Comment évaluer les impacts écologiques d'un projet en IA ?

Workshop frugalité en IA et en statistique
4 octobre 2024

Anne-Laure Ligozat

Pourquoi l’IA ?

IA ?

Intelligence artificielle

Apprentissage automatique
(*machine learning*)

Apprentissage profond
(*deep learning*)

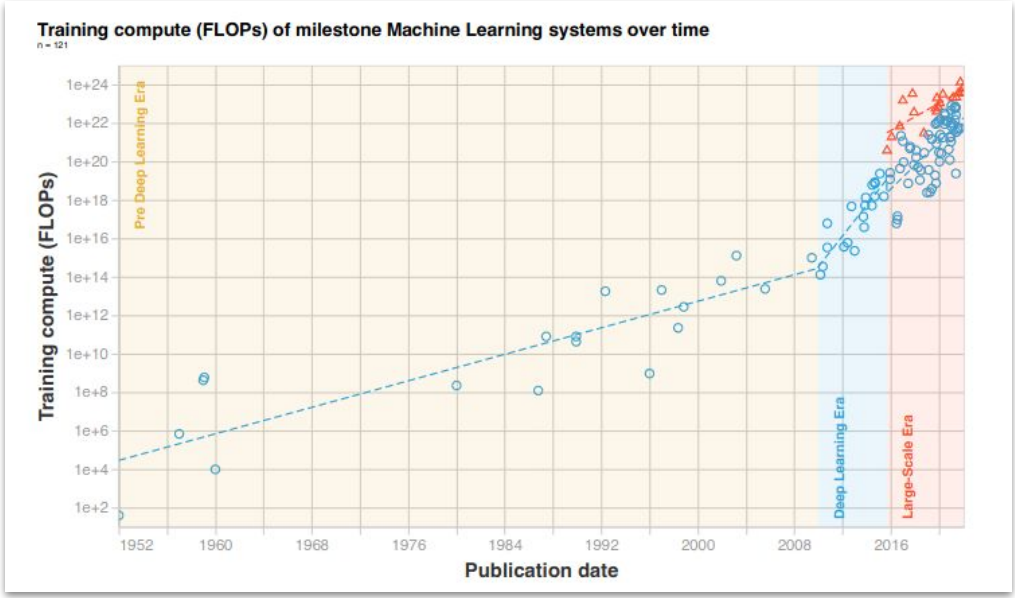
Pourquoi s’intéresser aux impacts de l’IA ?

impacts environnementaux potentiellement importants :

- grande quantité de **données**
- ressources de **calcul**

souvent présentée comme **solution**

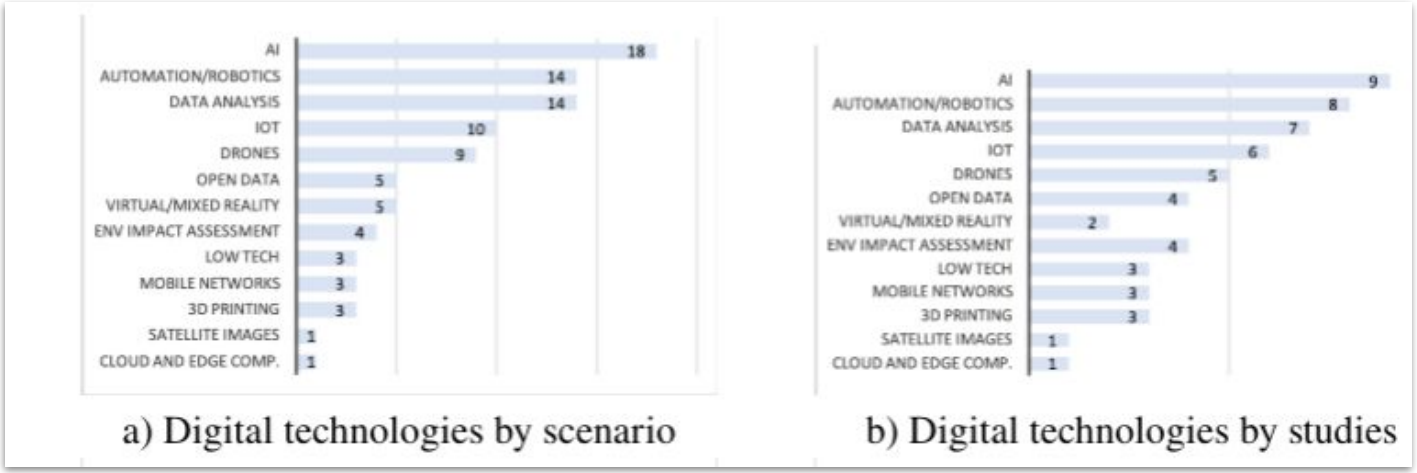
… sans prendre en compte ses impacts négatifs



(Sevilla et al., 2022)

L'IA comme solution aux problèmes environnementaux ?

Analyse d'études prospectives (Bugeau & Ligozat, 2023)



IPCC 2022

Ademe 2022

negaWatt 2021

EU green deal 2019

Eionet 2022

Arup 2019

DDC 2020

SNBC 2020

RTE 2022

Shift 2020

France 2072 2018

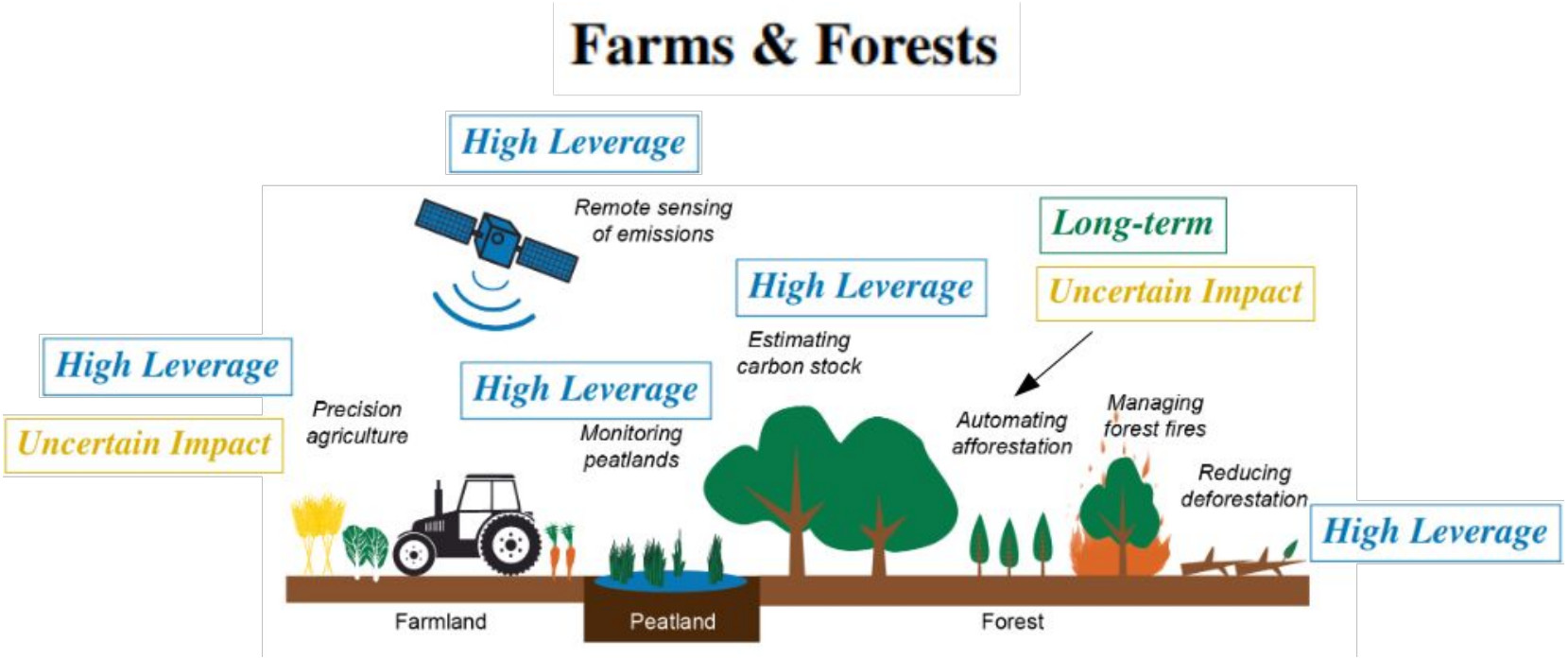
D&A 2022

CNIL 2021

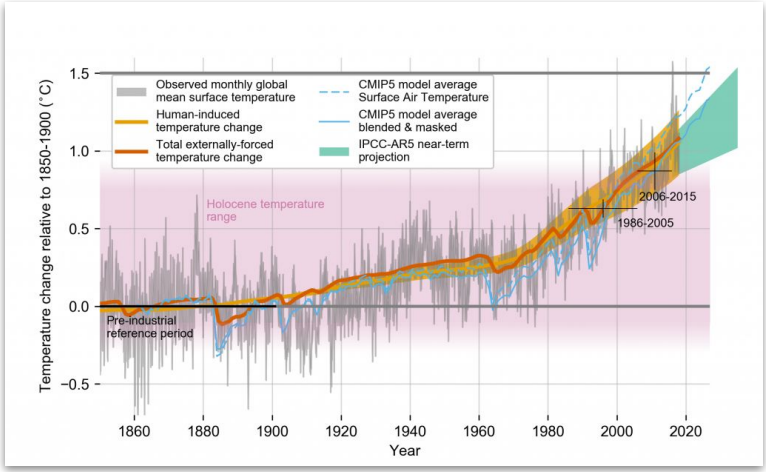
Digit. Challenge 2022

Applications de l'IA à des problématiques environnementales

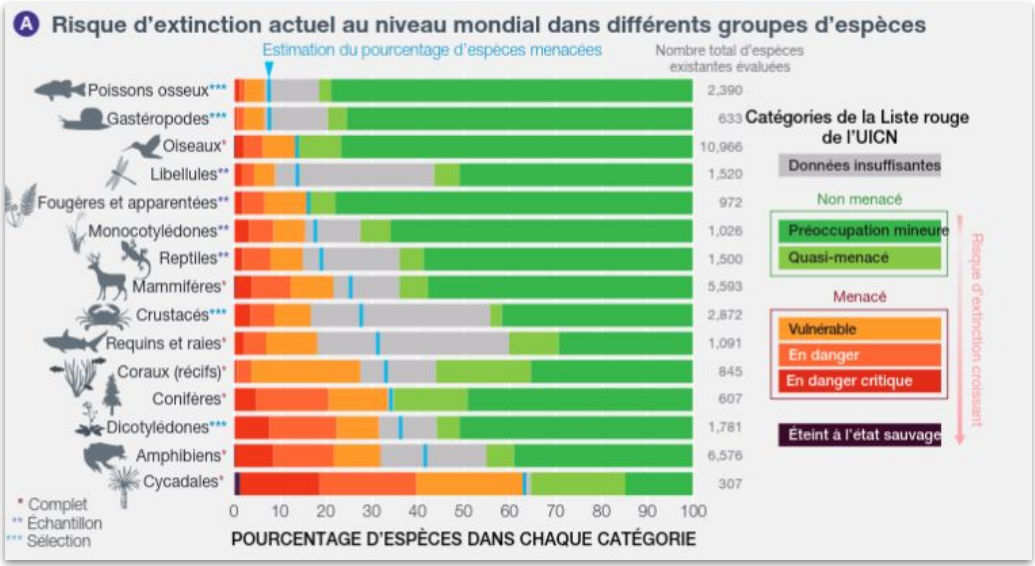
«Tackling Climate Change with Machine Learning» (Rolnick et al., 2019)



Contexte environnemental

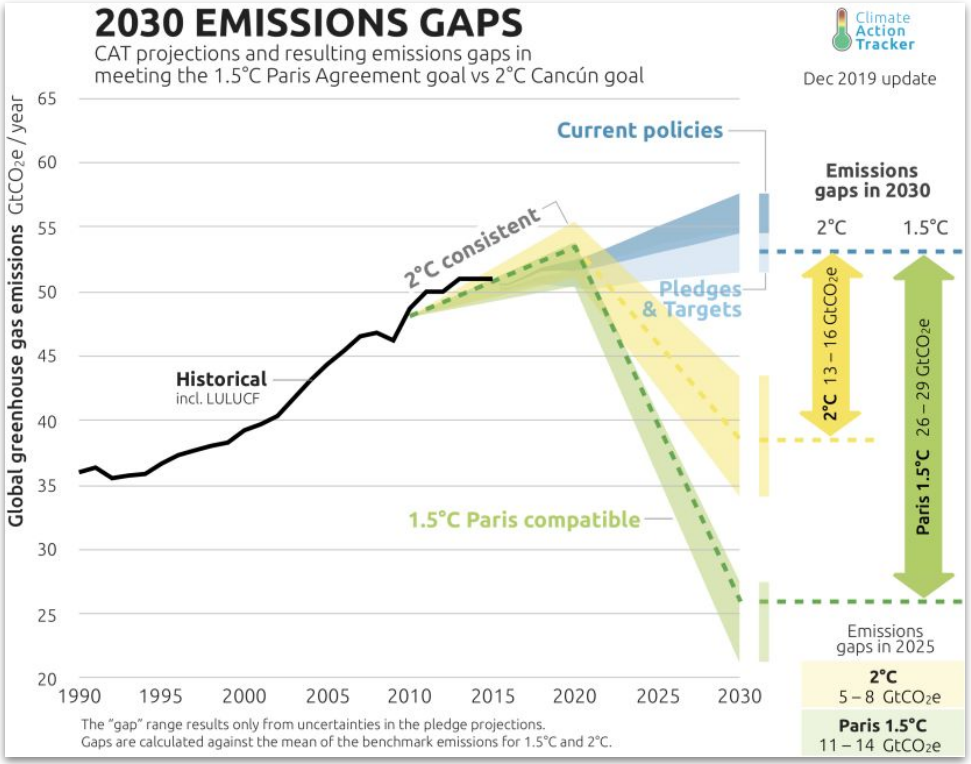


Source: GIEC



Source: IPBES

Contexte environnemental



Calcul des impacts d’un programme d’IA

Approche basique

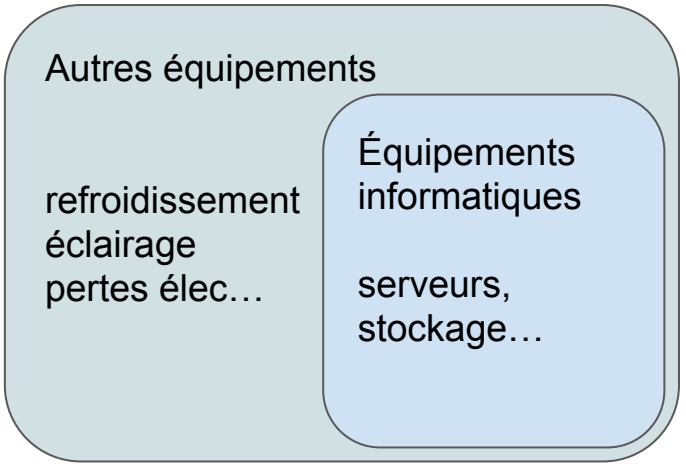
kWh → kg CO_2e

électricité → empreinte carbone

électricité = (CPU + GPU + RAM) x PUE

Efficacité énergétique du centre de calcul

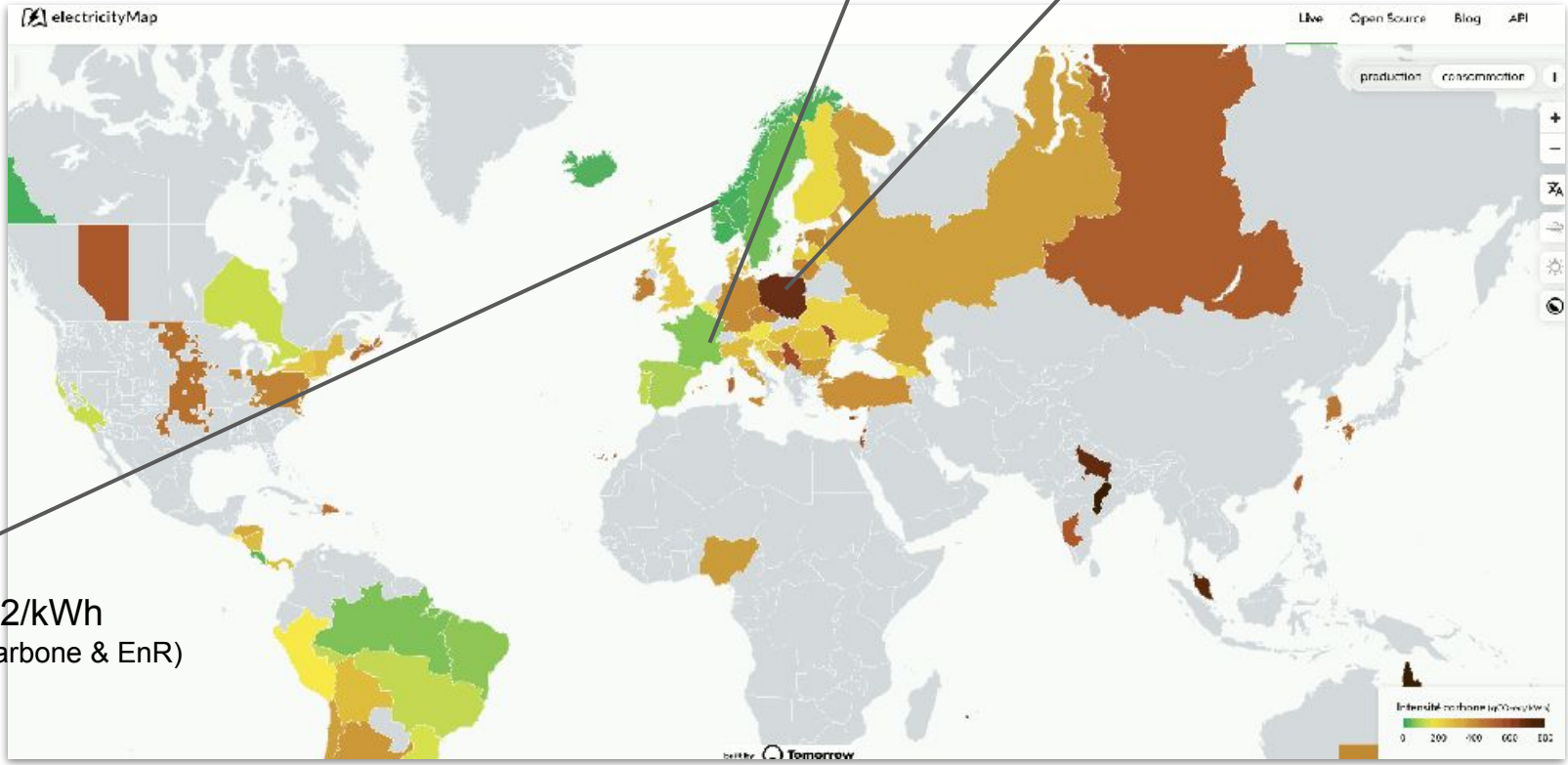
$$PUE = \frac{\text{consommation d'énergie du centre}}{\text{consommation des équipements informatiques}}$$



Facteur d'émission électricité

France : 101g éqCO2/kWh
(86% bas carbone, 13% EnR)

Pologne : 927g éqCO2/kWh
(13% bas carbone, 13% EnR)



Norvège :
22g éqCO2/kWh
(100% bas carbone & EnR)

source: [electricityMap]

Influence du facteur d'émission

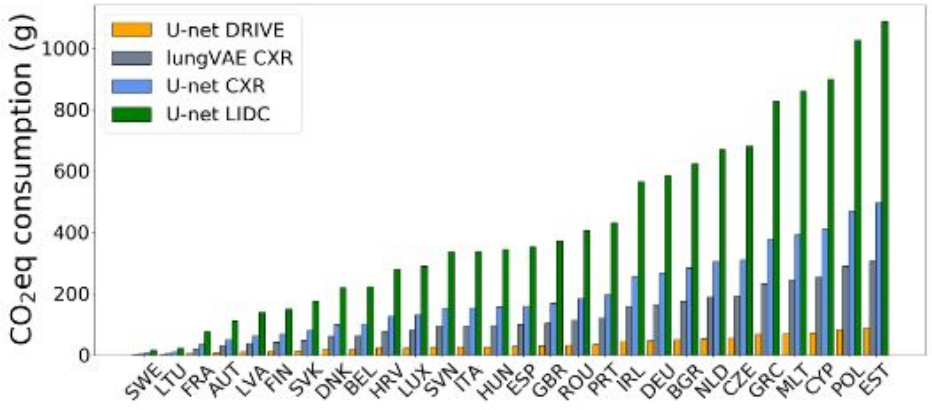
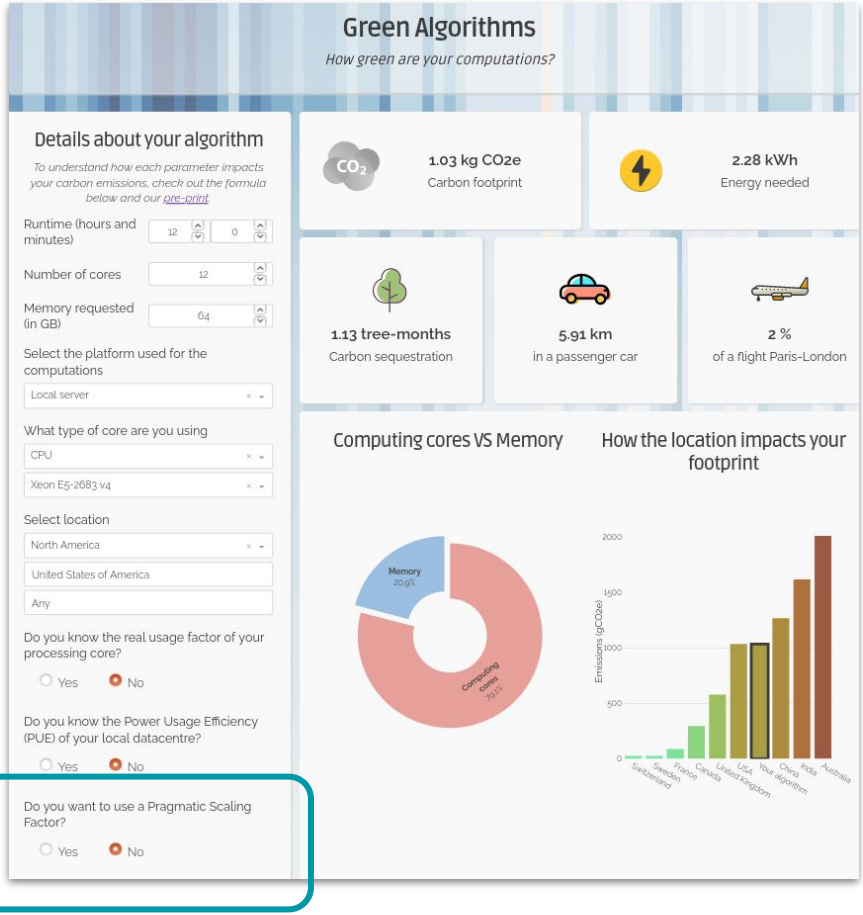


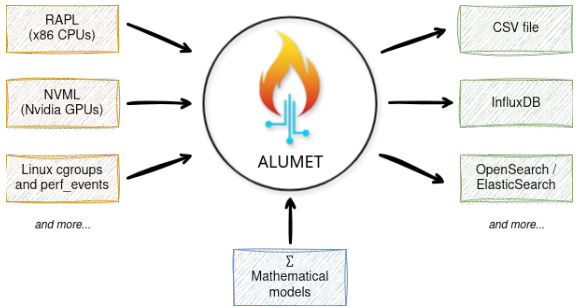
Figure 4. Estimated carbon emissions (gCO_2eq) of training our models (see Appendix B) in different EU-28 countries. The calculations are based on the average carbon intensities from 2016 (see Figure 8 in Appendix).

Source : (Anthony et al., 2020)



CODE CARBON

experiment-impact-tracker



Comparaison d'outils de mesure de consommation

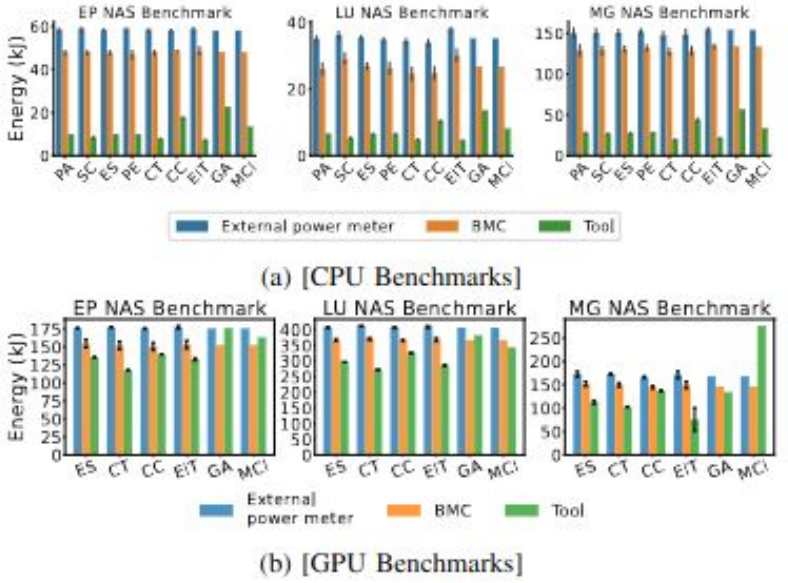
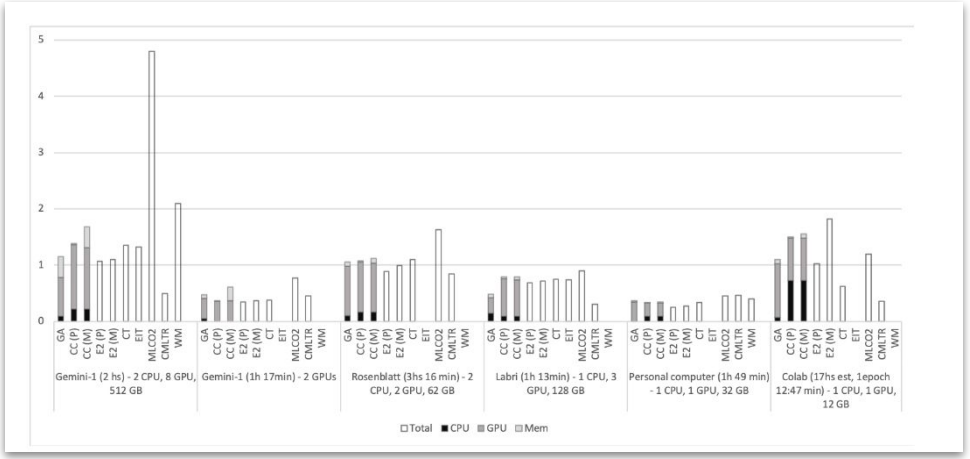


Fig. 2: Total energy consumed by the benchmarks as reported by the power meters. Tools: PowerAPI (PA), Scaphandre (SC), Energy Scope (ES), Perf (PE), Carbon Tracker (CT), Code Carbon (CC), Experiment Impact Tracker (EIT), Green Algorithm (GA), ML CO2 Impact (MCI)

source: (Jay et al., 2023)



source : (Bouza Heguerte et al., 2023)

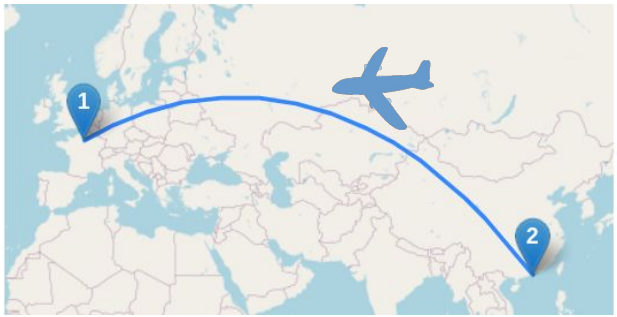
Exemple d’empreinte carbone (Strubell et al, 2019)

4 modèles de TAL/NLP état de l’art

mesure logicielle de la consommation

Résultats des entraînements

- quelques jours à quelques semaines
- émissions: entre 18kg CO2e et 284 000 kg CO2e
- modèle le plus utilisé : 652 kg CO2e, soit
  - un aller Paris-Hong Kong en avion
  - ou 2 500km en voiture



environ 58 GPU pendant 172 jours pour entraîner le modèle…

Précision vs CO2e (Parcollet et Ravanelli, 2021)

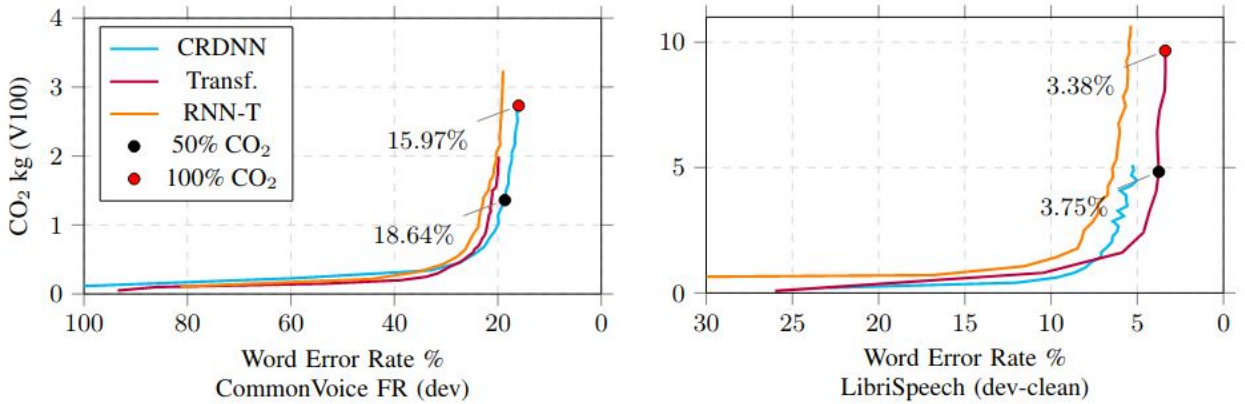
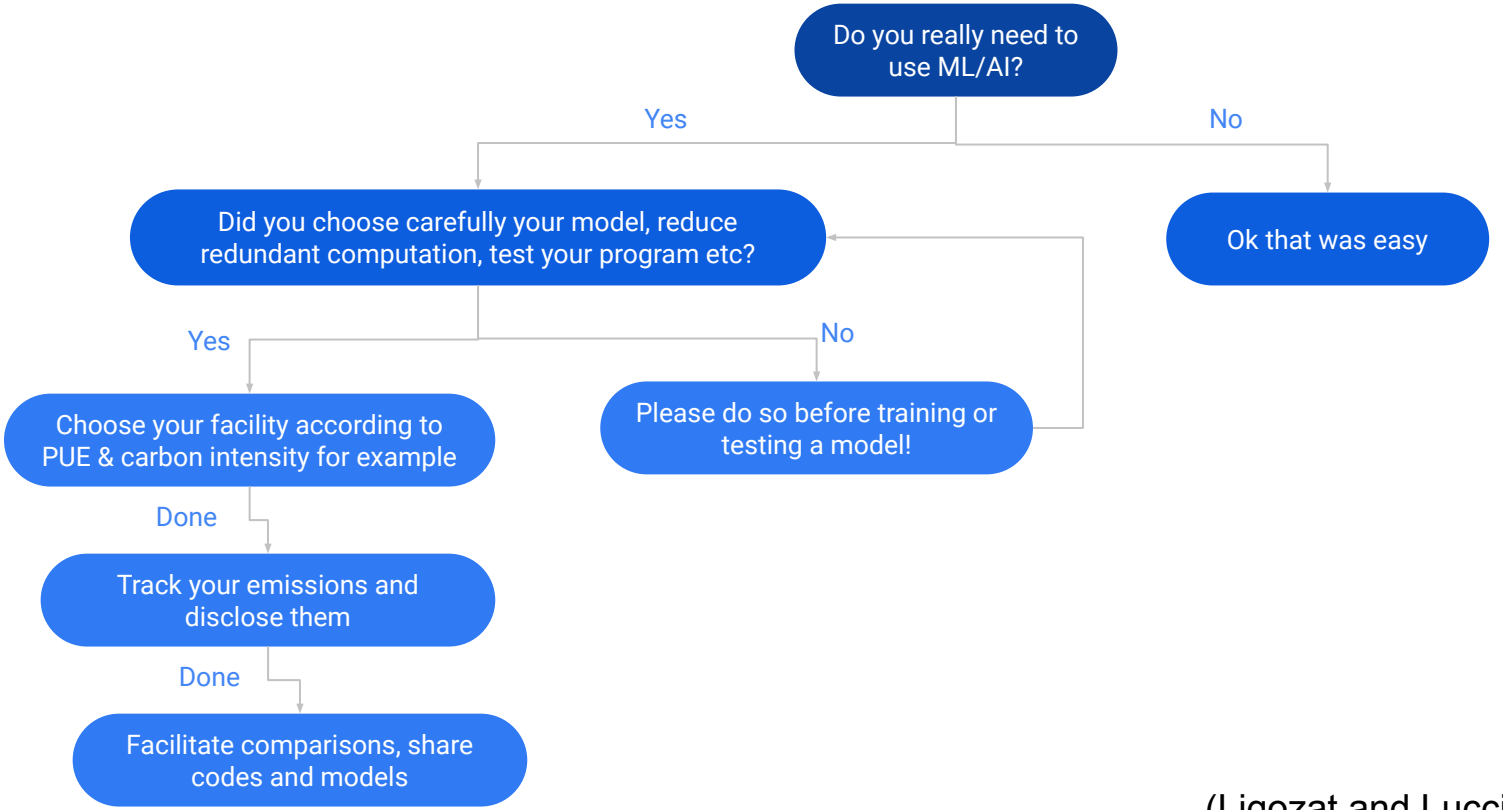


Figure 2: *CO_2 emitted in kg (in France) by different E2E ASR models with respect to the word error rate (WER) on the dev sets of LibriSpeech and CommonVoice. The curves exhibit an exponential trend as most of the training time is devoted to slightly reduce the WER. The black and red dots indicates the WER obtained with 50% and 100% of the emitted CO_2. On LibriSpeech, 50% of the carbon emissions have been dedicated to reach SOTA results with an improvement of 0.37%.*

Comment réduire l'empreinte carbone de mes calculs ?



(Ligozat and Luccioni, 2021)

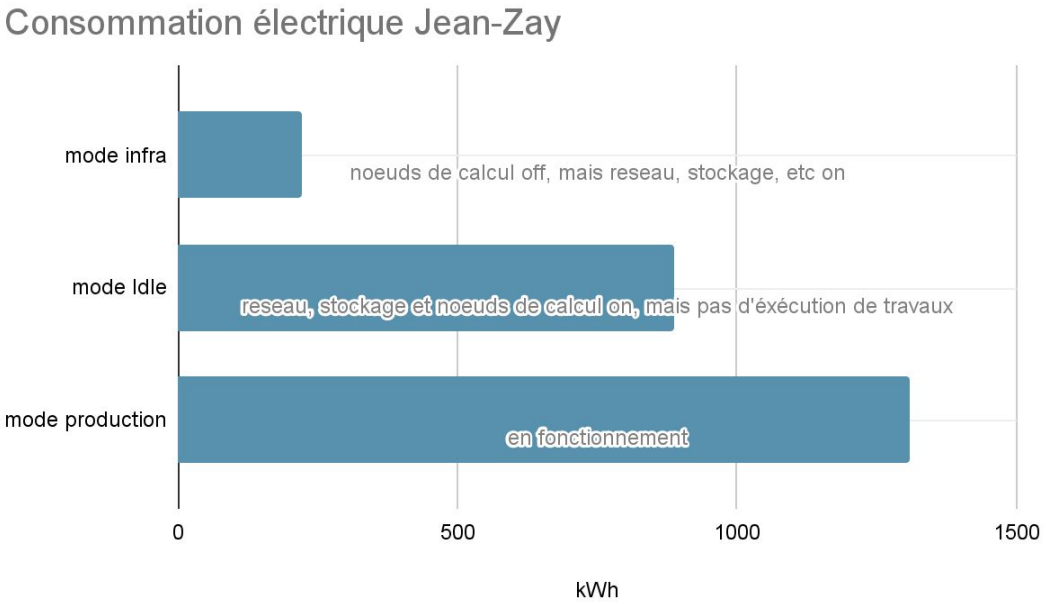
Données environnementales sur lesmodèles (Hershcovich et al, 2022)

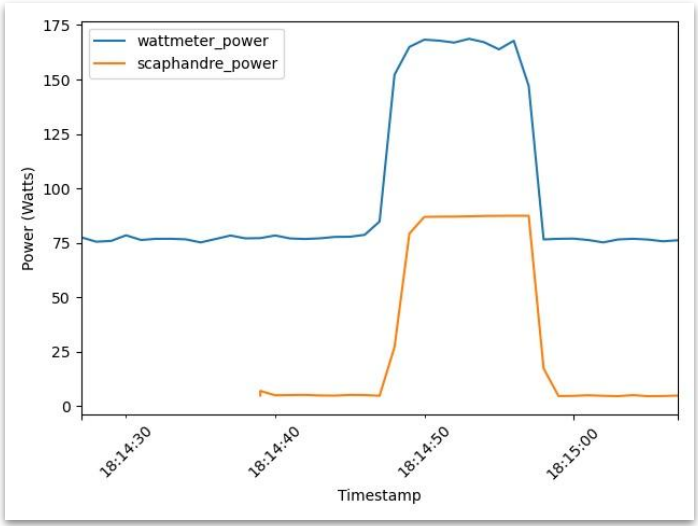
Minimum card	
Information	**Unit**
1. Is the resulting model publicly available?	Yes/No
2. How much time does the training of the final model take?	Time
3. How much time did all experiments take (incl. hyperparameter search)?	Time
4. What was the energy consumption (GPU/CPU)?	Watt
5. At which geo location were the computations performed?	Location

Extended card	
6. What was the energy mix at the geo location?	gCO2eq/kWh
7. How much CO2eq was emitted to train the final model?	kg
8. How much CO2eq was emitted for all experiments?	kg
9. What is the average CO2eq emission for the inference of one sample?	kg
10. Which positive environmental impact can be expected from this work?	Notes
11. Comments	Notes

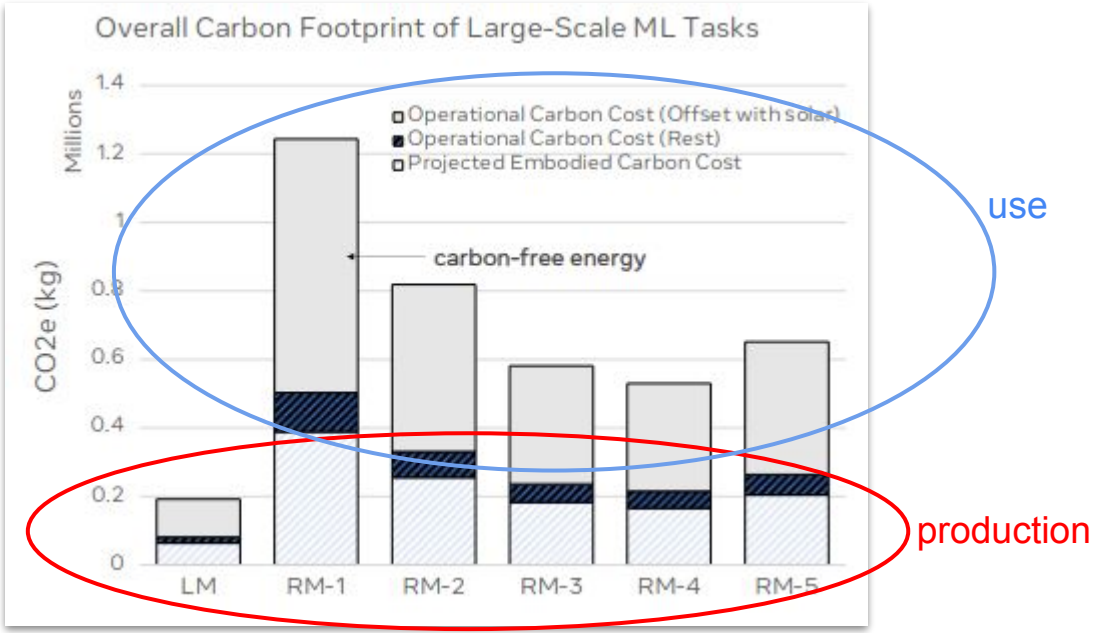
Mesure-t-on la totalité des impacts dus à l’exécution du programme ?

Consommation d’électricité dans un calculateur

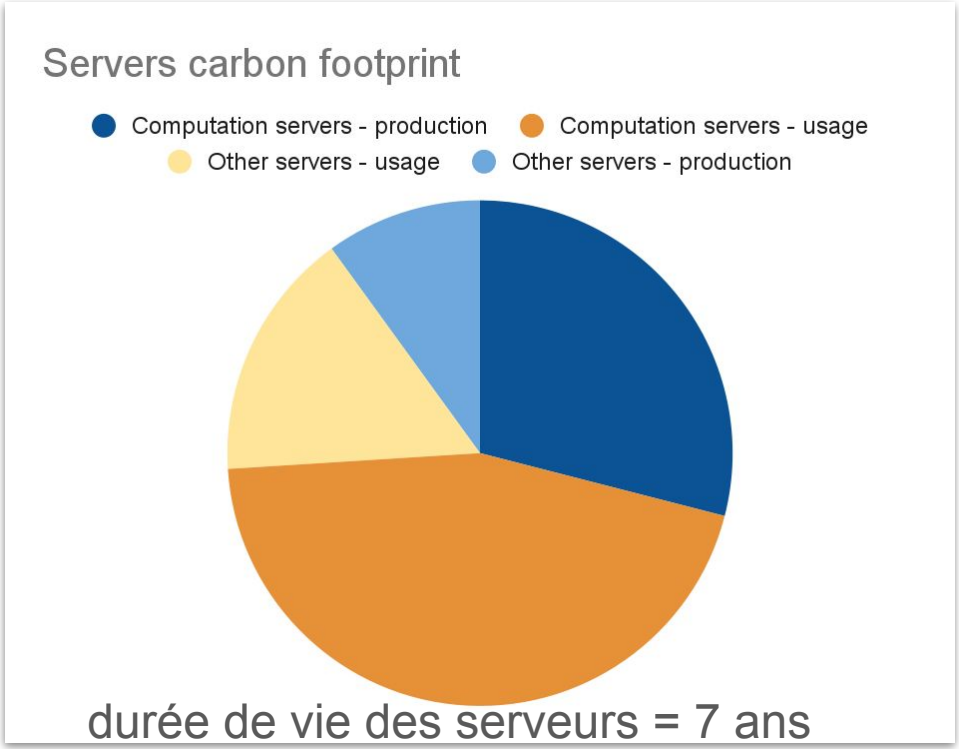




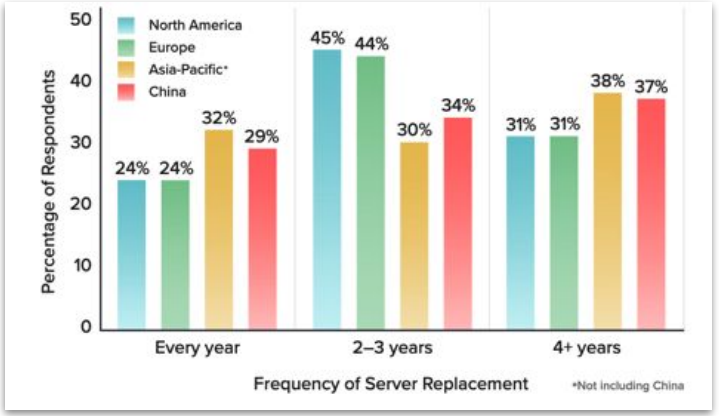
Empreintes carbone de production vs usage en Machine Learning (Wu et al., 2021)



Empreinte carbone des serveurs du GRICAD



Source: (Berthoud et al., 2020)

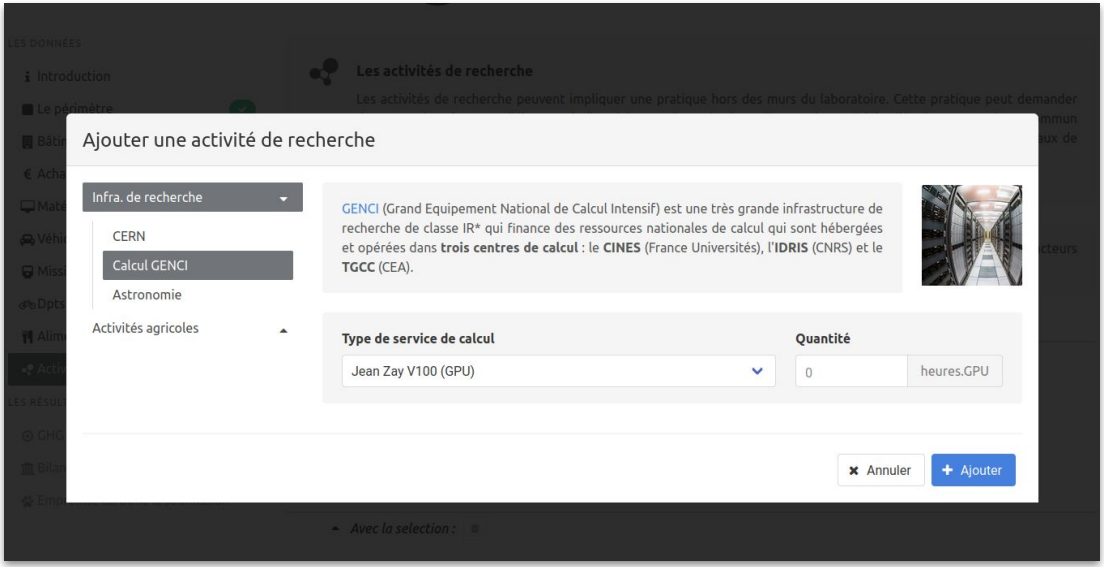


Source : https://www.supermicro.com/white_paper/DataCenters_and_theEnvironmentFeb2021.pdf

Facteurs d’émission des heures de calcul

Travail Labos 1point5 et EcoInfo

Approche top-down : empreinte totale / heures de calcul ⇒ x gCO2e / heure.coeur



Intégrer d’autres indicateurs environnementaux

	ADP	GWP	PE	Human toxicity	Water Consumption	...
Extraction	✓	✓	✓	✗	✗	✗
Manufacturing	✓	✓	✓	✗	✗	✗
Transport	✗	✗	✗	✗	✗	✗
Usage	✓	✓	✓	✗	✗	✗
End of Life	✗	✗	✗	✗	✗	✗

Modeling graphics card

Manufacturing impacts attribution

Infrastructure consumption

Putting impacts in perspective

source: (Morand, 2023)

Exemple : entraînement du modèle BLOOM

(Luccioni et al, 2023)

Empreinte carbone

- 59 t éq CO2
  - émissions annuelles de 59 personnes

Épuisement des ressources minérales

- 1,2 kg éq Sb
  - empreinte annuelle de 38 personnes

Équivalences = modèle des limites planétaires de (Sala et al, 2020)

Exemple : Stable Diffusion

Table 2: Environmental impact of Stable Diffusion for FU1 and FU2

FU	Abiotic Depletion Potential (kgSb eq)	Warming Potential ($kgCO^2$ eq)	Primary energy (MJ)
FU1 - Single use of service	$6.72e^{-08}$	$7.84e^{-03}$	$2.02e^{-01}$
FU2 - A year of service	$4.64e^{+00}$	$3.60e^{+05}$	$8.93e^{+06}$

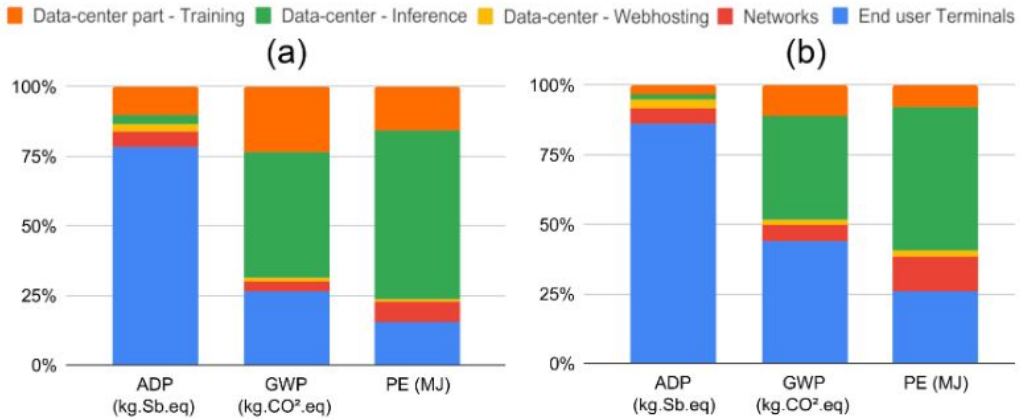


Figure 2: Impact distributions for (a) FU1 and (b) FU2

Source : (Berthelot et al., 2024)

Consommation d’eau

Sources de consommation d’eau (Li et al., 2023) :

- circuits ouverts de refroidissement à l’eau
- production d’électricité
- fabrication des équipements

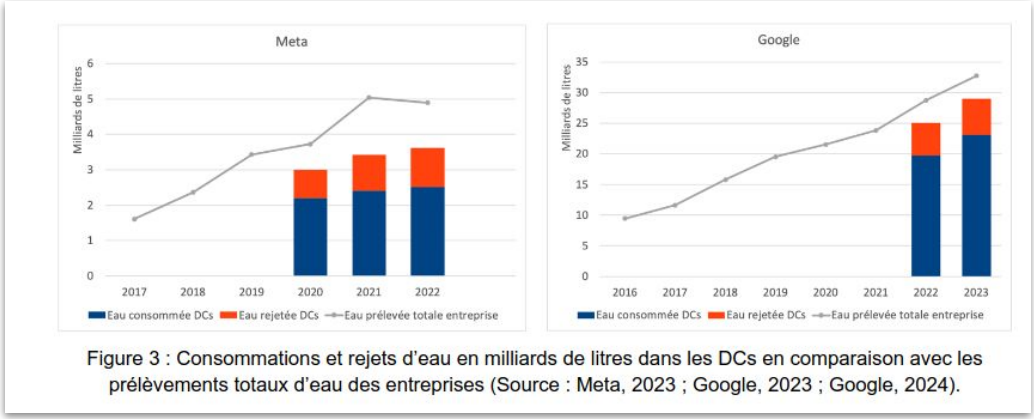
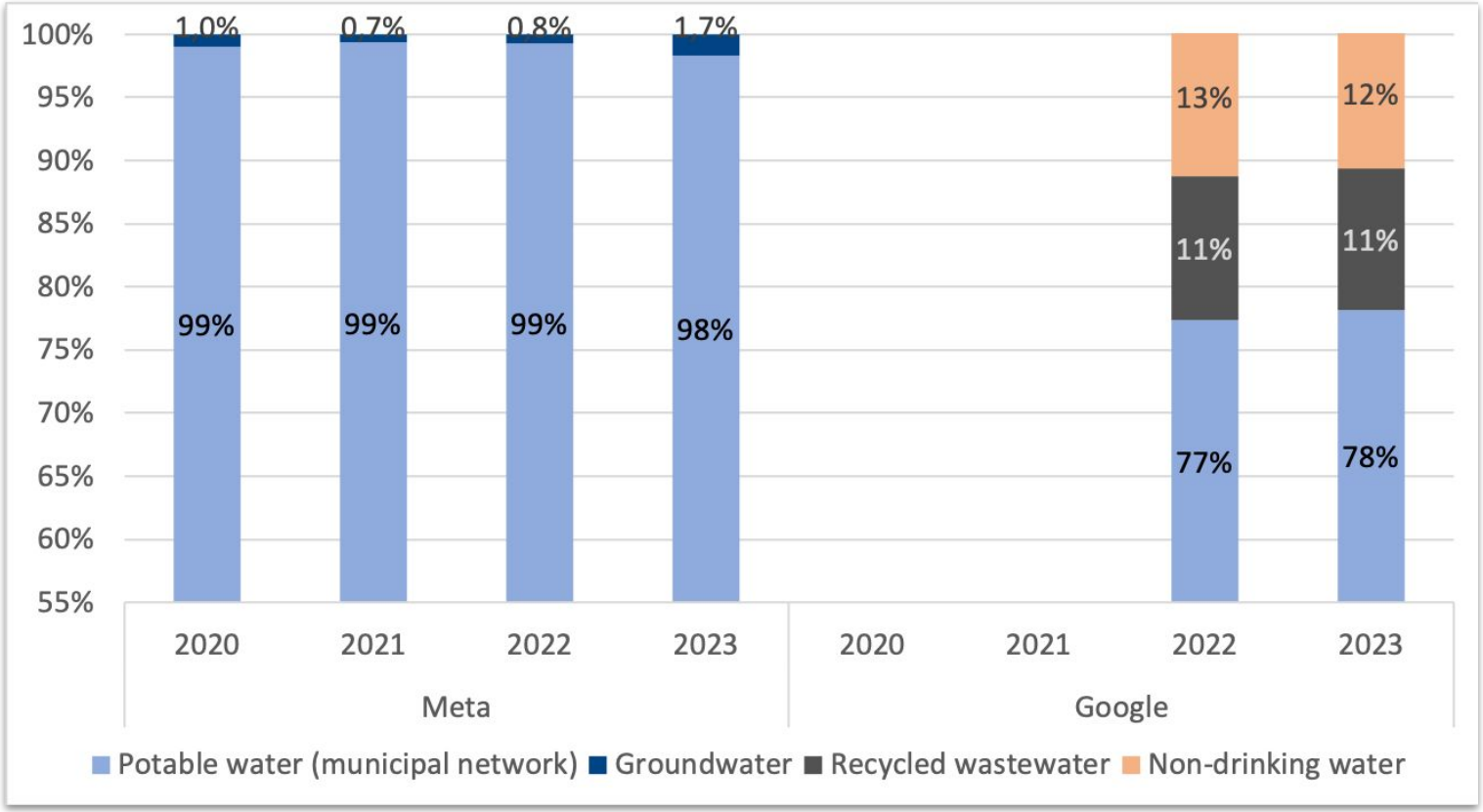


Figure 3 : Consommations et rejets d’eau en milliards de litres dans les DCs en comparaison avec les prélèvements totaux d’eau des entreprises (Source : Meta, 2023 ; Google, 2023 ; Google, 2024).

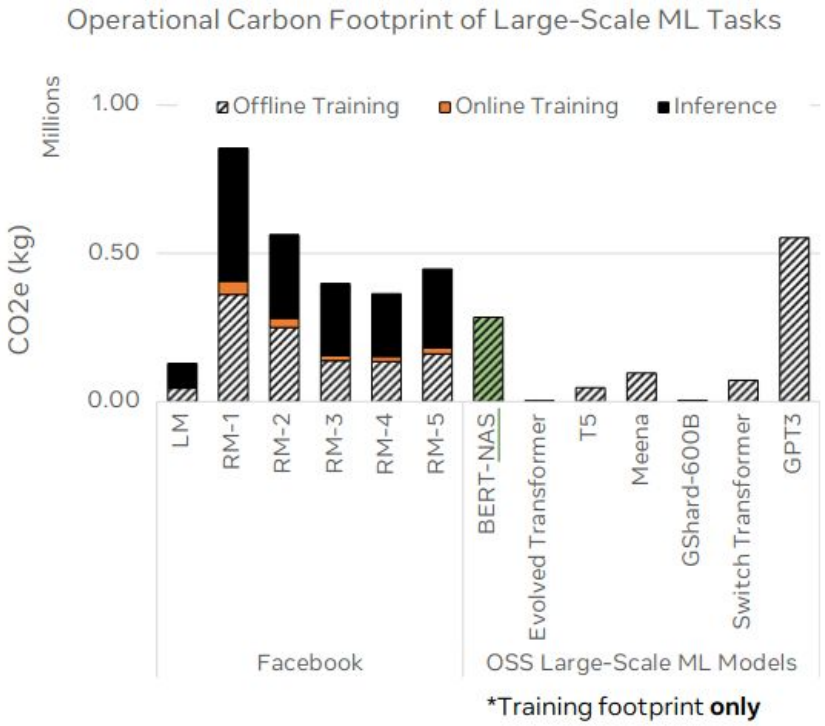
Source : (Bouveret et al., 2024)

Consommation d’eau

source : rapports environnementaux de Meta et Google, données compilées par Aurélie Bugeau



Entraînement vs inférence



Source : (Wu et al., 2021)

Inférence

Source : (Luccioni et Strubell, 2024)

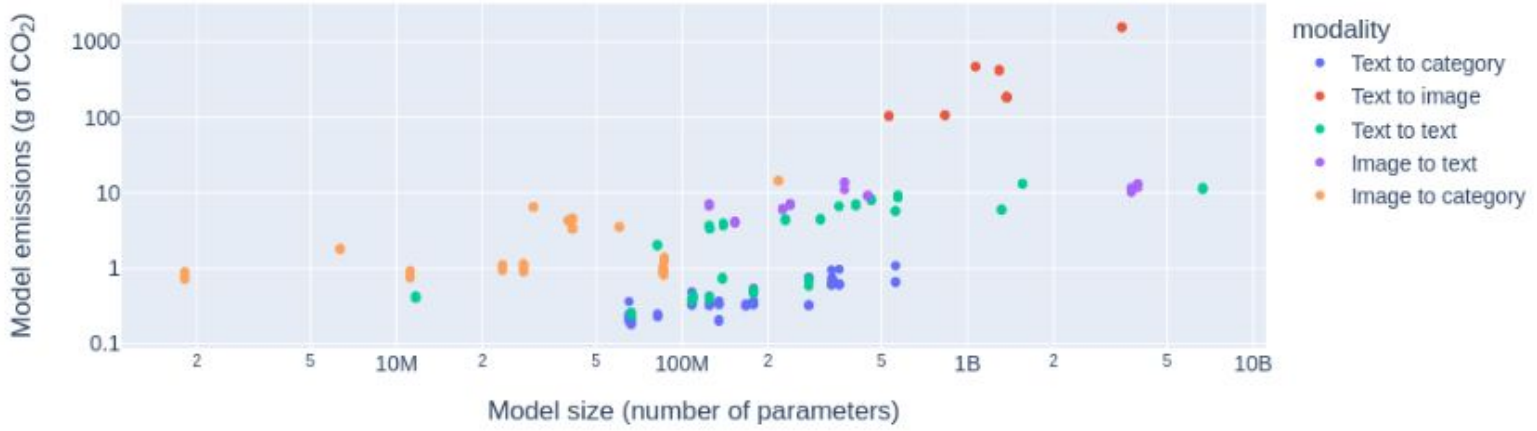
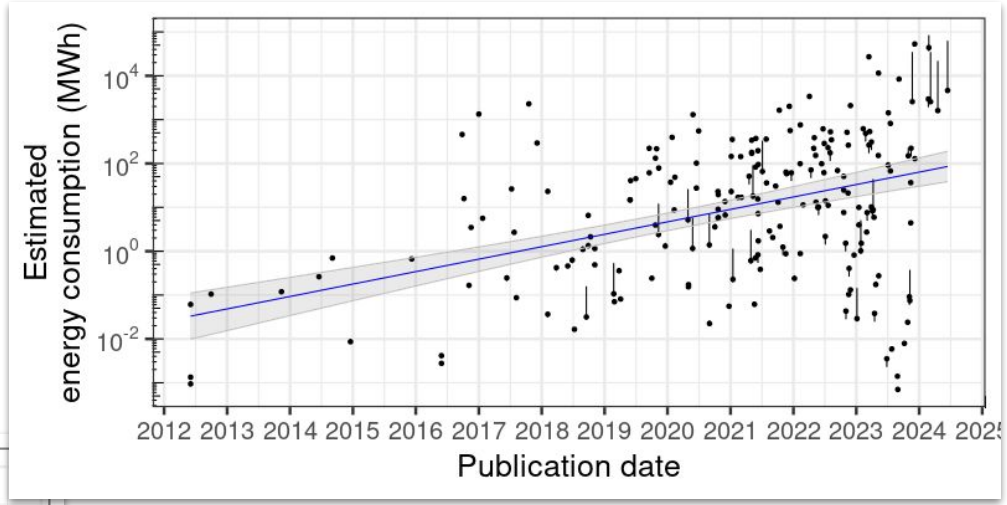


Figure 2: The 5 modalities examined in our study, with the number of parameters of each model on the x axis and the average amount of carbon emitted for 1000 inferences on the y axis. NB: Both axes are in logarithmic scale.

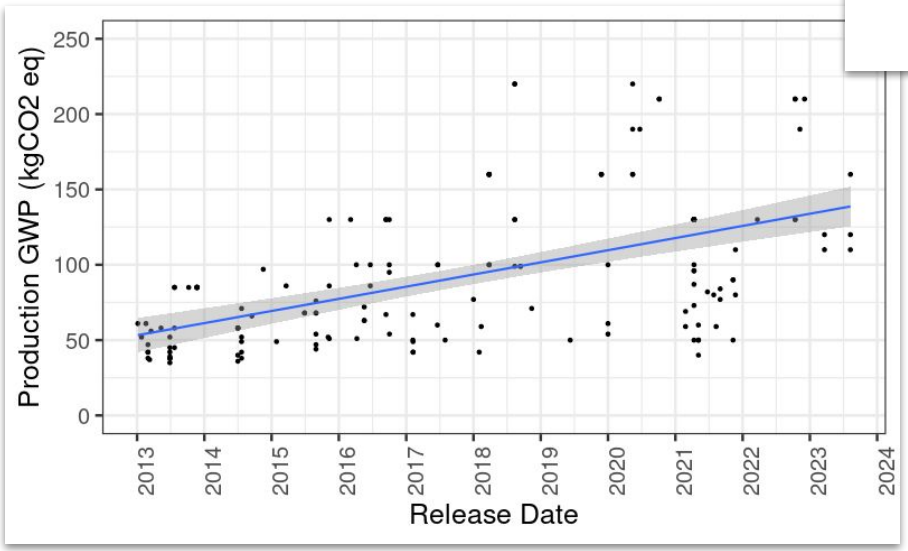
Évolution de l’IA

(Morand, à paraître)



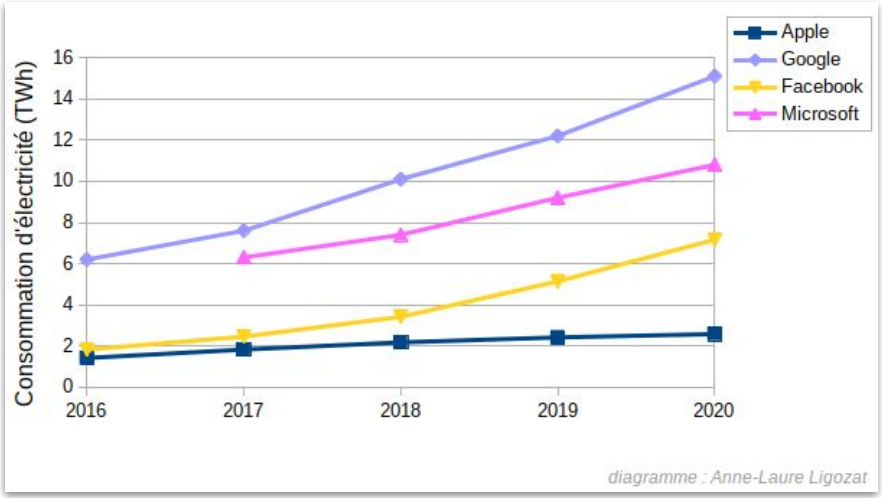
consommation électrique des modèles d’IA

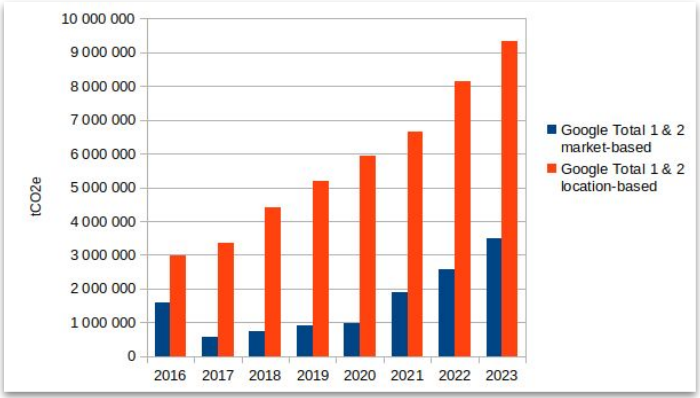
empreinte carbone de production des cartes graphiques



Influence de l'IA dans les empreintes carbone des GAFAM

Google : «As we further integrate AI into our products, **reducing emissions may be challenging** due to increasing energy demands from the **greater intensity of AI compute**, and the emissions associated with the expected increases in our technical infrastructure investment.»





source : rapports environnementaux, données compilées par Anne-Laure Ligozat

Réponse de Google à l'article de (Strubell et al., 2019)

The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink

David Patterson[1,2], Joseph Gonzalez[2], Urs Hölzle[1], Quoc Le[1], Chen Liang[1], Lluis-Miquel Munguia[1], Daniel Rothchild[2], David So[1], Maud Texier[1], and Jeff Dean[1]

Bonnes pratiques proposées :

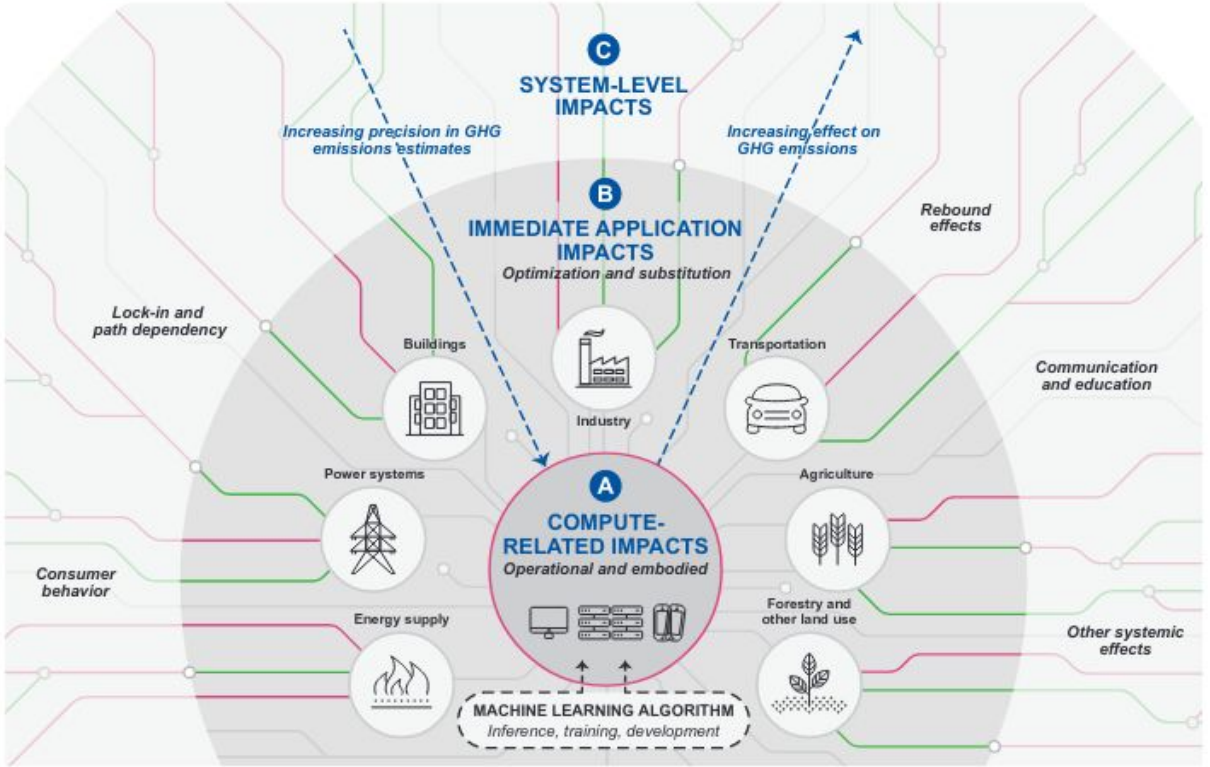
- Modèle efficient
- Processeurs optimisés pour ML
- Cloud pour efficacité énergétique
- Localisation avec mix électrique bas carbone

et pour finir «Google's renewable energy purchases further reduce the impact to zero»

mais :

- quid du cycle de vie ?
  - processeurs récents ⇒ empreinte carbone ↗
- quid de l'inférence ?
- énergie «carbon free» ou «net zero impact» ?
- empreinte carbone potentielle si tout optimisé, et non réelle

Impacts de 1er, 2e et 3e ordre de l'IA



Source : (Kaack et al., 2021)

Impacts indirects

impacts directs

impacts indirects

optimiser le trafic automobile ?

moins de consommation de carburant

priorité aux systèmes avec impacts significatifs ?

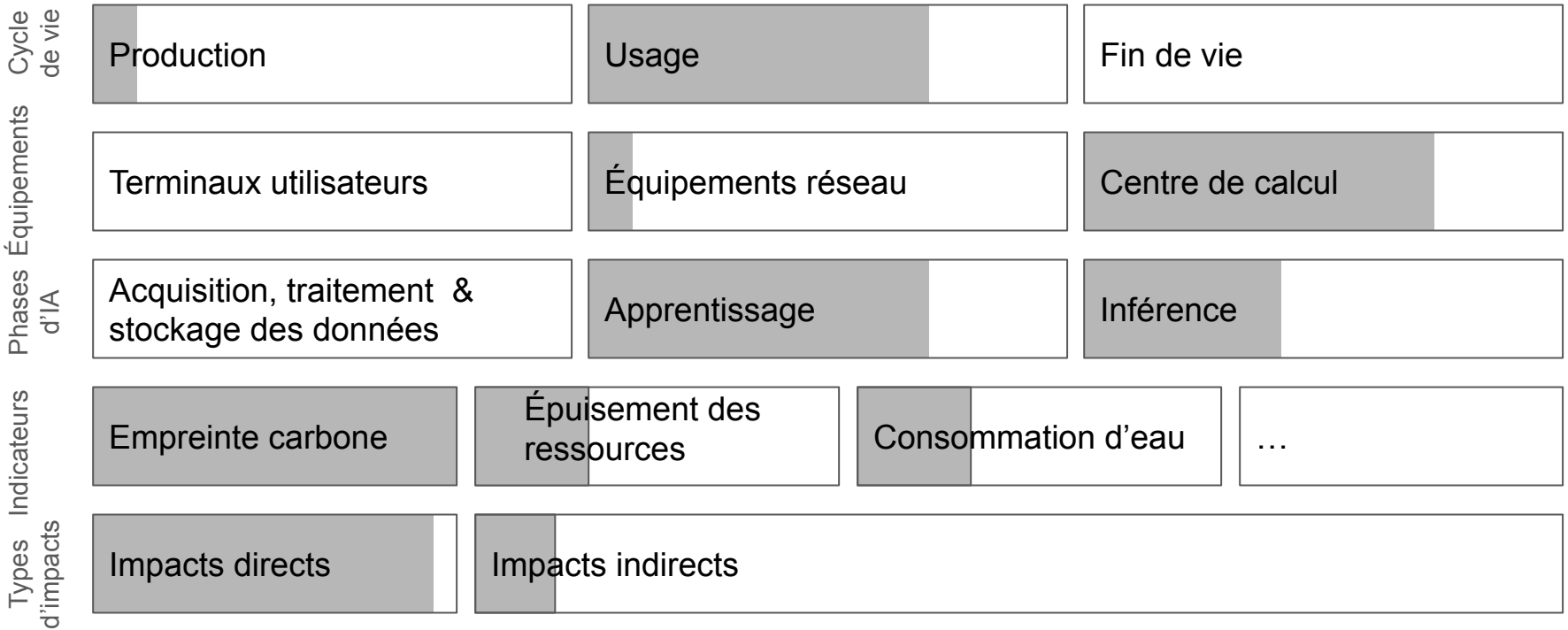
utilisation de nouveaux objets connectés, capteurs…

effet rebond
trafic plus fluide => gain de temps => éloignement du domicile => étalement urbain

dépendance de chemin
prolonge système actuel, vs transports en commun, mobilités actives…

Comment évaluer ?

Ce qu’on calcule actuellement



Spécification AFNOR 2024 IA frugale

Discussions sur :

- périmètre de la spécification (à qui s’adresse-t-elle)
- définition de frugalité (vs efficience): redéfinition des besoins et usages
    - service frugal : inclut IA nécessaire + usages et besoins questionnés et visent à rester dans limites planétaires

	Définition	Notions connexes	Raisonnement	Approche	Précisions
Efficience	Aptitude à optimiser les moyens alloués pour atteindre un résultat défini	Efficacité, optimisation	En relatif/par unité d'usage Le besoin prime : optimisation d'une solution jugée celle répondant le mieux au besoin	Recherche d'un optimum local ou d'un compromis sur un niveau de résultat fortement contraint	Prise en compte des effets de premier ordre pour les minimiser Prise en compte des parties prenantes de l'IA
Frugalité	Aptitude à se contenter d'un niveau de résultat jugé suffisant en redéfinissant les usages et les besoins	Sobriété (ou *Sufficiency* [10] en anglais)	En global La contrainte sur les ressources prime : recherche de la solution utilisant le moins de ressources possible et apportant une réponse satisfaisante au besoin	Recherche d'un optimum global ou d'un compromis large sur un niveau de résultat, ce qui nécessite d'élargir ou d'assouplir le besoin	Prise en compte des effets de premier ordre et de second ordre pour minimiser les impacts environnementaux négatifs Prise en compte de tous les acteurs au-delà des seules parties prenantes de l'IA

Spécification AFNOR 2024 IA frugale

Évaluation

- effets de 1er ordre (multi-indicateurs et multi-phases du cycle de vie)
  - suivant normes ACV ISO, recommandation ITU-T L.1410 et RCP (Référentiel par Catégorie de Produit) Services numériques
  - cf. outil MLCA (Morand et al., 2024)

- mais aussi 2e et 3e ordre
  - suivant recommandation ITU-T L .1480

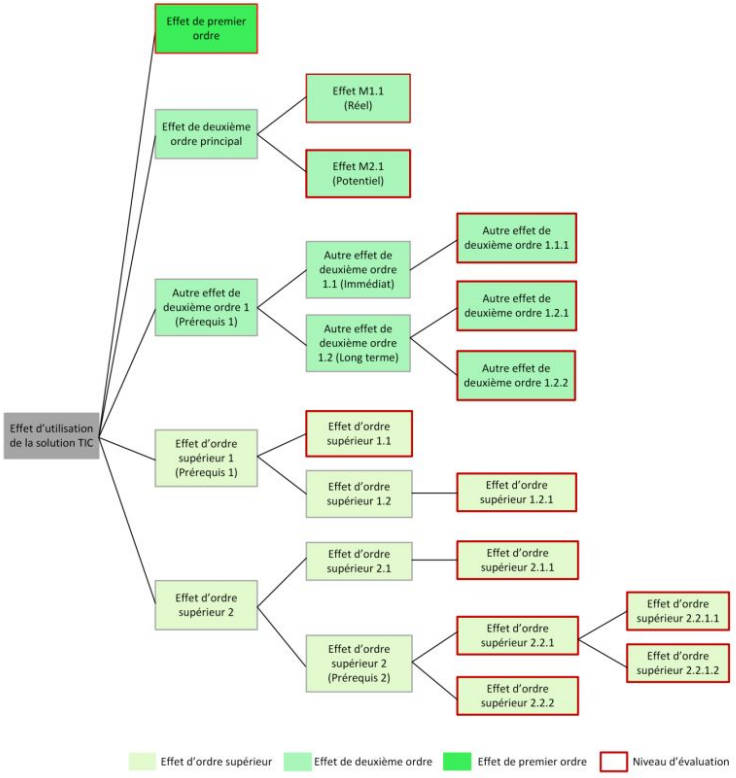
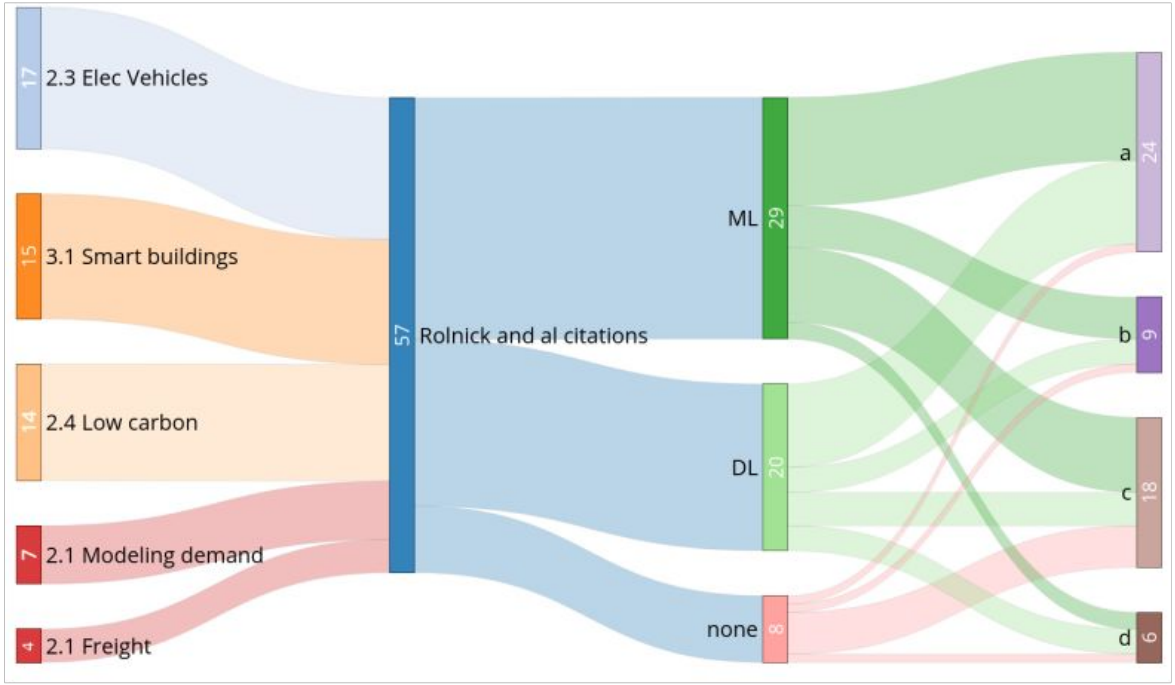


Figure 5 — Arbre de conséquences où l'utilisation du service d'IA entraine sur chaque effet une action ou événement qui a une incidence sur les impacts environnementaux

Evaluations dans applications de (Rolnick et al., 2019)



a. No mention of the environmental gain

b. General mention of the environmental gain

c. A few words about the environmental gain but no quantitative evaluation or only indirect estimation

d. Evaluation of the energy gain without taking the AI program into account

Biais des études d’impact (Rasoldier et al., 2022)

Périmètre

- pas de prise en compte du cycle de vie : (Ligozat et al., 2021) pour l'IA
- pas de prise en compte des effets indirects : 5G

Hypothèses

- comparaison à quel scénario de référence ?

Déconnexion de scénarios globaux

- bénéfices minimes + incertitudes mal gérées
- incompatibilité entre les mesures

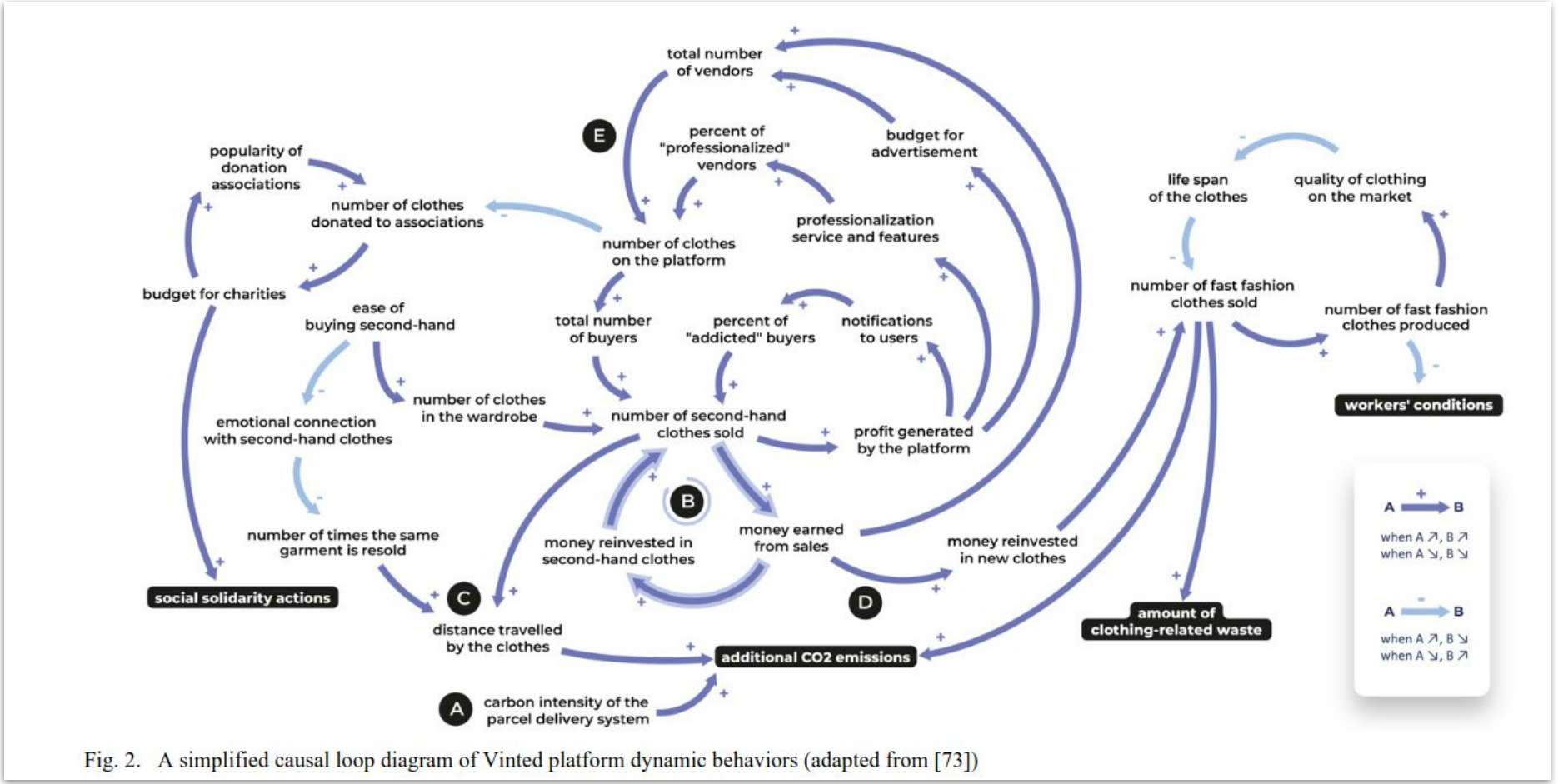
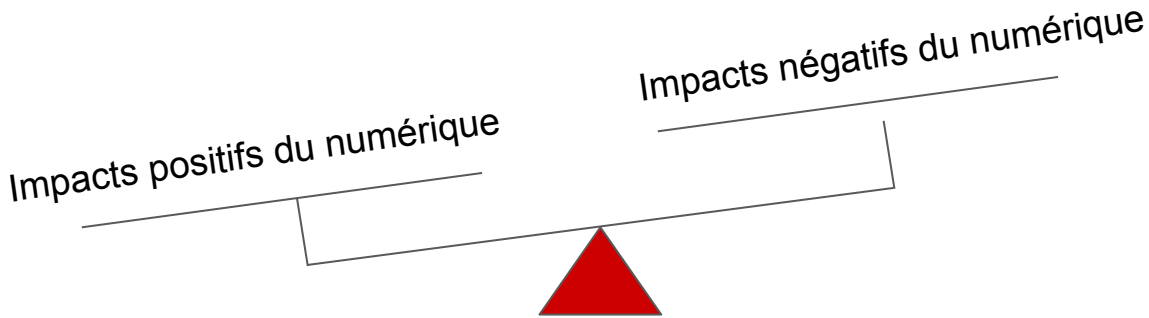


Fig. 2. A simplified causal loop diagram of Vinted platform dynamic behaviors (adapted from [73])

(Ekchajzer et al., 2024)

L’IA pour des applications environnementales



au moins avec des Analyses de Cycle de vie

en prenant en compte autant d’effets indirects que possible

Références

- Berthelot, A., Caron, E., Jay, M., Lefèvre, L. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. CIRP LCE 2024 - 31st Conference on Life Cycle Engineering, Jun 2024, Turin, Italy. pp.1-10. ⟨hal-04346102v2⟩
- Berthoud, F.; Bzeznik, B.; Gibelin, N.; Laurens, M.; Bonamy, C.; Morel, M.; Schwindenhammer, X. Estimation de l'empreinte carbone d'une heure.coeur de calcul. Research report, UGA - Université Grenoble Alpes ; CNRS ; INP Grenoble ; INRIA, 2020. https://hal.archives-ouvertes.fr/hal-02549565v4/
- Bouveret, S. Bugeau, A. Orgerie, A.-C., Quinton, S. De l'eau dans les nuages. Annales des Mines - Enjeux Numériques, 2024, 27, pp.40-47. ⟨hal-04698568⟩
- Lucia Bouza Heguerte, Aurélie Bugeau, Loïc Lannelongue. How to estimate carbon footprint when training deep learning models? A guide and review. Environmental Research Communications, 2023, ⟨10.1088/2515-7620/acf81b⟩. ⟨hal-04120582v2⟩
- Bugeau, A., Ligozat, A.-L. Analysing ICT in prospective scenarios to help reveal undone computer science. Undone Computer Science conference, Feb 2024, Nantes (France), France. ⟨hal-04486589⟩
- Ekchajzer, D., Bornes, L., Combaz, J., Letondal, C., Vingerhoeds, R. Decision-making under environmental complexity: the need for moving from avoided impacts of ICT solutions to systems thinking approaches. ICT4S 2024 : International Conference on ICT for Sustainability, Jun 2024, Stockholm, Sweden. ⟨hal-04637677⟩
- Jay, M., Ostapenco, V., Lefèvre, L., Trystram, D., Orgerie, A.-C., Fichel, B. An experimental comparison of software-based power meters: focus on CPU and GPU. CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing, May 2023, Bangalore, India. pp.1-13, ⟨10.1109/CCGrid57682.2023.00020⟩. ⟨hal-04030223v2⟩
- Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. arXiv preprint arXiv:2304.03271.
- Ligozat, A.-L., Lefèvre, J., Bugeau, A., & Combaz, J. (2021). Unraveling the hidden environmental impacts of AI solutions for environment. arXiv:2110.11822 [cs]. http://arxiv.org/abs/2110.11822 and Sustainability https://www.mdpi.com/2071-1050/14/9/5172/htm (preprint https://arxiv.org/abs/2110.11822)
- Luccioni, A. S., Viguier, S., Ligozat, A.-L. (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, JMLR, https://jmlr.org/papers/v24/23-0069.html (preprint https://arxiv.org/abs/2211.02001)
- Luccioni, S., Jernite, Y., & Strubell, E. (2024, June). Power hungry processing: Watts driving the cost of AI deployment?. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 85-99)., https://arxiv.org/pdf/2311.16863
- Clément Morand, Anne-Laure Ligozat, Aurélie Névéol. Empreinte carbone des expériences en TAL : les défis de la reproductibilité. journée d'étude Journée Éthique et TAL 2024, Karën Fort; Aurélie Névéol, Apr 2024, Nancy, France. ⟨hal-04579556⟩
- Morand, C. Névéol, A., Ligozat, A.-L. MLCA: a tool for Machine Learning Life Cycle Assessment. 2024 International Conference on ICT for Sustainability (ICT4S), Jun 2024, Stockholm, Sweden. ⟨hal-04643414⟩
- Parcollet, T., Ravanelli, M. The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. 2021. ⟨hal-03190119⟩
- Rasoldier, A., Combaz, J., Girault, A., Marquet, K., Quinton, S. How realistic are claims about the benefits of using digital technologies for GHG emissions mitigation?. LIMITS 2022 - Eighth Workshop on Computing within Limits, Jun 2022, Virtual, France. ⟨hal-03949261⟩
- Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., … (2019). Tackling Climate Change with Machine Learning. ArXiv:1906.05433 [Cs, Stat]. http://arxiv.org/abs/1906.05433
- S. Sala, E. Crenna, M. Secchi, and E. Sanyé-Mengual, "Environmental sustainability of european production and consumption assessed against planetary boundaries," Journal of Environmental Management, vol. 269, p. 110 686, 2020.
- Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute Trends Across Three Eras of Machine Learning. arXiv:2202.05924 [cs]. http://arxiv.org/abs/2202.05924
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ArXiv:1906.02243 [Cs]. http://arxiv.org/abs/1906.02243
- Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., … Hazelwood, K. (2021). Sustainable AI: Environmental Implications, Challenges and Opportunities. arXiv:2111.00364 [cs]. http://arxiv.org/abs/2111.00364