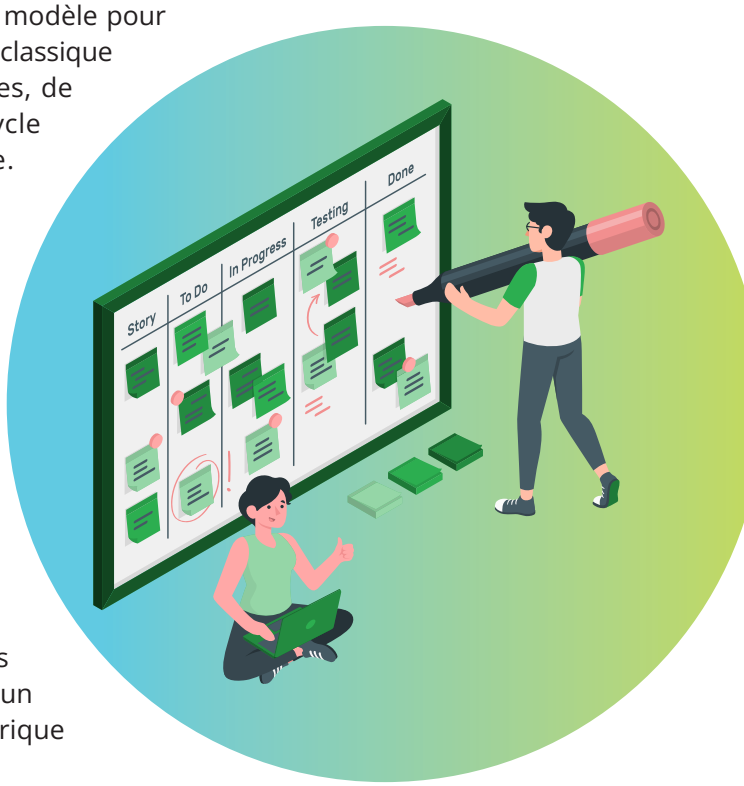
a. Utiliser un modèle déjà entraîné, dont les capacités de généralisation sont suffisantes pour permettre de réaliser la tâche attendue, d'où l'intérêt de définir avec précision cet usage attendu. Cette solution permet de fonctionner en mode « *zero-shot learning* », c'est-à-dire sans phase d'entraînement. Cette approche est

[174] http://proceedings.mlr.press/v70/jaderberg17a/jaderberg17a.pdf
[175] https://www.frontiersin.org/articles/10.3389/fncom.2017.00024/full
[176] https://bilans-ges.ademe.fr/

particulièrement intéressante pour les tâches généralistes : synthèse de texte, classification de documents, génération d'images à partir de prompts, etc. En un mot : réutiliser des modèles généralistes et open-source plutôt que d'entraîner des modèles spécifiques. Dans le domaine du NLP, les *LLM*, dont l'outil ChatGPT est issu, sont généralement de bons candidats pour une réutilisation immédiate, comme le montre ce comparatif [177] entre plusieurs modèles à l'état de l'art, datant d'avril 2023 :

	Latin						Non-Latin			
Language	English	French	Spanish	Portuguese	Italian	Deutsch	Chinese	Arabic	Japanese	Korean
Dolly	♥♠									
Alpaca	♥♠	♥	♥	♥	♥	♥				
Koala	♥♠	♥	♥	♥	♥	♥				
Baize	♥♠	♥	♥	♥	♥	♥				
Vicuna	♥♠	♥♠	♥♠	♥♠	♥♠	♥♠				
LuoTuo	♥	♥	♥	♥	♥	♥	♠			
Chinese-Alpaca	♥♠	♥	♥	♥	♥	♥	♠			
Guanaco	♥♠	♥	♥	♥	♥	♥♠	♠		♠	
BELLE	♥	♥	♥	♥	♥	♥	♥♠	♥♠	♥	♥
Phoenix	♥♠	♥♠	♥♠	♥♠	♥♠	♥♠	♥♠	♥♠	♥♠	♥♠
Latin Phoenix (Chimera)	♥♠	♥♠	♥♠	♥♠	♥♠	♥♠	♠	♠	♠	♠

b. Si le modèle pré-entraîné ne suffit pas à traiter le domaine d'application, une phase de *fine-tuning* peut être mise en place. L'enjeu pour en limiter l'impact environnemental sera de la rendre la plus efficace possible, baissant le nombre d'itérations d'apprentissage nécessaire au modèle pour obtenir des performances satisfaisantes. L'approche classique consiste à constituer un jeu de données labellisées, de la meilleure qualité possible afin que chaque cycle d'apprentissage soit le plus rentable possible. Labellisation par des experts, utilisation de labels positifs ou négatifs ou apprentissage contrastif [178] sont des méthodes qui permettent de donner autant d'indices que possible au modèle via le jeu de données qui lui sert pour apprendre.



Une fois cette première phase de conception terminée (affiner le *dataset* utilisé, trouver la bonne architecture ou le bon modèle pré-entraîné), suit généralement une longue phase d'entraînement qui aura pour objectif d'extraire tout le signal disponible dans les données mises à disposition.

C'est ici que l'infrastructure utilisée va avoir le plus gros impact, ces entraînements pouvant durer de quelques jours (dans le cas de *fine-tuning*) à plusieurs mois (pour un entraînement complet), avec une consommation électrique conséquente.

[177] https://arxiv.org/pdf/2302.13007.pdf
[178] https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607

2. Efficacité et impact des infrastructures

Les usages actuels préférant utiliser des infrastructures *cloud*, les conseils suivants vont principalement s'adresser aux utilisateurs de services *cloud*.

En s'inspirant du caractère itératif du travail d'assemblage de l'architecture du modèle, il est possible d'optimiser la métrique d'impact de l'infrastructure, représentée dans le tableau ci-dessous par la colonne d'émission en équivalent CO_2 pondéré par la valeur de PUE (*Power Usage Effectiveness*), représentant l'efficacité énergétique d'un *data center* (Luccioni et al. [179]).

Model name	Number of parameters	Datacenter PUE	Carbon intensity of grid used	Power consumption	CO_2eq emissions	CO_2eq emissions × PUE
GPT-3	175B	1.1	429 gCO_2eq/kWh	1,287 MWh	*502 tonnes*	552 tonnes
Gopher	280B	1.08	330 gCO_2eq/kWh	*1,066 MWh*	*352 tonnes*	380 tonnes
OPT	175B	*1.09*	*231 gCO2eq/kWh*	*324 MWh*	70 tonnes	*76.3 tonnes*
BLOOM	176B	1.2	57 gCO_2eq/kWh	433 MWh	25 tonnes	30 tonnes

Table 4 : Comparison of carbon emissions between BLOOM and similar LLM. Numbers in *italics* have been inferred based on data provided in the papers describing the models.

Avant de se lancer, une première estimation peut être faite en consultant les données publiées par les fournisseurs de services cloud, lorsqu'elles sont publiques. Une autre méthode d'estimation peut se faire via des outils de modélisation comme ML CO_2 Impact [180], qui permettent de comparer les serveurs, les *cloud providers* et les localisations de ces derniers.

D'autres tests, plus précis, peuvent être faits en exécutant directement le code instrumenté par des outils de mesure tels que Code Carbon [181] ou Carbon Tracker [182], par exemple.

En fractionnant la charge de travail (par l'intermédiaire d'un extrait du jeu de données, par exemple), il est possible d'évaluer l'impact de l'entraînement sur une configuration cible. Le type de machine virtuelle utilisée (mémoire demandée, quantité et modèle de GPU, espace de stockage disque), la région du *data center* et les échanges réseau nécessaires à l'entraînement seront les paramètres principaux influant sur la consommation électrique.

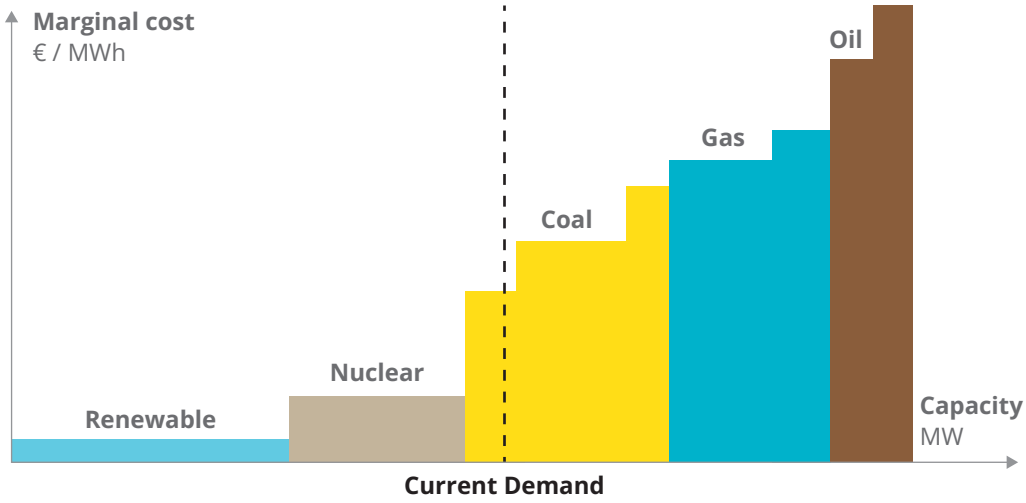
Une fois la configuration idéale trouvée, lancer l'entraînement complet avec ces outils de mesure de consommation énergétique et d'impact environnemental permet de constituer un historique, qui sera utile lors de ré-entraînements futurs, et donc de bâtir une expérience utile pour de futures optimisations.

Et à l'utilisation ?

Durant la phase d'utilisation des modèles, nous retrouverons les mêmes principes qu'à la phase de conception et de développement. La réduction de l'impact environnemental s'approxime assez bien avec un objectif de réduction des coûts.

[179] https://arxiv.org/pdf/2211.02001.pdf
[180] https://mlco2.github.io/impact/
[181] https://codecarbon.io/
[182] https://pypi.org/project/carbontracker/

En effet, l'infrastructure supportant l'exploitation des modèles peut être optimisée pour se faire au plus près de ses futurs utilisateurs. De la même manière que pour le service de contenus multimédias classiques, opérer un modèle proche des données utilisées permettra de minimiser les coûts liés au réseau, en plus de pouvoir sélectionner des *data centers* en fonction du mix énergétique par lesquels ils sont alimentés. Cocorico, la France proposant en Europe une des énergies les moins carbonées, il est donc optimal d'héberger une application intégrant de l'IA dans l'hexagone plutôt que de choisir le *data center* par défaut proposé par AWS ou Microsoft Azure.



Estimation de la source de puissance marginale, Electricity Map [183].

Pour des traitements discontinus, opérés sous forme de *batchs*, il est également possible de moduler le *data center* dans lesquels ils seront exécutés, et de les planifier à des périodes de temps où la demande en énergie est moins importante, afin d'optimiser la consommation marginale du traitement. Le graphique ci-dessus permet de jauger quel type d'énergie est disponible à un moment donné, couplé à la sélection de la géographie du *data center* dans lequel le traitement exécuté, l'impact peut être grandement réduit. **Cependant, il est important de noter qu'à grande échelle si ce comportement se généralise sans contrôle, cela peut avoir un impact négatif sur l'environnement.** Il est possible que le déplacement de la demande dans de nouvelles zones géographiques augmente la quantité d'équipement pour répondre à la demande (plus de *data centers* et plus de serveurs). Cette augmentation des usages peut alors entraîner une augmentation de l'impact du mix électrique puisque la demande augmente. Il est nécessaire d'avoir une stratégie globale de coopération entre tous les acteurs pour optimiser ces types d'usages.

Enfin, il est aussi possible de modifier l'usage de ces algorithmes, pour ne privilégier les calculs lourds qu'aux cas d'usages critiques, et préférer l'utilisation de résultats déjà calculés pour les usages moins importants. Un exemple de cette sobriété à l'usage est le site thispersondoesnotexist.com[184] qui jusqu'à avril dernier permettait de générer des portraits de personnes à la demande, avec une optimisation ingénieuse. En effet, dans le but de ne pas provisionner d'infrastructure trop coûteuse (et donc consommatrice d'électricité), le site ne générait qu'une seule photo par seconde, la servant à tous les utilisateurs simultanés. Vu l'affluence de ce site, l'impact de cette limitation fut non négligeable. D'autres optimisations du genre, comme le *caching* de

[183] https://www.electricitymaps.com/blog/marginal-carbon-intensity-of-electricity-with-machine-learning
[184] http://thispersondoesnotexist.com

résultats déjà calculés pour servir des demandes équivalentes peuvent être mises en place pour réduire l'impact d'un modèle.

Opportunités de réduction liées au rôle de la direction des entreprises

L'IA générative semble ouvrir de nouvelles possibilités créatrices de valeurs économiques : améliorer la durabilité des services, quantifier l'impact de son activité, favoriser l'innovation, la créativité, créer davantage de produits et de services, développer de nouveaux talents, etc. Elle attire autant qu'elle divise. Son utilisation et champ d'application semble aussi évident qu'inquiétant et comme pour chaque innovation, elle est suivie de toutes les projections les plus folles et révolutions annoncées.

L'IA générative arrive sur le même temps qu'un début de prise de conscience pour les entreprises des enjeux de réduction de l'empreinte écologique des sociétés.

Selon une étude Ekimetrics[185]* auprès de 314 dirigeants d'entreprises, l'IA générative semblerait laisser perplexe ce panel de décideurs. Des freins à l'utilisation d'une IA générative seraient portés par les besoins de :

- disposer de davantage d'informations sur son fonctionnement ;
- recevoir des garanties sur la protection de leurs données personnelles ;
- avoir des garanties sur le potentiel de l'IA générative à améliorer la performance économique de leur entreprise.

Ces freins semblent avant tout résulter d'une volonté de maintien des coûts économiques et de performance, de maîtrise des risques liés à la conformité et à la cybersécurité.

Mais l'IA générative est-elle compatible avec les enjeux écologiques de notre époque ? Les entreprises s'engagent dans des démarches RSE pour satisfaire parfois davantage à des réglementations qu'à des convictions.

La sobriété numérique appelée depuis plusieurs années et portée souvent par des équipes informatiques doit s'intégrer dans une stratégie globale RSE soutenue par les équipes dirigeantes et managériales, au risque de perdre la plupart des arbitrages entre la sobriété numérique et le « *business as usual* ».

La priorisation de la sobriété numérique ne peut pas s'inscrire dans un capitalisme responsable.

Les dirigeants sont certes de plus en plus sensibilisés positivement à l'aspect énergivore de l'IA et au ROI des projets. Mais fréquemment, cette sensibilisation aura d'abord été portée par les équipes de développement informatique déjà sensibles à ces sujets depuis plus longtemps.

Il est important de ne pas faire porter uniquement aux équipes DSI la responsabilité de penser à intégrer la réduction de l'empreinte écologique dans les processus de développement.

Il est nécessaire de former les dirigeants, décideurs, pour rentrer dans une économie régénérative et utiliser les nouvelles technologies selon des usages maîtrisés. La plupart des décideurs ont quitté un environnement d'apprentissage

[185] https://ekimetrics.com/fr/studies

depuis plusieurs années, et les formations dispensées il y a une vingtaine d'années ne reflétaient pas ces réflexions de sobriété. À titre d'exemple, le Groupe SNCF a mis en place un « label dirigeant numérique » auprès de l'ensemble de ses dirigeants et managers via des modules d'e-learning sur des sujets : conduite de projets, l'IoT, en passant par la cybersécurité et la sobriété énergétique.

Faire remonter au plus haut niveau la sensibilisation aux enjeux environnementaux des évolutions technologiques, devrait permettre aux dirigeants de s'interroger sur leur relation au numérique et usages associés. En supposant qu'une interrogation personnelle puisse conduire à des décisions d'entreprises qui permettraient aux comités exécutifs de mieux maîtriser leurs orientations numériques et leurs usages de technologie, telle que l'IA générative.

Faire remonter au plus haut niveau la sensibilisation des évolutions technologiques et des enjeux environnementaux, devrait permettre aux dirigeants de s'interroger sur leur relation au numérique et usages associés. En supposant qu'une interrogation personnelle puisse conduire à des décisions d'entreprises qui permettraient au COMEX de mieux maîtriser ses orientations numériques et ses usages de technologie, telle que l'IA générative.

La sobriété numérique n'appelle pas à renoncer à l'usage de l'IA générative, mais à la comprendre dans sa globalité pour mieux l'exploiter et replacer l'intérêt d'en disposer au cœur des réflexions, plus que la stricte efficacité énergétique.

Les recommandations de Data For Good pour les **décideurs**

Prendre en compte l'aspect environnemental dans l'encadrement des modèles génératifs. S'agissant de technologies nouvelles, il est encore difficile d'y voir clair concernant les gains et pertes en termes de performances et d'impact environnemental. Il faut s'assurer de la transparence des usages, notamment afin de pouvoir arbitrer avec des modèles moins gourmands.

Limiter la démultiplication des solutions et acteurs qui multiplie aussi l'impact environnemental. L'écosystème technique doit favoriser le développement de solutions moins gourmandes, de solutions partagées et éviter le gaspillage de ressources ou la réalisation de travaux redondants.

Rendre transparent l'usage des modèles à des fins publicitaires / politiques, notamment contre le *greenwashing*. Un des principaux risques engendrés par les modèles génératifs est celui de désinformation, par la répétition massive d'informations partiellement ou totalement erronées. Si nous avons vu que ce risque n'est pas propre à l'environnement (partie 1), les modèles génératifs pourraient jouer un rôle principal dans la montée du climato-scepticisme. A l'inverse, il faudrait privilégier les usages For Good.

Conclusion sur les enjeux environnementaux

L'urgence climatique actuelle nous oblige à repenser nos modes de vie, et ce dans toutes les dimensions qu'elles soient sociales, économiques ou techniques. Le secteur de l'intelligence artificielle générative ne fait pas exception, et contribue négativement au réchauffement climatique.

Nous avons vu que l'entraînement et l'utilisation des modèles d'IA générative consomment beaucoup d'énergie, de ressources informatiques et que la quantification précise des impacts environnementaux reste encore une tâche difficile. Les études d'impacts sont souvent trop simplistes, car elles ne prennent en compte qu'un périmètre très limité, se concentrent uniquement sur l'impact carbone en ne considérant qu'un seul critère, et négligent l'ensemble du cycle de vie. Elles prennent rarement en compte l'inférence des modèles génératifs qui, une fois passé à l'échelle de nombreux utilisateurs, peut être plus conséquente que la phase d'entraînement. De plus, elles ne tiennent pas compte des impacts indirects ou des effets rebond de ces technologies, qui ont tendance à accroître davantage l'empreinte environnementale de l'IA générative.

Cependant, il existe des méthodes permettant de réduire les impacts associés à ces technologies. Tout d'abord, les optimisations techniques peuvent être intégrées lors du développement de nouveaux services. L'accès aux technologies en *open-source* est un levier essentiel pour réduire l'impact en mutualisant les efforts et les résultats. Néanmoins, le meilleur moyen d'action consiste à limiter les utilisations, ou a minima à nous interroger sur l'intérêt de nos utilisations parfois compulsives. Ces technologies ont déjà prouvé leur grande utilité dans de nombreux cas d'application, mais sont-elles nécessaires dans tous les domaines ? Enfin, l'appareil législatif des gouvernements peut également jouer un rôle majeur dans la réduction de l'impact, à condition d'être utilisé de manière appropriée.

En somme, en adoptant des pratiques, des comportements responsables et en collaborant, nous pouvons façonner un avenir durable où l'IA générative contribue positivement à notre environnement et à la société.