RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Dares
Déchiffrer le monde du travail
pour **éclairer** le débat public

Drees

Colloque

TRAVAIL DE PLATEFORME ET USAGES DE LA PROTECTION SOCIALE

MARDI 4 OCTOBRE 2022

Direction de l'animation de la recherche, des études et des statistiques
Direction de la recherche, des études, de l'évaluation et des statistiques

RISQUES DISCRIMINATOIRES DES ALGORITHMES IMPLIQUÉS DANS DES SYSTÈMES D'IA

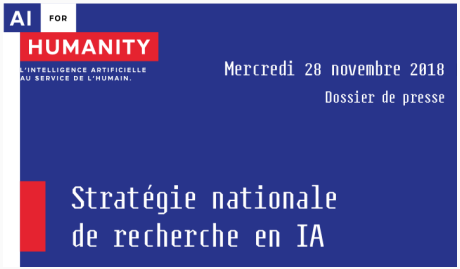
Philippe Besse

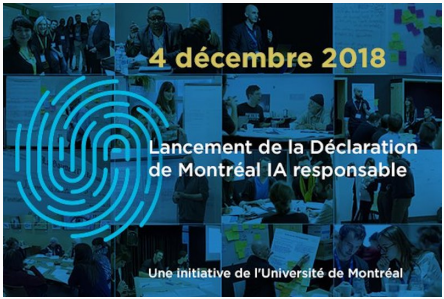
Université de Toulouse – INSA, IMT UMR CNRS 5219, ObvIA Université Laval

Introduction

De l'IA éthique (*soft law*) à l'obligation de conformité

Amazon, Apple, Facebook, Google, IBM, Microsoft... (2015)



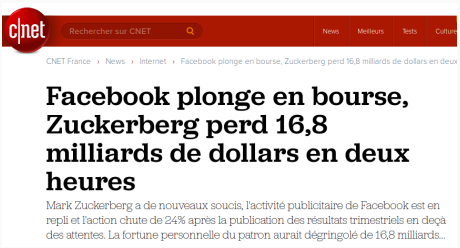


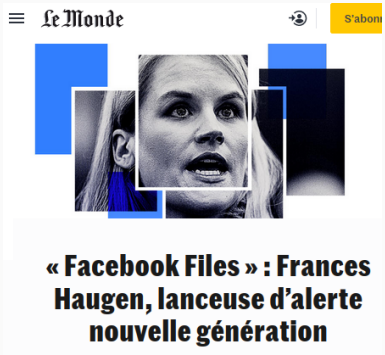
Éthique, Confiance, Acceptabilité

- Enjeux sociétaux & financiers considérables
- Acceptabilité des nouvelles technologies
- Pas de confiance ⇒ pas de données ⇒ pas d'IA
- Entreprises philanthropiques et altruistes ?

Faire confiance à la Loi plutôt qu'à l'Éthique

- RGPD inadapté, *éthical washing*
- Textes de loi *vs.* disruptions technologiques
- Livre blanc (CE 2020) IA: Une approche européenne axée sur l'excellence et la confiance
- Projets de réglementation européenne





Introduction

IA conforme

Projets de réglementation

1. *Digital Market Act* (2020) : risques d'entraves à la concurrence à l'encontre des entreprises européennes
2. *Digital Services Act* (2020) : hébergement, de plateforme en ligne et autres réseaux sociaux
3. *Data Governance Act* (2020) utilisations, réutilisations, des bases de données publiques que privées (fiducie des données) ;
4. *Artificial Intelligence Act* (2021) : proposition de règlement (85 articles) établissant des règles harmonisées sur l'intelligence artificielle.

Définitions

L'IA au sens de l'AI Act

Définition de l'IA (*AI Act* CE, 2021)

Systèmes d'IA (SIA) définis (Art. 3) par une liste d'algorithmes (annexe I)

- **(a)** Apprentissage supervisés, non-supervisé, par renforcement
- **(b)** Représentation de connaissances, systèmes experts, programmation inductive
- **(c)** Approches statistiques, bayésiennes, méthodes d'optimisation
- **(∅)** Algorithmes procéduraux

Définitions

Types d'Algorithmes

Algorithme

Ensemble de règles opératoires dont l'application permet de résoudre un problème énoncé au moyen d'un nombre fini d'opérations. Un algorithme peut être traduit, grâce à un langage de programmation, en un programme exécutable par un ordinateur. (Larousse)

- Procédural (calcul de prestations, impôts)
- Allocation de ressources par appariement (Parcoursup, plateformes)
- Apprentissage (non supervisé, supervisé, par renforcement)
  - Risque de défaut de paiement (score de crédit), comportement à risque (assurance)
  - Risque de rupture de contrat (marketing), récidive (justice), passage à l'acte (police)
  - Profilage automatique publicitaire, **professionnel (CV, vidéos)**
  - Risque de fraude (assurance, banque, fisc), défaillance système industriel
  - **Santé** : *e.g. Diagnostic en imagerie médicale (deep learning)*
- Calcul automatique de prix (*pricing* des courses)

Définitions

Systèmes d'IA à Haut Risque (AI Act)

Risque fonction du domaine d'application
AI Act (CE, 2021)

Applications prohibées (Art. 5) : manipulations, atteintes personnes vulnérables, score social, identification biométrique en temps réel...

Systèmes d'IA à haut risque (Art. 6) impactant des personnes physiques

- Annexe II : Véhicules, ascenseurs, dispositifs de santé
- Annexe III : Trafic, ressources, éducation, emploi, justice, police, crédit, droit d'asile...
  - 4. Emploi, gestion de la main-d'œuvre et accès à l'emploi indépendant :
    - (a) les systèmes d'IA destinés à être utilisés pour le recrutement ou la sélection...
    - (b) l'IA destinée à être utilisée pour la prise de décisions de promotion et de licenciement dans le cadre de relations professionnelles contractuelles, pour l'attribution des tâches et pour le suivi et l'évaluation des performances et du comportement de personnes dans le cadre de telles relations.

Définitions

Risques des impacts sociétaux

Risques des impacts sociétaux de l'IA (législation actuelle)
Principaux risques (interdépendants) : Besse et al. (2019), CE (2020)

1. Confidentialité des données : protection (RGPD, CNIL)
2. Erreur : qualité, robustesse, résilience, des prévisions (Besse 2021) : (néant)
3. Opacité : Explicabilité des algorithmes (RGPD flou, LIL3 inadaptée)
4. **Discrimination** ou biais des décisions algorithmiques (Loi stricte inapplicable)
5. Entraves à la concurrence : comparateurs, *pricing* automatique
6. Impacts environnementaux

Risques de discrimination algorithmique

Exemples

Exemples de discrimination algorithmique (apprentissage supervisé)

Les Pays-Bas contraints de stopper un «système de surveillance pour les pauvres»

Une cour néerlandaise contraint l'Etat à mettre fin à SyRI, un système chargé de débusquer les fraudeurs à l'aide sociale

Le Temps 5/02/2020

ACLU About Issues Our work News Take action Shop Donate

NEWS & COMMENTARY

How is Face Recognition Surveillance Technology Racist?

If police are authorized to deploy invasive face surveillance technologies against our communities, these technologies will unquestionably be used to target Black and

nature

Explore content About the journal Publish with us Subscribe

nature > news > article

NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

PROPUBLICA Journalism in the Public Interest

Receive our top stories daily

Email address

SUBSCRIBE

Home Investigations Data MuckReads Get Involved About Us

Search ProPublica

Machine Bias

Feature Stories

Read Our Investigation

Machine Bias

By Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica, May 23, 2016

There's software used across the country to predict future criminals. And it's biased against blacks. Read more.

Exemples de discrimination algorithmique (appariement)

Italian Supervisory Authority Fines Foodinho Over Its Use of Performance Management Algorithms

By Helena Milner-Smith, Giulia Romana Mele & Dan Cooper on July 13, 2021

POSTED IN GDPR

TIME SPOTLIGHT STORY FAMILIES FIGHT TO PRESERVE BEIRUT PORT SILOS SIGN IN SUBSCRIBE

TECH • ARTIFICIAL INTELLIGENCE

Uber Drivers Say a 'Racist' Algorithm Is Putting Them Out of Work

CORNELL RESEARCHERS FIND RATINGS SYSTEMS MAY DISCRIMINATE AGAINST UBER DRIVERS

Many new "sharing economy" companies, like Uber and Airbnb, use consumer-sourced ratings to evaluate their workers – but these systems can be fraught with difficulties, including bias

Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms

Akshat Pandey
George Washington University

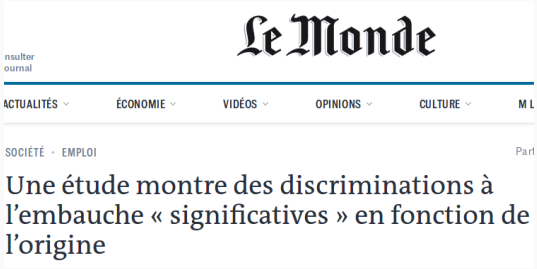
Aylin Caliskan
George Washington University

ArXiv 2021

Risques de discrimination algorithmique

Détecter une discrimination

Détection d'une discrimination humaine
Exemple : discrimination à l'embauche


nsulter
ournal

Le Monde

ACTUALITÉS ÉCONOMIE VIDÉOS OPINIONS CULTURE M L

SOCIÉTÉ · EMPLOI Part

Une étude montre des discriminations à l'embauche « significatives » en fonction de l'origine

- France – Testing : Comité National de l'Information Statistique, DARES, Économie, Sociologie (Riach et Rich, 2002)
  - USA – *Disparate Impact* : *four fith rule* (Barocas et Selbst, 2016)
    Civil Rights act & Code of Federal Regulations : Title 29 - Labor: Part 1607–Uniform Guidelines on Employee Selection Procedures (1978)

$$DI = \frac{\mathbb{P}(\widehat{Y} = 1|S = 0)}{\mathbb{P}(\widehat{Y} = 1|S = 1)}$$

  - CNIL & DDD (2012) Mesurer pour progresser vers l'égalité des chances

Détection d'une discrimination algorithmique

- Discrimination indirecte : critères statistiques
- Biais statistiques et indicateurs de discrimination : Zliobaité (2017), 70 sur `aif360.mybluemix.net`
- Critères, redondants, corrélés : Friedler et al. (2019), Verma et Rubin (2018)
- En priorité Trois niveaux de biais
  1. Effet disproportionné ou biais systémique : $DI = \frac{\mathbb{P}(\hat{Y}=1|S=0)}{\mathbb{P}(\hat{Y}=1|S=1)}$
  2. Taux d'erreur conditionnels : $\frac{\mathbb{P}(\widehat{Y}\neq Y|S=0)}{\mathbb{P}(\widehat{Y}\neq Y|S=1)}$
     Reconnaissance faciale, santé (Besse et al. 2019), emploi (De Arteaga et al. 2019)
  3. Égalité des cotes (*equali odds*) : $\frac{\mathbb{P}(\hat{Y}=1|Y=0,S=0)}{\mathbb{P}(\hat{Y}=1|Y=0,S=1)}$ et $\frac{\mathbb{P}(\widehat{Y}=0|Y=1,S=0)}{\mathbb{P}(\widehat{Y}=0|Y=1,S=1)}$
     Justice "prédictive" : Propublica & COMPAS, Score de crédit (Besse 2022)

Apprentissage supervisé et discrimination (Besse 2022)

- Reproduction du biais systémique présent dans les données
- Possibilité d'aggravation de ce biais (algorithmes naïfs)
- Supprimer l'information "sensible" ne change rien
- Conséquence : testing aveugle face à un algorithme
- Important : étudier les trois niveaux de biais
- Corrections possibles (*fair learning*) mais sans cadre juridique

Protection et certification européennes

Certification et marquage "CE"

Certification des Systèmes d'IA à Haut Risque

Conformité(*AI Act* CE, 2021) de non discrimination

- Documentation (Art. 11) exhaustive (annexe IV)
- Importance des données (Art. 10), biais, représentativité, données sensibles
- Précision, robustesse, résilience, interprétation des SIA (Art. 13, 15)
- Risque de discrimination : biais systémique des données (Art. 10), erreurs conditionnelles (Art. 13)
- Archivage du journal (Art. 12) tout au long du cycle de vie (*post market monitoring* art. 61)
- SIA à haut risque
  - Annexe II : Audit *ex-ante* par organisme de notification (GMED) désigné par autorité notifiante (ANSM)
    LNE (2021 – Référentiel de certification du Process IA), construction de normes
  - Annexe III : déclaratif, quel audit *ex-post* ?
- *NB* : rien sur atténuation ou correction des biais discriminatoires

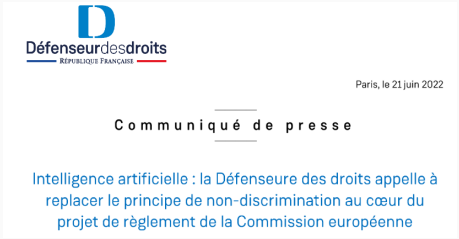
Conclusion

Limites de l'*AI Act*

AI Act et détection / preuve d'une discrimination

- Protection de l'utilisateur, pas de l'usager
  Harmonisation du marché de l'IA, sécurité des produits ou responsabilité du fait de produits défectueux
- Certification déclarative des SIA Annexe III
- Prouver une présomption de discrimination ?

- Qui accède à la documentation ?
- Qui accède au journal archivé ?
- Avec quelles compétences ?


Défenseurdesdroits
RÉPUBLIQUE FRANÇAISE

Paris, le 21 juin 2022

Communiqué de presse

Intelligence artificielle : la Défenseure des droits appelle à replacer le principe de non-discrimination au cœur du projet de règlement de la Commission européenne

30/08/2022

L'étude préconise une transformation profonde de la CNIL en autorité de contrôle nationale responsable de la régulation des systèmes d'IA, notamment publics, pour incarner [...] la protection des droits et libertés fondamentaux [...].

Références

- Barocas S. , Selbst A. (2016). Big Data's Disparate Impact, *California Law Review* (104), 671.
- Besse P. (2021). Médecine, police, justice: l'intelligence artificielle a de réelles limites, *The Conversation*, en ligne.
- Besse P. (2022). Statistique & Règlement Européen des Systèmes d'IA (AI Act), *Statistique et Société*, HAL-03253111, à paraître.
- Besse P., Besse-Patin A., Castets-Renard C. (2020). Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé, *Statistique & Société*, 3, pp 21-53.
- Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2019). L'IA du Quotidien peut elle être Éthique? Loyauté des Algorithmes d'Apprentissage Automatique, *Statistique et Société*, Vol 6 (3), pp 9-31.
- Besse P. del Barrio E. Gordaliza P. Loubes J.-M., Risser L. (2021). A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set, *The American Statistician*, DOI : 10.1080/00031305.2021.1952897.
- Défenseur des Droits, CNIL (2012). Mesurer pour progresser vers l'égalité des chances. Guide méthodologique à l'usage des acteurs de l'emploi.
- Commission Européenne (2016). Règlement Général sur la Protection des Données.
- Commission Européenne (2018). Lignes directrices pour une IA de confiance.
- Commission Européenne (2020). Livre blanc sur l'intelligence artificielle: une approche européenne d'excellence et de confiance.
- Commission Européenne (2021). Règles harmonisées concernant l'Intelligence Artificielle.
- Commission Européenne (2021). Règles harmonisées concernant l'Intelligence Artificielle, Annexes.
- Conseil d'État (2022). S'engager dans l'intelligence artificielle pour un meilleur service public, rapport d'étude mis en ligne le 30/08/2022.

- Défenseur des Droits (2019). Décision du Défenseur des droits n° 2019-021.
- Friedler S., Scheidegger C., Venkatasubramanian S., Choudhary S., Ha-milton E., Roth D. (2019). Comparative study of fairness-enhancing interventions in machine learning. in FAT'19, p. 329–38.
- LNE (2021). Référentiel de Certification d'un Processus d'IA, version 02 – juillet 2021.
- Lee P., Le Saux M., Siegel R., Goyal M., Chen C., Ma Y., Meltzer A. (2019). Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis and systematic review, *American Journal of Emergency Medecine*, 37(9), 1770-1777.
- Marty F. (2019). Plateformes Numériques, Algorithmes et Discrimination, *Revue de l'OFCE*, 2019/4 164, 47-86.
- Obermayer Z., Mullainathan S. (2019). Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People, *FAT 19, Proceedings of the Conference on Fairness, Accountability, and Transparency*.
- Riach P.A., Rich J. (2002). Field Experiments of Discrimination in the Market Place, *The Economic Journal*, Vol. 112 (483), p F480-F518.
- Verma S., Rubin J. (2018). Fairness Definitions Explained, ACM/IEEE International Workshop on Software Fairness.
- Villani C., Schoenauer M., Bonnet Y., Berthet C., Cornut A.-C., Levin F., Rondepierre B.(2018). Donner un sens à l'Intelligence Artificielle pour une stratégie nationale et européenne, *La Documentation Française*, rapport public.
- Zliobaité I. (2017). Measuring discrimination in algorithmic decision making, *Data Min Knowl Disc* 31, 1060–1089.

Annexe

Apprentissage supervisé

Principe de l'apprentissage supervisé

p Variables ou caractéristiques $\{X^j\}_{j=1,\dots,p}$ observées sur $i = 1, \dots, n$ individus

Y : Variable cible à modéliser ou prédire et observée sur le même échantillon

$$Y = \boldsymbol{f}\left(X^1 \quad X^2 \quad \cdots \quad X^j \quad \cdots \quad X^p \right)$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \widehat{\boldsymbol{f}}\left(\begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^j & \cdots & x_1^p \\ & & & \vdots & & \\ x_i^1 & x_i^2 & \cdots & x_i^j & \cdots & x_i^p \\ & & & \vdots & & \\ x_n^1 & x_n^2 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix} \right) + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\widehat{y_0} = \widehat{\boldsymbol{f}}\left(x_0^1 \quad x_0^2 \quad \cdots \quad x_0^j \quad \cdots \quad x_0^p \right)$$

$\widehat{y_0}$: prévision de Y après observation de $[x_0^1, x_0^2, \cdots, x_0^p]$