L’Intelligence Artificielle pour tous

Rachel Orti, IBM
Mélanie Shilpa Rao, IBM

Rachel Orti

Tech Lead/Dev
IBM Cloud and Cognitive Software
@rachel_orti

Mélanie Shilpa Rao

Cloud Engineer
IBM Cloud and Cognitive Software
@msr4cloud

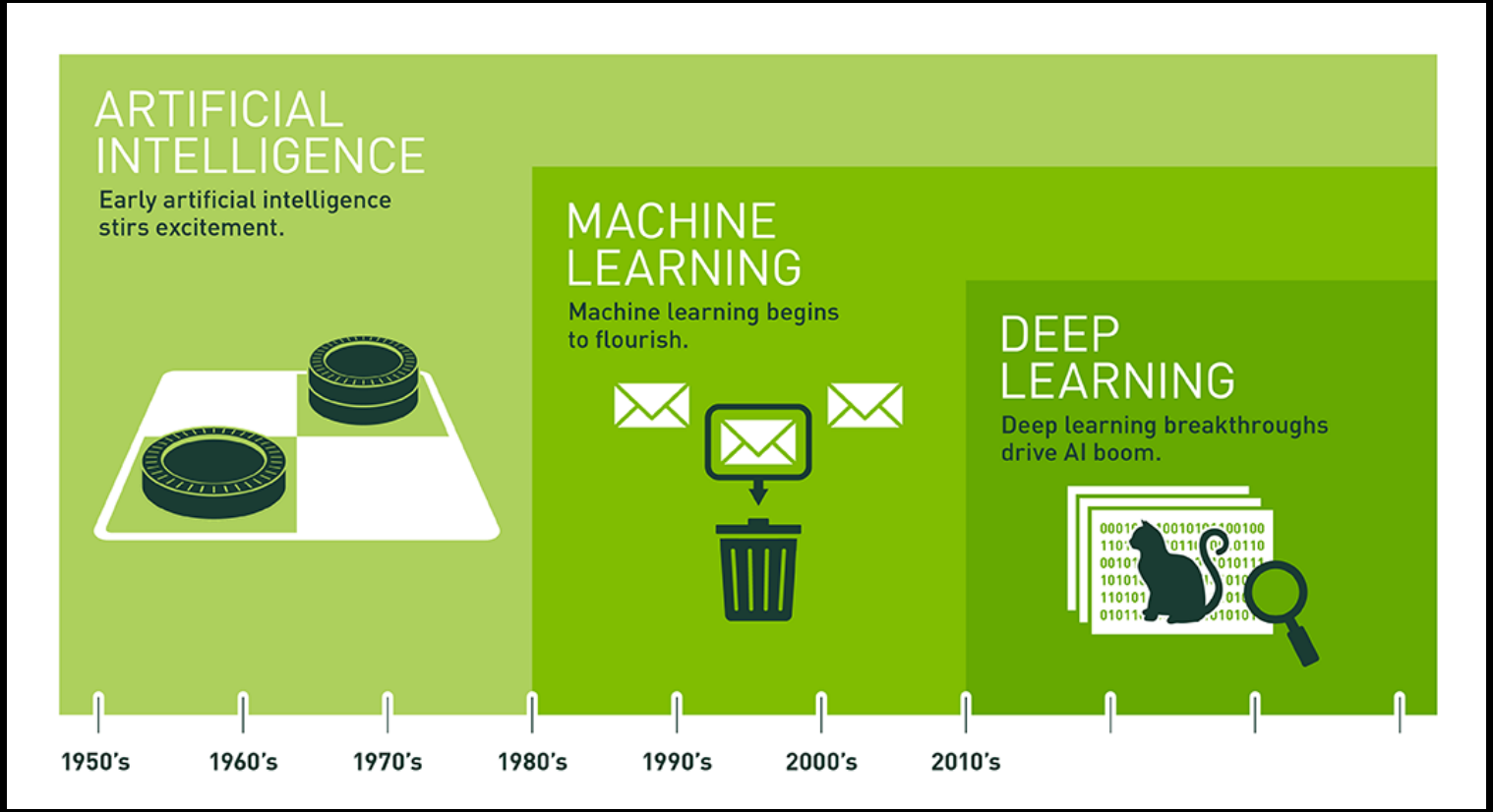
L’Intelligence Artificielle pour tous

L’Intelligence Artificielle pour tous

L’Intelligence Artificielle pour tous

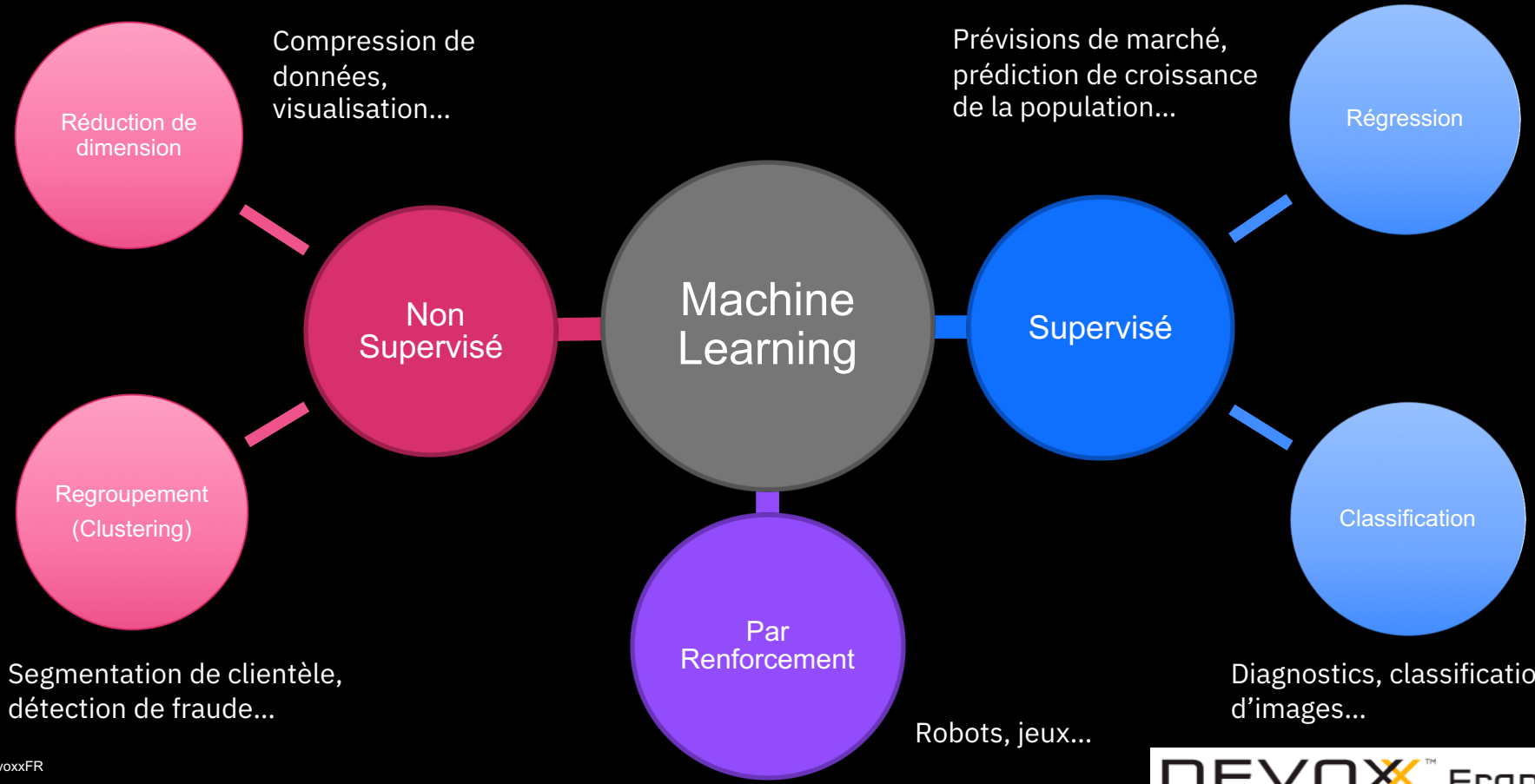
Une introduction pour tous

Un peu d'Histoire...



https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Les principaux types d'apprentissage



La démocratisation de l’IA

Plein de librairies open source...

Projet Norman

mit media lab

Test de Rorschach

Projet Norman

AI trained on COCO dataset sees:

“A couple of people standing next to each other.”

Projet Norman

AI trained on COCO dataset sees:

"A couple of people standing next to each other."

AI trained on subreddits about death, Norman, sees:

"Man jumps from floor window."

L'IA dans le recrutement : un futur annoncé comme prometteur

launchpadrecruits.com

HR Professionals are Using A.I. to Uproot Unconscious Bias

Written by Kirstie Kelly | Jun 14, 2016 7:30:00 AM

CIO US

FEATURE

How artificial intelligence can eliminate bias in hiring

AI and machine learning can help identify diverse candidates, improve the hiring pipeline and eliminate unconscious bias.

By Sharon Florentine

Senior Writer, CIO

DECEMBER 22, 2016 02:00 AM PT

How Companies Are Taking Unconscious Bias Out of the Hiring Equation

Maxwell Huppert February 21, 2017

Share Tweet Like 725

L'IA dans le recrutement : un présent mitigé

CNBC MENU

Silicon Valley is stumped: Even A.I. cannot always remove bias from hiring

Eric Rosenbaum | @erprose

Published 9:43 AM ET Wed, 30 May 2018

Updated 5:20 PM ET Wed, 30 May 2018

CNBC.com

QUARTZ

ROBOT INDEMNITY

Companies are on the hook if their hiring algorithms are biased

By Dave Gershgorn • October 22, 2018

REUTERS/KEVIN LAMARQUE

ReadWrite

AI Should be Reducing Bias, Not Introducing it in Recruiting

Lin Classon 2 months ago

La justice prédictive : COMPAS

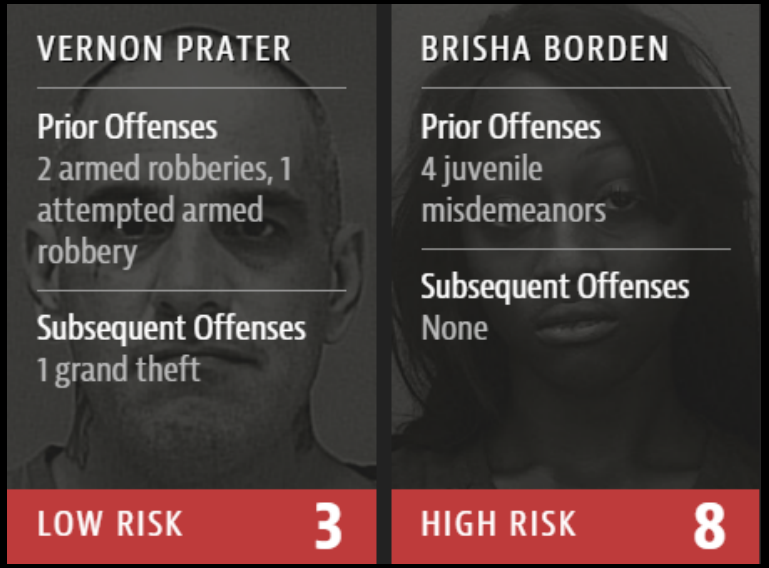
https://www.nature.com/articles/d41586-018-05469-3

La justice prédictive : COMPAS

PRO PUBLICA

Machine Bias

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



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

Comment développer des IAs sans biais ?

Comment développer des IAs sans biais ?

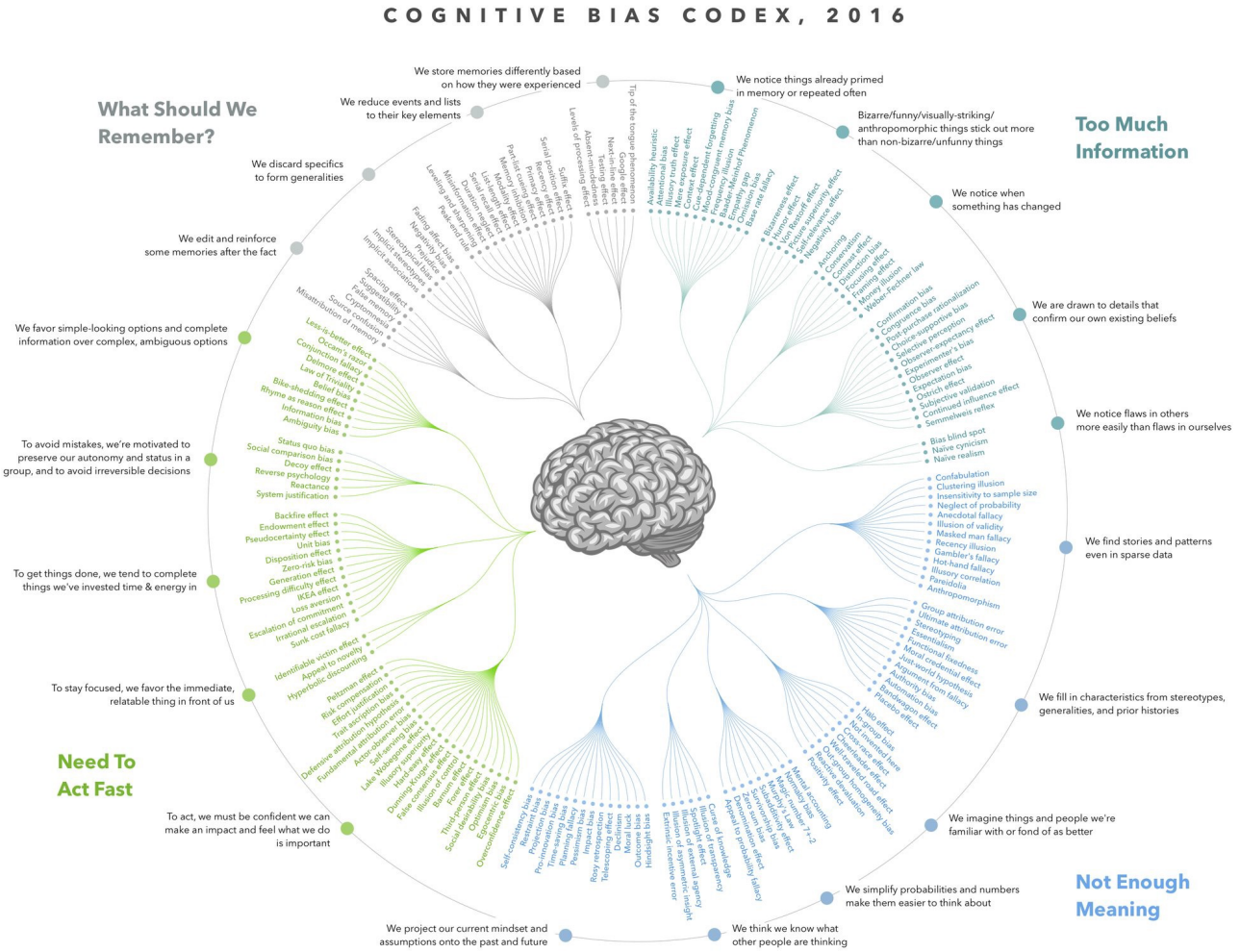
Le facteur humain

C’est quoi le biais humain ?

= Une distorsion que subit une information en entrant dans le système cognitif ou en sortant

C’est quoi le biais humain ?

= Une distorsion que subit une information en entrant dans le système cognitif ou en sortant



Débiaiser les humains : des mouvements pour tous

AINOW

FAIR BY DESIGN
Ending the extra costs of being poor

ALGORITHM WATCH

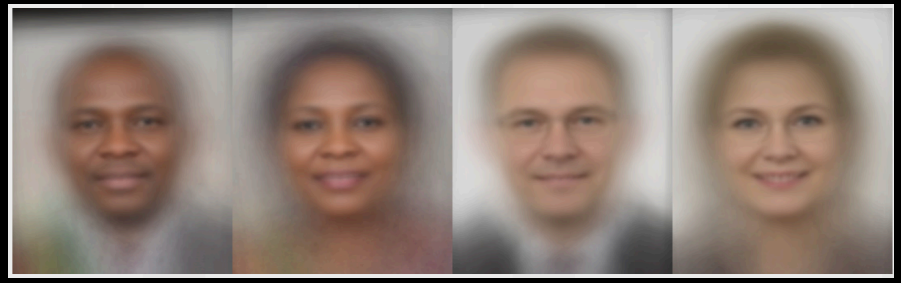
PARTNERSHIP ON AI

DATA FOR GOOD

ALGORITHMIC JUSTICE LEAGUE

La reconnaissance faciale : Gender Shades, l'étude

mit media lab

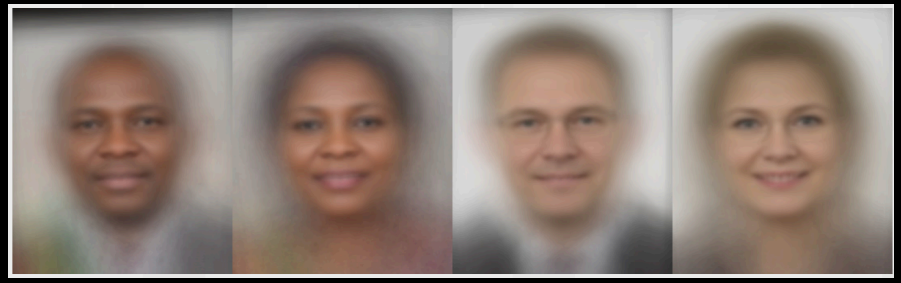


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

http://gendershades.org/

La reconnaissance faciale : Gender Shades, l'impact

mit media lab



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0% 99.67%	79.2% 98.48%	100% 100%	98.3% 99.66%	20.8% 1.52%
FACE++	99.3% 98.7%	65.5% 95.9%	99.2% 99.5%	94.0% 99%	33.8% 3.6%
IBM	88.0% 99.37%	65.3% 83.03%	99.7% 99.74%	92.9% 97.63%	34.4% 16.71%

https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf

Débiaiser les humains : des équipes soucieuses du sujet

"The Google People Operations team began building a workshop for employees - Unconscious Bias @ Work"

"At IBM, I've been researching and testing methods for integrating bias mitigation into traditional design thinking practices"

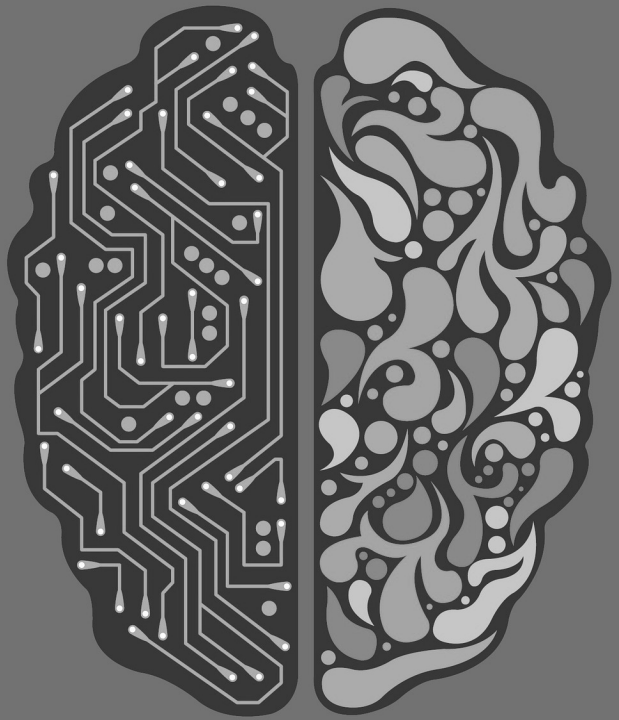
https://rework.withgoogle.com/guides/unbiasing-raise-awareness/steps/watch-unconscious-bias-at-work/
https://medium.com/design-ibm/the-origins-of-bias-and-how-ai-might-be-our-answer-to-ending-it-acc3610d6354

Débiaiser les humains : des équipes diverses

https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865

Comment développer des IAs sans biais ?

Les clés techniques



C’est quoi le biais algorithmique ?

= Une erreur faite systématiquement sur certains individus et qui les place à un désavantage systématique

Des critères protégés où les biais sont interdits

L'âge

Le sexe

La situation de famille

L'appartenance ou non à une nation

L'appartenance ou non à une ethnie

L'appartenance ou non à une religion déterminée

Le handicap

L'identité sexuelle

L'orientation sexuelle

La grossesse

...

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE LA JUSTICE

STOP-DISCRIMINATION.gouv.fr

RGPD : des critères protégés en plus

Article 22 paragraphe 4 — Interdit de prendre des décisions (sauf consentement ou motif d'intérêt public important) fondées sur :

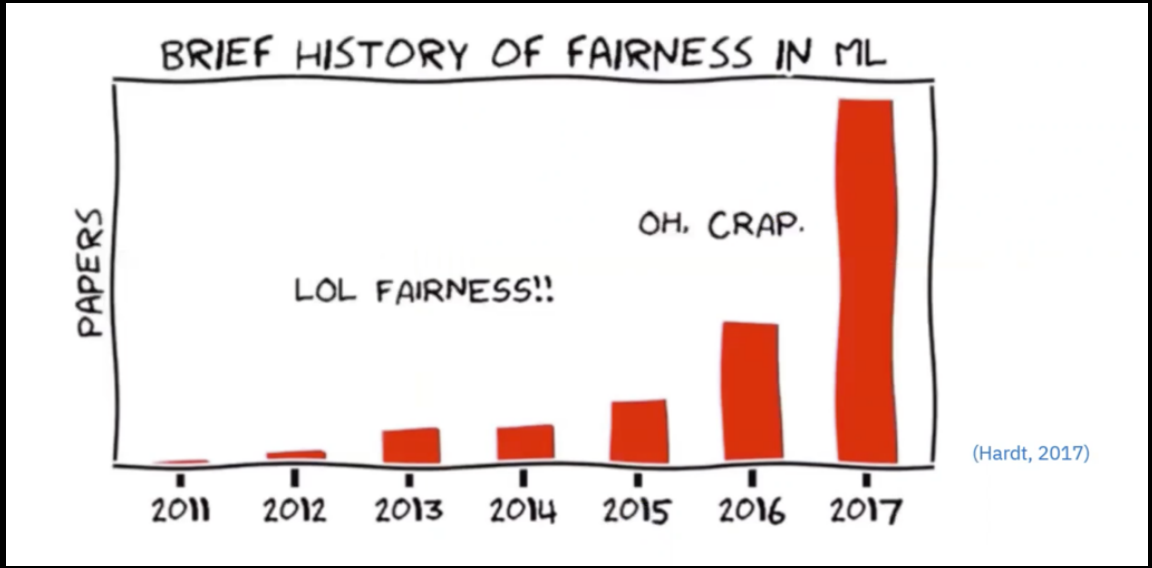
> Origine raciale ou ethnique, opinions politiques, convictions religieuses ou philosophiques ou appartenance syndicale, données génétiques, biométriques, sur la santé, sur la vie sexuelle ou l'orientation sexuelle *(article 9 paragraphe 1)*

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/

Equité de la procédure : l’approche ignorante

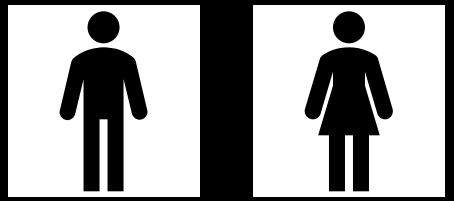
Ignorance
(Unawareness)

L’équité et les métriques d’équité : un sujet complexe



(Hardt, 2017)

Equité du résultat : l'approche informée



Equité individuelle
(Individual fairness)

Equité de groupe
(Group fairness)

La base des métriques d'équité : la matrice de confusion

		Prédiction	
		Oui	Non
Réalité	Oui	True Positive (TP)	False Negative (FN)
	Non	False Positive (FP)	True Negative (TN)

La base des métriques d'équité : la matrice de confusion

		Prédiction	
		Oui	Non
Réalité	Oui	True Positive (TP)	False Negative (FN)
	Non	False Positive (FP)	True Negative (TN)

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend (FP)	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend (FN)	47.7%	28.0%

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

Equité de groupe : quelques métriques

Noms	Définition	Critiques
Equal accuracy	Autant de prédictions correctes pour chaque groupe : TP + TN ≈	Intervention non-équitable sur les prédictions incorrectes

Equité de groupe : quelques métriques

Noms	Définition	Critiques
Equal accuracy	Autant de prédictions correctes pour chaque groupe : TP + TN ≈	Intervention non-équitable sur les prédictions incorrectes
Demographic parity = Statistical parity = Equal parity	Autant de prédictions favorables pour chaque groupe : - Favorable = positif : TP + FP ≈ - Favorable = négatif : TN + FN ≈	Intervention équitable mais ignore la notion de justesse

Equité de groupe : quelques métriques

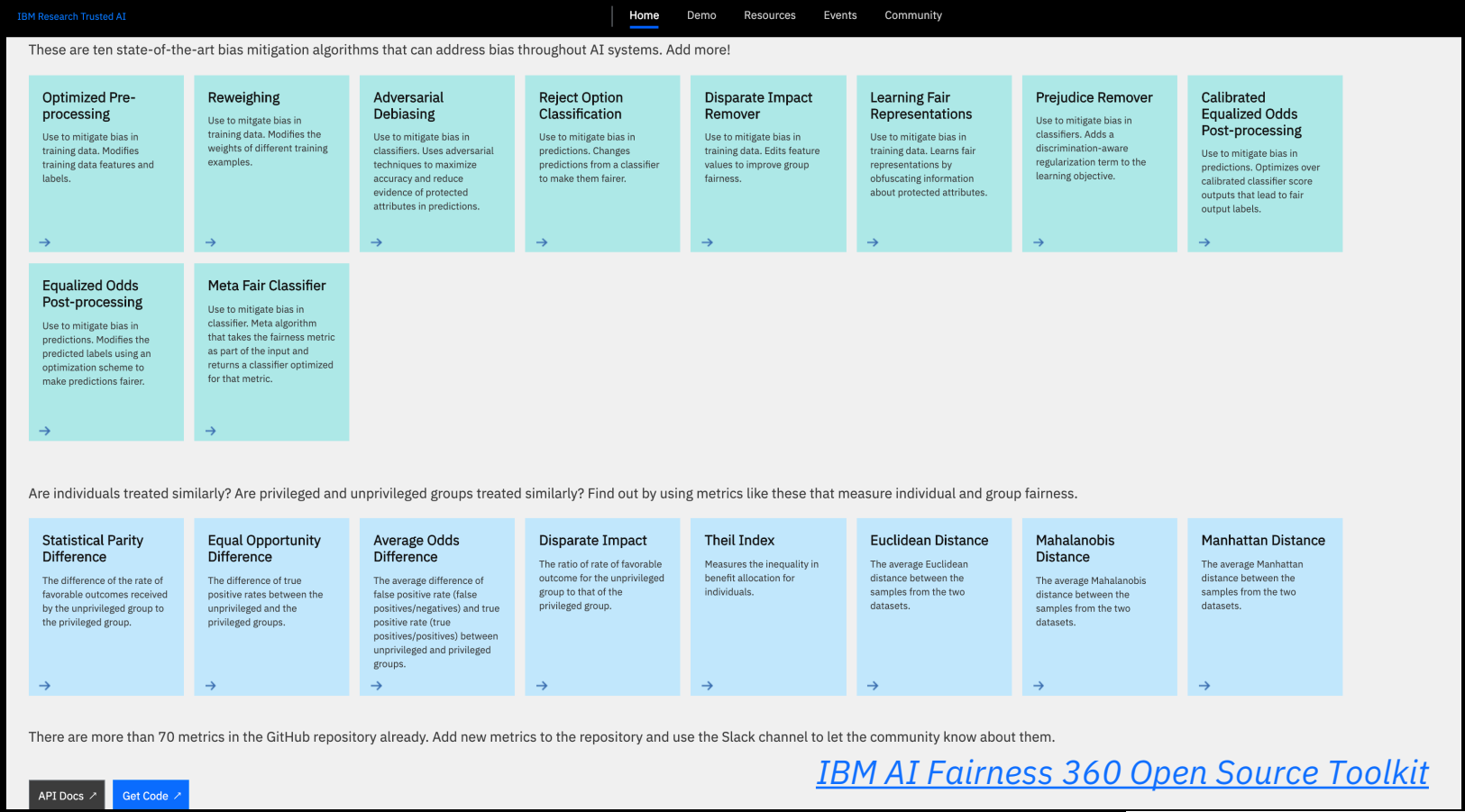
Noms	Définition	Critiques
Equal accuracy	Autant de prédictions correctes pour chaque groupe : TP + TN ≈	Intervention non-équitable sur les prédictions incorrectes
Demographic parity = Statistical parity = Equal parity	Autant de prédictions favorables pour chaque groupe : - Favorable = positif : TP + FP ≈ - Favorable = négatif : TN + FN ≈	Intervention équitable mais ignore la notion de justesse
Equal opportunity = Equalized odds = Independence = TP rate parity = FN rate parity	Autant de prédictions positives correctes pour chaque groupe : TP ≈ (et FN ≈)	Surtout utile pour les situations aidantes : important de ne pas se tromper négativement peu importe le groupe

Equité de groupe : quelques métriques

Noms	Définition	Critiques
Equal accuracy	Autant de prédictions correctes pour chaque groupe : TP + TN ≈	Intervention non-équitable sur les prédictions incorrectes
Demographic parity = Statistical parity = Equal parity	Autant de prédictions favorables pour chaque groupe : - Favorable = positif : TP + FP ≈ - Favorable = négatif : TN + FN ≈	Intervention équitable mais ignore la notion de justesse
Equal opportunity = Equalized odds = Independence = TP rate parity = FN rate parity	Autant de prédictions positives correctes pour chaque groupe : TP ≈ (et FN ≈)	Surtout utile pour les situations aidantes : important de ne pas se tromper négativement peu importe le groupe
FP Rate Parity = TN Rate Parity	Autant de prédictions positives incorrectes pour chaque groupe : FP ≈ (et TN ≈)	Surtout utile pour les situations punitives : important de ne pas se tromper positivement peu importe le groupe

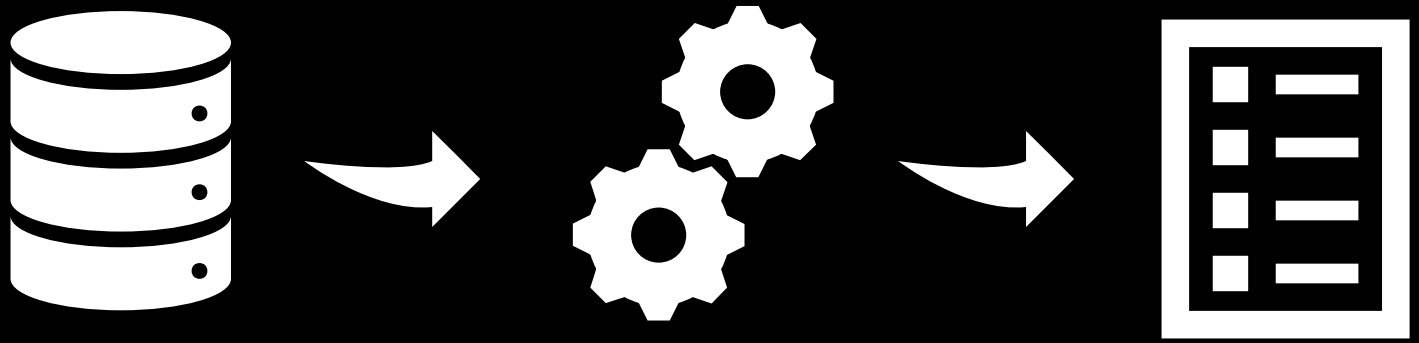
Débiaiser les IAs : agir sur les données “à la main”

Des algorithmes à la rescousse !

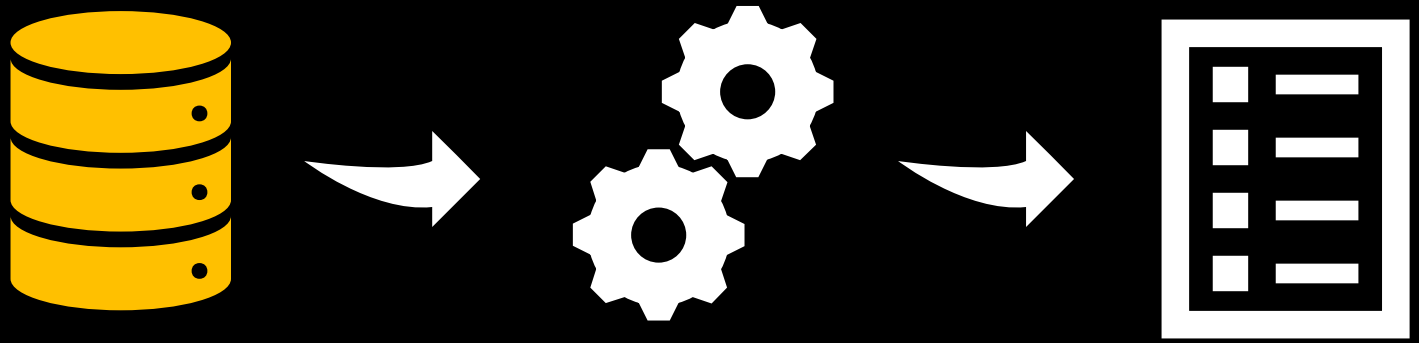


IBM AI Fairness 360 Open Source Toolkit

Débiaiser les IAs

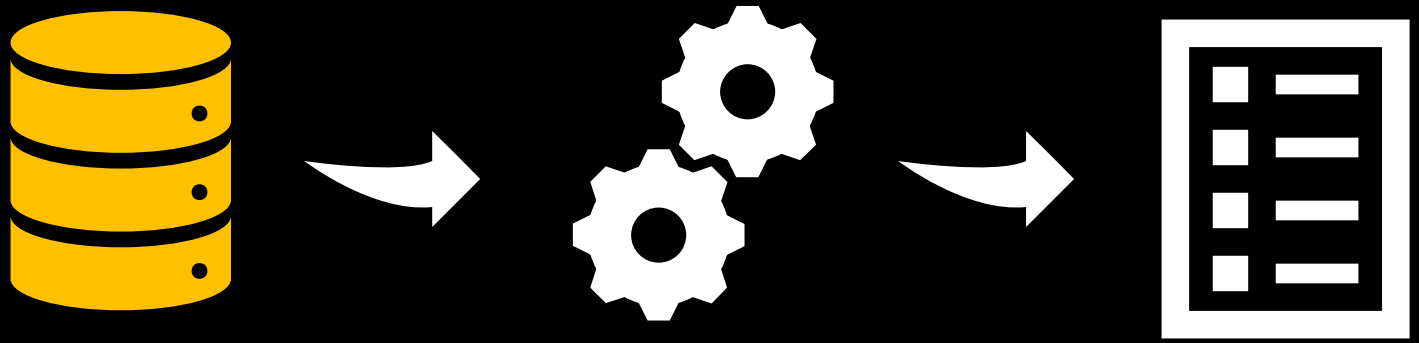


Débiaiser les IAs : agir sur les données



Optimized Pre-Processing

Débiaiser les IAs : agir sur les données



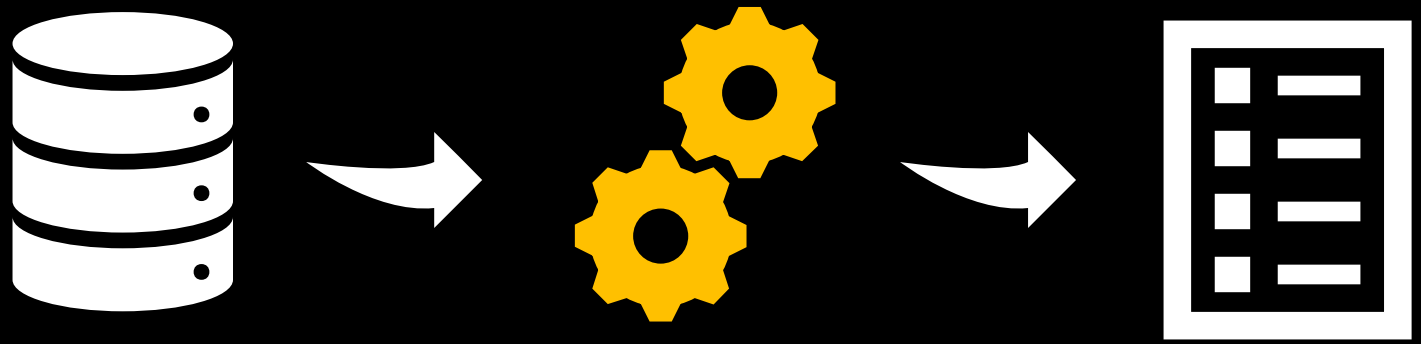
Optimized Pre-Processing
Reweighing

Débiaiser les IAs : agir sur le modèle

Optimized Pre-Processing
Reweighing

Prejudice Remover

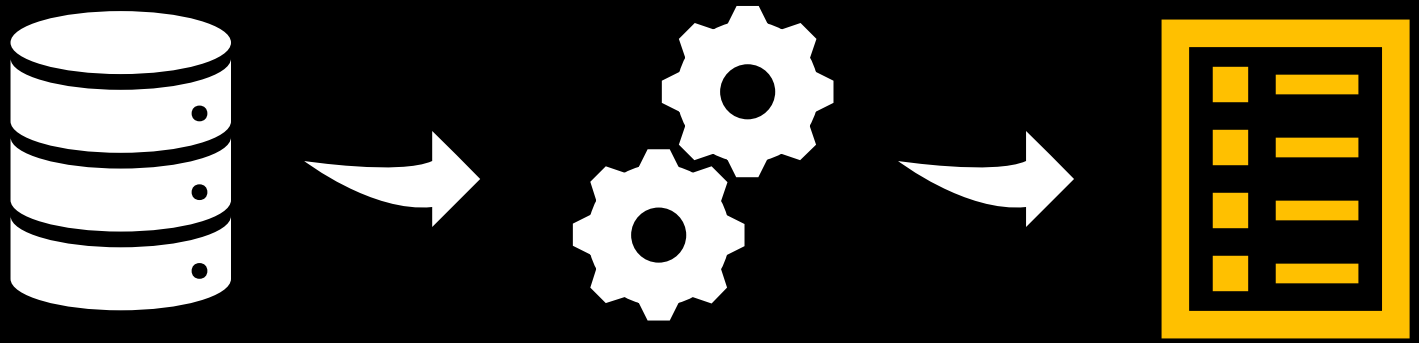
Débiaiser les IAs : agir sur le modèle



Optimized Pre-Processing
Reweighing

Prejudice Remover
Adversarial Debiasing

Débiaiser les IAs : agir sur les prédictions

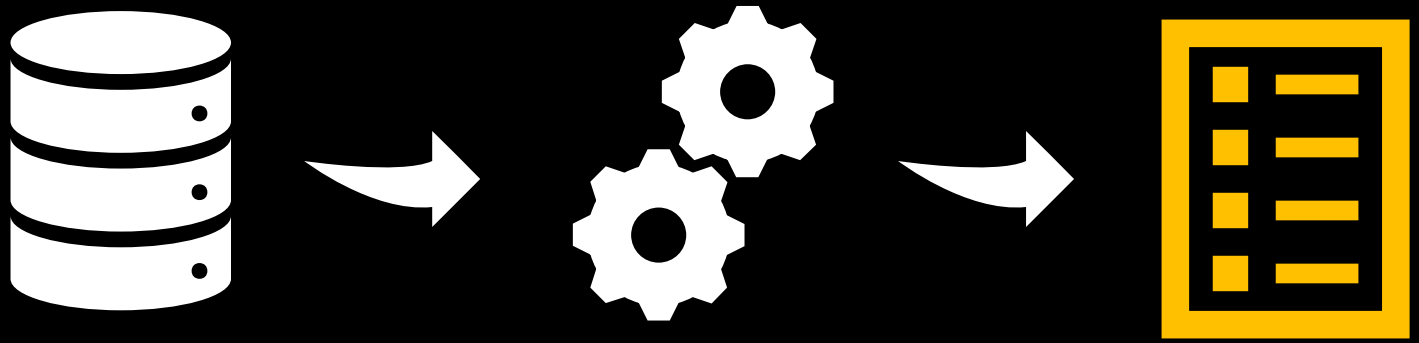


Optimized Pre-Processing
Reweighing

Prejudice Remover
Adversarial Debiasing

Equalized Odds Post-Processing

Débiaiser les IAs : agir sur les prédictions



Optimized Pre-Processing
Reweighing

Prejudice Remover
Adversarial Debiasing

Equalized Odds Post-Processing
Reject Option Classification

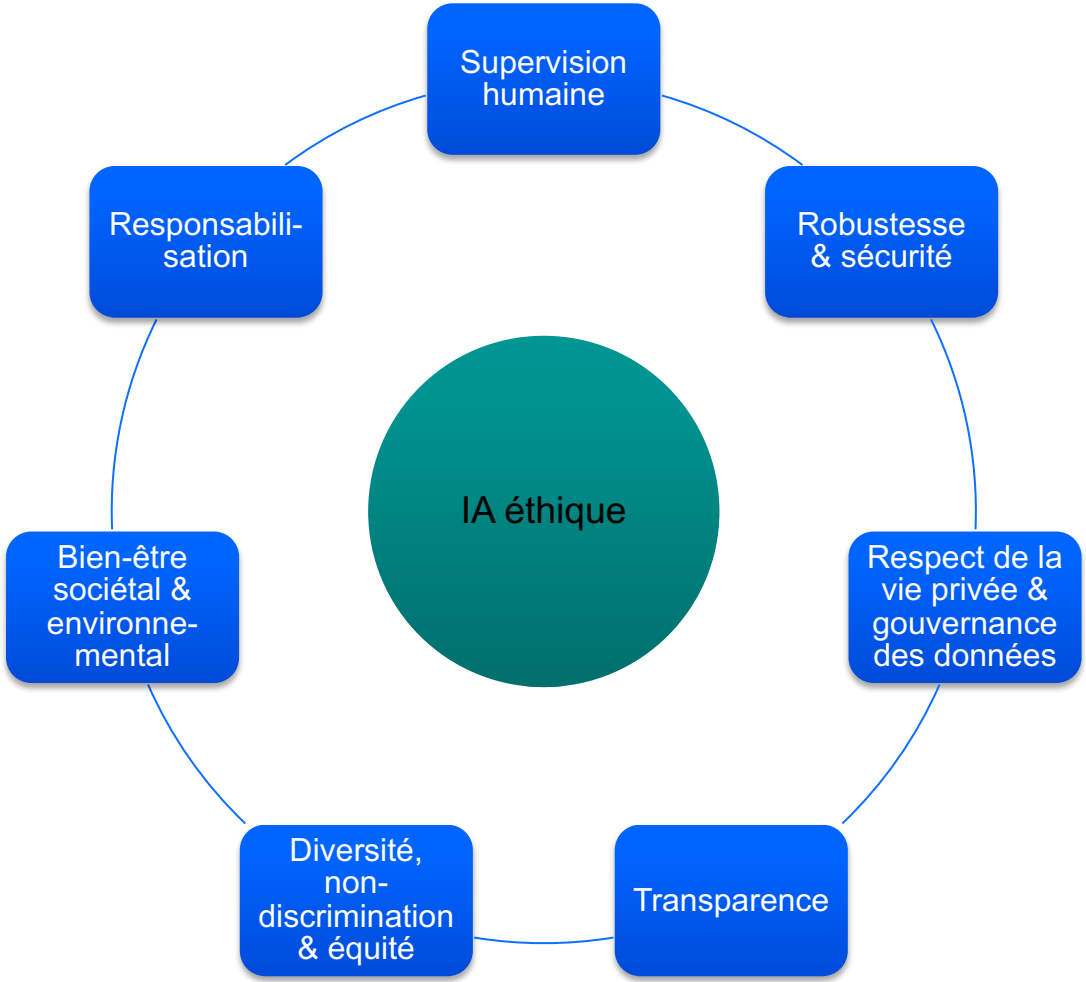
Quelques outils & frameworks actuels

Nom	Date de lancement	Outil/API	Disponi-bilité	Explicabilité	Estimation Equité	Atténuation de biais
Audit AI pymetrics	05/2018	API Python Notebooks Jupyter	Open source	Non	~10 métriques	Non
What-If Tool Google	09/2018	Outil graphique Intégration TensorBoard & notebooks Démo web Notebooks Jupyter	Open source	Oui : des données Non : de décisions	5 métriques	Variateur de seuil
AI Fairness 360 IBM	09/2018	API Python Démo web Notebooks Jupyter	Open source	Non	70+ métriques	10+ algorithmes
AI Openscale IBM	10/2018	API Python API REST Outil web	Offres gratuite/ payante	Oui	Oui : classification & régression structurées Non : classification texte/images	Oui : classification structurée Monitoring live
Teach & Test AI accenture	02/2018	?	Offre payante	Oui ?	Oui ?	?

Démo

https://github.com/rachel-orti/ai4all

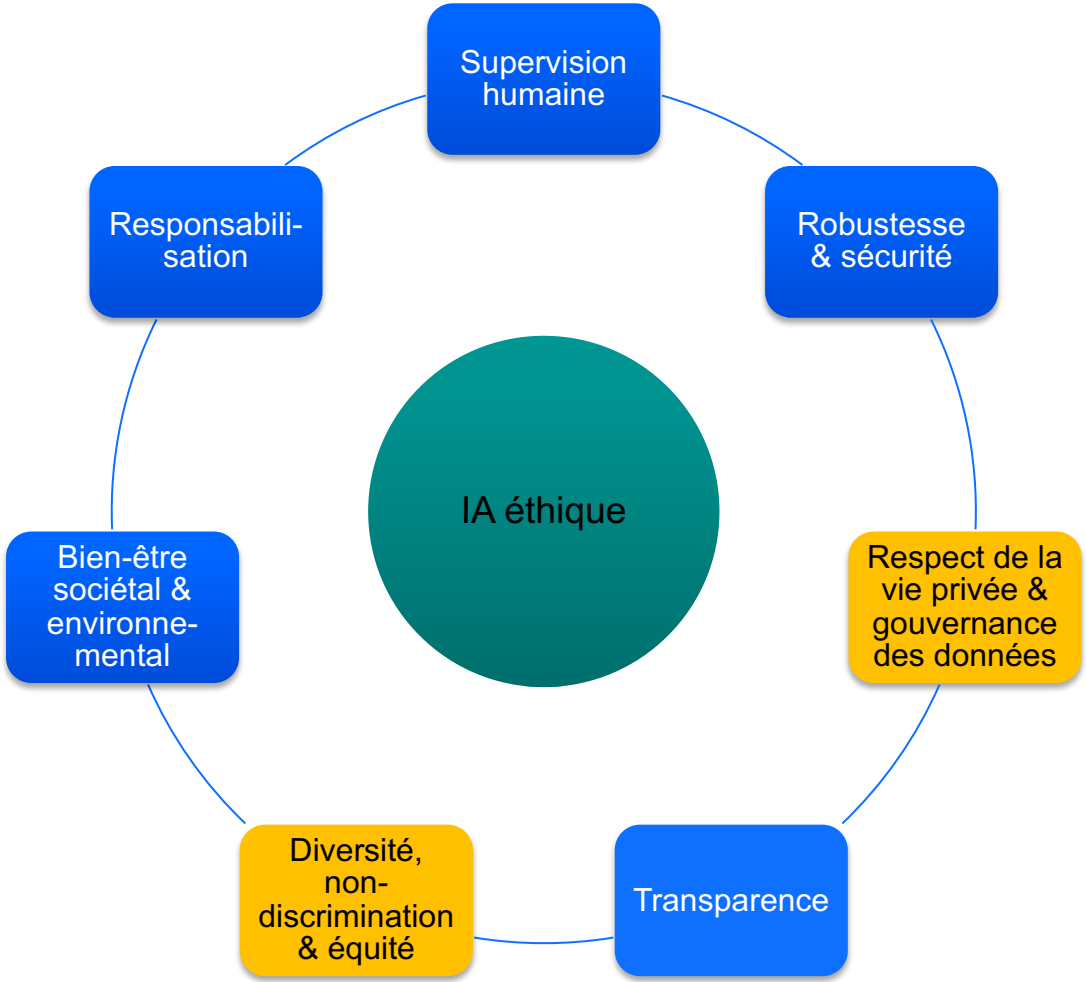
Conclusion



Pour une IA éthique

7

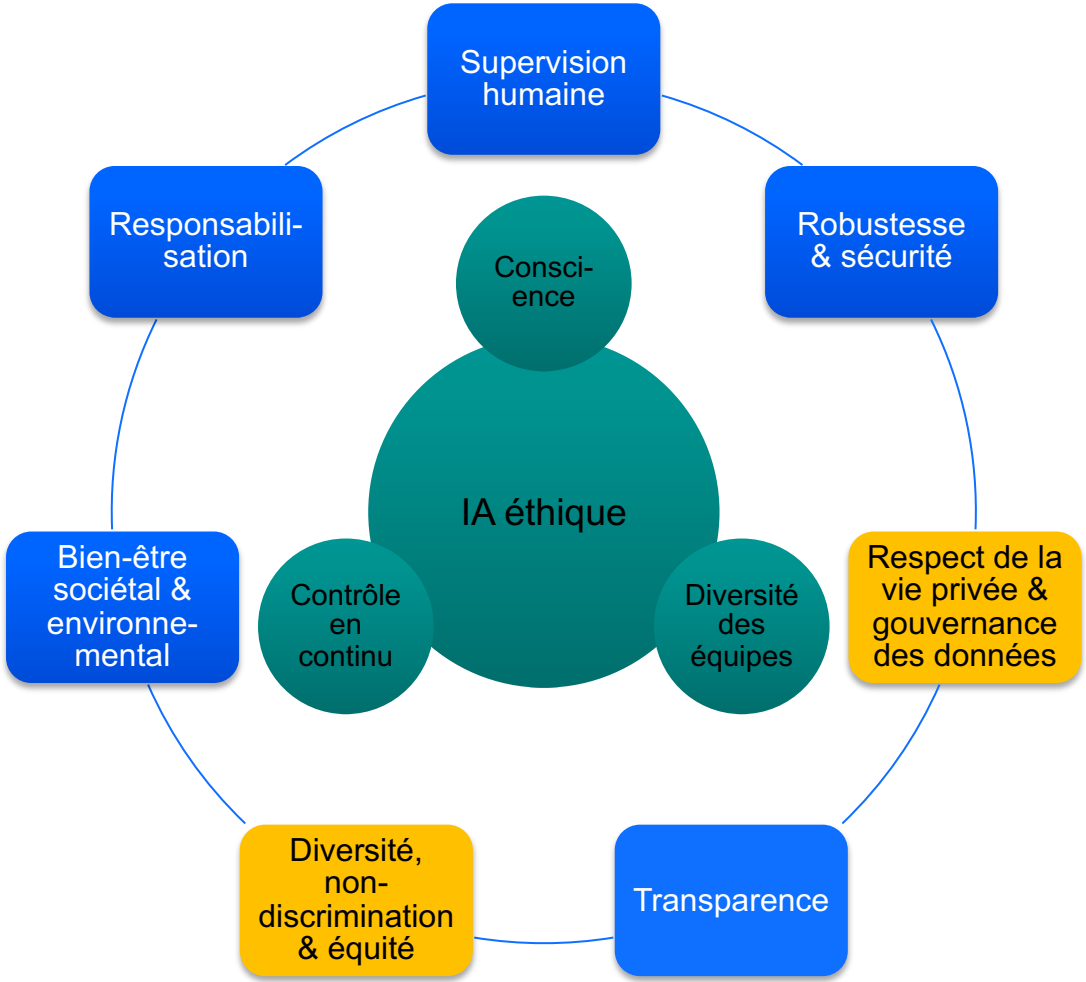
Principes
(selon la commission européenne)



Pour une IA éthique

7

Principes
(selon la commission européenne)



Pour une IA éthique

7

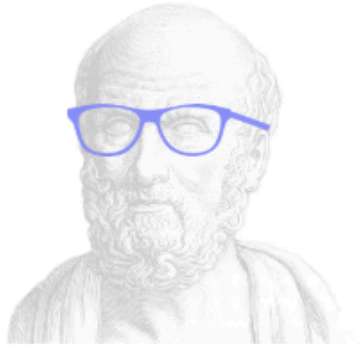
Principes
(selon la commission européenne)

3

Facteurs clés
pour tous

Data for Good propose un

Serment d'Hippocrate pour Data Scientist



… ou pour toute personne travaillant avec la donnée

En tant que **professionnel(le)** amené(e) à

collecter, stocker, traiter, modéliser, analyser des données et/ou à concevoir des algorithmes, des produits informatiques ou des interfaces,

je suis conscient(e) de **l'impact** que peut avoir mon travail sur des individus et sur la société dans son ensemble.

Merci !

https://github.com/rachel-orti/ai4all