Pour une intelligence artificielle inclusive et éthique

Vincent Lequertier

Campus du libre - 26 novembre 2022

Plan

Plan

Qu’est-ce que l’intelligence artificielle (IA) ?

Optimisation automatique de
paramètres selon un objectif donné

L'éthique en IA

- Pas de favoritisme ou de discriminations
- Les algorithmes traitent des informations personnelles
- Difficile de trouver une mesure quantitative

L'IA est utilisé pour des missions divers et importantes

- L'approbation des prêts
- La police prédictive
- Pour calculer des scores de crédits sociaux
- Les voitures autonomes
- En santé

Déterminer si une IA respecte l'éthique est complexe

Figure 1 : Il est difficile d'interpréter la sortie d'un modèle de prédiction [1]

1. https://xkcd.com/1838/

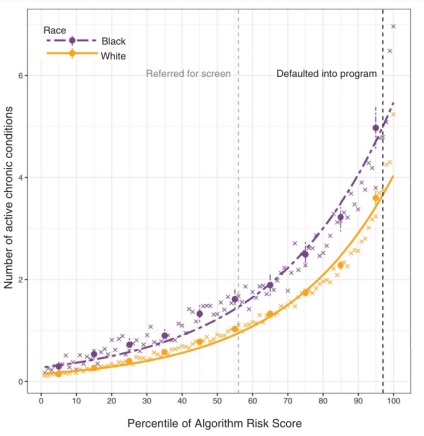


Figure 2 : Biais dans le calcul d'un score de risque score [2]

2. Ziad Obermeyer et al. « Dissecting racial bias in an algorithm used to manage the health of populations ». In : *Science* 366.6464 (25 oct. 2019), p. 447-453. issn : 0036-8075, 1095-9203. doi : 10.1126/science.aax2342.

Vocabulaire

- Vrai Positif (VP)
- Vrai Négatif (VN)
- Faux Positif (FP)
- Faux Négatif (FN)

		réalité	
		Vrai	Faux
prédit	Vrai	VP	FP
	Faux	FN	VN

Les problèmes d'éthique à cause des biais

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Figure 3 : Performance d'IA commercial pour la reconnaissance faciale. Elles sont particulièrement mauvaises pour les femmes de couleur noire. [4]

3. Joy Buolamwini et Timnit Gebru. « Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification ». In : *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Sous la dir. de Sorelle A. Friedler et Christo Wilson. T. 81. Proceedings of Machine Learning Research. PMLR, 23 fév. 2018, p. 77-91. url : `https://proceedings.mlr.press/v81/buolamwini18a.html`.

4. Buolamwini et Gebru, « Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification ».

Biais historique : score de recidive aux USA

Table 1 : Analyse de l'algorithme COMPAS, propublica.org, 2016 [5]

	Tous les accusés		Accusés noirs		Accusés blancs	
	Bas	Haut	Bas	Haut	Bas	Haut
Non récidive	2681	1282	990	805	1139	349
Récidive	1216	2035	532	1369	461	505
Taux FP	32.35		**44.85**		23.45	
Taux FN	37.40		27.99		**47.72**	

- Equivalent à une régression ou à des prédictions par des personnes lambda [6]
- L'algorithme n'utilisait pas la variable "race"

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

6. Julia Dressel et Hany Farid. « The accuracy, fairness, and limits of predicting recidivism ». In : *Science Advances* 4.1 (5 jan. 2018), eaao5580. issn : 2375-2548. doi : `10.1126/sciadv.aao5580`.

Biais dans la police prédictive

- Cercle vicieux si l'IA influence la donnée qu'elle obtient
- Prédiction du quartier dans lequel des crimes seront commis
- Si l'IA guide la police, elle trouve des crimes. Cela créé un cercle vicieux[7]
- Ce problème se retrouve aussi dans les systèmes de recommandation

7. Danielle Ensign et al. *Runaway Feedback Loops in Predictive Policing*. Number : arXiv:1706.09847. 21 déc. 2017. arXiv : `1706.09847[cs,stat]`.

Biais venant de la mesure

- Arrive si la mesure est biaisée
- L'échantillon à partir duquel une mesure est effectuée peut contenir du biais
- Exemple : prédiction des AVC à partir des diagnostics d'AVC précédent [8]

	Stroke	30 Day Mortality
Prior Stroke	.302*** (.012)	.041*** (.014)
Prior Accidental Injury	.285** (.095)	.007 (.101)
Abnormal Breast Finding	.224* (.092)	.162 (.110)
Cardiovascular Disease History	.218*** (.029)	−.017 (.034)
Colon Cancer Screening	.242 (.178)	−.475** (.222)
Acute Sinusitis	.220 (.155)	.056 (.166)

Figure 4 : Variables associées avec le fait d'avoir yun AVC sont biaisés car basé sur ceux pour lesquels un diagnostic a pu être établi

8. Sendhil Mullainathan et Ziad Obermeyer. « Does Machine Learning Automate Moral Hazard and Error ? » In : *The American Economic Review* 107.5 (mai [illegible] 0171084

Un biais peut arriver de plusieurs manières différentes

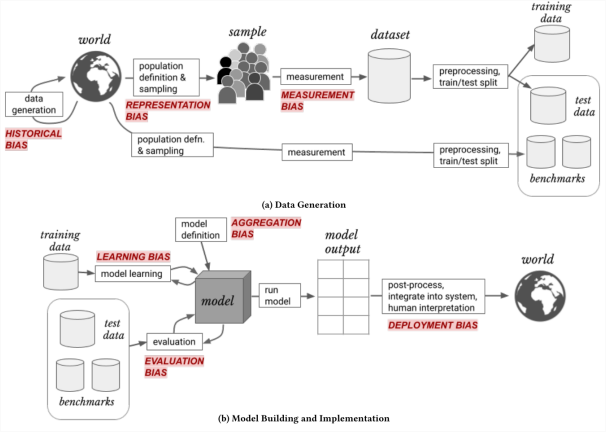


Figure 5 : Processus pour générer des données et développer des modèles d'IA et types de biais [10]

9. Harini Suresh et John V. Guttag. « A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle ». In : *Equity and Access in Algorithms, Mechanisms, and Optimization*. 5 oct. 2021, p. 1-9. doi : 10.1145/3465416.3483305. arXiv : 1901.10002[cs,stat].

Mais les humains sont injustes aussi !

- On tend à avoir une confiance aveugle envers les ordinateurs
- Une procédure automatique a moins de possibilités d'appels
- Une procédure automatique avec une IA traite beaucoup plus de dossiers qu'une procédure manuelle

Plan

Mesure de performance

Définition
Un algorithme est juste s'il n'inclut pas un attribut sensible dans les données d'entrées du processus de décision

Problème
L'hypothèse d'indépendance est souvent fausse dans la pratique

La couleur des voitures n'est pas un problème

Prédisons la vitesse des voitures

Table 2 : Les caractéristiques des voitures

Marque	Nombre de sièges	Année	Couleur	Vitesse (km/h)
A	5	2011	bleu	150
B	2	2012	noire	200
C	5	2010	grise	230

Le genre peut être un problème

Prédisons des salaires

Table 3 : Caractéristiques des employés

Genre	Hobby	Formation	Salaire
femme	Capitaine de l'équipe féminine de volley	Informatique	35K
homme	Capitaine de l'équipe de football	autodidacte	37K
homme	échec	Informatique	40K

Attention !
Il faut pensez aux corrélations !

Mesure de performance - la parité prédictive [11]

Définition
Un algorithme est éthique si il a la même précision pour toutes les valeurs d'un attribut sensible, c'est à dire le même ratio entre les VP et toutes les valeurs prédites "Vrais" (PPV) :

$$PPV = \frac{VP}{VP + FP}$$

11. Alexandra Chouldechova. *Fair prediction with disparate impact : A study of bias in recidivism prediction instruments.* Number : arXiv:1703.00056. 28 fév. 2017. arXiv : `1703.00056[cs,stat]`.

Mesure de performance - la parité prédictive étendue

Définition
Même PPV et NPV (pareil que PPV mais pour les Vrais Négatifs)

$$\text{PPV} = \frac{VP}{VP + FP}$$

$$\text{NPV} = \frac{VN}{VN + FN}$$

Mesure de performance - Egalité de l'opportunité [12]

Définition
L'algorithme est éthique s'il a le même toux de faux positifs et de vrais négatifs. Ce qui veux dire que les personnes avec VP ou VF doivent avoir la même performance de classification quelque soit la valeur d'un attribut sensible.

$$\text{TPR} = \frac{VP}{VP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

12. Moritz Hardt, Eric Price et Nathan Srebro. *Equality of Opportunity in Supervised Learning.* Number : arXiv:1610.02413. 7 oct. 2016. arXiv : 1610.02413[cs].

Définition formelle de l'éthique en IA

- Point de vue de l'individu et point de vue du groupe
- Du point de vue de l'individu, deux personnes similaires dans le contexte d'une tâche doivent avoir les mêmes prédictions [13]
- Comparaison de sous-groupes par rapport à un attribut sensible
- Compromis entre le point de vue du groupe et le point de vue de l'individu [14]

13. Cynthia Dwork et al. *Fairness Through Awareness*. Number : arXiv:1104.3913. 28 nov. 2011. arXiv : `1104.3913[cs]`.

14. Jon Kleinberg, Sendhil Mullainathan et Manish Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. Number : arXiv:1609.05807. 17 nov. 2016. arXiv : `1609.05807[cs,stat]`.

Compare la distance entre les attributs et les prédictions

Table 4 : Les caractéristiques des voitures et les prédictions

Attributes				Target	
Marque	Sièges	Année	Couleur	Vrai vitesse (km/h)	Vitesse prédite (km/h)
A	5	2011	bleu	150	150
B	2	2012	noire	200	140
C	5	2010	grise	230	210

Attention
Cela demande une connaissance précise de la donnée en question

L'éthique de groupe

Compare la distance entre les groupes

Table 5 : Les caractéristiques des voitures et les prédictions

Attributs				Cible	
Brand	Sièges	Année	Couleur	Vrai vitesse (km/h)	Vitesse prédite (km/h)
A	5	2011	bleu	150	150
B	2	2012	bleu	200	140
C	5	2010	rouge	230	210
B	2	2012	rouge	200	140
C	5	2010	verte	230	210
C	5	2010	verte	230	210

Il existe des solutions politiques

- Politique européenne *Artificial Intelligence Act*

Figure 6 : Échange de vues parlementaire conjoint entre les commissions du Marché intérieur et protection des consommateurs (IMCO) et la commission des libertés civiles, de la justice et des affaires intérieures (LIBE), 25/01/2022.

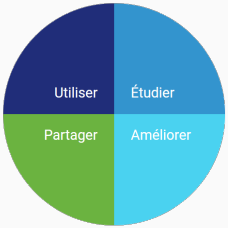
- Si l'IA reproduit et exacerbe les biais et discriminations de la société, rendre la société plus juste ne rendrait il pas l'IA plus juste elle aussi ?

Que peut on faire ?

- Détecter des biais possibles dans les données d'entrainements
- Analyser les corrélations entre les variables
- Comparer les résultats avec des prédictions manuelles ou des programmes plus simples
- Identifier comment les erreurs peuvent être découvertes a posteriori
- Analyser l'évolution des performances pour des sous-groupes

Plan

Les 4 libertés du logiciel libre [15]



Utiliser

Les Logiciels Libres peuvent être utilisés pour n'importe quel but et ne sont pas soumis à des restrictions telles que l'expiration de licence ou des limitations géographiques.

Étudier

Les Logiciels Libres peuvent être étudiés par n'importe qui, sans clauses de confidentialité ou restrictions similaires.

Partager

Les Logiciels Libres peuvent être copiés et partagés sans coût effectif.

Améliorer

Les Logiciels Libres peuvent être modifiés par quiconque, et ces améliorations peuvent être partagées avec tous.

15. *Logiciel libre - FSFE*. FSFE - Free Software Foundation Europe. url : `https://fsfe.org/freesoftware/freesoftware.html`.

Le libre appliqué à l'IA

- Les données d'entrainement
- Le programme d'entrainement du modèle
- Le modèle
- Les prédictions du modèle
- Améliorations ?

Le logiciel libre pour l'intelligence artificielle

- Rend l'IA plus accessible
  - Encourage l'innovation
  - Réduit les coûts
  - Facilite la maintenance
- Donne l'opportunité à tous d'inspecter le code (données, etc.)
- Permet de détecter les discriminations
- Tout le monde peut la rendre plus transparente
- Tout le monde peut la rendre plus éthique
- Donne plus de confiance et facilite l'adoption
- N'est pas une panacée
  - Données avec accès restreint, ou très volumineuses
  - L'entrainement demande de l'expertise et des ressources
  - Une IA sous licence libre n'est pas forcément éthique

- Résolution sur le logiciel libre [16]
- Résolution pas passée dans la régulation à ce stade
- Logiciel libre dans la recherche sur l'IA
- Document pour influencer le processus législatif européen [17, 18]
- Demande politique sur le logiciel libre et l'intelligence artificielle
- Comment définir l'IA et l'éthique dans un texte de loi ?

16. Special Committee on Artificial Intelligence in a Digital Age - Rapporteur Axel Voss. *REPORT on artificial intelligence in a digital age*. REPORT on artificial intelligence in a digital age. 7 juin 2022. url : https://web.archive.org/web/20220607153932/https://www.europarl.europa.eu/doceo/document/A-9-2022-0088_EN.html.
17. Alexander Sander et Lina Ceballos. *FSFE AI and Free Software*. Mars 2022. url : https://download.fsfe.org/campaigns/AIandFS/fsfe_AIandFreesoftware.pdf.
18. *Artificial Intelligence (AI) Act : Free Software is key ! - FSFE*. FSFE - Free Software Foundation Europe. url : https://fsfe.org/news/2022/news-20220330-01.html.

Déjà **33859 SIGNATURES** – signer la lettre ouverte maintenant !

Une campagne menée par la fsfe

PUBLIC MONEY
PUBLIC CODE

Pourquoi les logiciels financés par l'impôt ne sont pas publiés sous Licence Libre ?

Nous voulons une législation qui requiert que le logiciel financé par le contribuable pour le secteur public soit disponible publiquement sous une licence de Logiciel Libre et Open Source.
S'il s'agit d'argent public, le code devrait être également public.

Le code payé par le peuple devrait être disponible pour le peuple!

Conclusions

- Les problèmes d'éthiques peuvent avoir des conséquences désastreuses
- Le concept d'éthique est difficile à modéliser pour un ordinateur
- Des solutions techniques et politiques existent

Merci !

Lien vers les diapositives :

Références i

Obermeyer, Ziad et al. « Dissecting racial bias in an algorithm used to manage the health of populations ». In : *Science* 366.6464 (25 oct. 2019), p. 447-453. issn : 0036-8075, 1095-9203. doi : 10.1126/science.aax2342.

Buolamwini, Joy et Timnit Gebru. « Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification ». In : *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Sous la dir. de Sorelle A. Friedler et Christo Wilson. T. 81. Proceedings of Machine Learning Research. PMLR, 23 fév. 2018, p. 77-91. url : https://proceedings.mlr.press/v81/buolamwini18a.html.

Dressel, Julia et Hany Farid. « The accuracy, fairness, and limits of predicting recidivism ». In : *Science Advances* 4.1 (5 jan. 2018), eaao5580. issn : 2375-2548. doi : 10.1126/sciadv.aao5580.

Ensign, Danielle et al. *Runaway Feedback Loops in Predictive Policing*. Number : arXiv:1706.09847. 21 déc. 2017. arXiv : 1706.09847[cs,stat].

Mullainathan, Sendhil et Ziad Obermeyer. « Does Machine Learning Automate Moral Hazard and Error ? » In : *The American Economic Review* 107.5 (mai 2017), p. 476-480. issn : 0002-8282. doi : 10.1257/aer.p20171084.

Suresh, Harini et John V. Guttag. « A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle ». In : *Equity and Access in Algorithms, Mechanisms, and Optimization*. 5 oct. 2021, p. 1-9. doi : 10.1145/3465416.3483305. arXiv : 1901.10002[cs,stat].

Chouldechova, Alexandra. *Fair prediction with disparate impact : A study of bias in recidivism prediction instruments*. Number : arXiv:1703.00056. 28 fév. 2017. arXiv : 1703.00056[cs,stat].

Hardt, Moritz, Eric Price et Nathan Srebro. *Equality of Opportunity in Supervised Learning*. Number : arXiv:1610.02413. 7 oct. 2016. arXiv : 1610.02413[cs].

Dwork, Cynthia et al. *Fairness Through Awareness*. Number : arXiv:1104.3913. 28 nov. 2011. arXiv : 1104.3913[cs].

Références ii

Kleinberg, Jon, Sendhil Mullainathan et Manish Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. Number : arXiv:1609.05807. 17 nov. 2016. arXiv : 1609.05807[cs,stat].

Logiciel libre - FSFE. FSFE - Free Software Foundation Europe. url : https://fsfe.org/freesoftware/freesoftware.html.

Artificial Intelligence in a Digital Age - Rapporteur Axel Voss, Special Committee on. *REPORT on artificial intelligence in a digital age*. REPORT on artificial intelligence in a digital age. 7 juin 2022. url : https://web.archive.org/web/20220607153932/https://www.europarl.europa.eu/doceo/document/A-9-2022-0088_EN.html.

Sander, Alexander et Lina Ceballos. *FSFE AI and Free Software*. Mars 2022. url : https://download.fsfe.org/campaigns/AIandFS/fsfe_AIandFreesoftware.pdf.

Artificial Intelligence (AI) Act : Free Software is key! - FSFE. FSFE - Free Software Foundation Europe. url : https://fsfe.org/news/2022/news-20220330-01.html.