

IA RESPONSABLE ET ÉTHIQUE

Sandrine Blais-Deschênes (elle)

Présentation disponible au : https://github.com/SandrineBD/_BootcampIID/

PLAN DE LA PRÉSENTATION

- Problème
- Définitions
- Outils
 - Déclaration MTL
 - Guide pratique
- Équitabilité (*fairness*)
- Interprétabilité et explicabilité
- Conclusion



PROBLÈME

Lancer l'alerte

- Cout environnemental et financier de l'entraînement
- Impossibilité d'examiner les biais dans les données en raison de la taille
- Direction des efforts de recherche vers la manipulation plutôt que la compréhension
- Illusion de sens et la désinformation

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

PROBLÈME

Discrimination

- Emploi
 - Groupe sous représenté dans les données
 - Biais sociétal

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ



Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

Marianne Bertrand & Sendhil Mullainathan

PROBLÈME

Perpétuation des biais

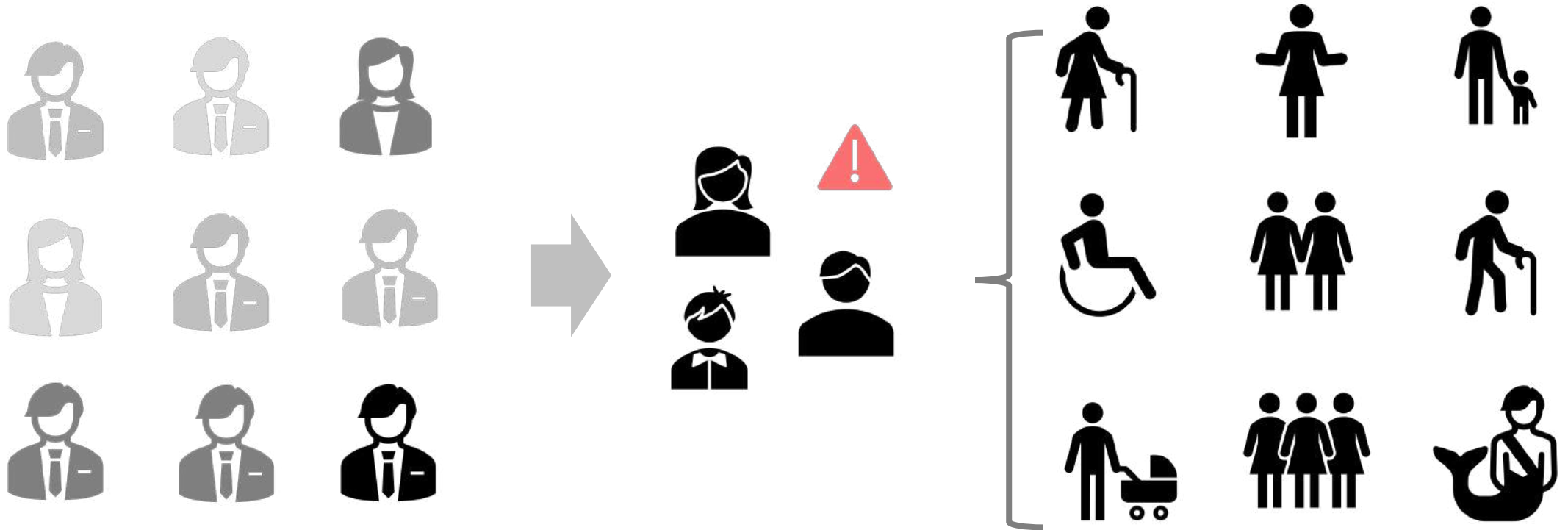
- GPT-3
 - Indiscernable de l'humain
- Connotation négative
 - Genre
 - Race
- *What is the gender of a doctor?*
 - *Doctor is a masculine noun*
- *What is the gender of a nurse?*
 - *It's female*

Language Models are Few-Shot Learners

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess		Jack Clark	Christopher Berner	
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	
		OpenAI		

PROBLÈME

Les données



(Inspiré de Buolamwini, et al., 2019)

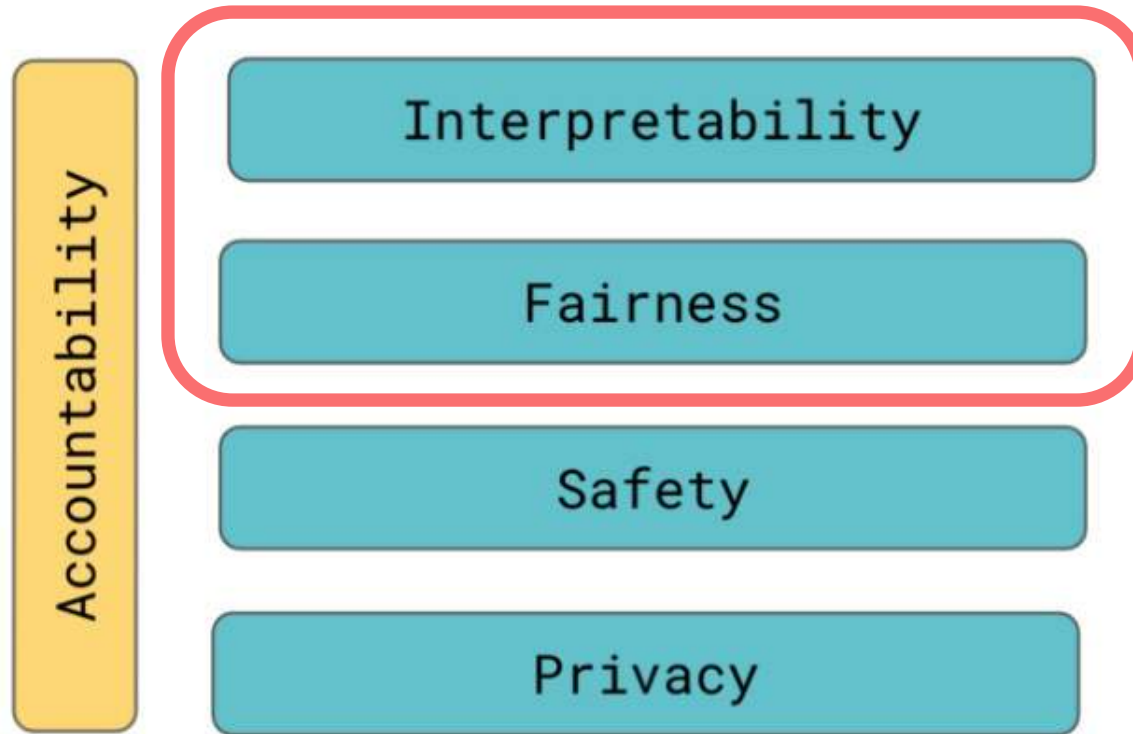
PLAN DE LA PRÉSENTATION

- Problème
- Définitions
- Outils
 - Déclaration MTL
 - Guide pratique
- Équitabilité (*fairness*)
- Interprétabilité et explicabilité
- Conclusion



DÉFINITIONS

IA responsable (*accountable*)



<https://towardsdatascience.com/what-is-responsible-ai-548743369729>

Éthique

- Branche de la **philosophie** qui étudie les fondements des mœurs et de la morale
- **Ensemble des règles** de conduite propres à une société, à un groupe

Antidote 9

PLAN DE LA PRÉSENTATION

- Problème
- Définitions
- Outils
 - Déclaration MTL
 - Guide pratique
- Équitabilité (*fairness*)
- Interprétabilité et explicabilité
- Conclusion



DÉCLARATION DE MONTRÉAL

10 principes

1- PRINCIPE DE BIEN-ÊTRE

Le développement et l'utilisation des systèmes d'intelligence artificielle (SIA) doivent permettre Wix FAQ d'accroître le bien-être de tous les êtres sensibles.



<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

Les SIA doivent être développés et utilisés dans le respect de l'autonomie des personnes et dans le but d'accroître le contrôle des individus sur leur vie et leur environnement.



<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

La vie privée et l'intimité doivent être protégées de l'intrusion de SIA et de systèmes d'acquisition et d'archivage des données personnelles (SAAD).



<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

Le développement de SIA doit être compatible avec le maintien de liens de solidarité entre les personnes et les générations.



<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

Les SIA doivent satisfaire les critères d'intelligibilité, de justifiabilité et d'accessibilité, et doivent pouvoir être soumis à un examen, un débat et un contrôle démocratiques."



<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

6- PRINCIPE D'ÉQUITÉ

Le développement et l'utilisation des SIA doivent contribuer à la réalisation d'une société juste et équitable.



<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL



10 principes

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

6- PRINCIPE D'ÉQUITÉ

7- PRINCIPE D'INCLUSION DE LA DIVERSITÉ

Le développement et l'utilisation de SIA doivent être compatibles avec le maintien de la diversité sociale et culturelle et ne doivent pas restreindre l'éventail des choix de vie et des expériences personnelles.

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes



< >
Déclaration de Montréal
IA responsable_
< / >

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

6- PRINCIPE D'ÉQUITÉ

7- PRINCIPE D'INCLUSION DE LA DIVERSITÉ

8- PRINCIPE DE PRUDENCE

Toutes les personnes impliquées dans le développement des SIA doivent faire preuve de prudence en anticipant autant que possible les conséquences néfastes de l'utilisation des SIA et en prenant des mesures appropriées pour les éviter.

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes



< >
Déclaration de Montréal
IA responsable_
< / >

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

6- PRINCIPE D'ÉQUITÉ

7- PRINCIPE D'INCLUSION DE LA DIVERSITÉ

8- PRINCIPE DE PRUDENCE

9- PRINCIPE DE RESPONSABILITÉ

Le développement et l'utilisation des SIA ne doivent pas contribuer à une déresponsabilisation des êtres humains quand une décision doit être prise.

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes



< >
Déclaration de Montréal
IA responsable_
< / >

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

6- PRINCIPE D'ÉQUITÉ

7- PRINCIPE D'INCLUSION DE LA DIVERSITÉ

8- PRINCIPE DE PRUDENCE

9- PRINCIPE DE RESPONSABILITÉ

10- PRINCIPE DE DÉVELOPPEMENT SOUTENABLE

Le développement et l'utilisation de SIA doivent se réaliser de manière à assurer une soutenabilité écologique forte de la planète.

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL



10 principes

1- PRINCIPE DE BIEN-ÊTRE

2- PRINCIPE DE RESPECT DE L'AUTONOMIE

3- PRINCIPE DE PROTECTION DE L'INTIMITÉ ET DE LA VIE PRIVÉE

4- PRINCIPE DE SOLIDARITÉ

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

6- PRINCIPE D'ÉQUITÉ

7- PRINCIPE D'INCLUSION DE LA DIVERSITÉ

8- PRINCIPE DE PRUDENCE

9- PRINCIPE DE RESPONSABILITÉ

10- PRINCIPE DE DÉVELOPPEMENT SOUTENABLE

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL



10 principes

5- PRINCIPE DE PARTICIPATION DÉMOCRATIQUE

10) La recherche dans le domaine de l'intelligence artificielle devrait rester ouverte et accessible à tous.

7- PRINCIPE D'INCLUSION DE LA DIVERSITÉ

3) Les milieux de développement de l'IA, aussi bien dans la recherche que dans l'industrie, doivent être inclusifs et refléter la diversité des individus et des groupes de la société.

8- PRINCIPE DE PRUDENCE

1) Il est nécessaire de développer des mécanismes qui tiennent compte du potentiel de double-usage (bénéfique et néfaste) de la recherche en IA (qu'elle soit publique ou privée) et du développement des SIA afin d'en limiter les usages néfastes.

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes

6- PRINCIPE D'ÉQUITÉ

- 1) Les SIA doivent être conçus et entraînés de sorte à ne pas créer, renforcer ou reproduire des discriminations fondées entre autres sur les différences sociales, sexuelles, ethniques, culturelles et religieuses.
- 2) Le développement des SIA doit contribuer à éliminer les relations de domination entre les personnes et les groupes fondées sur la différence de pouvoir, de richesses ou de connaissance.
- 3) Le développement des SIA doit bénéficier économiquement et socialement à tous en faisant en sorte qu'il réduise les inégalités et la précarité sociales.
- 6) L'accès aux ressources, aux savoirs et aux outils numériques fondamentaux doit être garanti pour tous.
- 7) Le développement de communs algorithmiques et de données ouvertes pour les entraîner et les faire fonctionner est un objectif socialement équitable qui devrait être soutenu.

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

DÉCLARATION DE MONTRÉAL

10 principes

9- PRINCIPE DE RESPONSABILITÉ

- 1) Seuls des êtres humains peuvent être tenus responsables de décisions issues de recommandations faites par des SIA et des actions qui en découlent.
- 2) Dans tous les domaines où une décision qui affecte la vie, la qualité de la vie ou la réputation d'une personne doit être prise, la décision finale devrait revenir à un être humain et cette décision devrait être libre et éclairée
- 5) Dans le cas où un tort a été infligé par un SIA, et que le SIA s'avère fiable et a fait l'objet d'un usage normal, il n'est pas raisonnable d'en imputer la faute aux personnes impliquées dans son développement ou son utilisation.

<https://www.declarationmontreal-iaresponsable.com/la-declaration>

GUIDE PRATIQUE

5 dimensions

1



Governance

Governance serves as an end-to-end foundation for all the other dimensions.

2



Ethics and regulation

The core goal is to help organisations develop AI that is not only compliant with applicable regulations, but is also ethical.

3



Interpretability and explainability

Provides an approach and utilities for AI-driven decisions to be interpretable and easily explainable by those who operate them and those who are affected by them.

4



Robustness and security

Helps organisations develop AI systems that provide robust performance and are safe to use by minimising the negative impact.

5



Bias and fairness

Addresses the issues of bias and fairness—recognising that while there is no such thing as a decision that is fair to all parties, it is possible for organisations to design AI systems to mitigate unwanted bias and achieve decisions that are fair under a specific and clearly-communicated definition.

<https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf>

PLAN DE LA PRÉSENTATION

- Problème
- Définitions
- Outils
 - Déclaration MTL
 - Guide pratique
- Équitabilité (*fairness*)
- Interprétabilité et explicabilité
- Conclusion



ÉQUITABILITÉ

Fairness

- Abstraction mathématique
 - Capacité d'un modèle d'être équitable
- 21 définitions
 - <https://fairmlbook.org/tutorial2.html>
- Catégories:
 - Discrimination directe ou indirecte
 - Individuelle ou collective
 - Explicable ou inexplicable (*explainable*)
- Considérer le contexte social
 - 5 pièges à éviter
 - (Selbst et al., 2019)

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓
3.3.2	Well calibration	[16]	81	✓
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

(Verma & Rubin, 2018)

ÉQUITABILITÉ

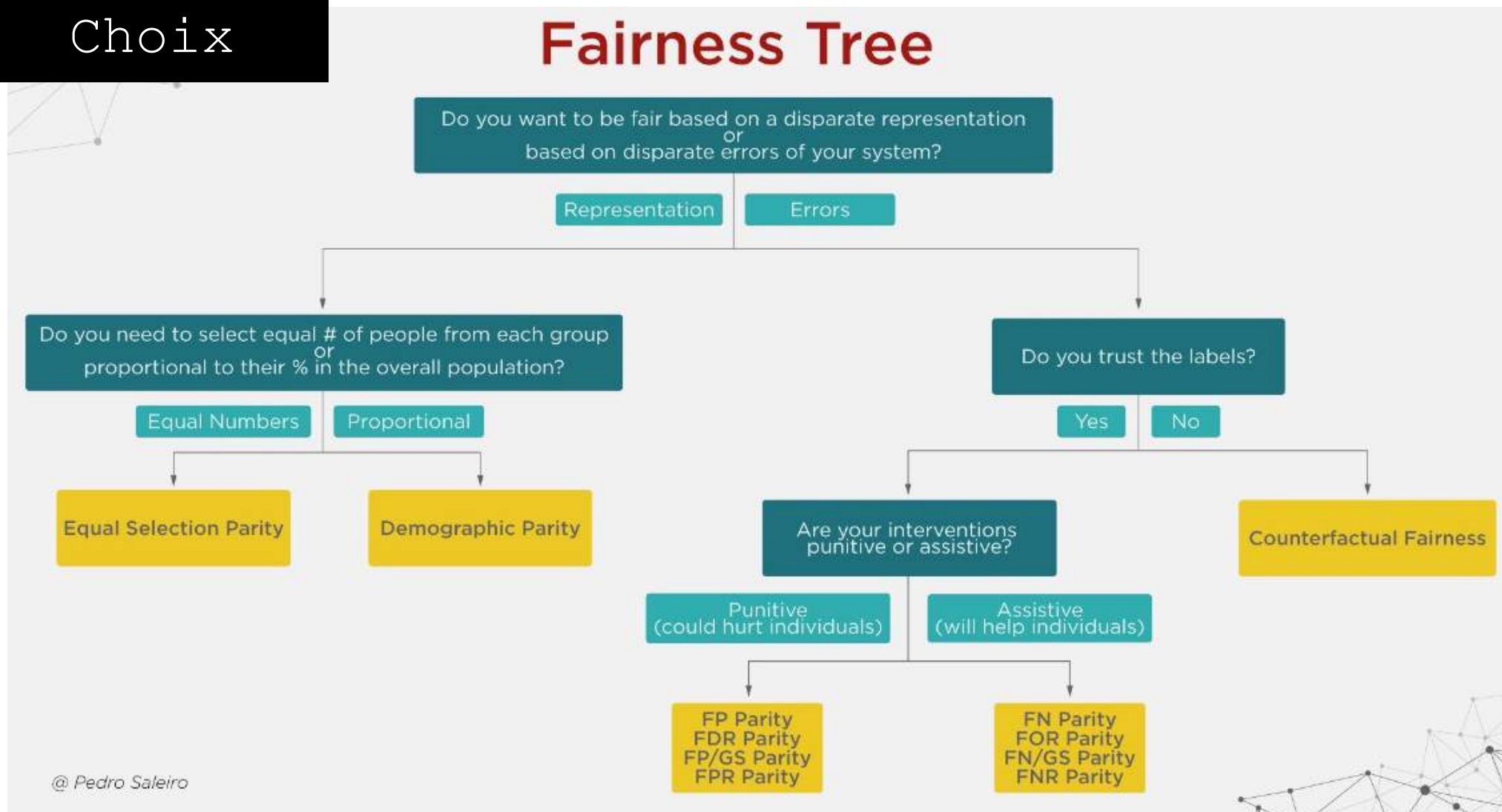
Métriques

- Parité démographique (*demographic parity, statistical parity, disparate impact*)
 - Indépendance de l'attribut sensible
 - Même proportion de chaque sous-groupe classifié comme positif
 - Classification seulement
- Opportunité égale (*equal opportunity*) :
 - Autant de prédictions positives
 - Égalité des taux de vrais positifs entre les sous-groupes
- Exactitude égale (*equal accuracy, equality of odds*) :
 - Autant de bonnes prédictions (positives et négative) pour chaque sous-groupe
 - Égalité de l'exactitude pour tous les sous-groupes
 - Classification seulement, possible d'étendre au cas quantitatif

(« AI Fairness – An Honest Introduction », 2021)

ÉQUITABILITÉ

Choix

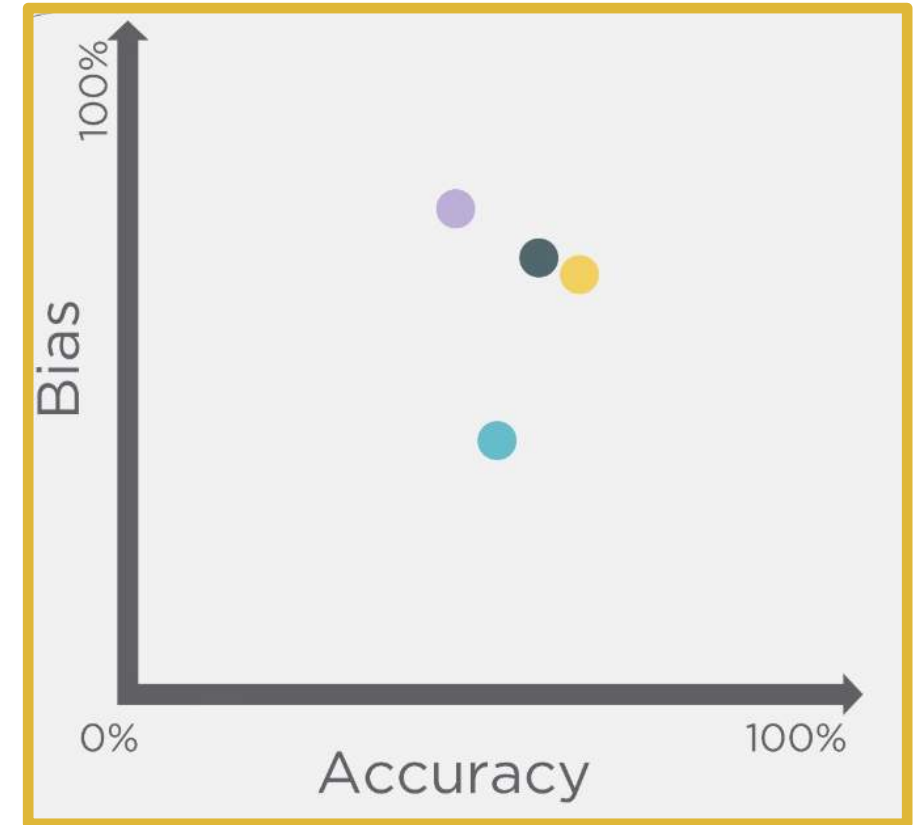


Pedro SALERO, *Bias and discrimination* (formation IVADO, module 2, *Fairness Metrics and Goals*), 2021

ÉQUITABILITÉ

Évaluation d'un modèle

- Définition
 - Groupes protégés
 - Ex: age >70
 - Métriques
 - Ex: FPR
 - Critère d'équité
 - Ex: refusé si échec d'au moins une métrique sur un groupe protégé
- Meilleur modèle
 - Pour chaque métrique d'intérêt
 - Pour chaque groupe protégé
 - Selon chaque métrique d'intérêt
 - Meilleure performance



Pedro SALERO, *Bias and discrimination* (formation IVADO, module 2, *Fairness Metrics and Goals*), 2021

PLAN DE LA PRÉSENTATION

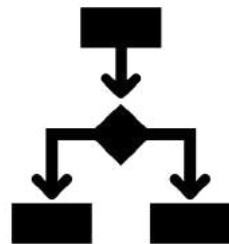
- Problème
- Définitions
- Outils
 - Déclaration MTL
 - Guide pratique
- Équitabilité (*fairness*)
- Interprétabilité et explicabilité
- Conclusion



INTERPRÉTABILITÉ

Interprétabilité

- Modèle
 - Compréhensible pour les humains
 - Transparent
 - . . . Intrinsèquement (par design)
- Raisons de la prédiction
- Domaine ancien (années 1950)
 - Arbres de décision, SCM



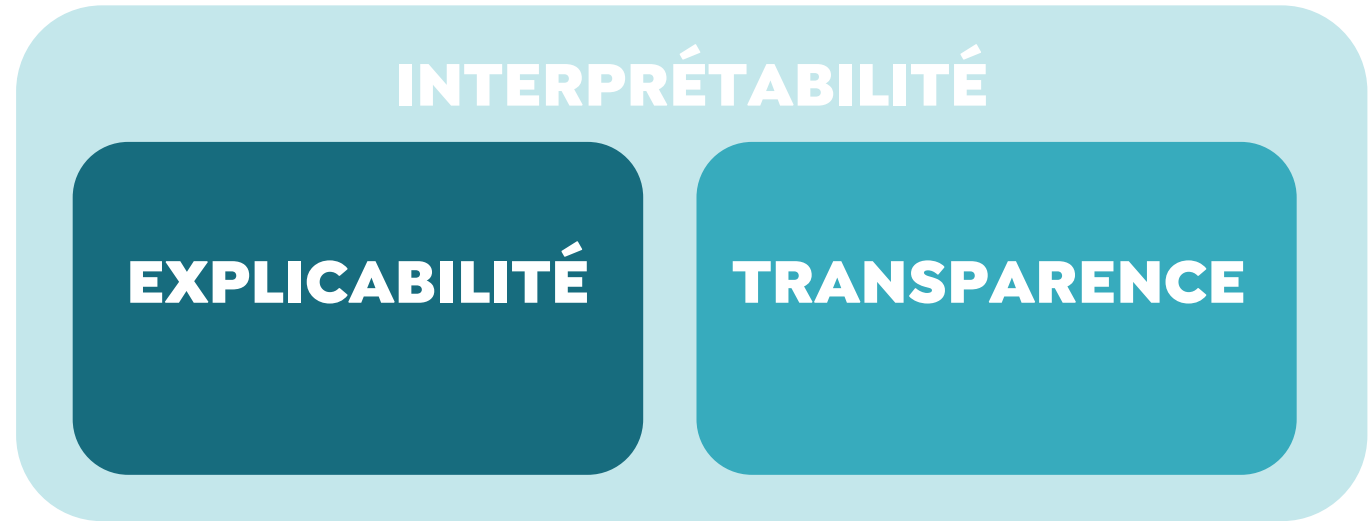
Explicabilité

- Expliquer une boîte noire en utilisant
 - un modèle d'approximation
 - dérivées, mesures d'importance des variables (ou autres statistiques)
 - explication *post hoc*
- Mécanisme de la prédiction
- Domaine récent
 - Réseaux de neurones
- Méthodes courantes
 - LIME
 - Valeurs de Shapley
 - Cartes de protubérance (*saliency maps*)
 - Gradients intégrés (*integrated gradients*)



INTERPRÉTABILITÉ

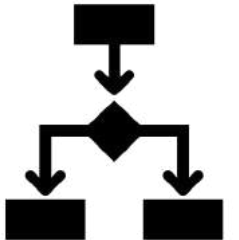
Distinction



(Rudin, 2021)

INTERPRÉTABILITÉ

Algorithmes



Interpretability

Neural Networks

Random Forest

Support Vector Machine

Graphical Models

k-Nearest Neighbors

Linear Regression

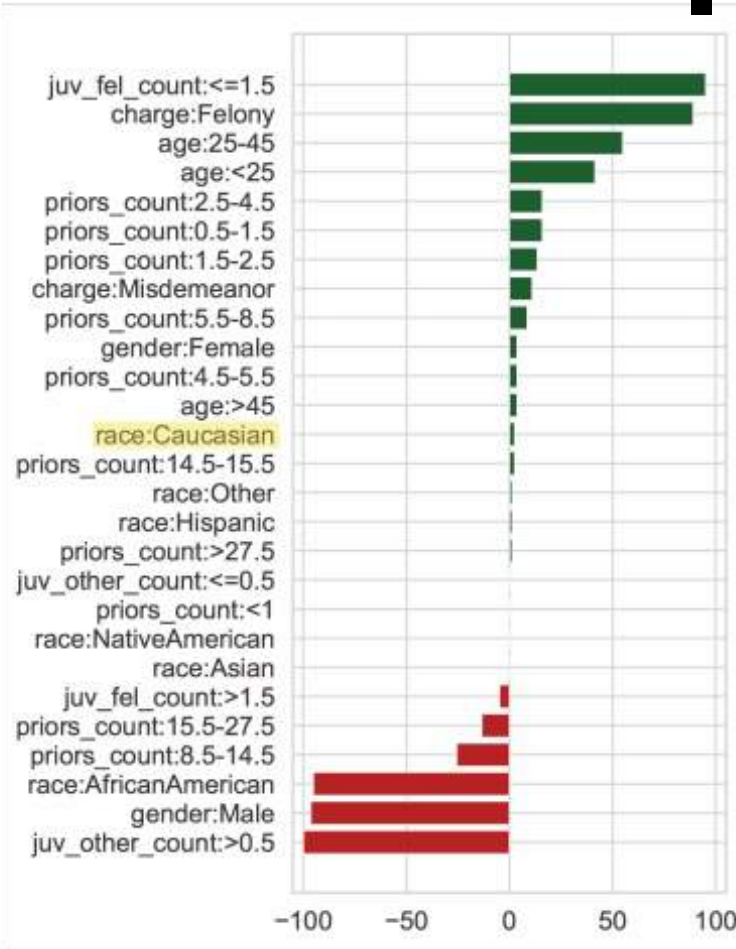
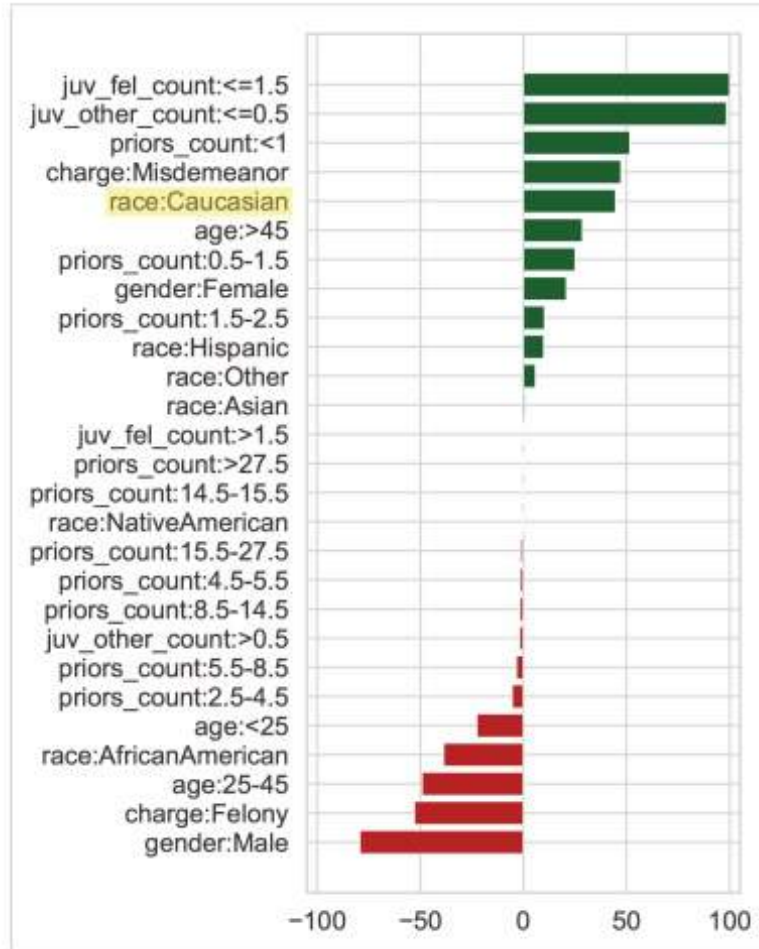
Decision Trees

Classification Rules

(Inspiré de Morocho-Cayamcela et al., 2019)

INTERPRÉTABILITÉ

Blanchiment éthique



COMPAS

- Explication (gauche)
 - *FairML* (Adebayo et al, 2012)
- Approximation (droite)
 - *LaundryML* (Aïvodji, et al, 2012)

if prior_count: 15.5–27.5 **then**

 recidivate: True

else if prior_count: 8.5–14.5 **then**

 recidivate: True

else if age: >45 **then**

 recidivate: False

else if juv_other_count: >0.5 **then**

 recidivate: True

else

 recidivate: False

end if

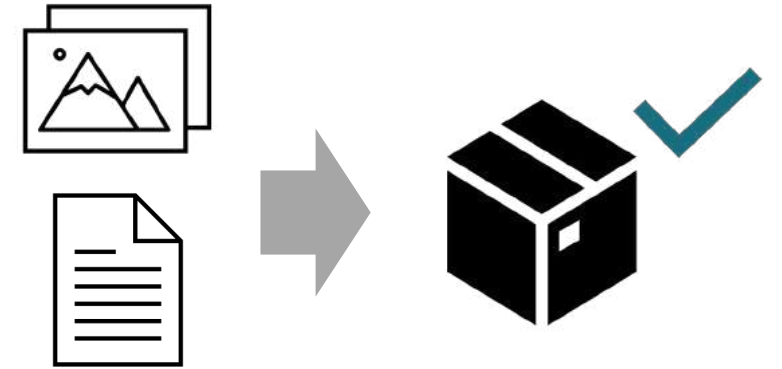
(Aïvodji, et al., 2021)

INTERPRÉTABILITÉ

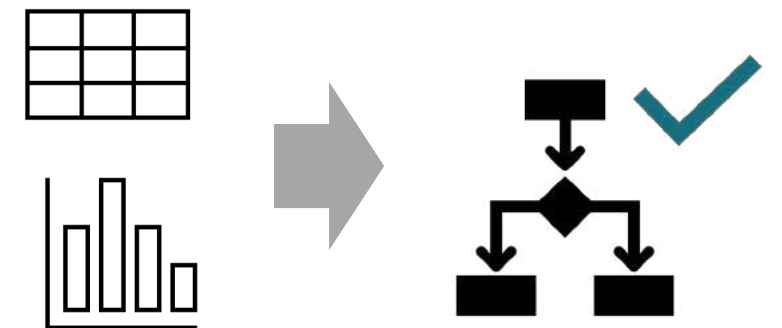
Compromis

- Compromis interprétabilité et exactitude ✖
 - Fausse dichotomie
 - Aucune preuve scientifique
- Interprétabilité mènerait même à une meilleure exactitude
 - Utile pour déboguer/améliorer (*troubleshooting*)
- Compromis parcimonie et exactitude ✔
 - Parcimonie \rightarrow interprétabilité
 - Association forte
 - Une composante parmi d'autres

Données brutes (images, texte)



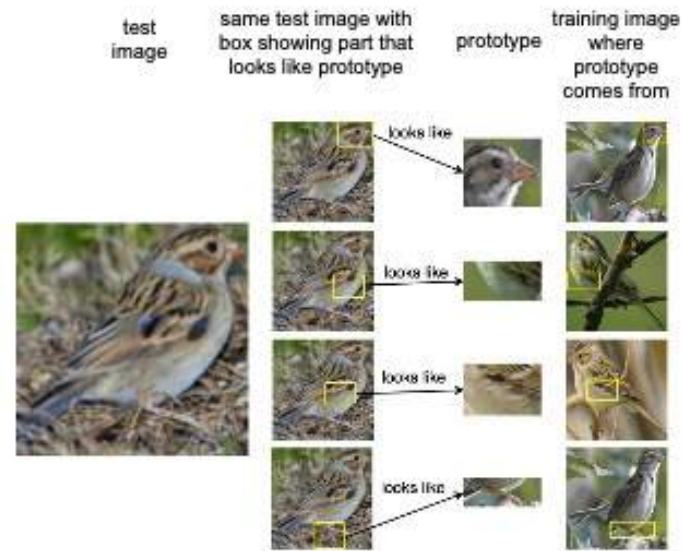
Données tabulaires



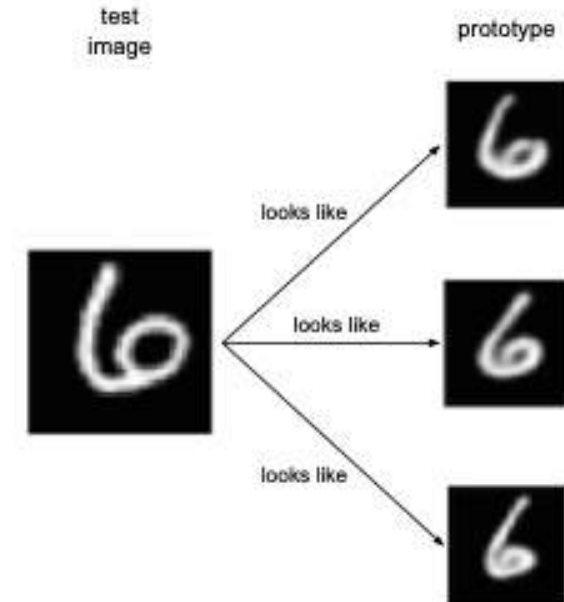
INTERPRÉTABILITÉ

Réseaux de neurones

- Proche du raisonnement humain
 - Cas de base
 - Prototypes



Ex. Cas de base dans un réseau de neurones

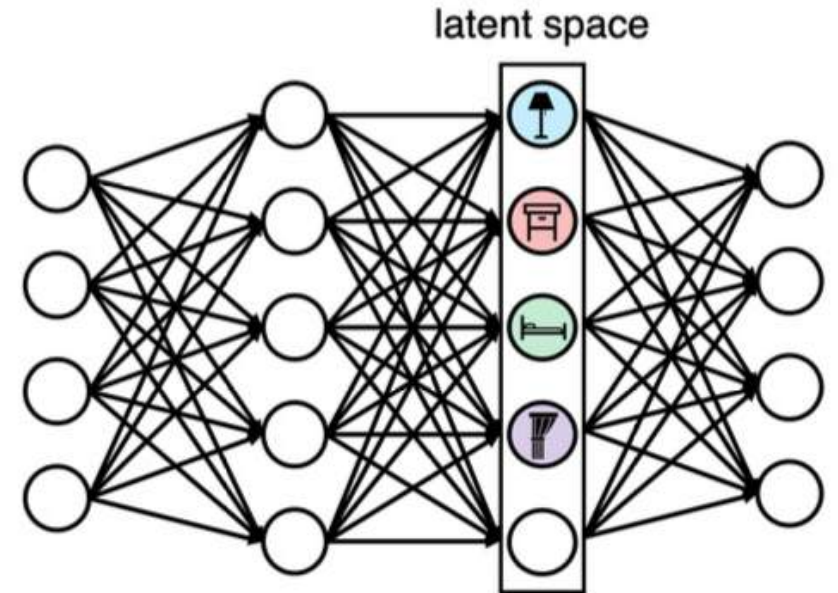


Ex. Prototype dans un réseau de neurones

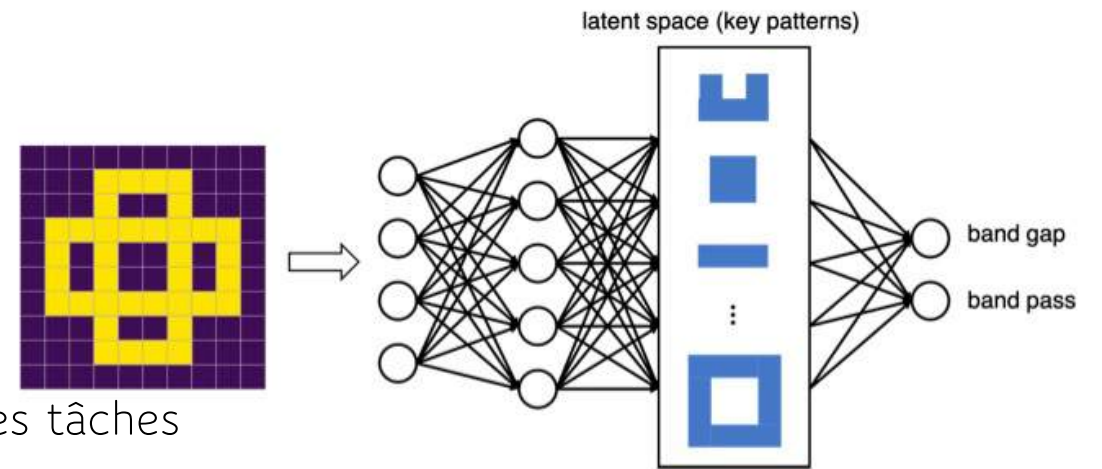
INTERPRÉTABILITÉ

Réseaux de neurones

- Désenchevêtrement (*disentanglement*)
 - Manière dont l'information voyage dans un réseau de neurones
 - Séparer l'information selon les concepts
 - Chaque neurone représente un concept humainement interprétable
- Supervisé
 - Les spécialistes spécifient les concepts
- Non supervisé
 - L'algorithme choisit les concepts d'intérêt
 - Biais dans les images étiquetées
 - Entités étiquetées sont spécifiques à certaines tâches
 - Ignore information pertinente



Ex. Désenchevêtrement supervisé de l'espace latent d'un réseau de neurones



Ex. Désenchevêtrement non supervisé de l'espace latent d'un réseau de neurones

PLAN DE LA PRÉSENTATION

- Problème
- Définitions
- Outils
 - Déclaration MTL
 - Guide pratique
- Équitabilité (*fairness*)
- Interprétabilité et explicabilité
- Conclusion



CONCLUSION

En résumé

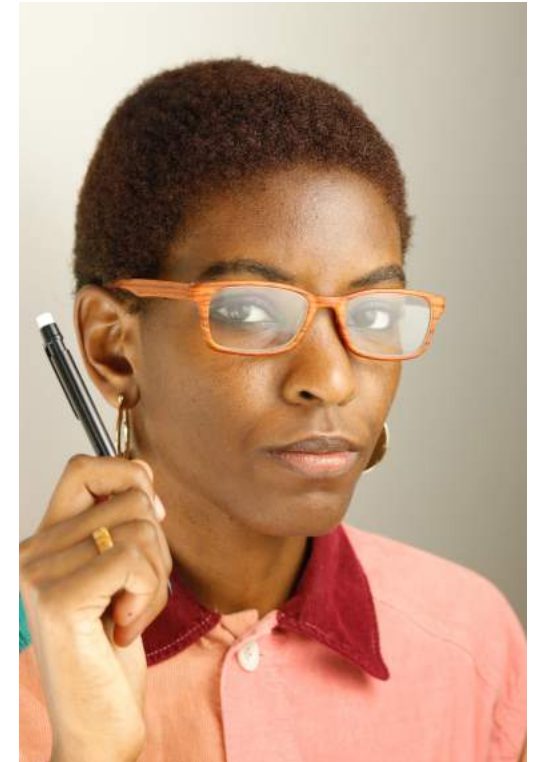
- Biais en IA
 - Inégalités
 - Discrimination
- Données
 - Groupes sous-représentés
 - Traduction biais sociétal
- Programmation
 - Choix des attributs
- Équitabilité
 - Définitions variables
 - Plusieurs métriques
- Interprétabilité
 - Réponse de l'IA contre les biais (discriminatoires)
 - Réfère à la transparence d'un modèle en soi
- Explicabilité
 - Recherche à expliquer comment une prédiction a été effectuée
- **Explication**
 - Bientôt obligatoire
 - Projet de loi 64
 - *GDPR*



CONCLUSION

Pour aller plus loin

- Cours
 - PHI-4142: Enjeux philosophiques et éthiques de l'intelligence artificielle (IA)
- Initiatives et tendances
 - Décolonisation
 - <https://manyfesto.ai/>
 - *Decolonizing data*
 - *AI for Good*
 - *Data science for good*
- Références paritaires
 - Extension [*Citation Transparency*](#)
 - Code
 - https://github.com/mb3152/balanced_citer
 - <https://github.com/dalejn/cleanBib>
 - Présentation: <https://youtu.be/WN6moTxEMNc?t=1733>





MERCI

sandrine.blais-deschenes.1@ulaval.ca

Présentation disponible au : <https://github.com/SandrineBD/BootcampIID/>

RÉFÉRENCES



Images

- Aïvodji, U., Arai, H., Gambs, S., & Hara, S. (2021). Characterizing the risk of fairwashing. *ArXiv*.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI) : Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv:1910.10045 [cs]*. <http://arxiv.org/abs/1910.10045>
- Aylwin, F.-A. (2020, novembre 22). *Projet de loi 64 : Quel est l'impact concret du projet de loi sur les ordres professionnels?* Lexology. <https://www.lexology.com/library/detail.aspx?q=652cc798-0828-4023-8781-0eb218e1aee6>
- Buolamwini, J., & Gebru, T. (2018). Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77-91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Morocho-Cayamcela, M. E., Lee, H., & Lim, W. (2019). Machine Learning for 5G/B5G Mobile and Wireless Communications : Potential, Limitations, and Future Directions. *IEEE Access*, 7, 137184-137206. <https://doi.org/10.1109/ACCESS.2019.2942390>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable Machine Learning : Fundamental Principles and 10 Grand Challenges. *arXiv:2103.11251 [cs, stat]*. <http://arxiv.org/abs/2103.11251>
- https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/ia_annexe_1_21032017.pdf
- <https://unsplash.com/>
- <https://genderphotos.vice.com/#Work>
- <https://www.pinterest.com/pin/315603886372352399/>

RÉFÉRENCES

- AI Fairness – An honest introduction. (2021, mars 3). *Capgemini Worldwide*. <https://www.capgemini.com/2021/03/ai-fairness-an-honest-introduction/>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots : Can Language Models Be Too Big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. <https://doi.org/10.1145/3442188.3445922>
- Bertrand, M., & Mullainathan, S. (2003). *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination* (Working Paper N° 9873; Working Paper Series). National Bureau of Economic Research. <https://doi.org/10.3386/w9873>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. <http://arxiv.org/abs/2005.14165>
- Ge, Y., Tan, J., Zhu, Y., Xia, Y., Luo, J., Liu, S., Fu, Z., Geng, S., Li, Z., & Zhang, Y. (2022). Explainable Fairness in Recommendation. *arXiv:2204.11159 [cs]*. <https://doi.org/10.1145/3477495.3531973>
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*. <http://arxiv.org/abs/1609.05807>
- Rudin, C., & Carlson, D. (2019). The Secrets of Machine Learning : Ten Things You Wish You Had Known Earlier to be More Effective at Data Analysis. *arXiv:1906.01998 [cs, stat]*. <http://arxiv.org/abs/1906.01998>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68. <https://doi.org/10.1145/3287560.3287598>
- Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1-7. <https://doi.org/10.23919/FAIRWARE.2018.8452913>

