# Statistique & Risques Juridiques (à venir) des Systèmes d'IA

Philippe Besse

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

# Introduction

## IA empirique

## Intelligence Artificielle au quotidien : IA empirique

- Pas de Science Fiction : transhumanisme, singularité technologique, lois d'Asimov
- Pas de Sociologie : destruction des emplois qualifiés, *big data big brother*
- Décisions algorithmiques ou aides automatiques à la décision (IA faible)
- Apprentissage statistique (*statistical learning*) entraînés sur des bases de données $\subset$ apprentissage automatique (*machine learning*) $\subset$ IA
  - 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, carrière)
  - Risque de fraude (assurance, banque), défaillance d'un système industriel
  - Diagnostic en imagerie médicale (*deep learning*)
  - Autres applications en Santé
  - ... 95% des applications de l'IA (Yan Le Cun)

## Principe de l'apprentissage statistique

$p$ 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 $[x_0^1, x_0^2, \cdots, x_0^p]$

# Introduction

## Éthique, Confiance, Acceptabilité

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




**Risques des impacts sociétaux des décisions algorithmiques**
(Besse et al. 2019)

**Quatre + deux questions Juridiques et / ou Éthiques**

1. Protection : propriété, confidentialité des données personnelles (RGPD, CNIL)
2. Qualité : performances, robustesse, résilience des prévisions (rien)
3. Explicabilité *vs.* opacité des algorithmes (flou, inadapté)
4. Discrimination des décisions algorithmiques (loi stricte mais inapplicable)

Risques interdépendants

- Entraves à la concurrence
- Impacts environnementaux

**De l'éthique (*soft law*) à la conformité réglementaire** (Besse, 2021)

# Réglementation à venir

## Rappel du RGPD


## Règlement Général sur la Protection des Données

- Considérant 71 : Afin d'assurer un **traitement équitable et transparent** à l'égard de la personne concernée [...], le **responsable du traitement** **devrait** utiliser des **procédures mathématiques ou statistiques** adéquates aux fins du profilage, appliquer les mesures techniques et organisationnelles appropriées pour faire en sorte, en particulier, que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et **que le risque d'erreur soit réduit au minimum**, et sécuriser les données à caractère personnel d'une manière qui tienne compte des risques susceptibles de peser sur les intérêts et les droits de la personne concernée et **qui prévienne, entre autres, les effets discriminatoires** à l'égard des personnes physiques fondées sur la l'origine raciale ou ethnique, les opinions politiques, la religion ou les convictions, l'appartenance syndicale, le statut génétique ou l'état de santé, ou l'orientation sexuelle, ou qui se traduisent par des mesures produisant un tel effet. La prise de décision et le profilage automatisés fondés sur des catégories particulières de données à caractère personnel ne devraient être autorisés que dans des conditions spécifiques

# Réglementation à venir

## Annonces européennes


## Lignes directrices en matière d'éthique pour une IA de confiance

Groupe d'experts indépendants de hauts niveaux sur l'Intelligence artificielle (2018–2020)

- (52) Si les **biais injustes** peuvent être évités, les systèmes d'IA pourraient même **améliorer le caractère équitable de la société**.
- (53) L'**explicabilité** est essentielle... les décisions – dans la mesure du possible – doivent pouvoir être expliquées.
- (69) Il est important que le système puisse indiquer le **niveau de probabilité de ces erreurs**.
- (80) **Absence de biais injustes**
  La persistance de ces biais pourrait être **source de discrimination et de préjudice (in)directs** Dans la mesure du possible, les **biais détectables et discriminatoires devraient être supprimés** lors de la phase de collecte.
- (106) (107) besoin de **normalisation**


## IA – Une approche européenne axée sur l'excellence et la confiance

*Livre blanc* — 19/02/2020

- IA, qui combine données, algorithmes et puissance de calcul
- Risques potentiels, tels que l'opacité de la prise de décisions, la discrimination
- Enjeu majeur : acceptabilité et adoption de l'IA nécessite une IA digne de confiance
- Fondée sur les droits fondamentaux de la dignité humaine et la protection de la vie privée
- Proposer les éléments clefs d'un futur cadre réglementaire
- Déceler et prouver d'éventuelles infractions à la législation
- Notamment aux dispositions juridiques qui protègent les droits fondamentaux, à cause de l'opacité des algorithmes


**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 établissant des règles harmonisées sur l'intelligence artificielle.


**Projets de réglementation de l'IA (*AI Act*)**

- Texte de 108 pages complété par 17 pages de 9 annexes
- 89 considérants, 85 articles structurés en 12 titres
- Objectifs
  - Commercialisation de systèmes d'IA sûrs, légaux et respectueux des droits fondamentaux
  - Développement d'un marché unique pour les applications d'IA licites, sûres et dignes de confiance
  - Meilleur équilibre entre bénéfices attendus et risques encourus
  - Logique de sécurité des produits basée sur législation relative au marché intérieur


## *AI Act :* **Considérants**

- (13) normes communes à tous les systèmes d'IA ... cohérentes avec la charte des droits fondamentaux de l'Union européenne ... non discriminatoires et conformes aux engagements commerciaux
- (44) haute qualité des données est essentielle ... afin de garantir que le système d'IA à haut risque fonctionne comme prévu et en toute sécurité et qu'il ne devienne pas une source de discrimination ... Des ensembles de données d'apprentissage, validation, test ... pertinents, représentatifs, exempts d'erreurs et complets ... protéger le droit d'autrui contre la discrimination ... traiter également des catégories spéciales de données à caractère personnel
- (47) remédier à l'opacité ... utilisateurs doivent être capables d'interpréter la sortie du système ... systèmes d'IA à haut risque devraient donc être accompagnés d'une documentation et d'instructions d'utilisation pertinentes et inclure des informations ... y compris en ce qui concerne les risques potentiels pour les droits fondamentaux et la discrimination
- (49) Systèmes d'IA ... fonctionner de manière cohérente tout au long de leur cycle de vie et répondre à un niveau approprié de précision, robustesse. Le niveau d'exactitude et les mesures d'exactitude doivent être communiqués aux utilisateurs


## *AI Act :* **Considérants – résumé**

- Demande de normes internationales indispensables
- Priorité au respect des droits fondamentaux dont la non-discrimination
- Représentativité statistique des ensembles de données
- Nécessité de documentations exhaustives notamment sur les performances
- Possibilités d'interprétation des sorties ou décisions en découlant
- Obligation de journalisation ou archivage des décisions et données afférentes


## *AI Act :* **Articles 1 – 11**

- **Article 3** Définitions : système d'IA de l'Annexe I : apprentissage, renforcement, systèmes experts, procéduraux, données d'apprentissage, validation, test...
- **Article 5** Applications prohibées : manipulations, atteintes aux personnes vulnérables, score social, identification biométrie en temps réel...
- **Article 6** Systèmes d'IA à haut risques
  - Annexe II : Véhicules, ascenseurs, dispositifs de santé
  - Annexe III : Trafic, ressources, éducation, emploi, justice, police, crédit, droit d'asile...
- **Article 9** système de gestion du risque toute la durée de vie d'un système d'IA, identifier, élimination atténuation des risques. Un système d'IA doit être testé afin d'identifier les meilleures mesures de risque.
- **Article 10** gouvernance des données, évaluation *a priori* documenté, (f) analyse des biais, représentativité, possibilité analyse données sensibles pour détection, correction biais sous réserve de confidentialité
- **Article 11** rédaction d'une **documentation** (annexe IV)


## *AI Act :* **Articles 12 – 85**

- **Article 12** archivage du journal pour la traçabilité tout au long du cycle de vie et article 61 (*post market monitoring*)
- **Article 13** Transparence et information des utilisateurs pour interpréter les résultats, instructions d'utilisations, niveau de précision, robustesse, cybersécurité, conditions d'utilisations abusives pouvant entraîner des risques (droits fondamentaux), performances concernant les groupes
- **Article 14** Surveillance par des personnes physiques, interpréter correctement les résultats ... tenant compte des outils et méthodes d'interprétation
- **Article 15** Précision, robustesse, cybersécurité, déclaration des mesures et niveaux de précision, résilience en ce qui concerne les erreurs, défauts ou incohérences, redondance technique, protections spécifiques des systèmes d'IA qui continuent à apprendre
- **Articles suivants** obligations, système de gestion de la qualité (données), **marquage "CE"** : certification par autorité notifiante & organisme de notification (Annexe II) ou déclaratif (Annexe III), base de données des systèmes d'IA à haut risques, sanctions.


## *AI Act :* **Articles – commentaires**

- Exigences essentielles du livre blanc : non-discrimination, environnement
- Définition pragmatique de l'IA (3, Annexe I)
- Systèmes d'IA à haut risque (6), documentation (11, Annexe IV) et marquage "CE"
  - Annexe II : **audit** *ex-ante* par organisme de notification avec renversement de la charge de la preuve
  - Annexe III : **déclaratif**
- Protection et information (13) de Utilisateur *vs.* Usager
  - Utilisateur : principe de sécurité des produits ou responsabilité du fait des produits défectueux
  - Usager : article 22 du RGPD soumis l'utilisateur (compétences, déontologie)
- Données (10) : rôle reconnu, données sensibles, analyse statistique exhaustive (représentativité, biais)
- Précision, robustesse, interprétation (13, 15)
- Non-discrimination : biais systémique des données (10), erreurs conditionnelles (13), normes ?
- Archivage du journal (12) : et confidentialité des données sensibles

# Anticiper l'*AI Act*

## Les données

## Documentation exhaustive de la gestion des données

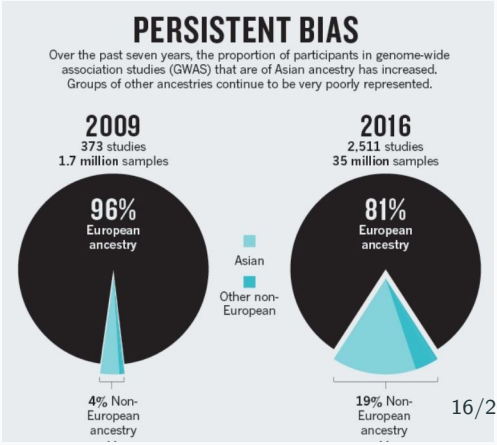
- Définition du domaine d'application, bénéfices attendus
- Confidentialité : pseudonymisation, anonymisation, simulations
- Représentativité des ensembles de données (apprentissage, validation, test)
- Préparation : nettoyage, enrichissement (*features*)
- Données manquantes : résilience
- Données atypiques : anomalies, robustesse
- **Biais systémiques** : données sensibles
- Journalisation : boucle de rétroaction & confidentialité

L'accord controversé de Google avec plus de cent cinquante hôpitaux aux Etats-Unis

**Le géant du numérique assure que le partenariat avec Ascension révélé par le « Wall Street Journal », qui lui donne accès aux données médicales de millions de patients sans leur consentement, est légal.**

Par Alexandre Piquard

Publié le 12 novembre 2019 à 14h50 · Mis à jour le 13 novembre 2019 à 10h29 · Lecture 5 min.

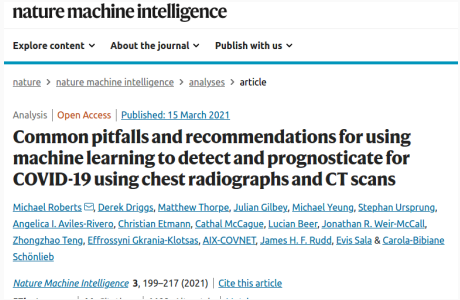


# Anticiper l'*AI Act*

## Qualité, robustesse, résilience des décisions algorithmiques

## Choix d'une métrique & précision, échantillon test

- Régression : variable cible $Y$ quantitative
  Fonction perte $L_2$ ou $L_1$
- Classification binaire : Taux d'erreur, AUC, score $F_\beta$, entropie...
- Multiclasse : Taux moyen, $F_\beta$ moyen...


nature machine intelligence

Explore content ~ About the journal ~ Publish with us ~

nature > nature machine intelligence > analyses > article

Analysis | Open Access | Published: 15 March 2021

**Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans**

Michael Roberts ✉, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

*Nature Machine Intelligence* **3**, 199–217 (2021) | Cite this article

## Robustesse

- Valeurs atypiques et choix de la fonction perte
- Détection des anomalies (*outliers*) de la base d'apprentissage & en exploitation

## Résilience

- Données manquantes de la base d'apprentissage, en exploitation

# Anticiper l'*AI Act*

## Explicabilité d'une décision

## Quelle niveau d'explication ? Pour qui ? (Barredo Arrieta et al. 2020)

426 références !

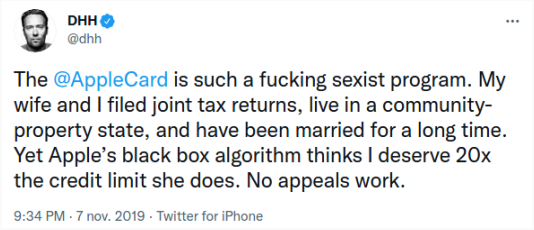
1. Fonctionnement général de l'algorithme, domaines de défaillances
   - Modèles linéaires, arbres *vs.* algorithme opaque : neurones, agrégation, SVM...
     - Approximation : linéaire, arbre, règles,...
     - Importance des variables
2. Décision spécifique
   - Concepteur : Expliquer une erreur, y remédier : ré-apprentissage
   - Utilisateur, usagers
     - Modèle linéaire, arbre de décision
     - Approximation locale : LIME, contre-exemple, règles,...
     - *a minima* : risque d'erreur

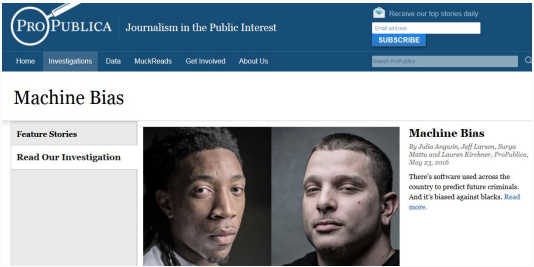
*Source : Davis Parkins, Nature, vol. 597, sept. 2021 p175*

# Anticiper l'*AI Act*

## Risques de discrimination

# Exemples de discrimination algorithmique




**Détection d'une discrimination humaine**

Exemple : discrimination à l'embauche


- 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)

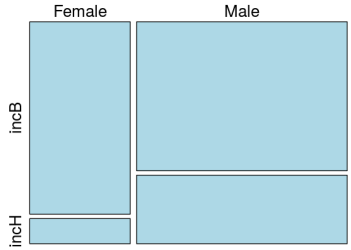
$$DI = \frac{\mathbb{P}(\widehat{Y} = 1 | S = 0)}{\mathbb{P}(\widehat{Y} = 1 | S = 1)}$$

## Détection d'une discrimination algorithmique : critères statistiques

- Pas de définition juridique de l'équité : absence de discrimination
- Indicateurs de discrimination : Zliobaité (2017), 70 sur `aif360.mybluemix.net`
- Critères, redondants, corrélés : Friedler et al. (2019), Verma et Rubin (2018)
- En pratique Trois niveaux de biais
  1. Effet disproportionné ou *demographic equality* : $DI = \frac{\mathbb{P}(\hat{Y}=1|S=0)}{\mathbb{P}(\hat{Y}=1|S=1)}$
  2. Taux d'erreur conditionnels (*overall error equality*) : $\frac{\mathbb{P}(\widehat{Y}\neq Y|S=0)}{\mathbb{P}(\widehat{Y}\neq Y|S=1)}$
     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 *vs.* equivant (COMPAS)
- Estimation par Intervalle de confiance (Besse et al. 2021) (Dépôt Github)

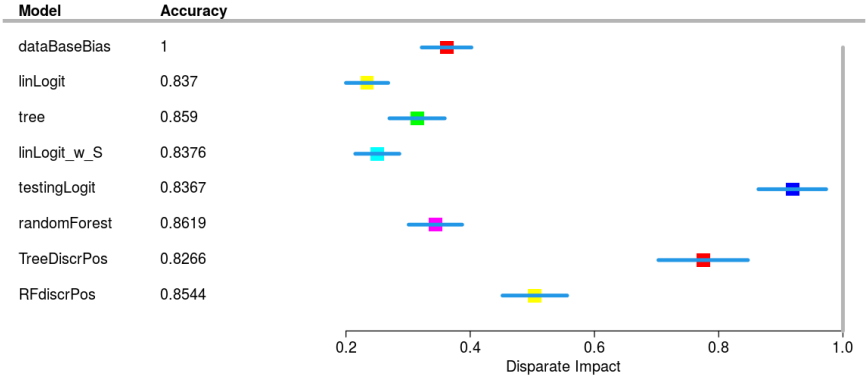
**Cas d'Usage illustratif :** *Adult Census Dataset*

- Codes R, Python, disponibles sur github/wikistat
- Données publiques de l'UCI
- 48 842 individus décrits par 14 variables issues d'un sondage aux USA (1994)
  - Genre, origine ethnique, niveau d'éducation, occupation, statut familial, nombre d'heures travaillées par semaine...
  - $Y$ : Seuil de Revenu inférieur ou supérieur à 50k$
  - Prévision de la classe ou "solvabilité"
  - Données largement biaisées selon le genre, biaisées selon l'origine




$$DI = \frac{\mathbb{P}(Y=1|S=0)}{\mathbb{P}(Y=1|S=1)} = 0.37$$

$$\mathbb{P}\left(DI \in [0.35, 0.38]\right) = 0.95$$




*Détection de la discrimination indirecte ($DI = \frac{\mathbb{P}(\widehat{Y}=1|S=0)}{\mathbb{P}(\widehat{Y}=1|S=1)}$) de différents algorithmes*

Erreurs : 0.07 *vs.* 0.19, TFP : 0.02 *vs.* 0.08, TFN : 0.50 *vs.* 0.45

Attention : atténuer un biais impacte les autres

## Conclusion provisoire – Avancées de l'*AI Act*

- Chartes insuffisantes : la nécessité de conformité se substitue à l'éthique
- Transparence et documentation exhaustive :
  - Analyse préalable des données
  - Recherche des biais (art. 10, 2, (f)), utilisation données sensibles (art. 10, 5)
  - Évaluation des performances, risques de défaillance, robustesse résilience
  - Capacités d'explication à la mesure des progrès scientifiques
  - Risques de biais (performances) vis-à-vis de groupes
- Contrôle humain de la gestion des risques en exploitation
- Enregistrement, traçabilité des décisions
- Marquage "CE" des systèmes d'IA Annexe II
  - Autorité notifiante : ANSM pour les dispositifs de santé
  - Organisme de notification : Référentiel de certification du Process IA (LNE 2021) GMED pour les dispositifs de santé
  - Évaluation et certification *ex-ante*

## Conclusion provisoire – Limites de l'*AI Act*

- Objectif : harmonisation des relations commerciales de l'UE
  Sécurité des produits ou responsabilité du fait de produits défectueux
  Cf. Exigences de la FTC (*federal trade commission*)
- Protection de l'utilisateur, pas celle de l'usager :
  - Quelles explications à l'usager (RGPD) des risques, d'une décision ?
  - Quid des biais systémiques ? De leur atténuation ?
  - Quelles normes ?
- Impacts environnementaux du numérique "oubliés"
- Marquage "CE" des systèmes d'IA Annexe III
  - Qui accède à la documentation ? l'usager ?
  - Possibilité de saisir la Défenseure Des Droits ?
  - Risque de se voir opposer le secret commercial ?

Risques interdépendants : recherche **documentée** d'une moins mauvaise solution

# Références


- Bachoc F., Gamboa F., Halford M., Loubes J.-M., Risser L. (2020). Entropic Variable Projection for Model Explainability and Intepretability, arXiv preprint : 1810.07924.
- Barocas S. , Selbst A. (2016). Big Data's Disparate Impact, *California Law Review* (104), 671.
- Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. (2020). Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, Vol. 58, pp 82-115.
- Besse P. (2020). Détecter, évaluer les risques des impacts discriminatoires des algorithmes d'IA, Contribution au séminaire Défenseur des Droits & CNIL, 28 mai 2020.
- Besse P. (2021). Statistique & Règlement Européen des Systèmes d'IA (AI Act), preprint, HAL-03253111.
- 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*, Vol6 (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.
- 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.
- 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. 32938.
- LNE (2021). Référentiel de Certification d'un Processus d'IA, version 02 – juillet 2021.
- Verma S., Rubin J. (2018). Fairness Definitions Explained, ACM/IEEE International Workshop on Software Fairness.
- Xu D., Yuan S., Zhang L., Wu X. (2018). FairGAN : Fairness-aware Generative Adversarial Networks, IEEE International Conference on Big Data, pp. 570-575.ala