Risques Éthiques & Juridiques des Impacts Sociétaux des Algorithmes d'IA

PHILIPPE BESSE

Intelligence Artificielle (IA) au quotidien

- 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
- 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*)
  - ... 95% des applications de l'IA (Yan Le Cun)
- NMF, MLG, Arbres binaires, SVM, *random forest, boosting*, *deep learning*...
- **Essentiel** : accès à des **données** fiables et **représentatives**

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

Confiance, Acceptabilité, Loi & Éthique

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

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

- Applicabilité des textes de loi *vs.* disruptions technologiques
- Auditabilité des algorithmes (Villani 2018)
- Capacité de détection des transgressions de la loi
- Attention à l'*éthical washing*
- Une loi applicable est préférable à des dizaines de chartes éthiques

Quels risques des impacts sociétaux des décisions algorithmiques ?

Cinq questions Juridiques et / ou Éthiques

1. Protection : propriété, confidentialité des données personnelles (RGPD, CNIL)
2. Entraves à la concurrence : comparateurs, *pricing* automatique
3. Qualité, robustesse des prévisions donc des décisions
4. Explicabilité *vs.* opacité des algorithmes
5. Biais & Discrimination des décisions algorithmiques

Jongler entre textes nationaux et RGPD

- Loi n˚78-17 du 6/01/1978 relative à l'informatique aux fichiers et aux libertés
- Loi n˚2015-912 du 24/07/2015 relative au renseignement
- Loi n˚2016-1321 du 7/10/2016 pour une République Numérique (Lemaire)
- Décrets d'applications (2017)
- RGPD Règlement Général pour la Protection des Données 05-2018
- Loi n˚2018-493 du 20 juin 2018 informatique et libertés (LIL 3)
- Code pénal
- Code des relations entre le public et les administrations
- Code de la Santé publique
- ...
- Conseil Constitutionnel Décision n˚2018-765 DC du 12 juin 2018

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

Règlement Général sur la Protection des Données

- Article 12 : Le **responsable du traitement** prend des mesures appropriées pour fournir toute information [...] ainsi que pour procéder à toute communication [...] en ce qui concerne le traitement à la personne concernée d'une **façon concise, transparente, compréhensible et aisément accessible, en des termes clairs et simples**, [...]
- Articles 14 et 15 : [...] le responsable du traitement fournit à la personne concernée les informations suivantes nécessaires pour garantir un **traitement équitable et transparent** à l'égard de la personne concernée : [...] l'existence d'une prise de **décision automatisée**, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, et, au moins en pareils cas, des **informations utiles concernant la logique sous-jacente**, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée.

Article 22 (RGPD) : Décision individuelle automatisée, y compris le profilage

1. La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un **traitement automatisé**, y compris le **profilage**, produisant des effets juridiques la concernant ou l'**affectant de manière significative** de façon similaire.
2. Le paragraphe 1 ne s'applique pas lorsque la décision :
   - *a* est nécessaire à la conclusion ou à l'exécution d'un **contrat** entre la personne concernée et un responsable du traitement ;
   - *b* est **autorisée par le droit** de l'Union ou le droit de l'État membre auquel le responsable du traitement est soumis et qui prévoit également des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ; ou
   - *c* est fondée sur le **consentement** explicite de la personne concernée.
3. Dans les cas visés au paragraphe 2, points a) et c), le responsable du traitement met en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée, au moins du droit de la personne concernée d'**obtenir une intervention humaine** de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision.
4. Les décisions visées au paragraphe 2 **ne peuvent être fondées** sur les catégories particulières de **données à caractère personnel** (cf. article 9 : biométriques, génétiques, de santé, ethniques ; orientation politique, syndicale, sexuelle, religieuse, philosophique) **sous réserve** d'un intérêt public substantiel et que des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ne soient en place.

Article 225-1 du code pénal

- Constitue une **discrimination** toute distinction opérée entre les personnes physiques sur le fondement de leur **origine**, de leur **sexe**, de leur situation de famille, de leur grossesse, de leur apparence physique, de la particulière vulnérabilité résultant de leur situation économique, apparente ou connue de son auteur, de leur patronyme, de leur lieu de résidence, de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle, de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leur capacité à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une **ethnie**, une Nation, une prétendue race ou une religion déterminée
- Constitue une **discrimination indirecte** une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour **des personnes par rapport à d'autres personnes**, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés.

Article 225-2 du code pénal

La **discrimination** définie aux articles 225-1 à 225-1-2, commise à l'égard d'une **personne physique** ou morale, est punie de **trois ans d'emprisonnement** et de **45 000 euros** d'amende lorsqu'elle consiste à :

1. refuser la fourniture d'un bien ou d'un service
2. entraver l'exercice normal d'une activité économique quelconque
3. refuser d'embaucher, à sanctionner ou à licencier une personne

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / 8 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Source : Agence Reuters

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

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

Chapitre III : **Liste d'évaluation pour une IA digne de confiance** (10 pages cf. *PIA*)

1. Action humaine et contrôle humain
2. Robustesse technique et sécurité (résilience, précision...)
3. Respect de la vie privée et gouvernance des données (qualité...)
4. Transparence (explicabilité, communication...)
5. Diversité, non-discrimination et équité
6. Bien-être sociétal et environnemental (durabilité, interactions...)
   Utilité & bien commun ? Balance bénéfice / risque
7. Responsabilité (auditabilité, recours...)

Qualité des décisions & vide juridique

- Algorithme d'apprentissage : erreur de prévision, qualité de décision, confiance
- Taux d'erreur de 3% en image *vs.* 30 à 40% pour le risque de récidive
- Considérant (71) du RGPD mais loi française muette
- *Ethical washing* & intérêt commercial : cf. publication des sondages d'opinion
- Ne pas confondre estimation / prévision d'une moyenne (*loi des grands nombres*) et celle d'un comportement individuel
- Éthique : Obligation de moyen, pas de résultat mais obligation de transparence
- Industrie et Santé : objectif de certification

Exemple de questions de la liste d'évaluation

2 *Robustesse technique et sécurité (résilience, précision...)*

- Avez-vous évalué le niveau de précision et la définition de la précision nécessaires dans le contexte du système d'IA et du cas d'utilisation concerné ?
- Avez-vous réfléchi à la manière dont la précision est mesurée et assurée ?
- Avez-vous mis en place des mesures pour veiller à ce que les données utilisées soient exhaustives et à jour ?
- Avez-vous mis en place des mesures pour évaluer si des données supplémentaires sont nécessaires, par exemple pour améliorer la précision et éliminer les biais ?

Précision & choix d'une métrique

- Régression : variable cible Y quantitative
  Fonction perte L_2 (quadratique) ou L_1 (valeur absolue)
- Classification binaire
  Taux d'erreur, AUC (*area under the ROC Curve*), score F_β, entropie...
- Multiclasse
  Taux d'erreur moyen, F_β moyen...

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

Santé : du buzz à la certification (Liu et al. 2019, *FDA*, HAS)

ARTIFICIAL INTELLIGENCE, DIAGNOSTICS, HEALTH TECH

Google's AI beats humans at detecting breast cancer — sometimes

A retrospective study published in Nature shows Google's DeepMind AI outperformed radiologists in detecting breast cancer. But it won't be replacing them anytime soon.

By ELISE REUTER

nature

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney, Marcin Sieniek, [...] Shravya Shetty

nature Subscribe

INNOVATIONS IN · 18 DECEMBER 2019

Rise of Robot Radiologists

Deep-learning algorithms are peering into MRIs and x-rays with unmatched vision, but who is to blame when they make a mistake?

Computer Science > Machine Learning

Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, Christopher Ré

(Submitted on 27 Sep 2019 (v1), last revised 15 Nov 2019 (this version, v2))

Exemple de questions de la liste d'évaluation

4 *Transparence (explicabilité, communication...)*

- Avez-vous évalué la mesure dans laquelle les décisions prises, et donc les résultats obtenus, par le système d'IA peuvent être compris ?
- Avez-vous veillé à ce qu'une explication de la raison pour laquelle un système a procédé à un certain choix entraînant un certain résultat puisse être rendue compréhensible pour l'ensemble des utilisateurs qui pourraient souhaiter obtenir une explication ?

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.* neurones, agrégation, SVM...
   - Approximation : linéaire, arbre, règles,...
   - Importance des variables, stress de l'algorithme et impact (Bachoc et al. 2020)
2. Décision spécifique
   - Concepteur : Expliquer une erreur, y remédier : ré-apprentissage
   - Personne concernée : client, patient, justiciable...
     - Interprétable : modèle linéaire, arbre de décision
     - Approximation locale : LIME, contre-exemple, règles,...
     - *a minima* : risque d'erreur

Quelques démos : https://aix360.mybluemix.net/

Détection d'une discrimination (in)directe : *Testing*

- Riach et Rich (2002) : économistes, sociologues
- Comité National de l'Information Statistique
- Observatoire des Discriminations (Paris 1)
- DARES (Direction de l'Animation, des Études, de la Recherche et des Statistiques)

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

Sur les 103 entreprises testées, les chercheurs identifient « entre 5 et 15 entreprises discriminantes » à l'encontre du « candidat présumé maghrébin ».

Le Monde avec AFP • Publié le 08 janvier 2020 à 18h30 - Mis à jour le 09 janvier 2020 à 07h41

Source : Le Monde

USA : Civil Rights act & Code of Federal Regulations

Title 29 - Labor : PART 1607—UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES (1978)

- D. **Adverse impact and the "four-fifths rule.**" A **selection rate** for any race, sex, or ethnic group which is **less than four-fifths (4/5) (or eighty percent)** of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. **Smaller differences** in selection rate may nevertheless constitute adverse impact, where they are **significant in both statistical and practical** terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant, or where special recruiting or other programs cause the pool of minority or female candidates to be atypical of the normal pool of applicants from that group.

Mesurer pour progresser dans l'égalité des chances

DdD & CNIL (2012) : Guide méthodologique à l'usage des acteurs de l'emploi

- 24 fiches 110 pages
- Mesurer des discriminations
- Question des statistiques ethniques : méthode patronymique

Tableau 3 : Probabilité d'être recruté suivant le type de fonction postulée, l'origine supposée et le niveau d'étude

Type de fonction postulée	Candidatures évoquant une origine...					
	« européenne »			« extra-européenne »		
	< Bac	Bac/Bac+2	≥ Bac+5	< Bac	Bac/Bac+2	≥ Bac+5
Fonctions en contact clientèle requérant un Bac/Bac+2	1%	**5%**	4%	0%	**2%**	5%
Fonctions d'encadrement requérant un Bac+5	0%	1%	**9%**	0%	0%	**6%**
Ensemble	1%	4%	8%	0%	2%	5%

Exemple de questions de la liste d'évaluation

5 *Diversité, non-discrimination et équité*

- Avez-vous prévu une définition appropriée de l'équité que vous appliquez dans la conception des SIA ?
- Avez-vous mis en place des processus pour tester et contrôler les biais éventuels au cours de la phase de mise au point, de déploiement et d'utilisation du système ?
- Avez-vous prévu une analyse quantitative ou des indicateurs pour mesurer et tester la définition appliquée de l'équité ?

Algorithmes d'apprentissage et discrimination

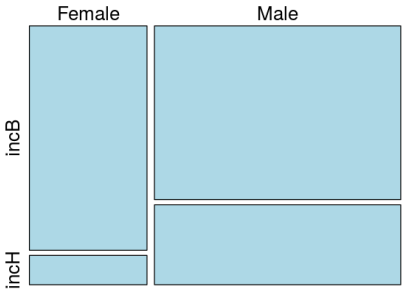
- Biais systémique ou social des bases de données
- Renforcement du biais et discrimination
- Variable sensible absente mais prévisible

Détection d'une discrimination de groupe ou indirecte : *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 estimés par IC (Besse et al. 2020) : Dépôt Github des fonctions en R et Python
  1. Effet disproportionné ou *Disparate Impact* (*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 $\frac{\mathbb{P}(\widehat{Y}=1|Y=1,S=0)}{\mathbb{P}(\widehat{Y}=1|Y=1,S=1)}$
     Justice "prédictive" : Propublica *vs.* equivant (Compas)

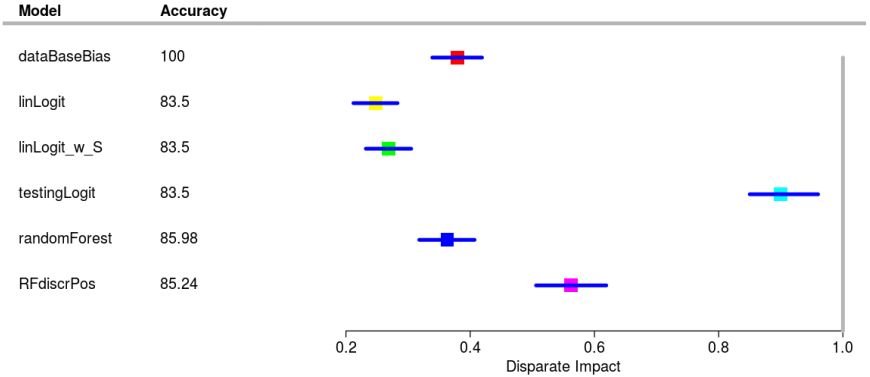
Cas d'Usage : *Adult Census Dataset*

- Code disponible 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*

Attention : impact de la correction de l'effet disproportionné sur les deux autres biais

Exemple : gestion des ressources humaines & recrutement prédictif

USA : *hiring tech* (Raghavan et al. 2019) *vs.* **France** : *easyrecrue* (Hemamou et al. 2018)

- Données : transcription lexicale d'une vidéo & évaluée par un recruteur humain
- $n = 305$ sélectionnés : favorable / réservé parmi 607
- $\ll$ *Le score d'aire sous la courbe ROC (0.69) est bien supérieur à celui qui serait obtenu avec l'aléatoire (0.5)* $\gg$
- Aucune évocation de biais potentiels

Fair learning : trois niveaux d'intervention

Friedler et al. (2019), 10 approches sur `aif360.mybluemix.net`

1. Pre-processing des données d'apprentissage :
   - *e.g.* Gordaliza et al. (2019) par transport optimal
     Approcher la parité statistique : $\mathbb{P}(\widehat{Y}=1|S=0)=\mathbb{P}(\widehat{Y}=1|S=1)$
     "Rapprocher" (Wasserstein) : $\mathcal{L}(\widetilde{X}|S=0)$ et $\mathcal{L}(\widetilde{X}|S=1)$
   - *Fairness-aware Generative Adversarial Networks* (Xu et al. 2018)
2. Modification de l'algorithme : *e.g.* Zafar et al. (2017)
   Contrainte d"équité" mais non convexe
3. Post-processing *e.g.* modification des seuils

- Sans connaître la ou les variables sensibles *e.g.* Romanov et al. (2019)
  Word embedding des noms et prénoms avec contrainte d'équité
- **Attention** : Intervenir sur un biais impacte les autres (Chouldechova, 2017).
- **Question politique** : quelle part de discrimination positive ?

Tout est lié et affaire de compromis

Utilité d'un système d'IA : équilibre bénéfice / risque

1. Confidentialité, protection ds données *vs.* connaissance de la variable sensible
2. Qualité, robustesse de la décision algorithmique
3. Explicabilité de la décision algorithmique
4. Types de bais donc risques de discrimination
   - Biais systémique, des erreurs, de leur asymétrie

En chantier

- Auditabilité et contrôle : liste d'évaluation & renversement de la charge de preuve
- Normes : ANSI, IEEE, ISO ?
- Certification en santé : *FDA*, HAS et dans l'industrie : projet DEEL, ANITI
- Tutoriel dépôt Fair-ML-4-Ethical-AI

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. del Barrio E. Gordaliza P. Loubes J.-M., Risser L. (2020). A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set, à paraître.
- Besse P. Besse-Patin A., Castets-Renard C. (2019). Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé, soumis, hal-02424285.
- Chouldechova A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *Big Data*, Special issue on Social and Technical Trade-offs, Vol. 5, No. 2, pp 153-163.
- CNIL, Défenseur des Droits (2012). Mesurer pour progresser vers l'égalité des chances, Guide méthodologique à l'usage des acteurs de l'emploi.
- Code of Federal Regulations (1978). Title 29 - Labor PART 1607—UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES (1978), CHAPTER XIV - EQUAL EMPLOYMENT OPPORTUNITY COMMISSION.
- Commission Européenne (2016). Règlement Général sur la Protection des Données.
- Commission Européenne (2018). Lifgnes directrices pour une IA de confiance.
- Commission Européenne (2020). Livre blanc sur l'intelligence artificielle: une approche européenne d'excellence et de confiance.
- Défenseur des Droits, CNIL (2012). Mesurer pour progresser vers l'égalité des chances. Guide méthodologique à l'usage des acteurs de l'emploi.
- De-Arteaga M., Romanov A. et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, in FAT'19, pp 120–128.

Références suite

- 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.
- Gordaliza P., Barrio E.D., Fabrice G., Loubes J.-M. (2019). Obtaining Fairness using Optimal Transport Theory, Proceedings of the 36th International Conference on Machine Learning, in PMLR 97 :2357-2365.
- Hemamou L., Wajntrob G., Martin J.-C., Clavel C. (2018). Entretien vidéo différé: modèle prédictif pour la pré-sélection de candidats sur la base du contenu verbal, Workshop sur les Affects, Compagnons Artificiels et Interactions.
- Raghavan M., Barocas S., Kleinberg J., Levy K. (2020) Mitigating bias in Algorithmic Hiring : Evaluating Claims and Practices, in FAT* 20, pp 469–481.
- Riach P.A., Rich J. (2002). Field Experiments of Discrimination in the Market Place, *The Economic Journal*, Vol. 112 (483), pp F480-F518.
- Romanov A., De-Arteaga M., Wallach H., Chayes J., Borgs C., Chouldechova A., Geyik S., Kenthapadi K., Rumshisky A., Kalai A. (2019). What's in a name? Reducing bias in bios without access to protected attributes, in annual conference of the north american chapter of the association for computational linguistics, pp 4187-4195.
- 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.
- Xu D., Yuan S., Zhang L., Wu X. (2018). FairGAN: Fairness-aware Generative Adversarial Networks, IEEE International Conference on Big Data, pp. 570-575.
- Zafar M., Valera I., Rodriguez M., Gummadi K. (2017). Fairness Constraints: Mechanisms for Fair Classication, in International Conference on Articial Intelligence and Statistics (AISTATS), vol. 5.
- Zliobaité I. (2017). Measuring discrimination in algorithmic decision making, *Data Min Knowl Disc* 31, 1060–1089.

HAS
HAUTE AUTORITÉ DE SANTÉ

Liste d'évaluation de la Haute Autorité de Santé

- Évaluer les technologies de Santé
- Dépôt d'un dossier auprès de la Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé(CNEDiMTS)
- p43 : Informations descriptives spécifiques à fournir pour les fonctionnalités du dispositif médical s'appuyant sur des procédés d'apprentissage automatique (technologies relevant du champ de l'intelligence artificielle)
- Grille descriptive de 42 questions
- Accès au document

Liste évaluation provisoire de la Commission Européenne

- Lignes directrices en matière d'éthique pour une IA digne de confiance
- Groupe d'experts de haut niveau sur l'intelligence artificielle
- Chapitre III p32 : Liste d'évaluation pour une IA digne de confiance (version pilote)
- Accès au document