

IA & Éthique

Là où le bât blesse... (et ce qu'on peut y faire)

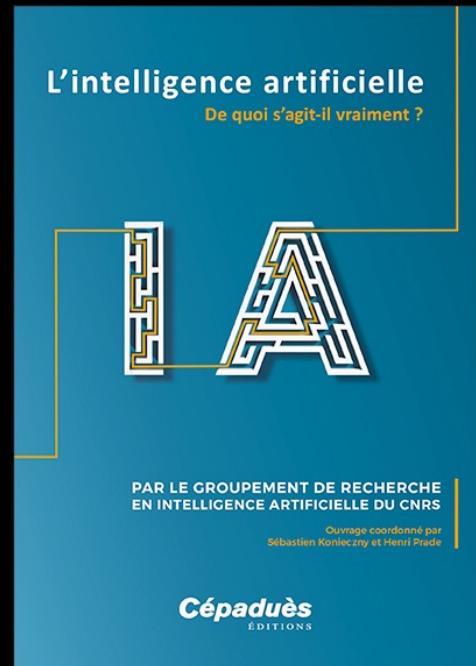
Elisa Fromont

Séminaire « éthique et maths » – 11/10/2024

**DES DÉFINITIONS PAS CLAIRES
POUR TOUT LE MONDE
(LIENS IA - NUMÉRIQUE)**

Définition personnelle

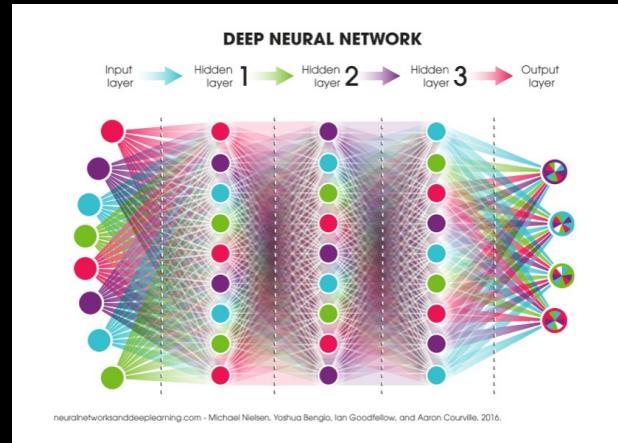
L'intelligence artificielle (IA) est une science qui explore la construction et l'étude d'algorithmes qui permettent à des machines de résoudre des tâches humaines que les ordinateurs ne savent, pour le moment, pas bien résoudre/simuler (e.g. apprendre, raisonner, décider).



🤔 Frontière entre **algorithmique** et IA très floue...

L'IA du grand public

- Est **apprise** (sur des données)
- ... par **deep learning** (réseaux de neurones)



- ... est **générateive** (what else?)

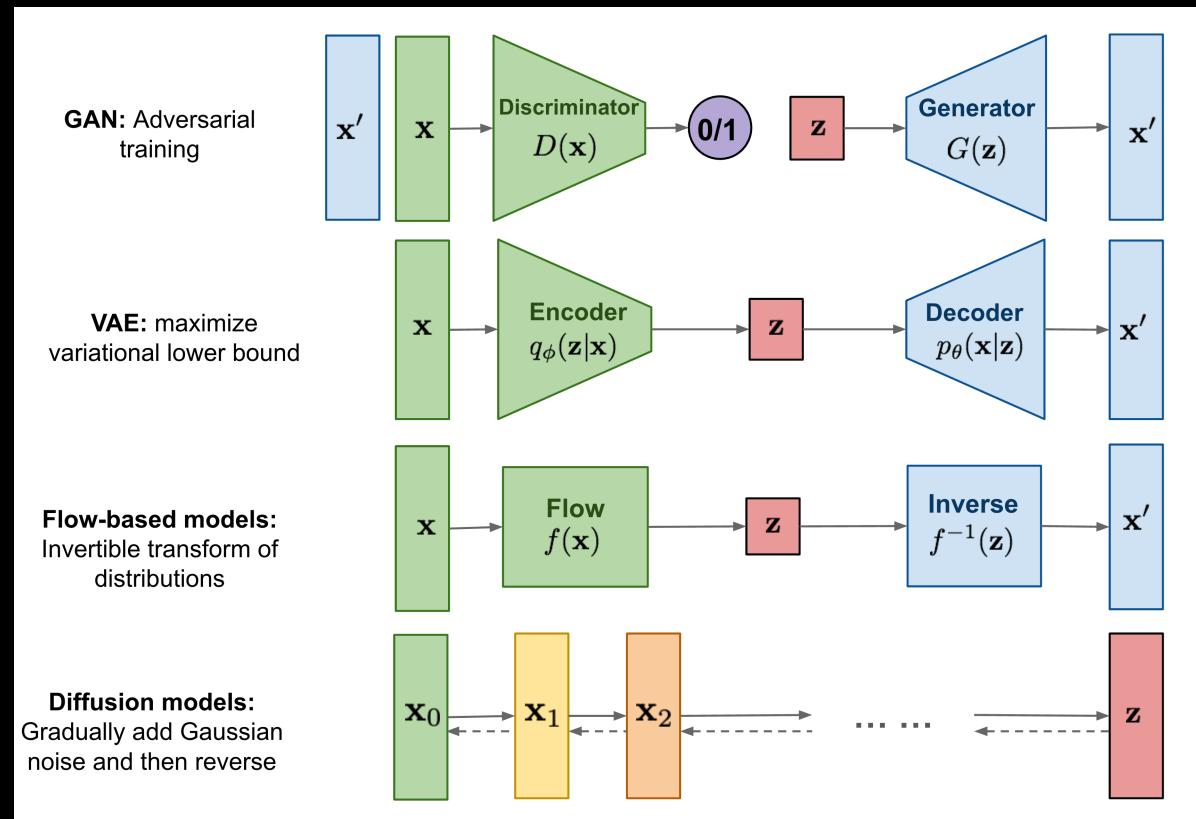
Ce n'est pas (toute) l'IA de la recherche...

IA GENERATIVE ?

Réseaux profonds génératifs

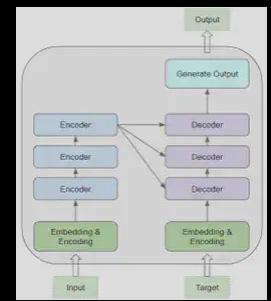
Génératif = capable de créer de la nouvelles données

- Généralement appris de manière « **non supervisée** » (i. e. pas pour la classification)
- But: estimer la distribution inconnue $p(X)$ des données → échantillonner à partir de cette distribution estimée = générer de nouveaux exemples



IA GENERATIVE ?

Le cas des “Transformer”



Attention Is All You Need. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, NeurIPS 2017

L'architecture Transformer peut être UTILISÉE de manière générative mais elle n'est pas intrinsèquement générative

- Pas de distribution de probabilité explicite sur l'espace de sortie.
 - Mais GPT est entraîné en utilisant des objectifs de modélisation du langage, tels que prédire le token suivant dans une séquence ou reconstruire des tokens manquants. Ces objectifs encouragent le modèle à apprendre une distribution de probabilité sur les séquences de tokens possibles.
- Pas de mécanisme d'échantillonnage intégré pour générer de nouveaux points de données à partir de l'espace de représentation appris
 - Mais, pendant la génération, ChatGPT échantillonne une distribution de probabilité sur les tokens suivants possibles. Ce processus d'échantillonnage permet au modèle d'explorer différentes possibilités et de générer des réponses diverses.
- Elles sont principalement conçues pour la génération conditionnelle, où elles prédisent une séquence de sortie étant donnée une séquence d'entrée.
 - Mais GPT utilise un processus de décodage autoregressif pour générer du texte un token à la fois (la prédiction dépend du mot précédent prédit)
- Sorties déterministes alors que les modèles génératifs produisent généralement des sorties stochastiques.
 - Cf. échantillonage



ChatGPT

L'exemple de Chat-GPT

A screenshot of a dark-themed ChatGPT interface. On the left is a green square icon containing the white 'G' logo. To its right is a text input field with the placeholder "Bonjour! Comment puis-je vous aider aujourd'hui?". On the far right are three small icons: a clipboard, a thumbs-up, and a thumbs-down.

- ChatGPT est un **agent conversationnel**
 - Basé sur un réseau de neurones (GPT : « Generative Pre-trained Transformer »)
 - Entraîné (par apprentissage supervisé) à **prédir le prochain mot d'un texte**
 - ...à partir de zettabytes (**10²¹**) de données textuelles...
 - 180 milliards de paramètres... (GPT 3)
 - Affiné (policé) avec des « vraies » interactions humaines (au Kenya) + vos interactions passées, par apprentissage par renforcement
- ChatGPT fait de la **génération de texte « plausible »**. Il n'est pas « fait » pour répondre de manière exacte à une question.

Une bonne référence pour en savoir plus : <https://www.youtube.com/watch?v=R2fjRbc9Sa0>

Les “IA” populaires

(50% Transformer; 50% diffusion models; 100% considérées génératives)

Top 100 IA Populaires

Rank	Model	Icon	Change
1.	ChatGPT +1301		
2.	MidJourney V6.1 +820		
3.	Stable Diffusion +727		
4.	Perplexity AI +715		
5.	Gemini AI +517		
6.	Playground OpenAI +493		
7.	Adobe Firefly 3 +423		
8.	Microsoft Bing +410		
9.	HeyGen AI +407		
10.	Adcreative AI +383		

Top 100 Tendance [24H]

Rank	Model	Icon	Change
1.	Photoshop IA (+84)		
2.	Quill (+60)		
3.	Illuminate by Google (+56)		
4.	AutoGen (+32)		
5.	Adobe Firefly Video (+32)		
6.	Pixtral12B (+32)		
7.	Remini (+24)		
8.	Riddle Me This (+24)		
9.	G1 Llama Meta (+20)		
10.	SlideSpeak AI (+20)		

Top 100 Tendance [7 jours]

Rank	Model	Icon	Change
1.	Luma AI (+640)		
2.	Photoshop IA (+592)		
3.	Pixtral12B (+348)		
4.	Imagen 3 by Google (+308)		
5.	Adobe Firefly Video (+248)		
6.	Illuminate by Google (+244)		
7.	OpenAI o1 (+240)		
8.	SlideSpeak AI (+200)		
9.	AI MiniMax (+176)		
10.	Quill (+128)		

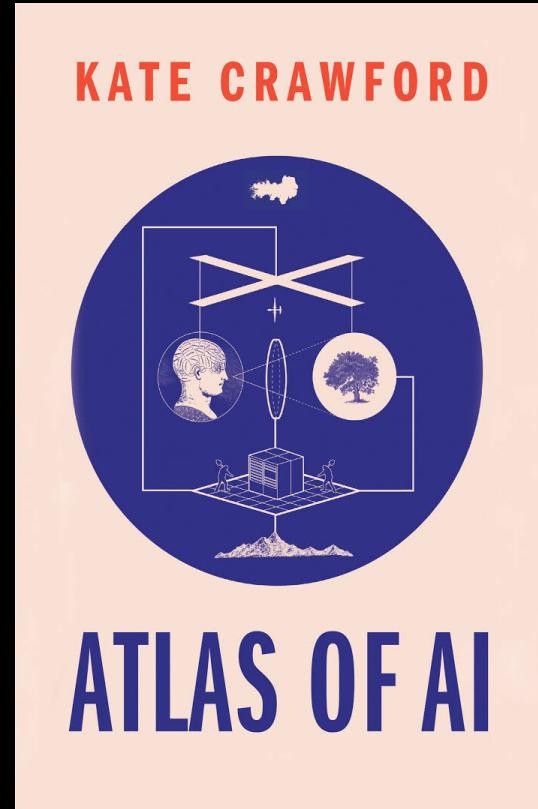
<https://www.aixploria.com/top-100-ia> (au 19/09/2024)



LES PROBLÈMES ETHIQUES...

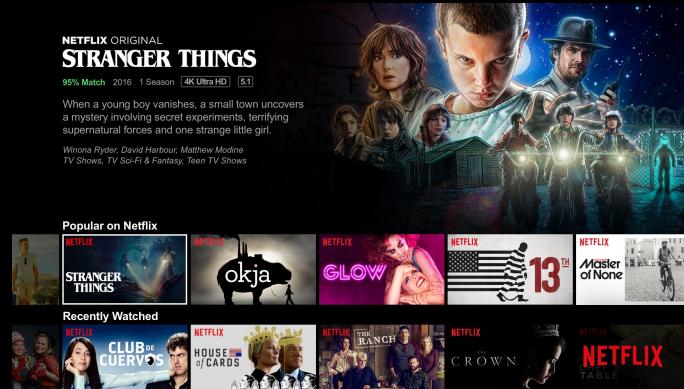
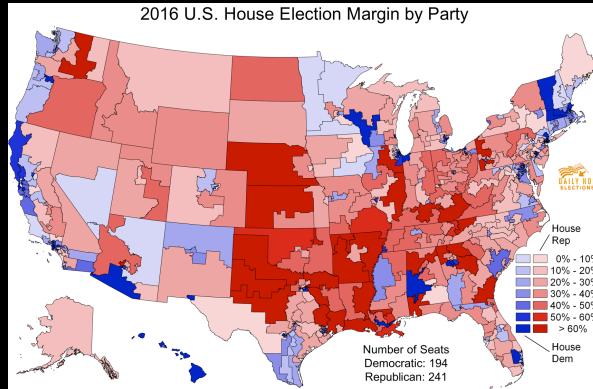
ATLAS OF AI

- Chapitre 1 : l'extraction de minerais nécessaire pour alimenter l'informatique contemporaine
- Chapitre 2 : comment l'intelligence artificielle est faite de travail humain
- Chapitre 3 : le rôle des données (non plus personnelles, mais « d'infrastructure »)
- Chapitre 4 : comment la classification automatique peut être vraiment choquante
- Chapitre 5 : la reconnaissance automatique de des émotions
- Chapitre 6 : affaires d'Etat (surveillance, manipulation, ...)
- ...



2021 (pré Chat-GPT)
<https://www.katecrawford.net/>

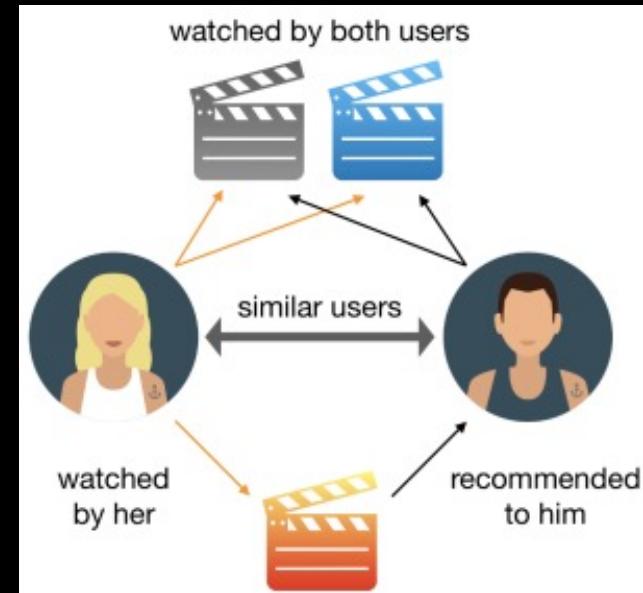
Quand la recommandation à large échelle menace la démocratie



Les systèmes de recommandation sont partout (réseaux sociaux, achats en ligne, publicités...).

Ils peuvent :

- être manipulés (faux comptes, empoisonnement...)
- être biaisés (intentionnellement ou non, cold start)
- vous manipuler (incitation à la haine, vote,)



Des biais partout... (algorithmique ou dans les données)



<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

UK ditches exam results generated by biased algorithm after student protests



Students protest outside the constituency office for Education Secretary Gavin Williamson in Codsall



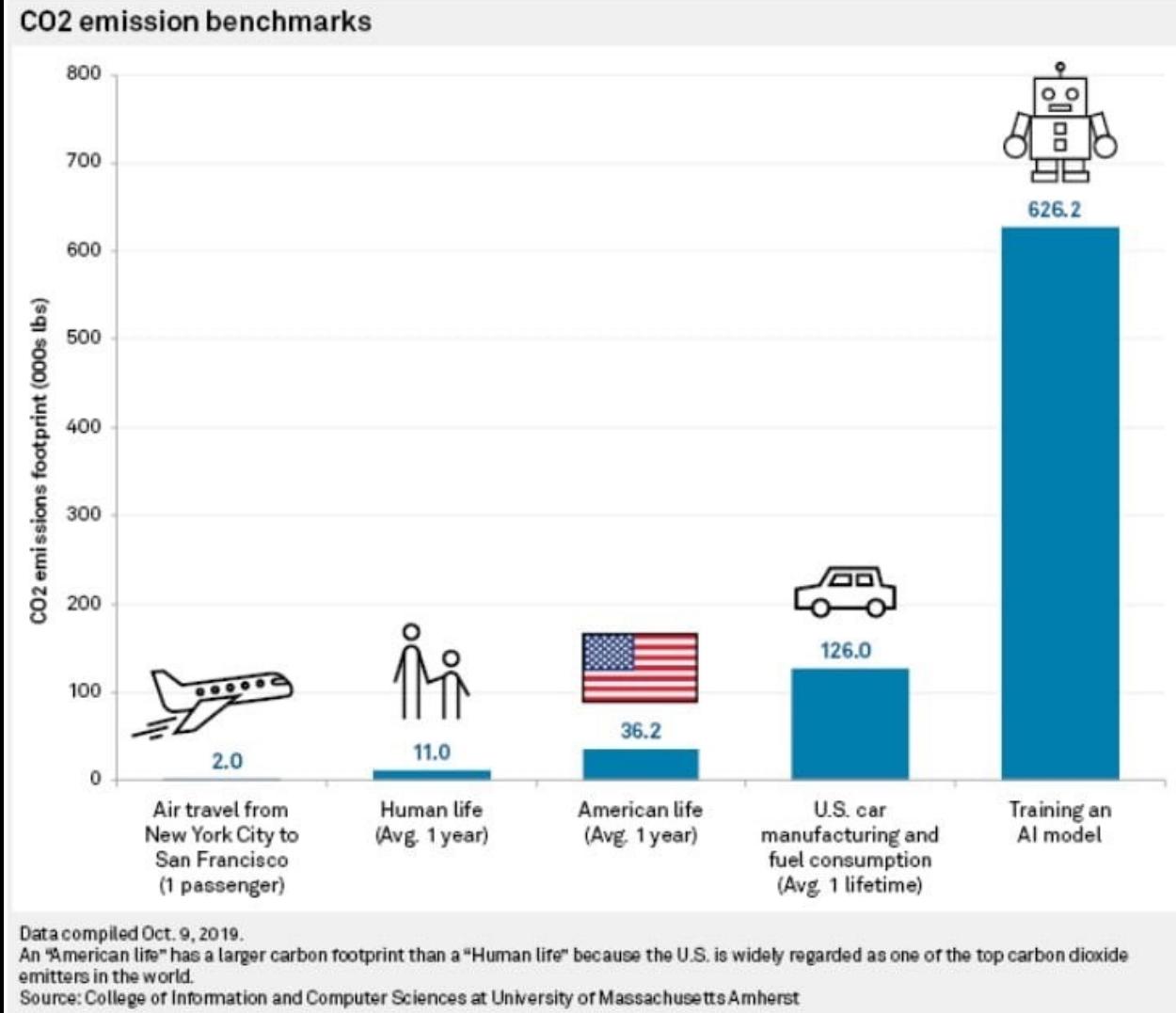
Catastrophe écologique



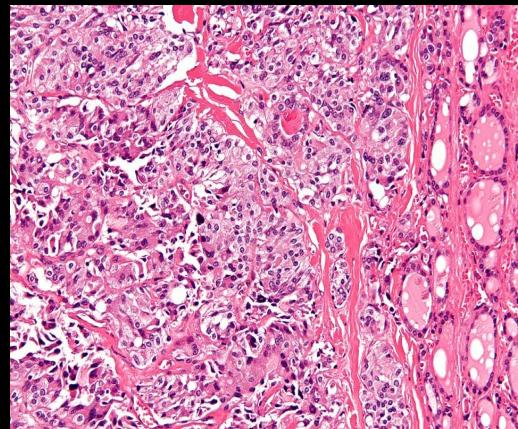
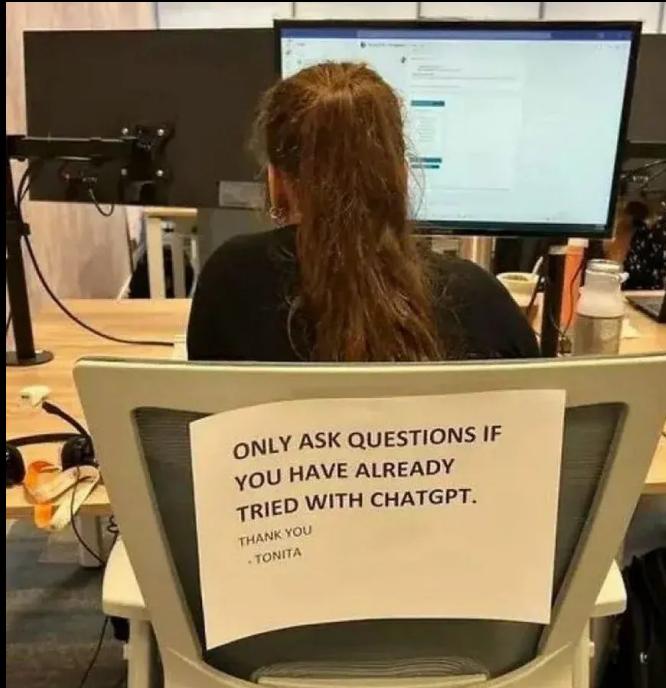
Data centers :
2% de la
consommation
mondiale
d'électricité

Industrie du
numérique
(03/2024) : 4%
des émissions
mondiales de
gaz à effet de
serre

2019 →



Quand l'IA nous remplace (plutôt bien)



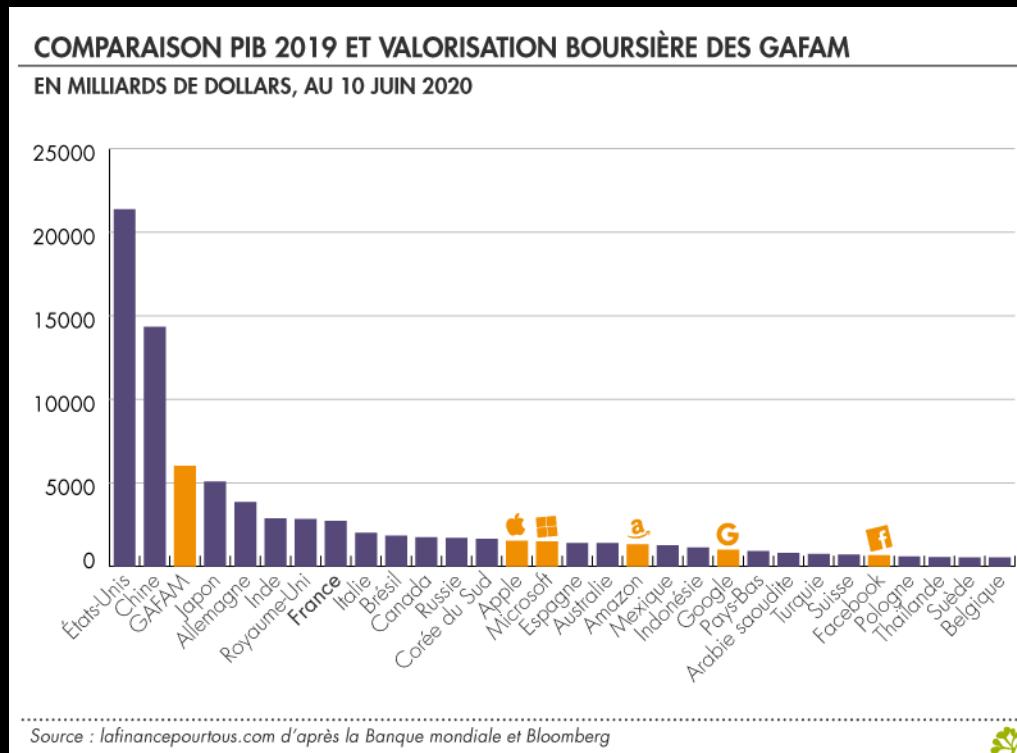
Analyse d'images
médicales

- Telemarketing
- Commis comptable
- Gestionnaires de la rémunération et des avantages sociaux
- Réceptionnistes
- Coursiers
- Correcteurs d'épreuves
- Spécialistes de l'assistance informatique
- Analystes d'études de marché
- Vendeurs en publicité
- Vendeurs au détail
- ...

<https://blog.hubspot.com/marketing/jobs-artificial-intelligence-will-replace>



GAFA, GAFAM ou NATU : les nouveaux maîtres du monde ?



Google plus de 90 % des requêtes sur internet dans le monde.

YouTube (diffuseur de vidéos) racheté par Google en 2006 : bien plus que n'importe quelle chaîne de télévision : chaque jour, plus d'un milliard d'heures de vidéos sont ainsi visionnées.

Facebook : en octobre 2020, plus de 2,7 milliards d'utilisateurs actifs mensuels

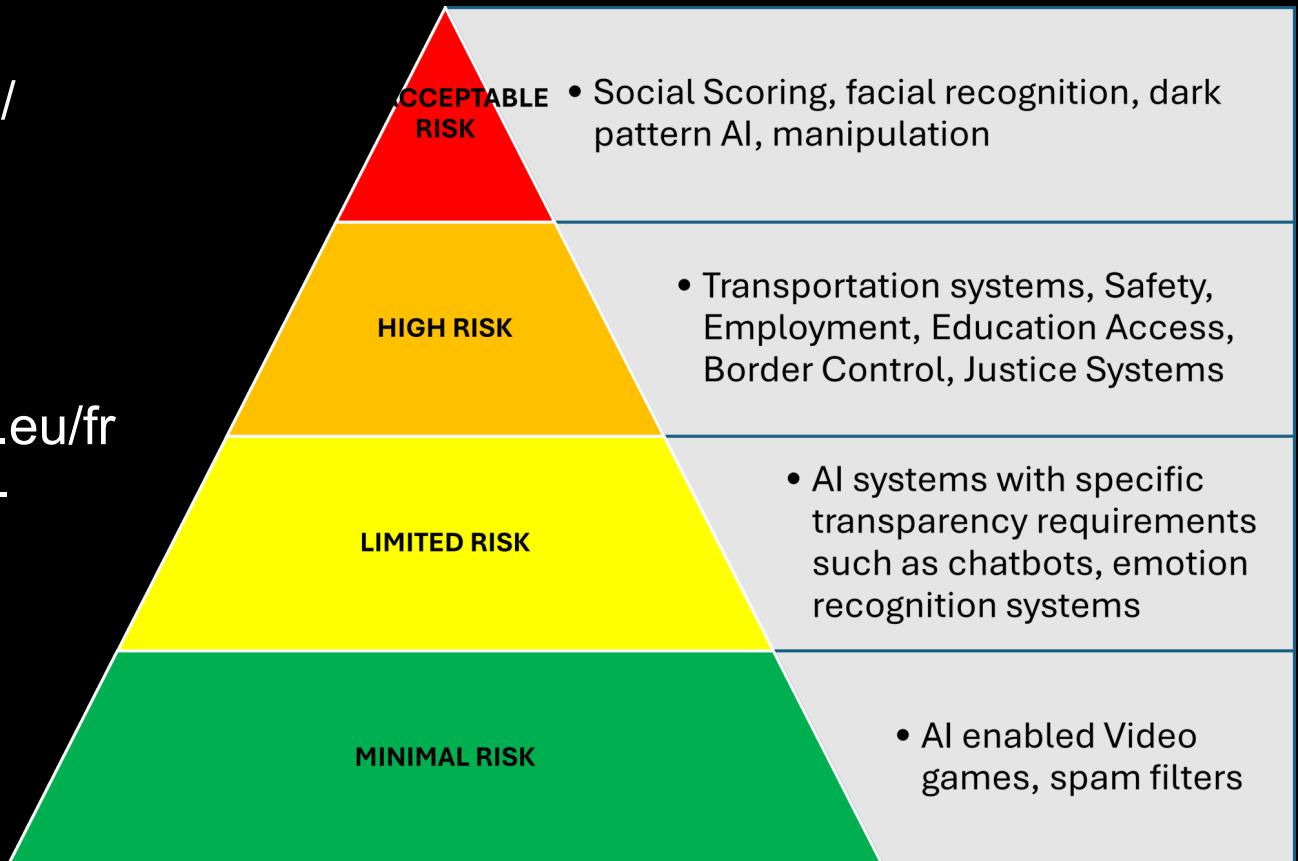
Législation (trop faible ?)

- RGPD

<https://gdpr-info.eu/>

- AI Act

<https://digital-strategy.ec.europa.eu/fr/policies/european-approach-artificial-intelligence>





QUELQUES DEFIS ACTUELS DE LA RECHERCHE EN IA

Raisonner avec “intelligence”

The screenshot shows a conversation between a user and ChatGPT. The user asks for a recipe for Breton far au lait de canard. ChatGPT responds positively and provides a detailed list of ingredients.

You

You

Bonjour. Peux tu m'expliquer comment faire un far breton au lait de canard ?

ChatGPT

Bonjour ! Bien sûr, je serais ravi de vous expliquer comment faire un far breton au lait de canard. Le far breton est un dessert traditionnel de la région de Bretagne en France.

ChatGPT

Bonjour ! Bien sûr, je serais ravi de vous expliquer comment faire un far breton au lait de canard. Le far breton est un dessert traditionnel de la région de Bretagne en France.

• Voici une recette adaptée pour utiliser du lait de canard :

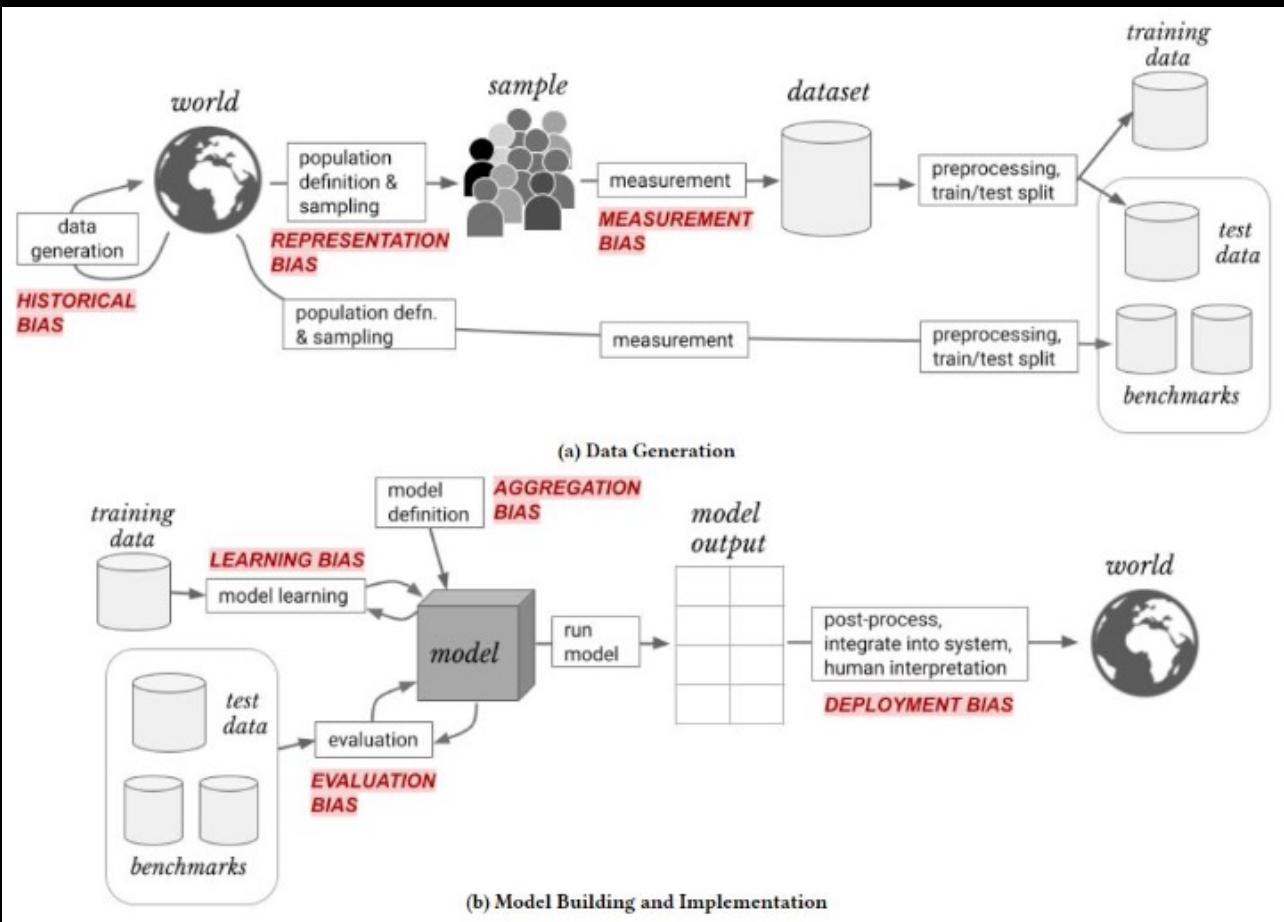
- 4 œufs
- 1 pincée de sel
- 1 litre de lait de canard
- 1 cuillère à soupe de rhum (facultatif)
- Beurre pour le moule
- 1 litre de lait de canard

Instructions :

Créer des modèles de confiance



Confiance : identification et atténuation des biais



Équité algorithmique

- Équité statistique

→ suivre les disparités statistiques entre les prédictions algorithmiques de différents groupes.

- Équité individuelle

→ les individus en situation similaire devraient être traités de la même manière

- Équité contrefactuelle

→ les questions relatives à la justice et à la discrimination sont de nature causale (découvrir les causes)

Confiance: calculer frugalement



Low-complexity distributed and frugal methods

[Learn More](#)

ABOUT US

Edge Intelligence is a research program of the **MIAI institute** that works on efficient resource management and distributed/federated Machine Learning. For instance, we promote locally distributed computations related to AI to avoid the waste of energy lost during data transfers. Thanks to a small data center close to the edge of the network, data produced on mobile devices can be pre-computed in low latency networks. Further computations may then be executed on a regular data center over the Cloud.

FIELDS OF INTEREST

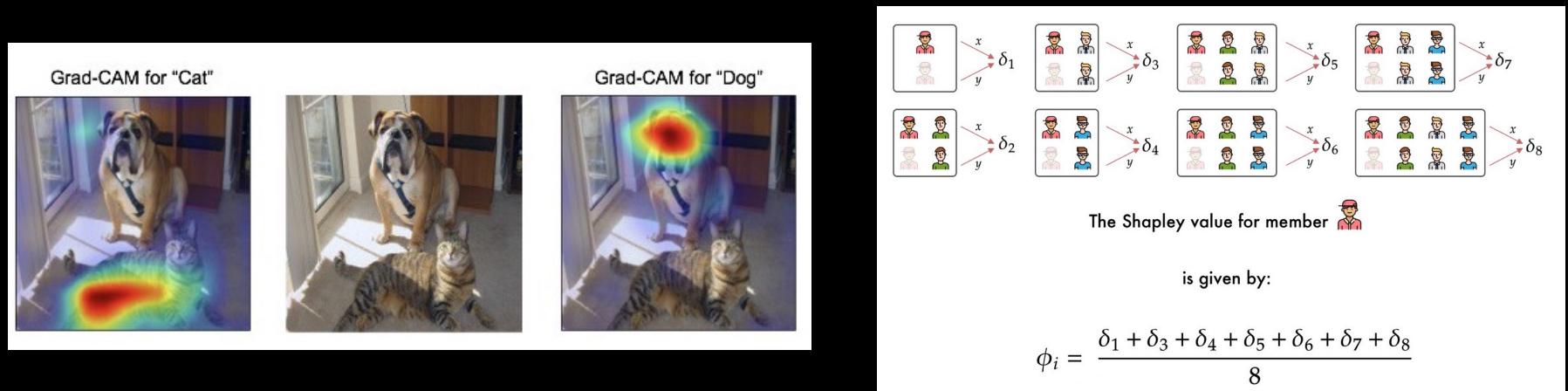
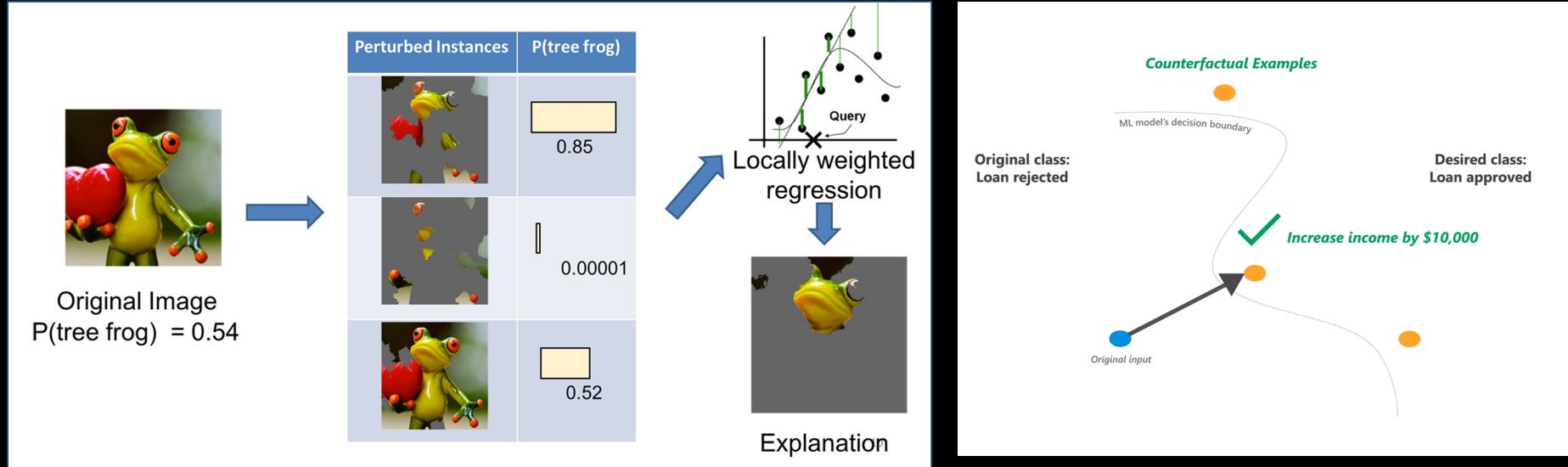
Distributed Learning	Federated learning	Task management	Online Learning
Distribute the computation for better efficiency. (E.g. computing power, workload, storage management)	Collaborate with the intent to keep the data of each agent heterogeneous, local, and private. Federated learning can be centralized, decentralized, and semi-centralized.	Efficient allocation and execution of the tasks in the appropriate heterogeneous and dynamic distributed computing devices connected at the edge level.	Predict better for future data. In online settings, data becomes available in sequential order.

<https://edge-intelligence.imag.fr/>



- Réutilisation des modèles existants (transfert, adaptation, apprentissage continue ...)
- Apprendre mieux avec peu de données (étiquetées)
- Apprendre mieux avec des modèles moins complexes (= moins de ressources)

Confiance : XAI



Démêler le vrai du faux

Un problème de chats et de souris...

- Deepfakes (e.g. « fake news », cyber-harcèlement, ...)
- Cyber-arnaques (e.g. usurpation d'identité, love scam, ...)
- Plagiats (e.g. rapports, articles scientifiques)
- Empoisonnement des données et des modèles
- ...



Cybersécurité de l'IA

Conclusion

- Le problème vient-il de l'IA ?
ou du numérique...
- L'utilisation de l'IA (comme tout le numérique)
pose des problèmes éthiques
- La recherche en IA s'empare de ces sujets
- L'IA peut-être utile mais son utilisation **doit**
être réfléchie

Pour aller plus loin...

<https://intelligence-artificielle.univ-rennes.fr/semaine-IA2024>

Revoir les conférences

- ✓ Élisa Fromont Introduction + Intelligence Artificielle : de quoi parle-t-on ?
- ✓ Mélanie Gornet Doctorante - Gouvernance de l'Intelligence artificielle (IA)
- ✓ Gaël Varoquaux IA: des promesses aux enjeux
- ✓ J-M Loubes Les biais de l'intelligence artificielle (IA)
- ✓ Carina Prunkl Les biais de l'intelligence artificielle (IA)
- ✓ Margarida Romero Apprendre à l'ère de l'intelligence artificielle: rapport au savoir et pratiques co-créative
- ✓ Marc Cuggia Les applications de l'IA en santé : Perspectives, Cas d'Usage, et Défis
- ✓ Denis Bonnay Machines, vision, créativité



Questions ?