

le
campus
numérique
in the ALPS

Intelligence Artificielle: questions éthiques

5/3/24

Éthique ?



[Charles Levy](#): Atomic Cloud Rises Over Nagasaki,

Éthique de conviction :

on agit en fonction de principes supérieurs, de valeurs auxquelles on croit.

Éthique de responsabilité :

on est prêt à assumer les conséquences de ses actes, en fonction des effets que l'on peut prévoir

- Législation
- Anonymisation
- Modèles génératifs
- Biais et discrimination
- Transparence
- Impact sociétal
- Impact Environnemental

Réglementation européenne de l'IA

1ere initiative mondiale de régulation



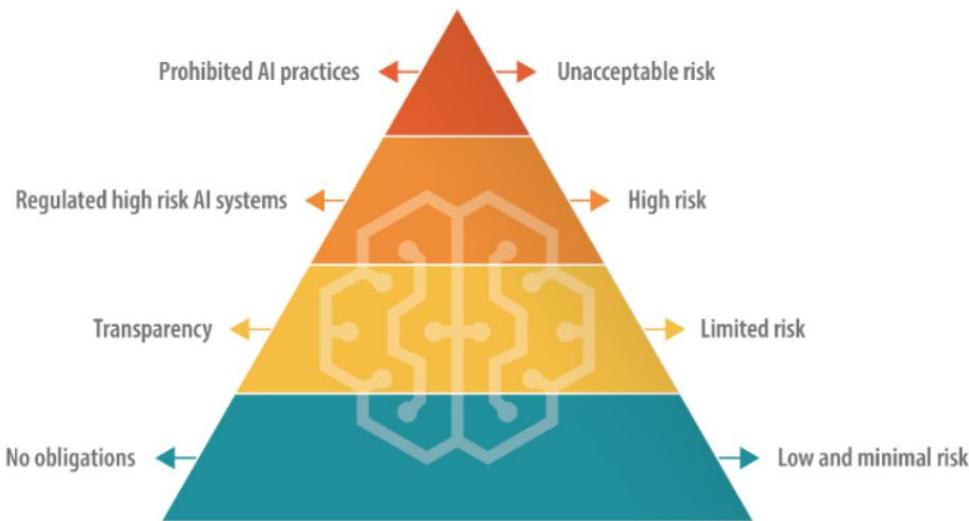
En cours : vote du parlement européen 14 juin 2023 puis négociations avec le Conseil de l'Union européenne et la Commission européenne pour une implémentation en 2024.

Difficulté de réguler un secteur qui bouge beaucoup, illustration avec Chat GPT.

Approche par le niveau de risque et des domaines d'application sensibles (mais comment les anticiper ?)

Bientôt un RGPD de l'IA en UE ?

Régulation par le niveau de risque :



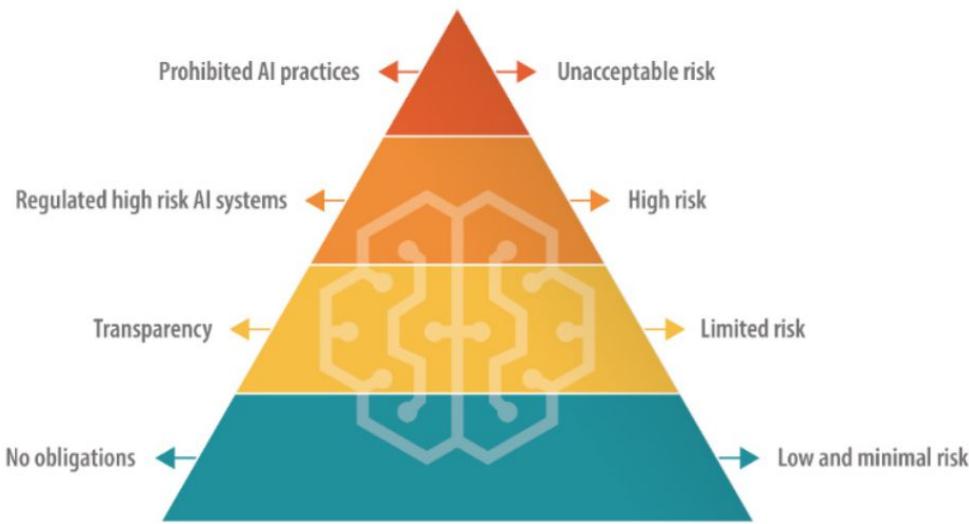
Risque inacceptable :

- Influencent de manière subliminale le comportement d'une personne en vue de lui causer ou de causer à un tiers un dommage
- Systèmes de notation sociale
- Identification biométrique à distance en temps réel dans des espaces accessibles au public, sauf exception (sécurité nationale)
- Identification biométrique et catégorisation des personnes

<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

Bientôt un RGPD de l'IA en UE ?

Régulation par le niveau de risque :



Risque haut :

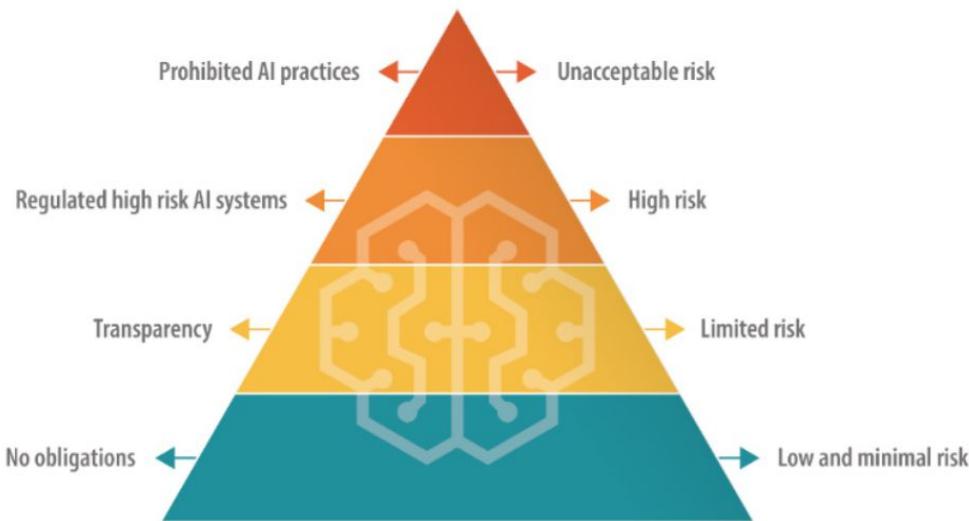
- Tous les produits soumis aux normes de sécurité et de santé de l'UE
- Infrastructures critiques
- Éducation et formation professionnelle
- Emploi, gestion des travailleurs
- Services privés essentiels et des services et prestations publics
- Forces de l'ordre
- Gestion des migrations, de l'asile et du contrôle des frontières
- Assistance à l'interprétation juridique et à l'application de la loi.

=> Exigences en termes de gestion des risques, tests, robustesse, formation aux données et gouvernance des données, transparence, surveillance humaine et cybersécurité.

<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

Bientôt un RGPD de l'IA en UE ?

Régulation par le niveau de risque :



Risque limité :

- Transparence (à définir)
- Pour les modèles génératifs :
 - Mentionné dans le contenu généré
 - Pas de possibilité de générer du contenu illégal
 - Publier des résumé des sources utilisées pour l'entraînement du modèle

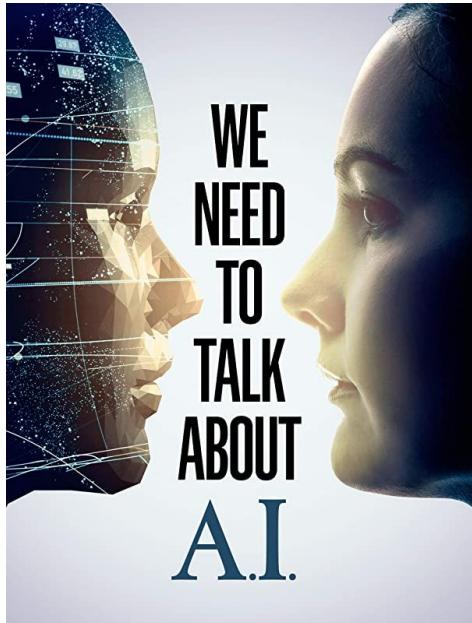
<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

Les discussions en cours

Seraient interdits les systèmes d'IA qui

- Influencent de manière subliminale le comportement d'une personne en vue de lui causer ou de causer à un tiers un dommage
- Mettent en service des systèmes de notation sociale
- Permettent l'identification biométrique à distance en temps réel dans des espaces accessibles au public, sauf exception.

(Les pays de l'UE avaient obtenu que les systèmes d'IA développés à des fins militaires ou policières soient retirés du champ d'application du règlement (coucou les drones IA aux JO). Mais l'interdiction a été étendue par les députés européens aux modèles d'IA de catégorisation biométrique, à la police prédictive et à la récupération d'images faciales pour la constitution de bases de données.)



Les discussions en cours



- Les logiciels de reconnaissance des émotions sont interdits dans les domaines du maintien de l'ordre, de la gestion des frontières, du travail et de l'éducation.
- Les systèmes d'IA développés dans le seul but de la recherche et du développement scientifique sont eux exclus du champ d'application.
- Seraient exclus également les IA sans domaine particulier, à usage général. La responsabilité incombe alors à ceux qui les incorporent dans leurs produits dans tel ou tel domaine.
- Les modèles d'IA générative tels que ChatGPT devraient quant à eux signaler chaque fois qu'un texte est généré par l'IA et fournir un résumé détaillé des données d'entraînement couvertes par la loi sur le droit d'auteur.

Les données personnelles



Des identifiants directs :

- Nom, prénom, pseudonyme, date de naissance ;
- photos, enregistrements sonores de voix...
- numéro de téléphone fixe ou portable, adresse postale, adresse e-mail...
- empreinte digitale, réseau veineux ou palmaire de la main, empreinte rétinienne, haleine...
- numéro de plaque d'immatriculation, numéro de sécurité sociale, numéro d'une pièce d'identité

Des quasi-identifiants :

- adresse IP, identifiant de connexion informatique ou identifiant de cookie...
- Localisations...
- Historiques de navigations...
- données d'usage d'une application, commentaires...

« Toute information se rapportant à une personne physique identifiée ou identifiable (...); est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale »

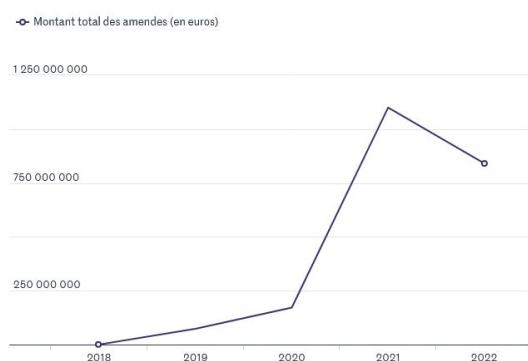
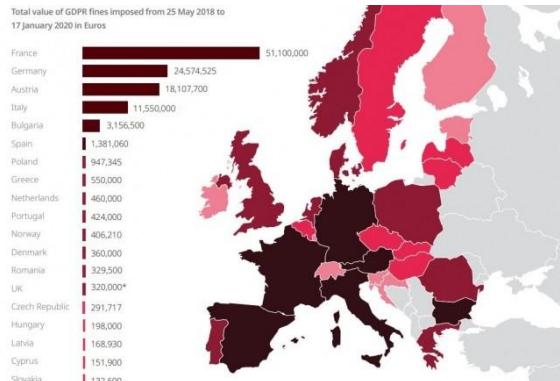
Art. 4 du RGPD

RGPD (règlement général de protection des données)

- Au moment de la collecte de données, les entreprises doivent obtenir un **consentement** clair et éclairé de la part des personnes.
- L'entreprise doit garantir le **droit des personnes** : droit d'accès, droit de rectification, droit à l'oubli, droit à la portabilité, droit à la limitation des traitements.
- Les entreprises doivent mettre en place les mesures adéquates pour assurer un niveau de **sécurité** optimal des données personnelles : pseudonymisation, analyses d'impact, tests d'intrusion...
- Toutes les entreprises doivent tenir un **registre des traitements** et le mettre régulièrement à jour.
- Certaines entreprises ont pour obligation de nommer un **Data Protection Officer** ou DPO.



Du côté des données personnelles, on encaisse



Décembre 2023 :

- Amazon : 32 millions pour surveillance excessive de ses salariés

Juin 2023 :

- Criteo : 40 millions pour non respect du RGPD (consentement, transparence, droit d'accès)

Mai 2023 :

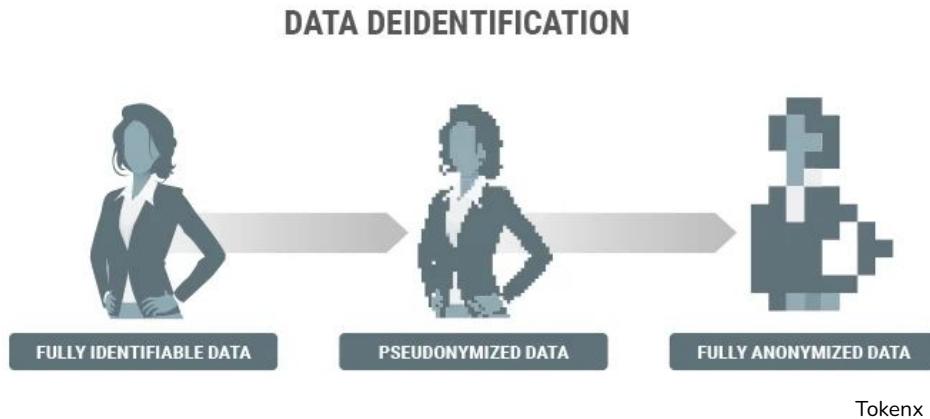
- Meta : 1.2 milliards par le régulateur irlandais pour transfert de données personnelles de citoyens européens vers les Etats-Unis

Décembre 2022

- Microsoft : 60 millions pour dépôt de cookie de traçage sans consentement de l'utilisateur sur le moteur de recherches Bing

- Législation
- Anonymisation
- Modèles génératifs
- Biais et discrimination
- Transparence
- Impact sociétal
- Impact Environnemental

Pourquoi s'intéresser à l'anonymisation



Il y a des enjeux à **transférer ou publier** des données personnelles à des tiers pour la recherche, pour la santé, ...

C'est un frein au développement de l'**open data** ou aux bénéfices du **big data**.

Le RGPD l'impose si on veut transférer des données, utiliser des données au-delà de la durée légale de conservation, ou encore, l'envoi des données à des partenaires situés en zone hors UE.

Pour des raisons éthiques : trop de données personnelles identifiantes circulent

Pour des raisons économiques : il y a un marché (cf essor des "data clean room")

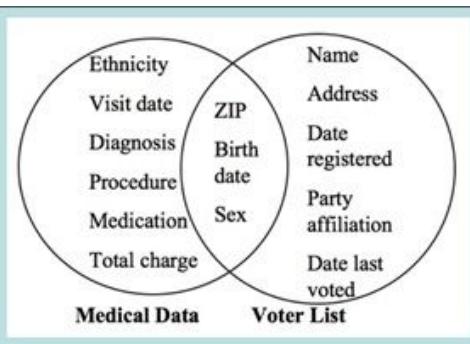
1996 La chercheuse et le gouverneur



Le 18 mai 1996 le Gouverneur William Weld s'effondre lors d'une réunion publique et est hospitalisé dans la foulée. Latanya Sweeney, chercheuse au MIT, croise alors des données “anonymisées” sur les hospitalisations et les données de liste électorale. Les deux dataset avait des éléments communs : code postal, date de naissance et le sexe.

Dans la liste des électeurs de Cambridge, 6 personnes avaient la même date de naissance que lui, mais seulement 3 d'entre eux étaient des hommes, et il était le seul dans son code postal à 5 chiffres. Avec l'heure approximative d'hospitalisation, elle accède à son dossier médical.

La ré-identification par individualisation



Latanya Sweeney l'a mis en évidence aux Etats-Unis en 2000 en croisant deux bases de données, une base de données médicale pseudonymisée et une liste électorale avec des données nominatives.

Le croisement a été effectué non pas sur des champs directement identifiants, mais sur un triplet de valeurs : code postal, date de naissance et sexe. Cela permettrait la réidentification de 87 % de la population des Etats-Unis !

Des chercheurs ont déterminé en 2019 que 99,98 % des Américains seraient correctement ré-identifiés dans n'importe quelle base de données en utilisant 15 attributs démographiques ([Rocher, L., Hendrickx, J.M. & de Montjoye, Y. Nat Commun 10, 3069 -2019](#)).

2008 Netflix et la corrélation



A 4x4 grid of 16 icons arranged in four rows and four columns. The icons are as follows:

- Row 1: Thumbs Up, Blank, Thumbs Down, Thumbs Up
- Row 2: Thumbs Up, Thumbs Up, Thumbs Down, Thumbs Up
- Row 3: Thumbs Up, Blank, Thumbs Down, Blank
- Row 4: Thumbs Up, Thumbs Up, Thumbs Down, Thumbs Up

Anonymized NetFlix data



A 4x4 grid containing eight thumbs up icons and four thumbs down icons. The icons are arranged as follows: Row 1: Two thumbs up on the left, one thumbs up on the right. Row 2: One thumbs up in the center. Row 3: Two thumbs up on the left, one thumbs down in the center. Row 4: One thumbs down in the center, two thumbs down on the right.

Public, incomplete **IMDB** data

500 000 notations de film produite par des utilisateurs sont publiées en supprimant les données directement identifiantes et en mettant du bruit sur les autres (modification de dates, de notes) : 1 millions \$ à qui améliorerait le système de recommandation

Arvind Narayanan et Vitaly Shmatikov ont pu réidentifier à partir de la base Internet Movie Database (IMDB) en comparant les notes et les dates : avec 6 à 8 notes et dates, ils obtiennent une probabilité de réidentification de 90%

2006 AOL et l'inférence

A sample of Thelma Arnold's search data released by AOL						
4417749	swing sets	2006-04-24	15:39:30	4	http://www.buyswingset.com	
4417749	swing sets	2006-04-24	15:39:30	9	http://www.buychoice.com	
4417749	swing sets	2006-04-24	15:39:30	10	http://www.creativeplaythings.com	
4417749	swing sets	2006-04-24	15:39:30	5	http://www.childlife.com	
4417749	swing sets	2006-04-24	15:39:30	6	http://www.planitplay.com	
4417749	that do not shed	2006-04-28	9:05:54	2	http://www.gopetsamerica.com	
4417749	dog who urinate on everything	2006-04-28	13:24:07	6	http://www.dogdaysusa.com	
4417749	walmart	2006-04-28	14:07:32	1	http://www.walmart.com	
4417749	womens underwear	2006-04-28	14:12:28	10	http://www.bizrate.com	
4417749	jcpenny	2006-04-28	14:16:05			
4417749	jcpenny	2006-04-28	14:16:49	1	http://www.jcpenney.com	
4417749	tortus and turtles	2006-04-29	13:12:47			
4417749	manchester terrier	2006-05-02	9:05:31	1	http://www.manchesterterrier.com	
4417749	delta	2006-05-02	11:49:26			
4417749	fingers going numb	2006-05-02	17:35:47			
4417749	dances by laura	2006-05-02	17:59:32			
4417749	dances by lori	2006-05-02	17:59:57			
4417749	single dances	2006-05-02	18:00:18	1	http://solosingles.com	
4417749	single dances in atlanta	2006-05-02	18:01:13			
4417749	single dances in atlanta	2006-05-02	18:01:50			
4417749	dry mouth	2006-05-06	16:49:14	2	http://www.mayoclinic.com	
4417749	dry mouth	2006-05-06	16:49:14	8	http://www.wrongdiagnosis.com	
4417749	thyroid	2006-05-06	16:53:34			
4417749	thyroid	2006-05-06	16:55:44			
4417749	competitive market analysis of homes in liburn	2006-05-14	12:14:52			
4417749	competitive market analysis of homes in liburn	2006-05-14	12:16:17			
4417749	competitive market analysis of homes in liburn	2006-05-14	12:16:43			

Source : AOL. AOL has released the search history of 650,000 AOL users, including Arnold, from March 2006. AOL says it has removed identifying information such as names and IP addresses.

The New-York Times



Publication des requêtes de plus de 650 000 utilisateurs (+ de 20 requêtes) sur 3 mois

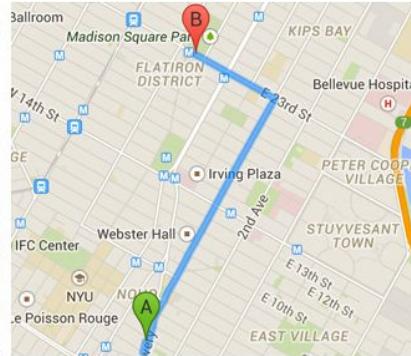
Chaque identifiant (nom d'utilisateur AOL, adresse IP...) avait été remplacé par un nombre choisi aléatoirement.

Mais l'historique de navigation est très identifiant : nom, âge, lieu, goûts

Les taxis new-yorkais exposés



JESSICA BIEL



MAY 6, 2013 • 2:52 PM - 3:07 PM
342 BOWERY TO 8 E. 23RD ST.
\$11.50 FARE • CASH; UNKNOWN TIP • @SPLASH
[Gawker](#)

medallion,hack_license,vendor_id,rate_code,store_and_fwd_flag,pickup_datetime,dropoff_datetime,passenger_count,trip_time_in_secs

6B111958A39B24140C973B262EA9FEA5,D3B035A03C8A34DA17488129DA581EE7,VTS,5,2013-12-03 15:46:00,2013-12-03 16:47:00,1,3660,22.71,-73.813927,40.698135,-74.093307,40.829346

<https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1>

2014 : l'autorité de régulation des taxis libère un fichier “anonymisé” de 173 millions de trajets de taxis new-yorkais.

Identifiants conducteur et véhicule sont hashés.

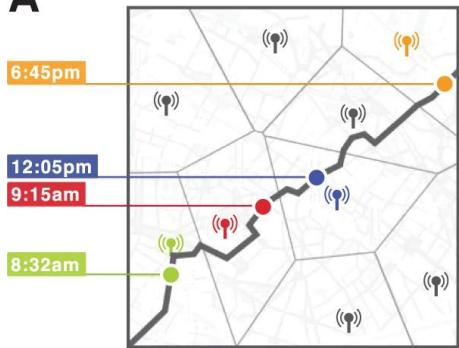
Mais : une ligne d'erreur du fichier permet de déterminer la fonction de Hash MD5

Connaissant la forme des identifiants
retro-engineering possible sur 22 millions de combinaisons.

Conducteurs identifiés, comme des célébrités...

Anonyme mais unique dans la foule

A



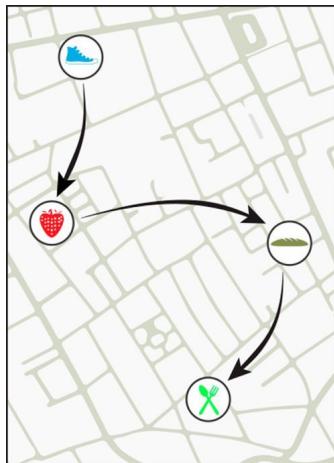
B



[Unique in the Crowd: The privacy bounds of human mobility](#), Nature, 2013

Sur un dataset de données de mobilité d'1.5 million d'utilisateurs sur une période de 15 mois, une équipe de chercheurs du MIT a prouvé en 2013 que 4 coordonnées géo-temporelle (fournies par un appel, un sms, un tweet...) suffisent dans 95% des cas à réidentifier la trace unique d'un individu, même dans un ensemble de données de mobilité clairsemées, à grande échelle et grossières.

Les métadonnées de la CB

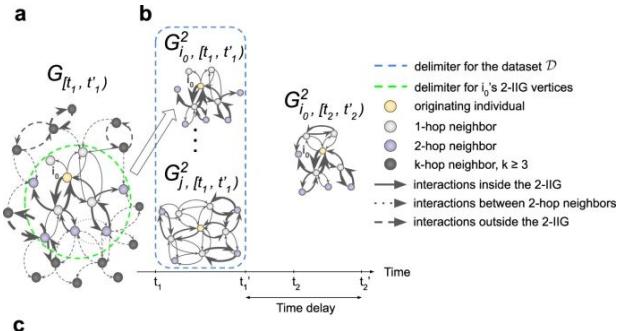


shop	user_id	time	price	price_bin
👟	7abc1a23	09/23	\$97.30	\$49 – \$146
🍓	7abc1a23	09/23	\$15.13	\$5 – \$16
🛒	3092fc10	09/23	\$43.78	\$16 – \$49
🏔️	7abc1a23	09/23	\$4.33	\$2 – \$5
🏊	4c7af72a	09/23	\$12.29	\$5 – \$16
🏔️	89c0829c	09/24	\$3.66	\$2 – \$5
🍴	7abc1a23	09/24	\$35.81	\$16 – \$49

Unique in the shopping mall: On the reidentifiability of credit card metadata,
Science, 2015

L'équipe du [Computational Privacy Group](#) a travaillé sur un dataset des transactions CB de 1.1 millions d'utilisateurs dans 10 000 commerces sur une période de 3 mois. Avec 4 points spatio-temporels ils identifient 90% des historiques d'achat à une personne unique. Et en ajoutant la variable prix on augmente encore la capacité de réidentification.

Les graphes d'interactions sont identifiants



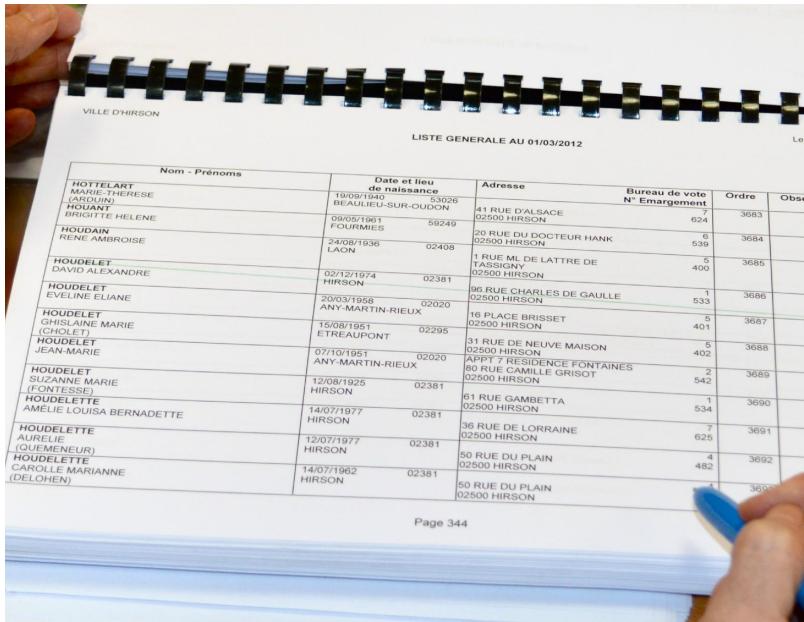
Party A	Party B	Timestamp	Type	Direction	Duration (secs)
aSG64X	rxJKc9	2020-11-05 20:00:05	call	in	325
gvuQjU	dPefYb	2020-11-06 10:23:11	text	in	-
gvuQjU	dPefYb	2020-11-06 10:25:13	text	out	-
gvuQjU	LURKAK	2020-11-06 10:47:20	call	out	124
⋮	⋮	⋮	⋮	⋮	⋮

En entraînant un modèle de réseau de neurones sur les interactions téléphoniques de 40 000 personnes (appel, sms...), pseudonymisées, identification de 52 % des profils en comparant les graphes d'interactions...

<https://www.nature.com/articles/s41467-021-27714-6>

Science, 2022

En France aussi



Selon le Rapport sur la gouvernance et l'utilisation des données de santé (2013) **89 % des patients hospitalisés en 2008** sont identifiables à partir de :

- l'hôpital d'accueil,
- le code postal du domicile,
- le mois et l'année de naissance,
- le sexe,
- le mois de sortie et la durée du séjour

Les listes électorales sont accessibles dans leur intégralité (nom, adresse personnelle, date et lieu de naissance),.

Les failles de l'anonymisation



Dès qu'il est possible d'isoler des données d'un individu dans la base (ex le gouverneur)
= **individualisation**

Dès qu'il est possible de lier des caractéristiques relatives à une personne en utilisant des données déjà disponibles ailleurs (ex Netflix Prize)
= **corrélation**

Dès que l'on peut déduire des informations concernant un individu à partir de la base de données (ex AOL)
= **inférence**

Si ces 3 critères sont parfaitement vérifiés pour un jeu de données, les autorités de protection des données européennes estiment que cela assure qu'il est **véritablement anonyme**.

Psudonymisation vs anonymisation

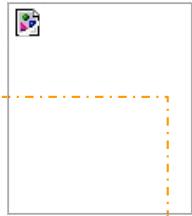


PSEUDONYMISATION

- Processus réversible
- Entre dans le cadre du RGPD : information et consentement nécessaire de l'utilisateur
- Les identifiants directs remplacés par un pseudonyme “data masking”

Exemples d'utilisation :

- Études nécessitant des données fines à l'échelle individuelle (e.g. étude de parcours de soin individuel)



ANONYMISATION

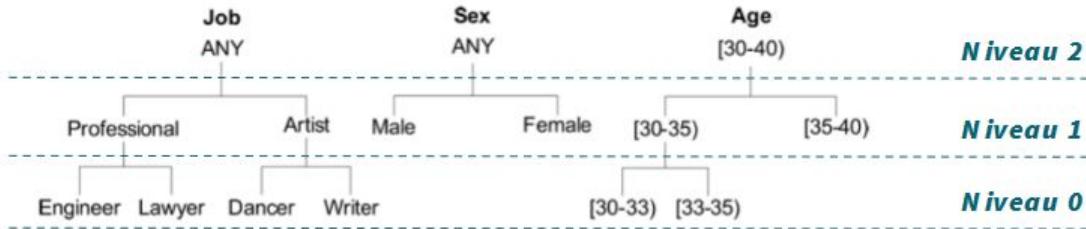
- Un processus irréversible
- Sortie du cadre des lois de protection des données personnelles
- Complexité des mesures à mettre en œuvre : supprimer ou transformer les valeurs d'un attribut peut ne pas suffire

Exemples d'utilisation :

- Statistiques génériques (e.g. populationnelles)
- Partage libre de données, pour assurer la reproductibilité de la recherche par exemple

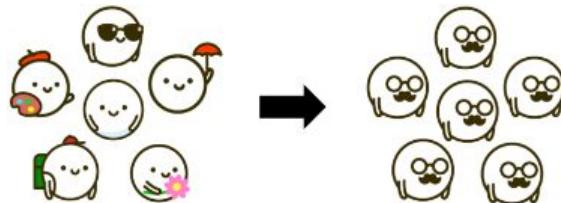


Techniques de généralisation : K-anonymat



Empêcher qu'une personne puisse être isolée en la regroupant avec, au moins, k autres individus -> K-anonymat

→ Remplacer chaque valeur d'un quasi-identifiant par des valeurs moins spécifiques jusqu'à obtenir une base de données où chaque entrée se confond avec au moins k autres entrées.



Ressources

Générales

Techniques d'anonymisation - Benjamin Nguyen

43 Techniques d'anonymisation

Thèse de Feten BEN FREDJ (Docteur du Conservatoire National des Arts et Métiers) :

Méthode et outil d'anonymisation des données sensibles

Attaques sur les modèles de machine learning pour reconstruire le training set (entre autres):

<https://github.com/stratosphereips/awesome-ml-privacy-attacks>

Confidentialité différentielle

Modèles de machine learning avec confidentialité différentielle

<https://github.com/IBM/differential-privacy-library>

<https://github.com/tensorflow/privacy>

Podcast Post Mortem : comment Google anonymise vos données personnelles avec la Differential Privacy Docteur Damien Desfontaines - Senior Software Engineer, Privacy pour Google <https://podcast.ausha.co/postmortem/12>

K-Anonymat

Librairies python

<https://github.com/glassonion1/anonypy>

<https://github.com/kaylode/k-anonymity>

- Législation
- Anonymisation
- Modèles génératifs
- Biais et discrimination
- Transparence
- Impact sociétal
- Impact Environnemental

Fake news, comment s'y retrouver ?

AFP Factuel Follow

Trump arrêté, Obama et Merkel à la plage, Macron éboueur, Poutine agenouillé devant Xi Jinping...

Les images générées par intelligence artificielle ancrées dans l'actualité se multiplient en ligne, semant parfois la confusion. Comment les reconnaître ? #AFP 1/8

4:20 PM · Mar 25, 2023

Read the full conversation on Twitter

724 Reply Copy link

Read 20 replies

Discussion

Midjourney

Rowan Atkinson from Version 1 through Version 5 of Midjourney's render.

The same prompt was used, except for replacing the version number at the end of each one. The most 'Rowan'-like image from each group of four was selected.



À vous de dire !

You against the machine: Can you spot which image was created by A.I.?



<http://bit.ly/3nE9Z2G>

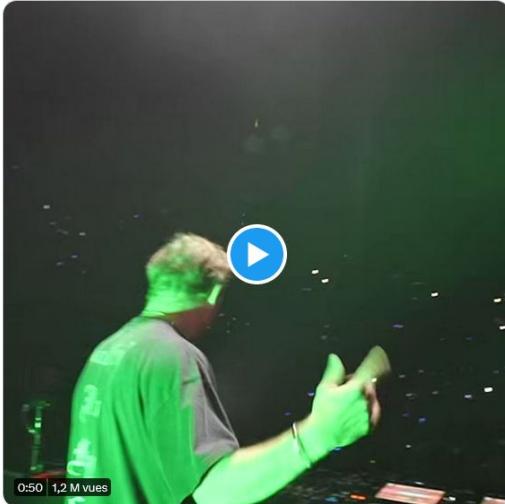
Droit d'auteur



David Guetta

@davidguetta

Let me introduce you to... Emin-AI-em ::



9:23 PM · 3 févr. 2023 · 2,1 M vues

1 483 Retweets 902 citations 8 098 J'aime 692 Signets



- Procès New York Times contre OpenAI (ChatGPT) pour non respect du droit d'auteur
- Pour la musique ? Pour les images ?
- Le “style” ou “à la manière de” n'est pas une contrefaçon, mais se plaide...
- Prix Goncourt au Japon gagné grâce à ChatGPT

Prompter c'est tricher ?



Des outils pour vérifier...

<https://gptzero.me/>

<https://contentatscale.ai> (aide aussi à masquer l'IA...)

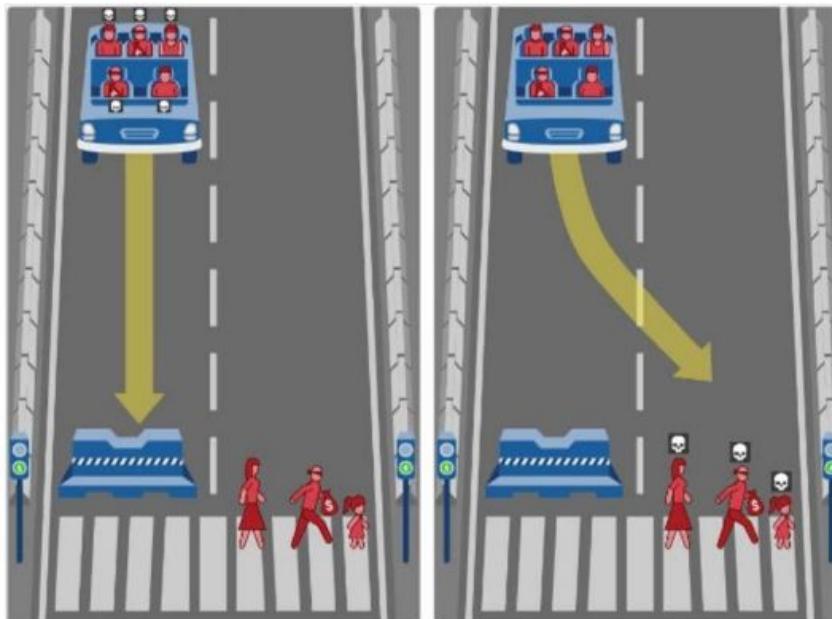
<https://openai-openai-detector.hf.space/> (déjà
dépassé...)

- Est-ce que l'IA flatte notre paresse ?
- Doit-on laisser nos apprentissages à des machines ?
- Comment structurer notre pensée en interaction avec un langage automatisé, dressé à partir de données du passé ?

Oblige à repenser la pédagogie, peut-être moins évaluer le livrable lui même que le moyen d'y parvenir

<https://help.openai.com/en/collections/5929286-educator-faq>

Responsabilité juridique des machines autonomes



<https://www.moralmachine.net/hl/fr>

Moral Machine

A platform for public participation in and discussion of the human perspective on machine-made moral decisions

 Moral Machine

- Législation
- Anonymisation
- Modèles génératifs
- **Biais et discrimination**
- Transparence
- Impact sociétal
- Impact Environnemental

Des algorithmes éthiques ?

Egalité/Équité de traitement

Procédure la même pour tous ou les individus aux profils proches seront traités à l'identique, sans défavoriser arbitrairement un groupe en particulier.

Transparence

Description claire du process et des priorités

Intelligibilité

Les personnes concernées comprennent

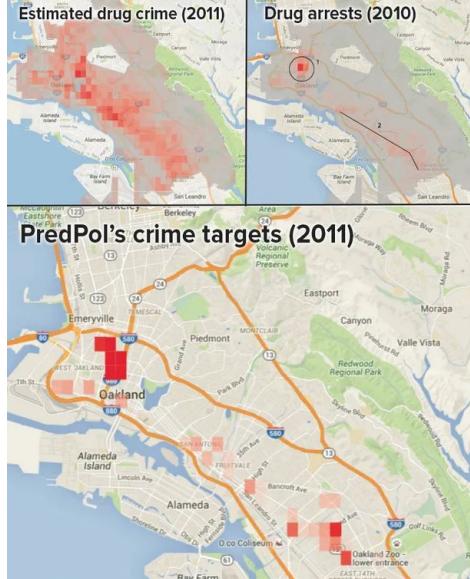
Loyauté

Algo conforme à la description, no dark pattern

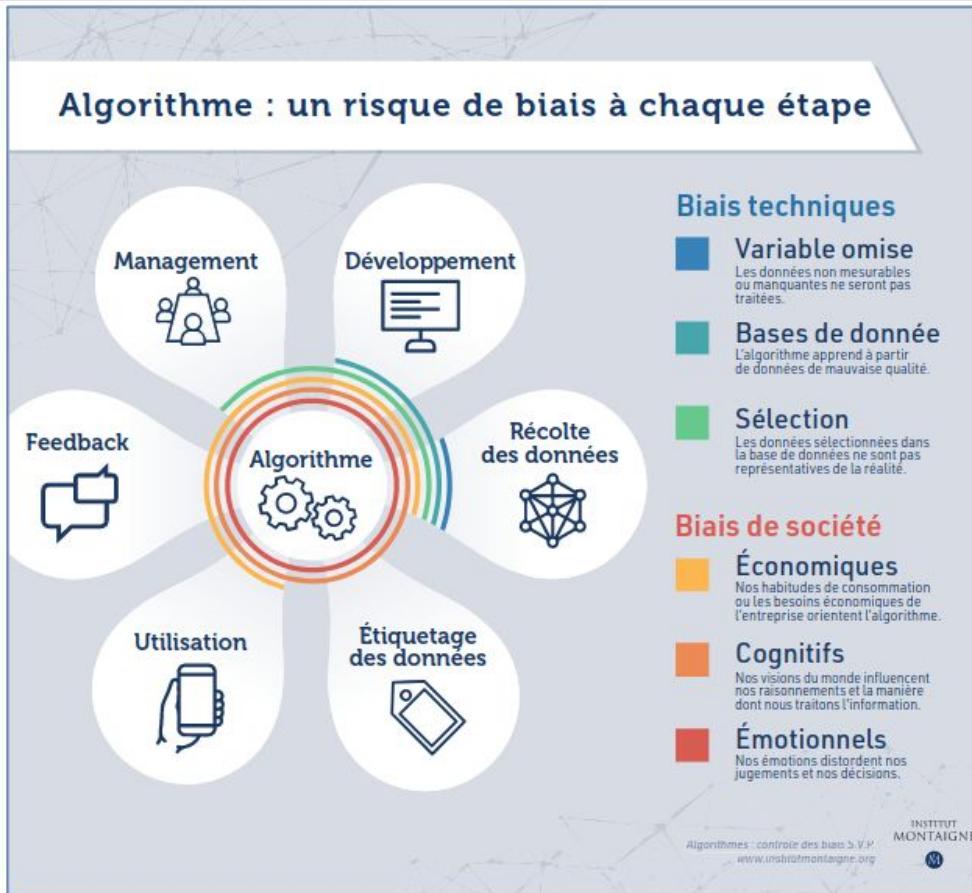
Biais des algorithmes et des données

Cas avérés

- ❑ Police prédictive : [Weapons of Math Destruction | Cathy O'Neil | Talks at Google](#)
- ❑ Recrutement chez Amazon: préfiltrage des profils
 - Femme dévaluées pour les postes de cadres
- ❑ Reconnaissance faciale
 - 10x à 100x plus d'erreur en moyenne sur les personnes d'Afrique/Asie de l'Est
- ❑ ...



Biais des algorithmes et des données



Biais des algorithmes et des données

Caractéristiques d'un algorithme biaisé à risque

- ❑ Les biais sont en général déjà présents dans les données, mais aussi dans la tête des concepteurs
 - 25 critères de discriminations protégés par la loi Française : Origine, sexe, âge, situation familiale, grossesse, apparence physique, situation économique, patronyme, état de santé, perte d'autonomie, handicap, mœurs, caractéristiques génétiques, orientation sexuelle, genre, opinions politiques ou philosophiques, langues, appartenance réelle ou supposée à une ethnie, nation ou prétendue race, lieu de résidence, domiciliation bancaire.
 - Sauf exceptions “objectives et raisonnables” (assurances, ...)
- ❑ Automatisation à grande échelle de la reproduction des discriminations
 - Pour l'instant les implications sont principalement aux États-Unis, mais arrivent rapidement en France

Etude d'impact et audit des algorithmes

Overall Impact Risk		Overall Direction			
		Positive	Mostly Positive	Mostly Negative	Negative
Scope	Very Narrow	Very low	Very low	Low	Moderate
	Limited/Narrow	Very low	Low	Moderate	Significant
	Substantial	Low	Moderate	Significant	High
	Broad/wide ranging	Moderate	Significant	High	Extreme

Appropriate Use Risk Score		Reputation and Perception		
		Supportive	Mixed	Not Supportive
Consistency and Compatibility	Yes	Low	Low	Medium
	Somewhat	Low	Medium	High
	Unknown	Medium	Medium	High
	No	Medium	High	High

Accountability Risk		Automation Score		
		Low - human mediated	Medium - algorithm mediated	High - algorithmically determined
Accessibility Score	Accessible	Low	Low	Medium
	Some concerns	Low	Medium	High
	Major concerns	Medium	High	High

Overall Technical Bias Risk		Methodology Risk		
		Low	Medium	High
Representativeness and Inaccuracy Risk Score	Low	Low	Low	Medium
	Medium	Low	Medium	High
	High	Medium	High	High

- ❑ Auditer le risque, l'impact et la conformité réglementaire = contrôle qualité de l'algorithme
- ❑ En interne et en externe (si pas de contre-indication commerciale)
- ❑ En amont, pendant la conception et après le déploiement.
- ❑ Mettre en place des corrections ou atténuations (comités citoyens, comités d'auditeurs experts, intervention humaine, outils de test automatisés, etc.)

Un serment d'Hyppocrate du data scientist ?

C'est pourquoi je m'engage à respecter les 5 principes suivants :

1

Intégrité scientifique et rigueur

J'exploiterai les données avec toute la rigueur requise et en conformité avec les meilleurs standards de ma profession.

2

Transparence

J'informeraï de façon compréhensible et précise toutes les parties prenantes sur les finalités, les modalités et les implications potentielles de mon utilisation des données.

3

Équité

Je veillerai à toujours m'assurer que des individus ou des groupes ne soient pas discriminés par rapport à des critères illégaux ou illégitimes, de façon directe ou indirecte, sur la base de mes travaux sur les données.

4

Respect

J'exercerai mon activité professionnelle en respectant la vie privée et la dignité des personnes dans toutes leurs dimensions.

5

Responsabilité et indépendance

J'assumerai mes responsabilités en cas manquements ou de conflits d'intérêt et je donnerai l'alerte si des actes illégaux liés à des données sont constatés.

Data science responsable et de confiance - Référentiel d'évaluation

Le [référentiel d'évaluation](#) ci-dessous est le fruit du travail participatif initié au printemps 2019 par Labelia Labs (ex- Substra Foundation) et en cours depuis. Il procède de l'identification des [risques](#) que l'on cherche à prévenir en visant une pratique responsable et de confiance de la data science, et des bonnes pratiques qui permettent d'y faire face. Il regroupe également pour chaque sujet des ressources techniques qui peuvent être de bons points d'entrée pour les organisations intéressées.

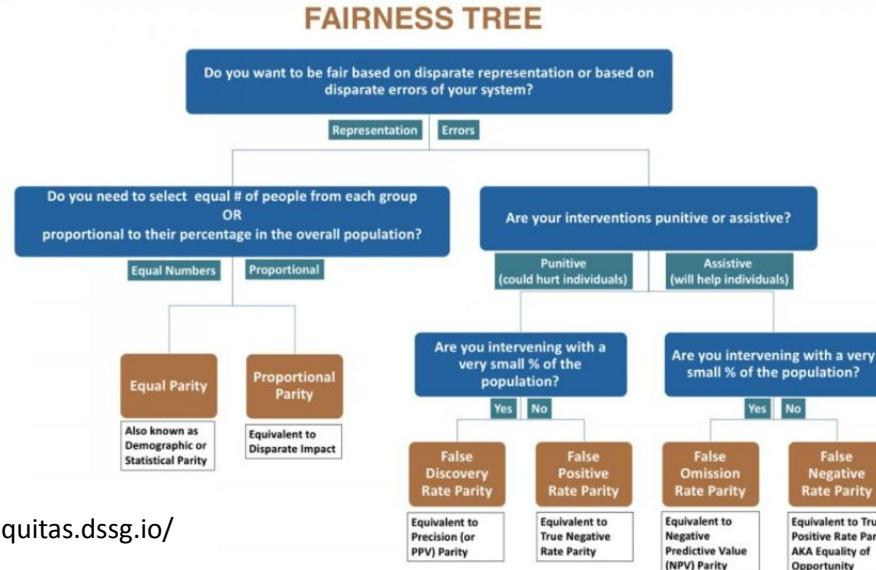
Dernière mise à jour : 2ème semestre 2022.

Référentiel d'évaluation de la maturité d'une organisation

L'évaluation est composée des 6 sections suivantes :

- Section 1 - Protéger les données personnelles ou confidentielles
- Section 2 - Prévenir les biais, élaborer des modèles non discriminatoires
- Section 3 - Évaluer la performance de manière rigoureuse
- Section 4 - Assurer la reproductibilité des modèles et en établir la chaîne de responsabilité
- Section 5 - Utiliser des modèles en confiance et de manière responsable
- Section 6 - Anticiper, suivre et minimiser les externalités négatives de l'activité data science

En pratique : la classification



En pratique : la classification

- Choix de la mesure d'équité
 - Selon l'application
 - Généralement mutuellement exclusives [1.2], si plusieurs nécessaires, à discuter en comité d'éthique
- Si vous ne disposez pas des variables sensibles pour tester :
 - Appliquer les algos sur des bases de données publiques similaires
 - Collecter ces données de façon optionnelle sur consentement de l'utilisateur dans le seul but de débiaiser
- Débiaiser les modèles, plusieurs approches :
 - Pre-processing : Subsampling, reweighing, ...
 - In-processing : Adversarial debiasing, regularization, ...
 - Post-processing : Decision threshold optimization, model selection, ...

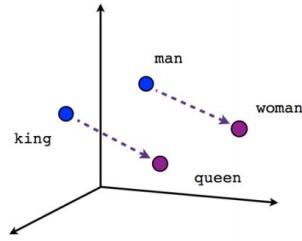
En pratique : la classification

Outils disponibles

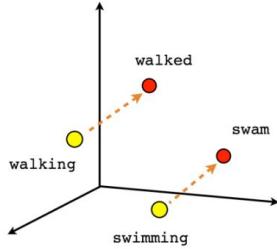
- ❑ Évaluer les métriques d'équité sur vos modèles :
 - <http://aequitas.dssg.io/upload.html>
 - <https://github.com/dssg/aequitas>
- ❑ Corriger les biais de vos modèles :
 - Fairlearn
 - https://fairlearn.org/main/user_guide/fairness_in_machine_learning.html
 - Doc ++, modèles de régression
 - IBM AI Fairness 360
 - <https://aif360.readthedocs.io/en/latest/index.html>
 - Doc -, fonctionnalités ++
- ❑ Visualisation
 - Google What If Tool
 - <https://pair-code.github.io/what-if-tool/>

Autres domaines du ML

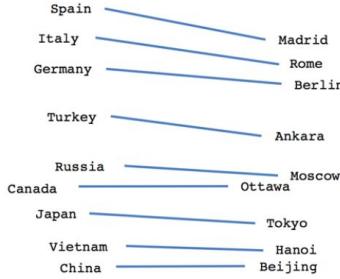
Natural language processing : word embeddings



Male-Female



Verb tense



Country-Capital

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Biais :

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker.}}$$

Traitement du language : google traduction

The screenshot shows the Google Translate interface. At the top, it displays language pairs: "Anglais - DéTECTé" and "Français" on the left, and "Français" and "Anglais" on the right. Below this, there are two main translation boxes. The left box contains the English sentence "The nurse shaved his beard" and has a delete button ("X") to its right. The right box contains the French translation "L'infirmière lui a rasé la barbe" and includes a star icon for favoriting. Both boxes have small icons at the bottom: a microphone for audio, a speaker for playback, and a pencil for editing. The center of the screen shows the progress "26 / 5 000".

"The nurse arrived"

"The nurse shaved her beard"

"The nurse shaved his beard"

"The nutritionist arrived"

"The nutritionist is clever"

"The nutritionist is beautiful"

Traitements du langage : ChatGPT



How did you come to the conclusion that "she" refers to the paralegal?



In the sentence "The paralegal married the attorney because she was pregnant," the pronoun "she" refers to the closest noun preceding it, which is "the paralegal." Therefore, "she" in this sentence refers to the paralegal. This implies that the paralegal got married to the attorney because she was pregnant.



Isn't "the attorney" the closest noun preceding it?



You are correct. I apologize for my mistake. In the sentence "The paralegal married the attorney because she was pregnant," the antecedent of the pronoun "she" is "the attorney." Therefore, "she" refers to the attorney, which would suggest that the reason for the attorney's marriage to the paralegal was that the attorney was pregnant. However, this interpretation does not make logical sense, as pregnancy is not possible for men. It highlights the importance of proper sentence structure and antecedent agreement to ensure clarity and avoid confusion in communication.

Solutions pas seulement techniques



“Fairness is not just about data sets, and it’s not just about math. Fairness is about society as well, and as engineers, as scientists, we can’t really shy away from that fact.”

Timnit Gebru, Ex-co-directrice de l'équipe Ethical AI de Google, Juin 2020



Facebook (2018)

- ❖ 4% Black workers
- ❖ 5% Hispanic workers

Microsoft (2018)

- ❖ 4% Black workers
- ❖ 6% Latinx workers

Google (2018)

- ❖ 2.5% Black workers
- ❖ 3.6% Latinx workers

[Discriminating Systems: Gender, Race, and Power in AI. AI Now, 2019]

Biais des algorithmes et des données

Préconisations

- ❑ Conduire des tests d'équité et débiaiser
 - Nécessite les données relative aux 25 critères, parfois difficile en France
- ❑ Mettre de la diversité
 - Dans les équipes de développement
 - Dans les équipes projet : associer les sciences sociales (géographes, sociologues, etc.)
- ❑ Mettre en place un comité d'éthique dans votre structure impliquant le management
 - Introduction d'une nouvelle variable, arbitrage à réaliser entre performance et équité, définition d'un critère d'équité à évaluer, ...
- ❑ Évaluer l'impact de l'algorithme
 - Fort impact selon la Commission Européenne
 - Secteur défini comme à risque (santé, transport, police...)
 - Présente des effets juridiques ou un risque d'atteinte physique ou matérielle pour la personne ou l'entité visée
 - Ethics and Algorithm toolkit : <http://ethicstoolkit.ai>
 - Audit interne et externe de l'impact
 - Pour les algorithmes à fort impact :
 - Droits de recours
 - Transparence
 - Labellisation
 - <https://dataethics.eu/data-ethics-labels-are-coming-up/>
 - Audit par des tiers

Biais des algorithmes et des données

Ressources

- ❑ [Algorithmes : contrôle des biais SVP](#)
- ❑ [Concrètement, comment rendre les algorithmes responsables et équitables ?](#)
- ❑ <http://ethicstoolkit.ai/> : pour évaluer l'impact de l'algo et les risques associés aux données
- ❑ [Substra Assessment](#) : outil pour une data science responsable et de confiance
- ❑ [ACM FAccT Conference](#) : ACM Conference on Fairness, Accountability, and Transparency
- ❑ [Aequitas – Center for Data Science and Public Policy](#) : outil open source pour auditer son algo
- ❑ <https://dataethics.eu/data-ethics-labels-are-coming-up/> : labels sur l'éthique des données
- ❑ [Investigating the Impact of Gender on Rank in Resume Search Engines](#) : exemple d'audit externe
- ❑ [Algorithmic Equity Toolkit | ACLU of Washington](#) : outil plutôt orienté associations, citoyens
- ❑ [Fairness in Machine Learning — Substra Foundation](#) : Définitions et applications de l'équité
- ❑ [Coded Bias | Netflix trailer](#) : documentaire sur les biais algorithmiques

50

Articles scientifiques :

- ❑ [The ethics of algorithms: key problems and solutions \[1.1\]](#)
- ❑ [Inherent Trade-Offs in the Fair Determination of Risk Scores \[1.2\]](#)
- ❑ [Fairness Definitions Explained \[1.3\]](#)
- ❑ [Man is to Computer Programmer as Woman is to Homemaker?Debiasing Word Embeddings \[1.4\]](#)

- Législation
- Anonymisation
- Modèles génératifs
- Biais et discrimination
- Transparency
- Impact sociétal
- Impact Environnemental

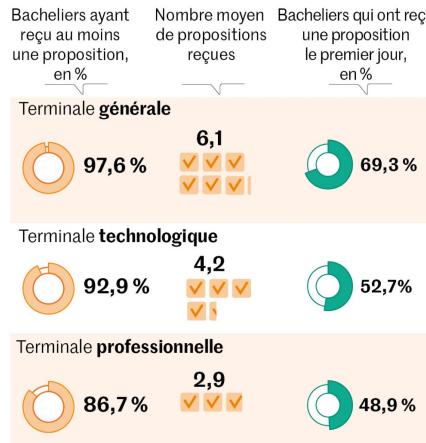
Transparence, explicabilité et interprétabilité



Le cas parcoursSup

Comment ça marche ?

Au moins une proposition pour 95 % des bacheliers



Dates-clés pour 2023

-  Du 18 janvier au 9 mars
Inscription et formulation des vœux
-  Du 10 mars au 6 avril
Confirmation des candidatures
Les lycéens doivent compléter leur dossier et confirmer la liste finale de leurs vœux avant le 6 avril.
-  A partir du 7 avril
Examen des dossiers par les formations
-  Du 1^{er} juin au 13 juillet
Phase principale d'admission

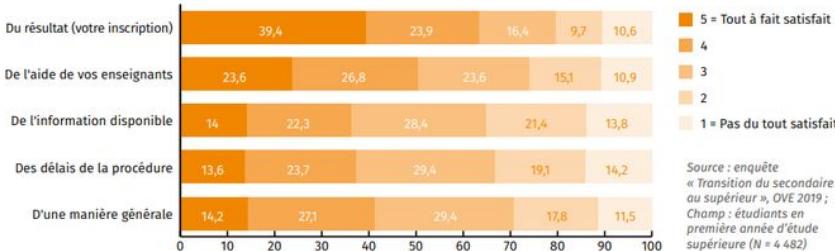
Infographie : *Le Monde*
Source : Parcoursup, campagne 2022, Traitement SIES

- Apparier 940 000 candidats, près de 12 millions de voeux pour 21 000 formations d'enseignement supérieur
- Remplace en 2018 le décrié APB : plus de tirage au sort et intervention humaine
- Introduction d'un critère social avec un quota de boursiers proportionnel au nombre de candidats pour chaque formation
- Les établissements d'enseignement supérieur utilisent des OAD (algorithmes locaux) avec chacun leurs critères spécifiques
- Publication du code la veille de la première mise en ligne + comité éthique associé

Il y a un bug ?

GRAPHIQUE 2 : ÉVALUATION DE LA PROCÉDURE PARCOURSUP PAR LES ÉTUDIANTS SELON DIFFÉRENTS CRITÈRES (EN %)

Dans quelle mesure êtes-vous satisfait(e) de la procédure d'orientation Parcoursup ?



- ❑ Dans les perceptions, défaut de compréhension et d'information du système, stress évoqué par 82% des lycéens
- ❑ Primauté de la note brute : le système de notation est-il le même partout ?
- ❑ Critère social ? Renforcement des inégalités avec la fixation des objectifs de taux de boursiers et en déléguant la sélection finale aux établissements
- ❑ Plus d'obligation de proximité : favorise les classes aisées informées
- ❑ Utilisation de critères spécifiques à chaque établissement quand les demandes excèdent le nombre de place
- ❑ Licences de science politique de Paris 1 et Sciences Po Bordeaux ont travaillé leurs algos pour s'ouvrir plus

Histoire d'une transparence sélective

=> Utilité du débat public sur la conception algorithme

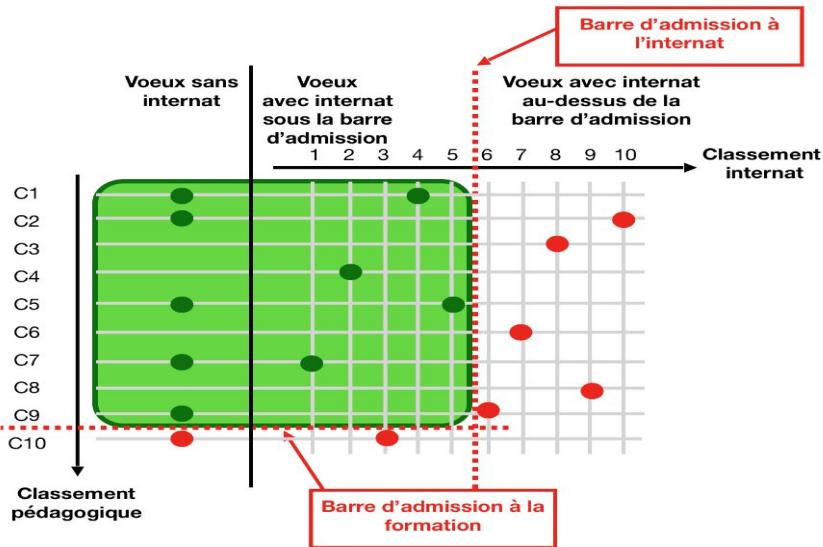


Schéma officiel de l'exécution de l'algorithme concernant la sélection pour les places d'internat

- Mars 2018 : la Loi ORE instaurant ParcourSup et la transparence de son algorithme, préserve le secret des délibération des jurys locaux et donc de leurs outils d'aide à la décision,
- Avril - juin 2018 : débats au sénat sur l'opacité des algorithmes locaux
- Janvier 2019 : CADA et Défenseur des droits incitent le gouvernement à faire publier ces algorithmes locaux
- Mars 2019 : "les algorithmes locaux n'existent pas" Frédérique Vidal, ministre de l'enseignement supérieur
- Février 2020 la Cour des Comptes recommande de rendre public les algorithmes locaux
- Avril 2020 : Conseil Constitutionnel impose la publication des algorithmes locaux suite à un recours d'un syndicat étudiant
- Septembre 2022 : question écrite d'un sénateur demandant que soit appliquée la décision du CC
- Janvier 2023 : ouverture des voeux, nouvel onglet « Comprendre les critères d'analyse des candidatures », à la discrétion des formations...

Transparence, explicabilité et interprétabilité



Face Detection

Model Card v0 Cloud Vision API

The model analyzed in this card detects one or more faces within an image or a video frame, and returns a box around each face along with the location of the faces' major landmarks. The model's goal is exclusively to identify the existence and location of faces in an image. It does not attempt to discover identities or demographics.

On this page, you can learn more about how well the model performs on images with different characteristics, including face demographics, and what kinds of images you should expect the model to perform well or poorly on.

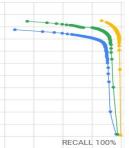
MODEL DESCRIPTION



Input: Photo(s) or video(s)

Output: For each face detected in a photo or video, the model outputs:

PERFORMANCE



Explore

Object Detection

About Model Cards



Limitations

The following factors may degrade the model's performance:



Overview

Limitations

Trade-offs

Performance

Test your own images

Provide feedback

Explore

Object Detection

About Model Cards

Face Size: Depending on image resolution, faces that are distant from the camera (a *Euclidean distance* of < 10px) might not be detected. Not designed for estimating the size of a crowd.

Faces greater than 90% of image height or width might not be detected.

Facial Orientation: Needs visible facial landmarks such as eyes, noses, and mouths to work correctly. Faces that are looking away from the camera (pan > 90°, roll > 45°, or tilt > 45°) might not be detected.

Lighting: Poorly illuminated faces might not be detected.



Occlusion: Partially hidden or obstructed

Blur: Blurry faces might not be detected.

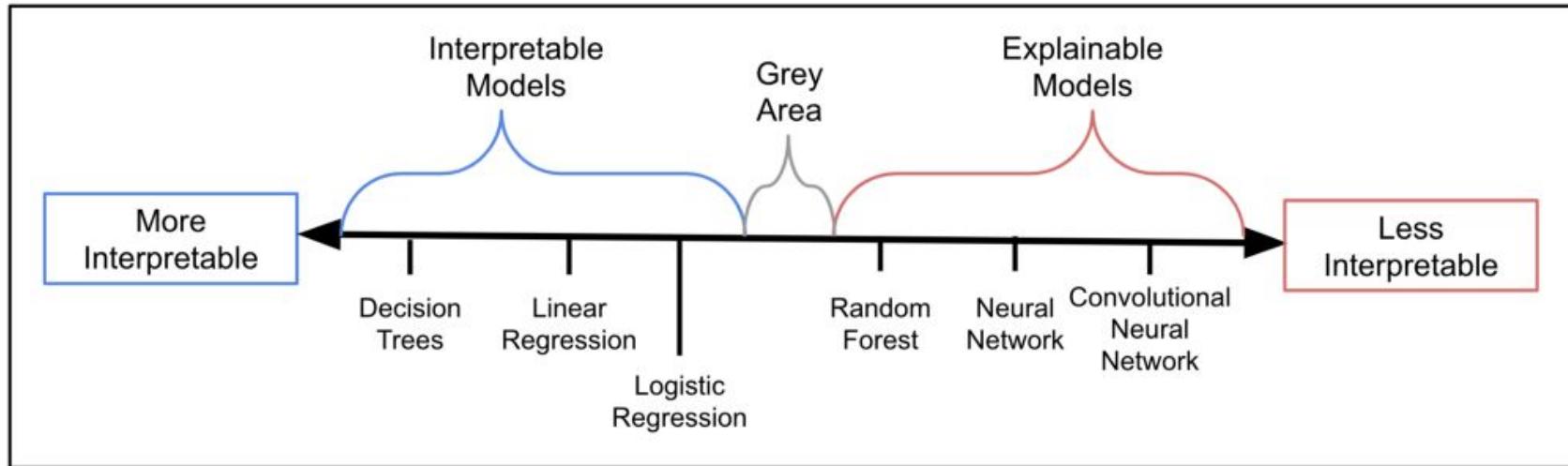
Motion (in video): Rapid movement

L'algorithme a bon dos : les critères de choix derrière l'algorithme sont éminemment humains (politiques, sociologiques, bureaucratiques, etc.)

Préconisations

- ❑ Publier le code si possible
- ❑ À minima publier les features utilisées
- ❑ [Google Cloud Model Cards](#)
- ❑ Mettre en place un comité d'éthique (bis)
- ❑ Évaluer l'impact de l'algorithme (bis)
- ❑ Machine learning : utiliser des modèles interprétables

Explicabilité vs Interprétabilité



Explicabilité vs Interprétabilité

- ❑ Modèle interprétable (white box) : qui est compris et avec lequel un humain peut raisonner
- ❑ Favoriser les modèles interprétables dans la mesure du possible
 - Analyse de l'importance des features directe (Decision trees, linear models)
 - Les modèles explicables ne le sont jamais totalement (sinon on pourrait les remplacer par leurs explications) : explicabilité est par nature limitée [2.2]
 - Explicabilité donne l'importance des features, mais pas le sens qui en est fait
 - Possible baisse de performances (surtout en NLP et en CV)
- GA2MS :
$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$
 - <https://blog.fiddler.ai/2019/06/a-gentle-introduction-to-ga2ms-a-white-box-model>

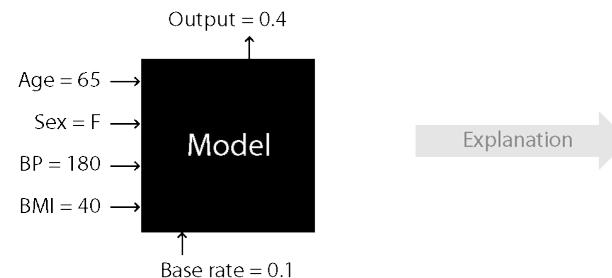
Feature importance des modèles “black box”: Explicabilité locale (pour 1 exemple)

SHAP (SHapley Additive ExPlanations)

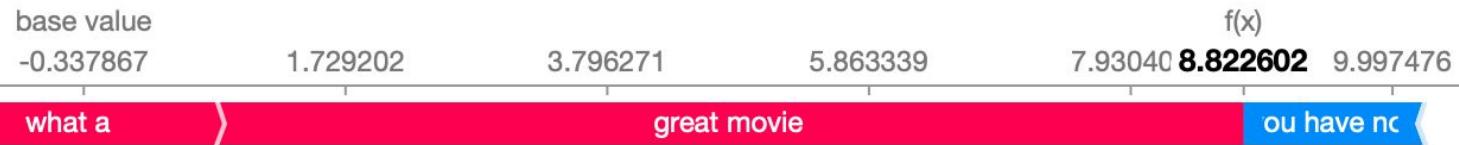
- ❑ Feature importance pour tout modèle
- ❑ Pour une prédiction particulière
- ❑ Contribution de chaque feature à la différence d'une prédiction par rapport à la prédiction moyenne
- ❑ <https://github.com/slundberg/shap>



SHAP



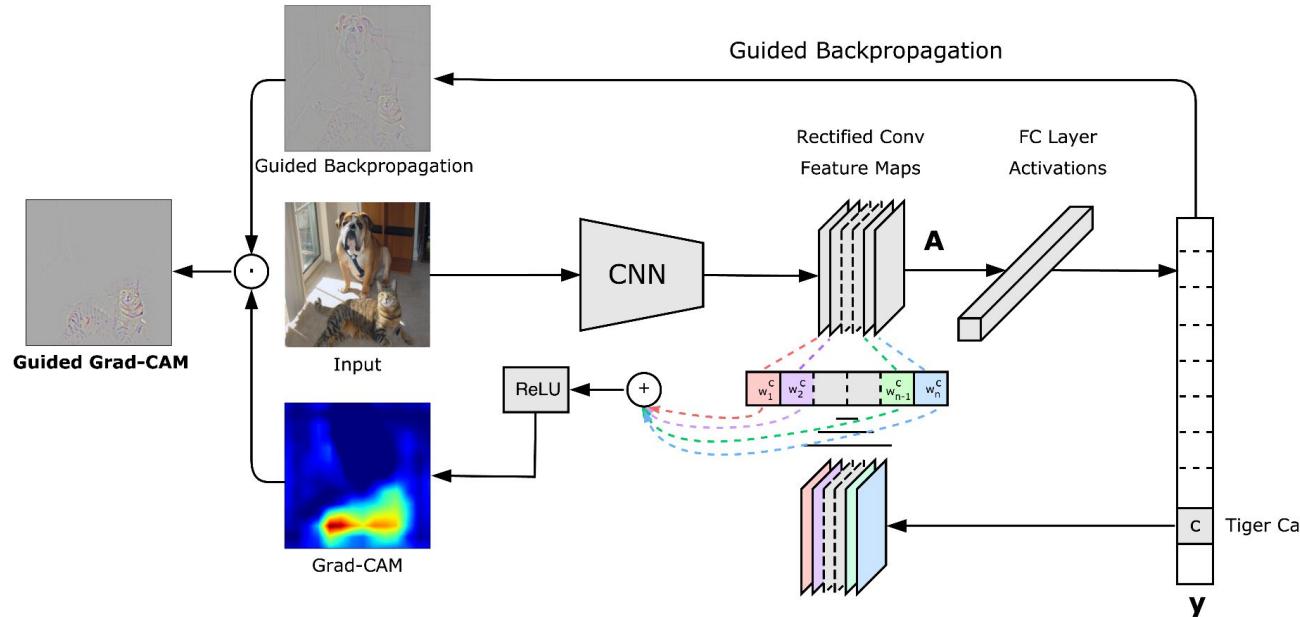
Explicabilité



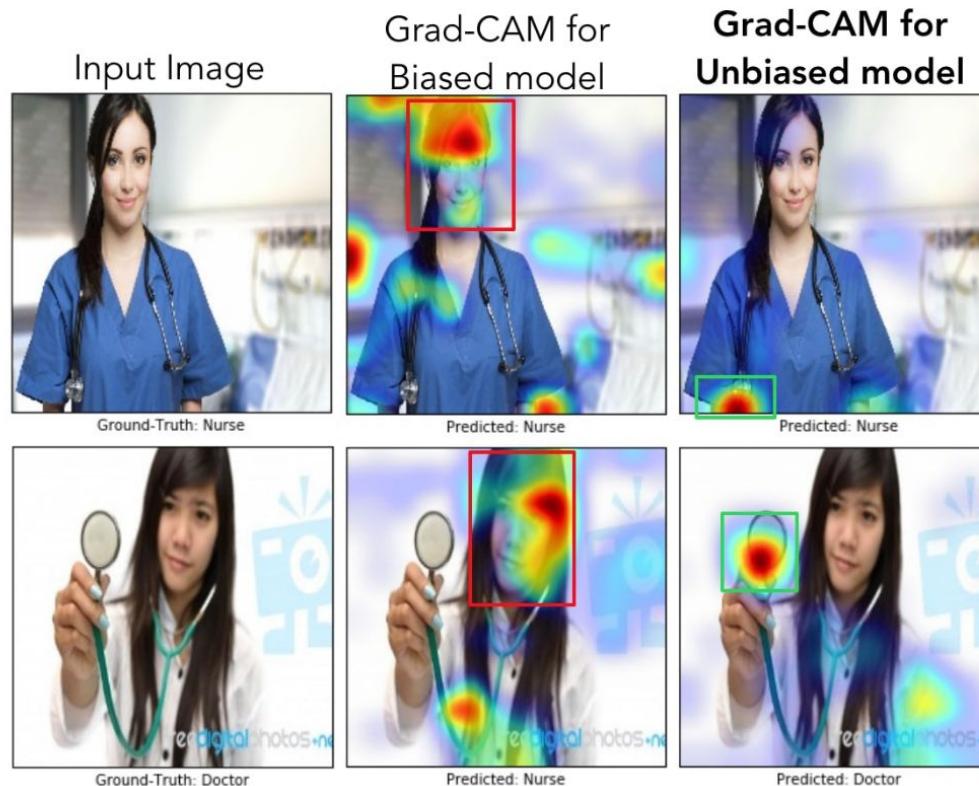
what a great movie! . . . if you have no taste .

Feature importance des modèles “black box”: Explicabilité locale (pour 1 exemple)

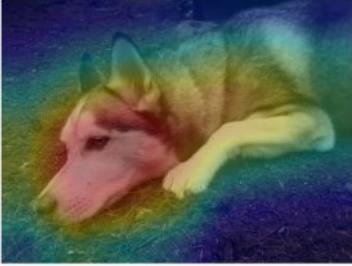
Computer Vision : gradient-based saliency maps : Grad-CAM [2.3]



Explicabilité pour les modèles de Computer Vision



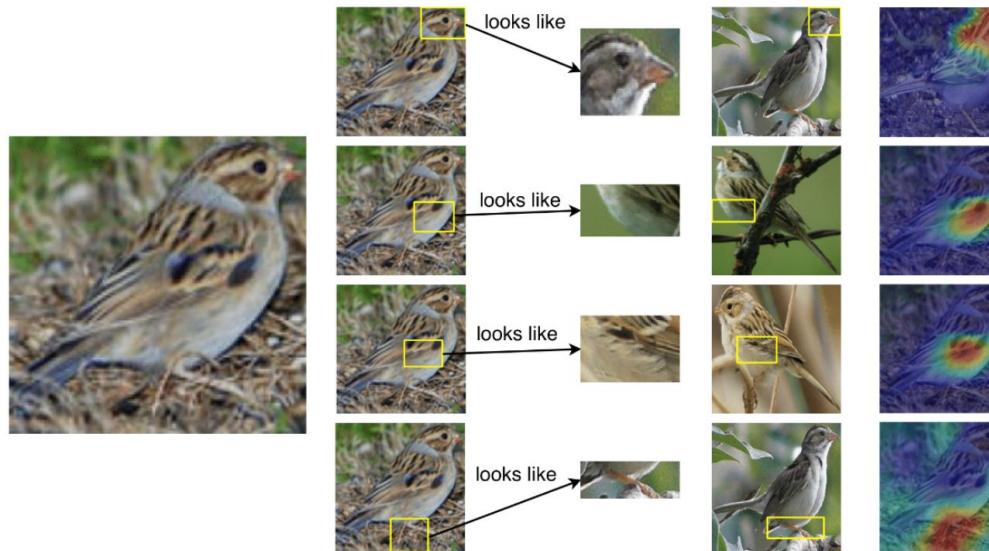
Explicabilité pour les modèles de Computer Vision

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

L'explicabilité a ses limites ...

Explicabilité pour les modèles de Computer Vision

Prototypical Part Networks (ProtoPNet)



Feature importance des modèles “black box”: explicabilité globale

Sensitivity analysis

- ❑ Principe : générer aléatoirement des échantillons X pour évaluer l'impact de chaque feature sur la variance de la prédiction \hat{y}
- ❑ Nécessite uniquement la fonction de prédiction
- ❑ Plusieurs méthodes : Morris, Sobol, FAST, ...
 - Différentes métriques selon les méthodes
 - Plusieurs niveaux d'interactions entre les features selon les méthodes
 - <https://salib.readthedocs.io/en/latest/> : SALib

Feature importance des modèles “black box”: explicabilité globale

Sensitivity analysis

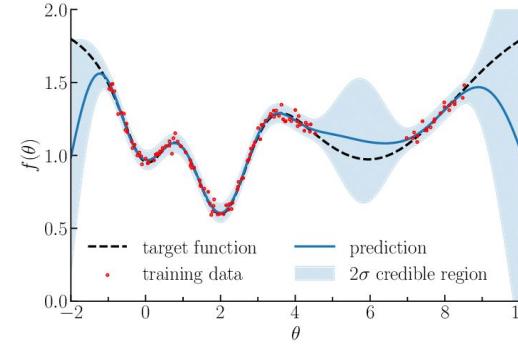
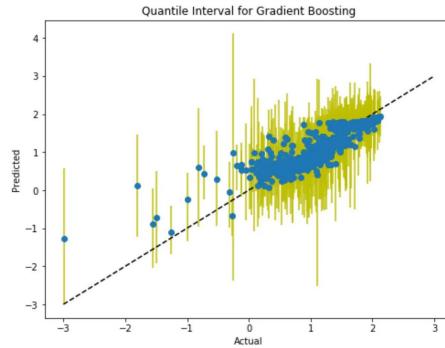
- ❑ Principe : générer aléatoirement des échantillons X pour évaluer l'impact de chaque feature sur la variance de la prédiction \hat{y}
- ❑ Nécessite uniquement la fonction de prédiction
- ❑ Plusieurs méthodes : Morris, Sobol, FAST, ...
 - Différentes métriques selon les méthodes
 - Plusieurs niveaux d'interactions entre les features selon les méthodes
 - <https://salib.readthedocs.io/en/latest/> : SALib
 -

Une seule librairie pour toutes les approches (glassbox, locales et globales) :

- ❑ InterpretML : <https://github.com/interpretml/interpret>

Parenthèse estimation d'incertitude

Produire un intervalle de prédiction pour les modèles à fort impact



- ❑ Modèles bayésiens : Bayesian ridge regression, Gaussian processes
- ❑ Apprentissage ensembliste : Random forests, Boosting, ...
- ❑ Méthodes d'estimation pour n'importe quel modèle
 - Jackknife, conformal predictions, ...
 - MAPIE : <https://mapie.readthedocs.io>
- ❑ Intervalles à valider en calculant le coverage réel sur un set de validation !
- ❑ <https://medium.com/analytics-vidhya/uncertainty-quantification-ac33ca98ba3e>
- ❑ <https://www.bryanshalloway.com/2021/03/18/intuition-on-uncertainty-of-predictions-introduction-to-prediction-intervals/>

Ressources

Parcours Sup

- [Parcoursup : archéologie des « algorithmes locaux » et de leur dissimulation](#)
- [Que vaut la data map qui géolocalise les voeux des candidats sur Parcoursup ?](#)
- [Parcoursup ou la sélection par les algorithmes](#)
- [Document de présentation des algorithmes de Parcoursup](#)
- [Parcoursup / algorithmes-de-parcoursup](#)
- [Parcoursup 2020 - vœux de poursuite d'études et de réorientation dans l'enseignement supérieur et réponses des établissements](#)
- [Parcoursup : infox et premières conséquences de la réforme](#)

Explicabilité & Interprétabilité

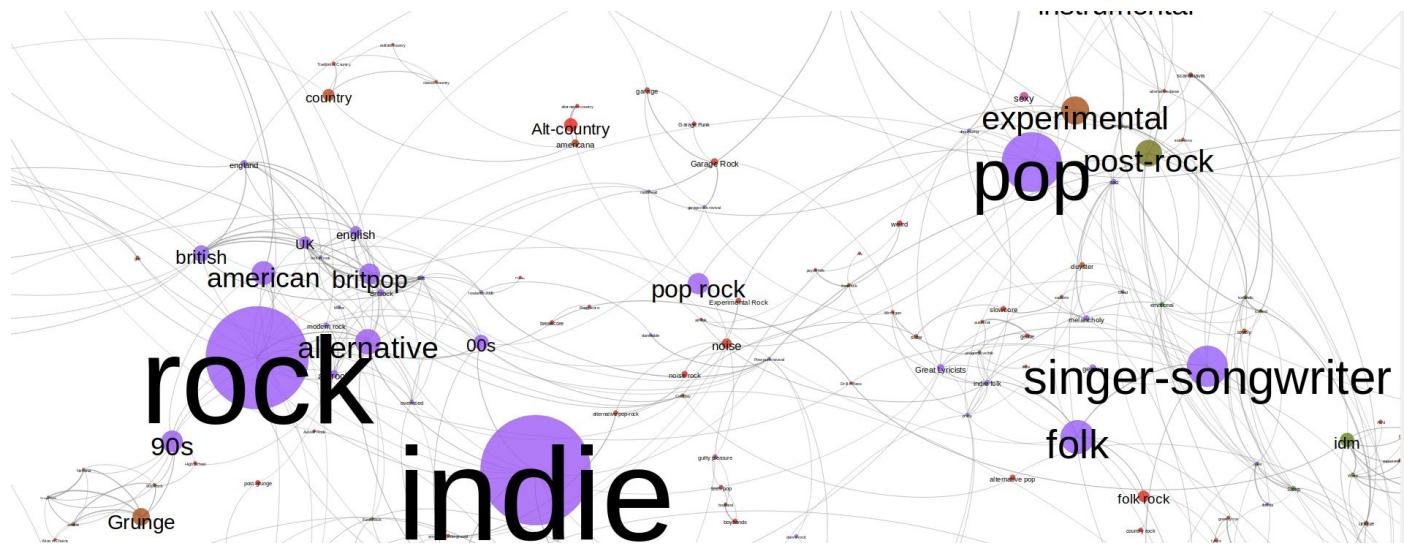
- [Interpretable vs Explainable Machine Learning | by Conor O'Sullivan](#)
- [A Gentle Introduction to GA2Ms, a White Box Model](#)
- <https://github.com/slundberg/shap> : SHAP (SHapley Additive ExPlanations)
- <https://github.com/interpretml/interpret> : InterpretML : train interpretable glassbox models and explain blackbox systems

Articles scientifiques :

- [Explainable AI: A Review of Machine Learning Interpretability Methods \[2.1\]](#)
- [Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead \[2.2\]](#)
- [Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization \[2.3\]](#)
- [This Looks Like That: Deep Learning for Interpretable Image Recognition \[2.4\]](#)
- [Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges](#)

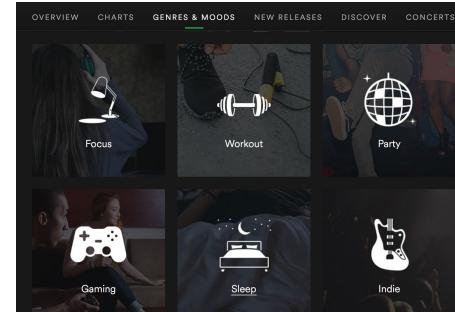
- Législation
- Anonymisation
- Modèles génératifs
- Biais et discrimination
- Transparence
- Impact sociétal
- Impact Environnemental

Impact sociétal, intérêt général : bulles de filtres, influence, ...



Le cas des algorithmes de recommandation musicale

Diversité des données des algorithmes de recommandation musicale



- Analyse du signal : tempo, fréquences, séquençage,...
- Annotations humaines : genre,
- Analyse des paroles, pochettes, livrets

=> peut aller jusqu'à plus de 300 champs descriptifs de métadonnées

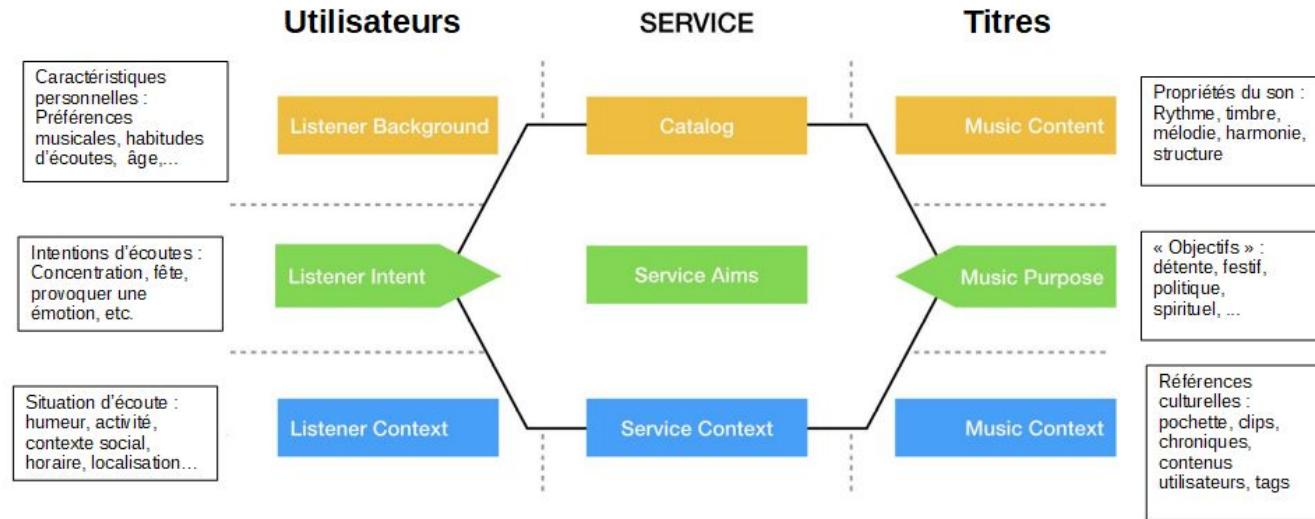
- Profilage des utilisateurs
- habitudes d'écoutes
- Recherche
- top titres
- l'activité déclarée
- l'émotion recherchée,
- la personnalité
- le contexte culturel

...

Catégorisation du contexte : l'heure de la journée, la localisation, la météo, le trafic routier, contexte social, particularité culturelle de certaines régions,...

Les sources : les capteurs du smartphone, les questions directes

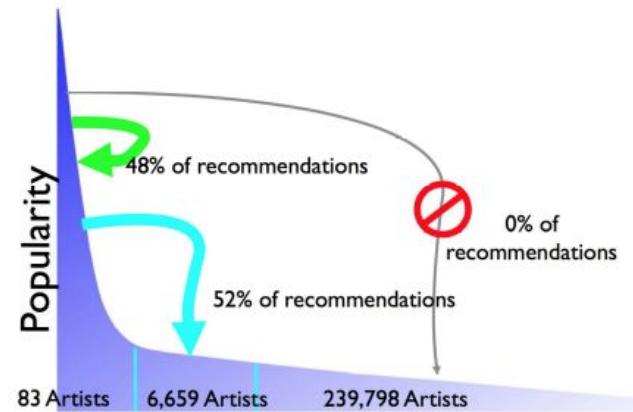
Matrice des “filtres collaboratifs”



Sophistication et recherche importante dans le domaine

Coincés dans la bulle

Mais... peu de diversité



(CF-based recommendations, Last Fm data)

Et on comprend pas toujours ...

Vous aimez
Bernard Herrmann ?



Vous aimerez
« Gimme Some More »
de Busta Rhymes



Mais pourquoi !?

Parce qu'il a samplé Herrmann...

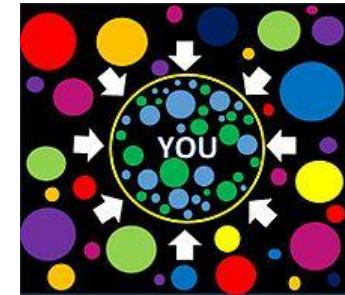


Si l'on transpose aux algorithmes de recommandation de contenus sur les réseaux sociaux, plateformes vidéos ou d'information...

Risques de l'hyper-personnalisation

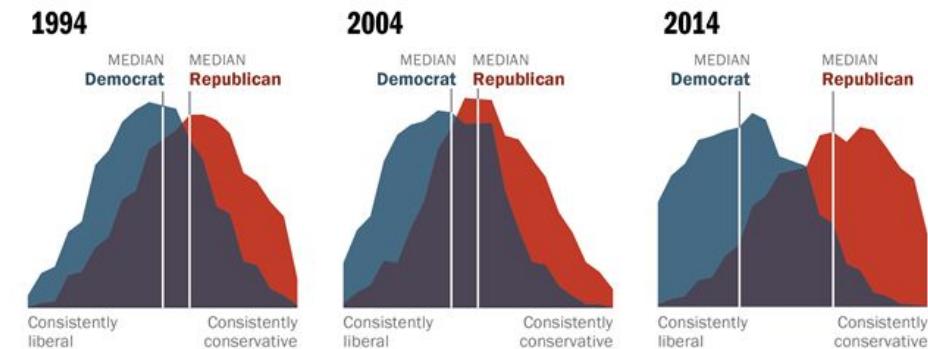
Bulles de filtres dans les newsfeed, vidéos youtube, ...

- Maximisation du nombre de (publicités) vues
- Les bulles de filtres créent des “chambres d’écho”
- Biais cognitifs
 - Biais de confirmation amplifié
 - Biais de simple exposition



Democrats and Republicans More Ideologically Divided than in the Past

Distribution of Democrats and Republicans on a 10-item scale of political values



Source: 2014 Political Polarization in the American Public

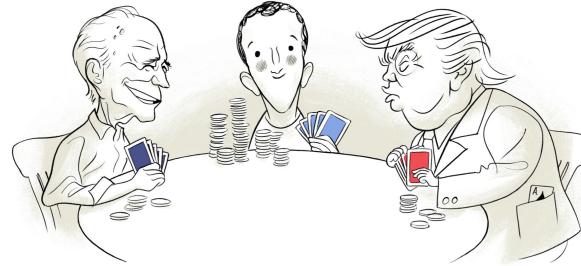
Notes: Ideological consistency based on a scale of 10 political values questions (see Appendix A). The blue area in this chart represents the ideological distribution of Democrats; the red area of Republicans. The overlap of these two distributions is shaded purple. Republicans include Republican-leaning independents; Democrats include Democratic-leaning independents (see Appendix B).

En pratique : faire éclater les bulles de filters

Aller au delà de l'accuracy et du RMSE

- ❑ Nombreuses métriques dans la littérature des systèmes de recommandation
 - Diversity : Dissimilarité des recommandations
 - Coverage : Pourcentage des items recommandés
 - Serendipity : Caractère surprenant des recommandations pertinentes
 - Novelty : Caractère surprenant des recommandations en général
 - Relevancy : Pertinence des recommandations
- ❑ [Recommender Systems — It's Not All About the Accuracy](#)
- ❑ Pas de méthodes ni de librairies unifiées
 - Malgré des concepts qui existent depuis 2001 dans la communauté !
- ❑ En pratique :
 - Cross-valider sur ces métriques
 - Faire sa littérature pour trouver les modèles qui biaisent vers la métrique voulue
 - [Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems](#)
 - [Diversity in Recommender Systems](#)

Impact sociétal, intérêt gé



Sujets connexes

- ❑ Manipulations électorales
 - Élections présidentielles US 2016 : Cambridge-Analytica
 - Publicités politiques bannies sur facebook une semaine avant les élections US 2020 pendant 4 mois
- ❑ Manipulations des émotions : expériences de Facebook 2014
 - Suppression des contenus positifs ou négatifs pour 689,003 utilisateurs
 - Les posts des utilisateurs qui ne voyaient plus de contenu positifs étaient plus négatifs, et vice-versa
- ❑ Fake news
 - Les fake news circulent plus vite que les autres news
- ❑ Réseaux sociaux et santé mentale
 - Amplification du biais de comparaison sociale

Ressources

Bulles de filtres et systèmes de recommandations

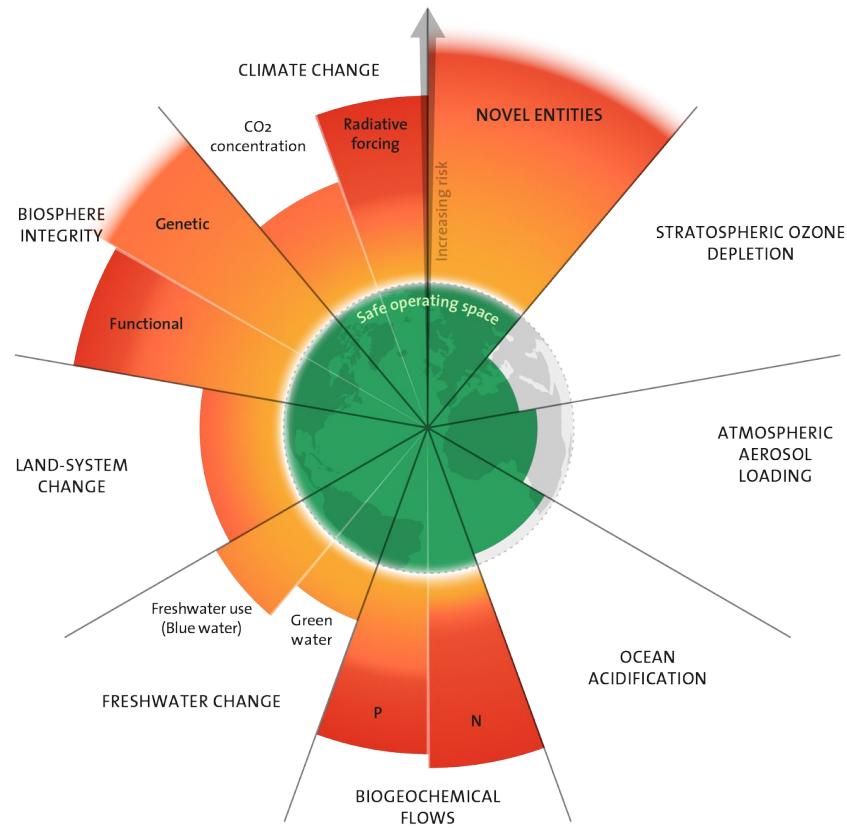
- ❑ [Les algorithmes peuvent-ils manipuler la société ? - Dominique Cardon - 10/05/17](#) : conférence de Dominique Cardon à propos de son ouvrage “[A quoi rêvent les algorithmes](#)”
- ❑ [Overview and new challenges of music recommendation research in 2018](#)
- ❑ [Biais de confirmation - Wikipedia](#)
- ❑ [Effet de simple exposition - Wikipedia](#)
- ❑ [Chambres d'écho \(media\) - Wikipedia](#)
- ❑ [Bulles de filtre et chambres d'écho](#)
- ❑ [Political Polarization in the American Public](#)
- ❑ [Recommender Systems — It's Not All About the Accuracy](#)
- ❑ [Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender System](#)
- ❑ [Diversity in Recommender Systems](#)

Sujets connexes

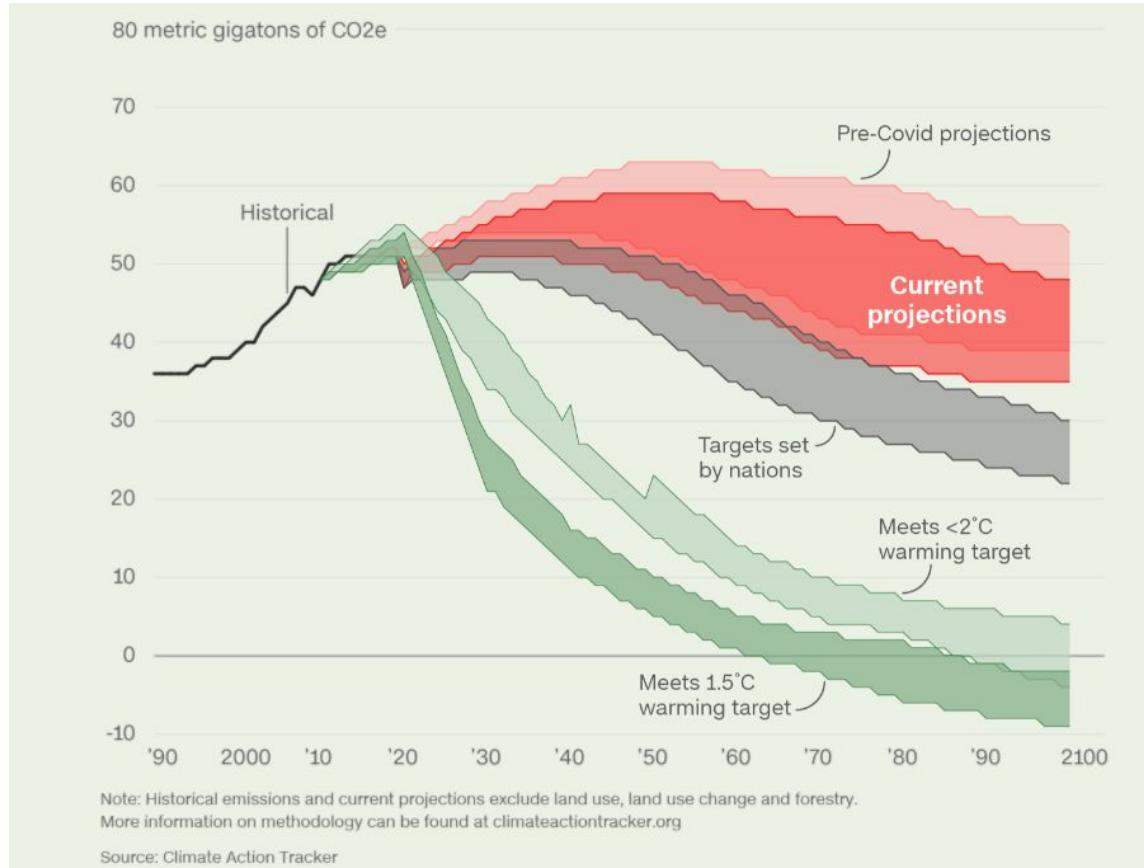
- ❑ [Everything You Need to Know About Facebook's Controversial Emotion Experiment](#)
- ❑ [Facebook–Cambridge Analytica data scandal](#)
- ❑ [Study: On Twitter, false news travels faster than true stories](#)
- ❑ [Social media and the effects on American adolescents](#)
- ❑ [The Social Dilemma | Official Trailer | Netflix](#)

- Législation
- Anonymisation
- Modèles génératifs
- Biais et discrimination
- Transparence
- Impact sociétal
- Impact Environnemental

Impact environnemental : les limites de la planète



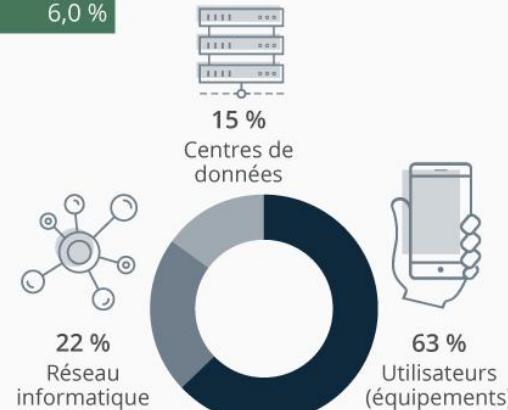
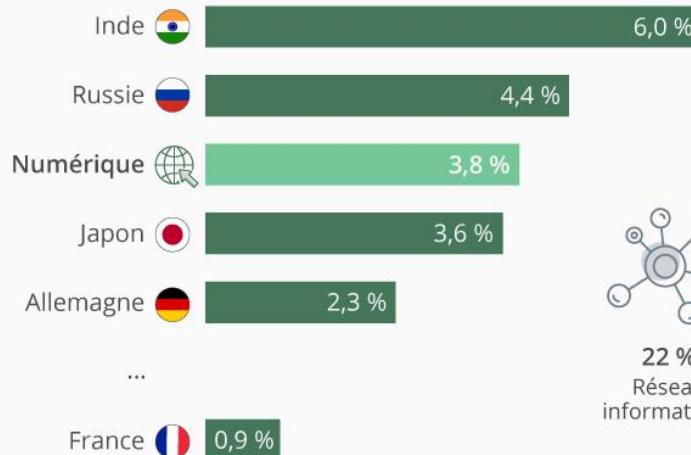
Impact environnemental : le climat



Impact environnemental : Le Numérique

Le numérique émet 4 fois plus de CO2 que la France

Part des émissions mondiales de CO2 du numérique comparée à celle des pays sélectionnés *



Répartition des émissions du numérique selon la source

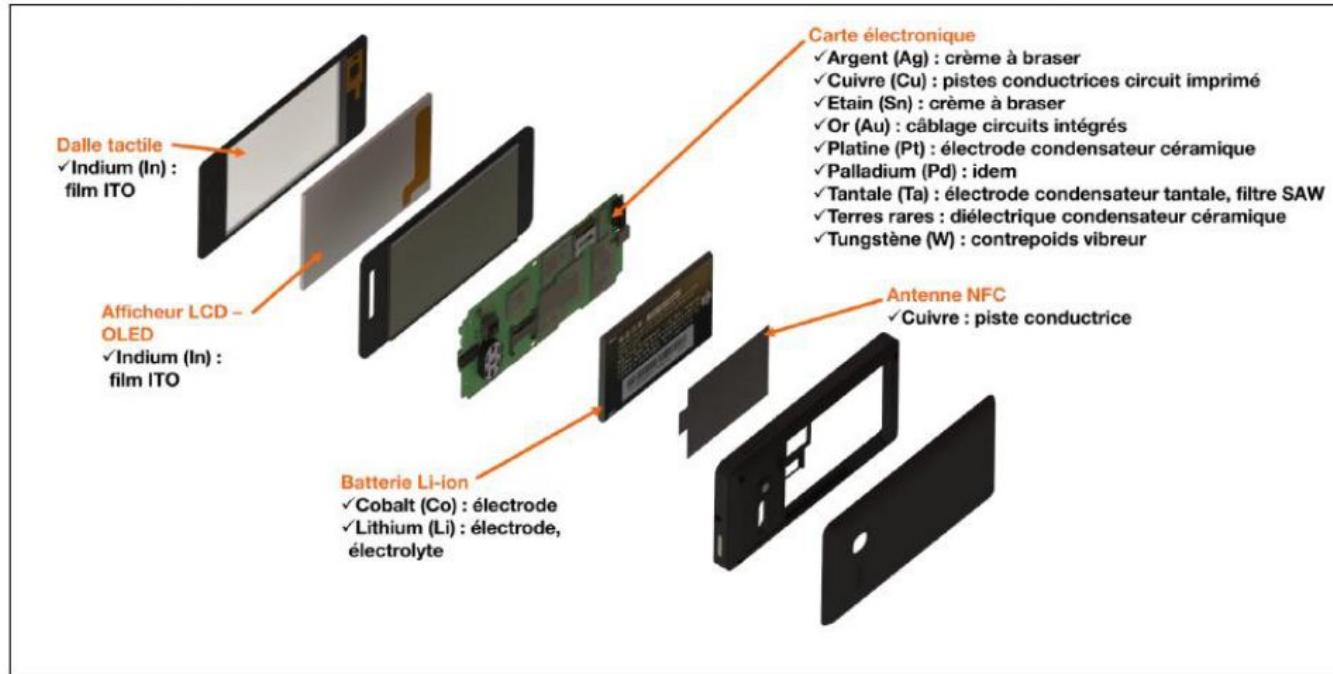


* en 2019 pour le numérique, 2018 pour les pays.

Sources : GreenIT, Agence internationale de l'énergie

Impact environnemental : Le Numérique

Figure 9 – Principaux éléments dans les composants d'un téléphone



- Pollution due à l'extraction minière
- Très peu de recyclage
- Épuisement de certaines ressources
 - Possible pic d'extraction du Cobalt, Cuivre et Nickel d'ici 2050

Impact environnemental

Cas concret : la 5G

“La 5G permet, lorsque l'on transporte des données, de le faire avec moins d'énergie. En revanche, elle augmente considérablement les débits et permet donc un usage beaucoup plus important, donc de transporter davantage de données, ce qui est beaucoup plus consommateur. Il est donc erroné d'affirmer que la 5G permettra des efforts en matière d'énergie.”

Olivier Roussat, Président de Bouygues Télécom, audition au Sénat, Juin 2020

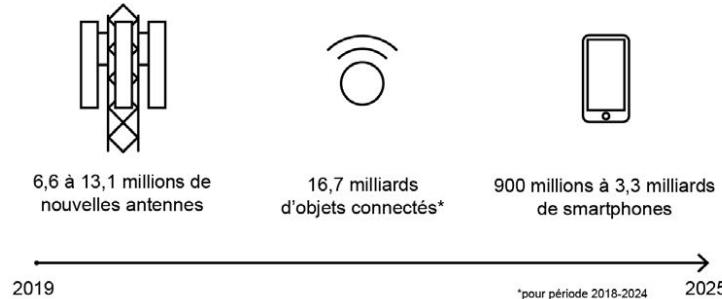
- ❑ 1 antenne 5G consomme 3 fois plus qu'une antenne 4G
- ❑ Débit 10 à 15 fois supérieur
- ❑ Les antennes 5G viendront s'ajouter aux 2G/3G/4G
- ❑ Housse de la consommation du réseau estimée à 35% entre 2020 et 2025

Traffic Load	4G	5G	4/5G Energy Consumption Ratio
100%	1044.72W	3674.85W	$5G/4G=3.5$
50%	995.06W	2969.97W	$5G/4G=3$
30%	949.22W	2579.83W	$5G/4G=2.7$
0%	837.21W	2192.57W	$5G/4G=2.6$

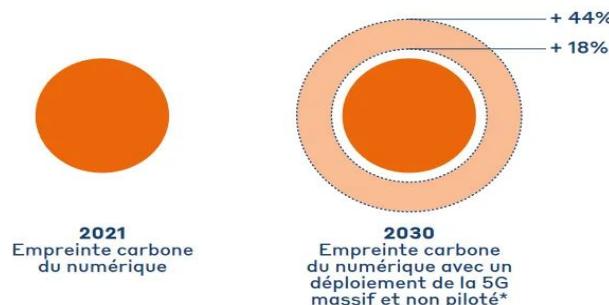
Huawei

Impact environnemental

L'effet rebond de la 5G



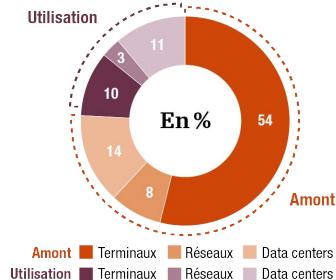
Estimation de la fabrication nette d'équipements liés à la 5G, de 2019 à 2025



* D'après le Haut Conseil pour le Climat, le déploiement de la 5G devrait engendrer une augmentation de 18 à 44 % de l'empreinte carbone du numérique à horizon 2030 (HCC, 2020a) (Citizing - Virtus Management, 2020).

- Durée d'usage moyenne d'un smartphone en France : entre 2 et 3 ans
- Nouveaux usages des voitures autonomes, du cloud gaming, de la réalité virtuelle,...

Répartition de l'impact carbone de la 5G, par postes d'émissions en 2030 dans l'évaluation haute



SOURCE : TRAITEMENTS HCC 2020 D'APRÈS CITIZING

Impact environnemental

Concepts clés

- ❑ Utilité finale
- ❑ Empreinte environnementale
 - Empreinte carbone
 - Extraction de ressources
 - ...
- ❑ Effet rebond
 - Autres exemples :
 - Compression vidéo pour le streaming (résolutions d'image toujours plus grandes)
 - Virtualisation des serveurs (facilité d'utilisation augmente le volume)
- ❑ Pour aller plus loin : Méthodologie d'évaluation de l'empreinte environnementale d'un projet numérique :
 - Déployer la sobriété numérique, Rapport du shift project
 - https://theshiftproject.org/wp-content/uploads/2020/10/Deployer-la-sobriete-numerique_Rapport-complet_ShiftProject.pdf

Impact environnemental : Le Machine Learning

Common carbon footprint benchmarks

in lbs of CO₂ equivalent

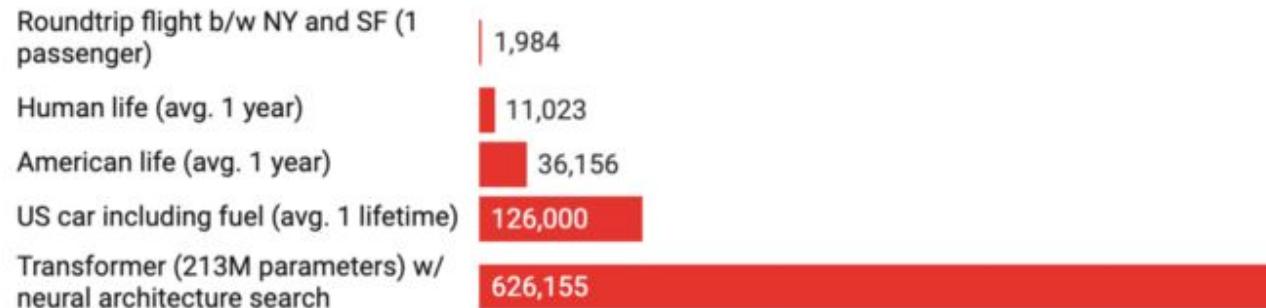
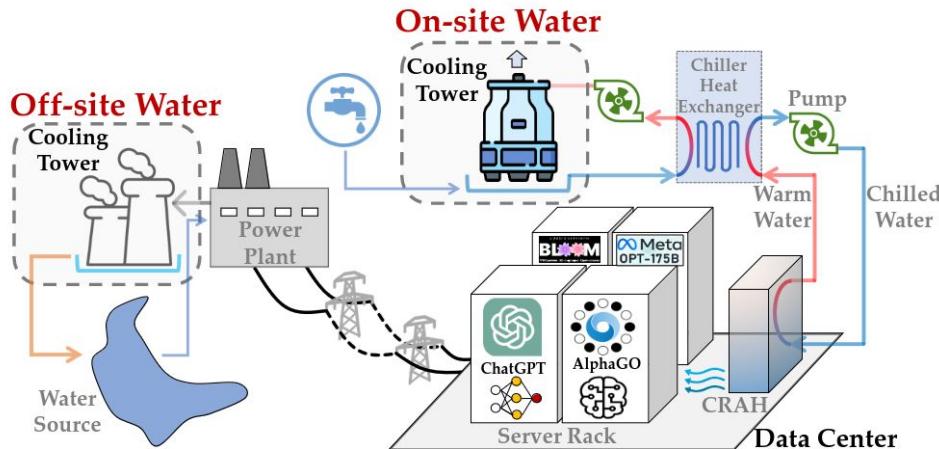


Chart: MIT Technology Review • Source: Strubell et al. • [Created with Datawrapper](#)

- ❑ Proportion du machine learning dans le numérique difficile à estimer
- ❑ Tendance : quantité de calculs pour le deep-learning double tous les 3,4 mois => x300.000 de 2012 à 2019
- ❑ En pratique l'inférence consomme plus que le training (mais reste proportionnelle à la taille du réseau)

ChatGPT : CO₂ et eau



L'estimation de l'empreinte carbone journalière de ChatGPT est de 43.2 kgCO₂e. Ce qui équivaut à 15.8 tCO₂e/an.

<https://piktochart.com/blog/carbon-footprint-of-chatgpt/>

500mL consommé par conversation de 20-50 questions, 50 millions de visite par jour.

(Making AI Less “Thirsty”: Uncovering and Addressing the Secret Water Footprint of AI Models (Li et al. 2023))

Impact environnemental : Le Machine Learning

Préconisations

- ❑ Utiliser des modèles moins coûteux à performance quasi-égale
- ❑ Éviter de relancer le même training de nombreuses fois
 - Sauvegarder checkpoints du modèle
 - Tester son implémentation sur des petits jeux de données
- ❑ Rapporter le coût en CO₂ du training de vos modèles (dans vos rapports, publications)
 - Outils comme carbontracker : <https://github.com/lfwa/carbontracker>
- ❑ Hyperparameter tuningAutomated ML with Bayesian Optimization
 - Successive Halving Grid Search
 - Randomized hyper-parameter search
 - Tirage aléatoire avec une distribution sur les hyperparamètres
 - { 'C' : loguniform(1,100) } instead of { 'C' : [1, 10, 100] }
 - Atteint performance similaire à grid-search en moins d'essais
 - Automated ML with Bayesian Optimization
 - Similaire, mais utilise les score de validation des tirages précédents pour guider les tirages suivants
https://scikit-optimize.github.io/stable/auto_examples/sklearn-gridsearchcv-replacement.html
- ❑ [Transfer learning]
 - Inférence coûteuse pour un gros modèle
- ❑ [Louer un serveur GPU dans un pays à mix bas carbone (Norvège, France, ...)]
 - En heures creuses !

Resources

Numérique et environnement

- ❑ <https://nourrituresterrestres.substack.com/p/6-dataviz-pour-mieux-saisir-lurgence>
- ❑ https://fr.wikipedia.org/wiki/Impact_environnemental_du_num%C3%A9rique
- ❑ [https://fr.wikipedia.org/wiki/Effet_rebond_\(%C3%A9conomie\)](https://fr.wikipedia.org/wiki/Effet_rebond_(%C3%A9conomie))
- ❑ <https://www.greenit.fr/2014/02/19/l-effet-rebond-dans-le-numerique-est-il-evitable/>
- ❑ https://theshiftproject.org/wp-content/uploads/2020/10/Deployer-la-sobriete-numerique_Rapport-complet_ShiftProject.pdf
- ❑ <https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/fs-2020-dt-consommation-metaux-du-numerique-juin.pdf>
- ❑ La controverse de la 5G gauthierroussilhe.com

Machine learning et environnement

- ❑ <https://towardsdatascience.com/deep-learning-and-carbon-emissions-79723d5bc86e>
- ❑ Energy and Policy Considerations for Deep Learning in NLP, Strubell et al., 2020, <https://arxiv.org/pdf/1906.02243.pdf>
- ❑ Carbontracker : Tracking and Predicting the Carbon Footprint of Training Deep Learning Models, Anthony et al., 2020, <https://arxiv.org/pdf/2007.03051.pdf>
 - <https://github.com/lfwa/carbontracker>
- ❑ https://scikit-learn.org/stable/modules/grid_search.html#randomized-parameter-search
- ❑ <https://towardsdatascience.com/automated-machine-learning-hyperparameter-tuning-in-python-dfda59b72f8a>



Fafiec =

CAMPUS
RÉGION®
DU NUMÉRIQUE



le
campus
numérique
in the ALPS

Le campus numérique in the Alps

www.le-campus-numerique.fr
contact@le-campus-numerique.fr