

L'Intelligence Artificielle pour tous

Rachel Ortí, IBM

Mélanie Shilpa Rao, IBM



Rachel Ortí

Tech Lead/Dev
IBM Cloud and Cognitive Software
@rachel_ortí



Mélanie Shilpa Rao

Cloud Engineer
IBM Cloud and Cognitive Software
@msr4cloud



L'Intelligence Artificielle pour tous



L'Intelligence Artificielle pour tous



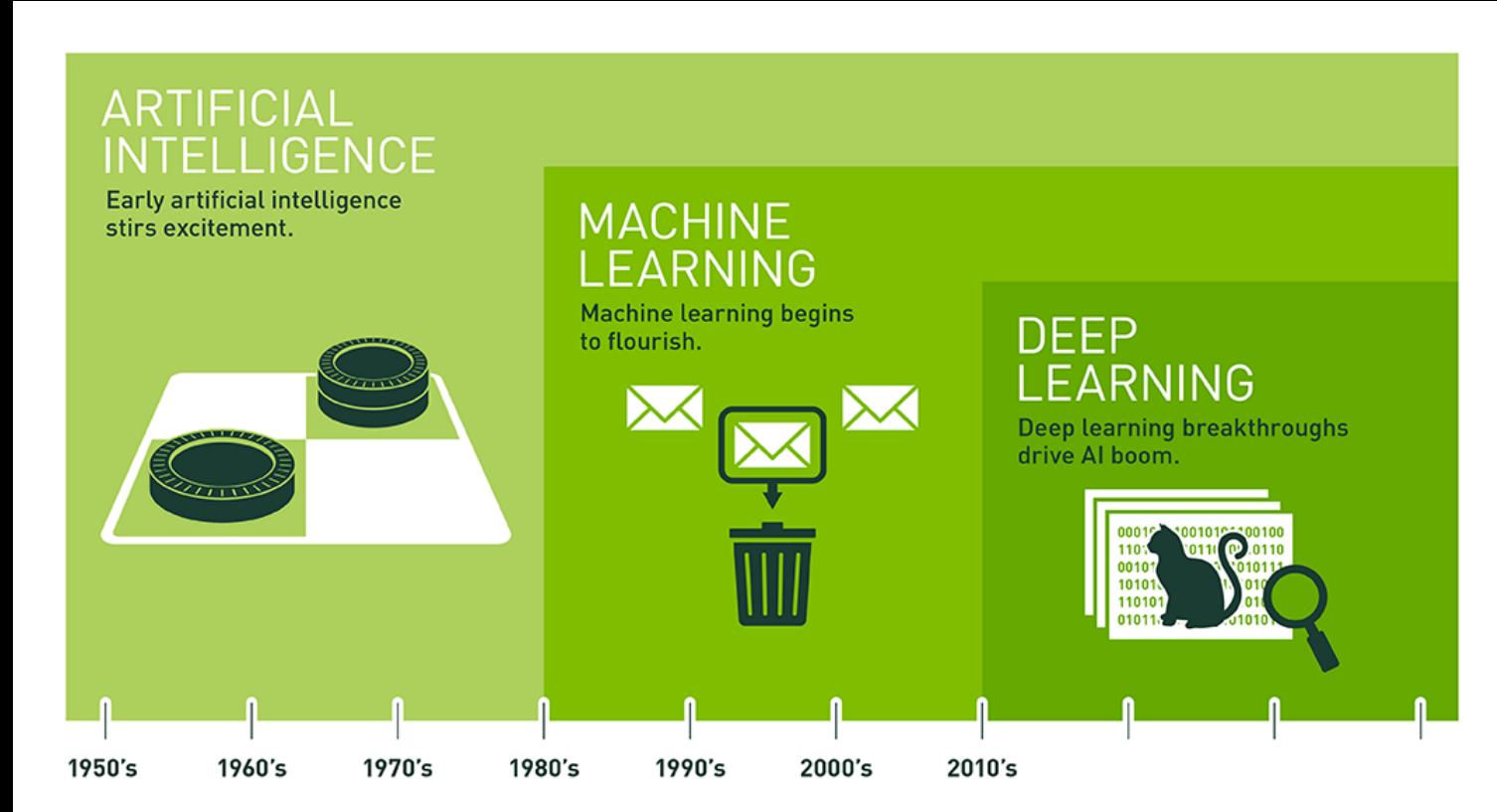
L'Intelligence Artificielle pour tous





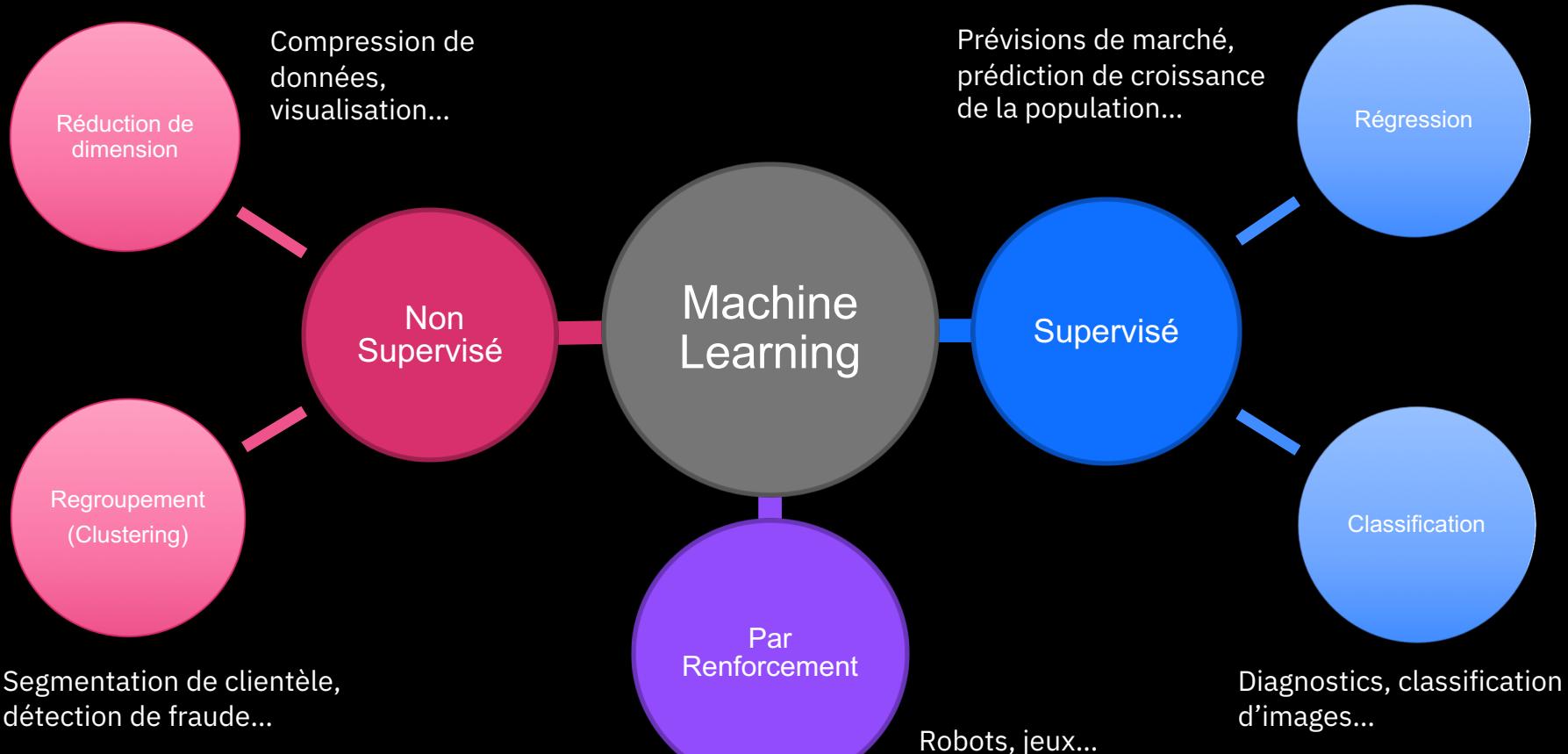
Une introduction pour tous

Un peu d'Histoire...



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

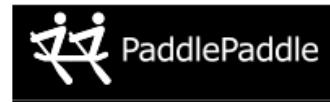
Les principaux types d'apprentissage





La démocratisation de l'IA

Plein de librairies open source...



Projet Norman



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

in

How Companies Are Taking Unconscious Bias Out of the Hiring Equation

 Maxwell Huppert February 21, 2017

Like 725

L'IA dans le recrutement : un présent mitigé



CNBC

A photograph showing a man in a dark suit and a woman in a light blue blazer shaking hands over a table covered with a blue cloth. On the table are several blue plastic cups and a white coffee cup. In the background, there are more people and a blue banner. The image is used as the main visual for an article about algorithmic bias in hiring.

A top-down photograph of a person's hands working on a silver laptop keyboard. To the left is a small white teacup with a gold rim containing tea. Next to it is a whole golden pineapple. To the right is a white plate with a gold rim containing strawberries, blueberries, and lemon slices. The background is a light-colored marble surface.

La justice prédictive : COMPAS



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

Machine Bias

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

VERNON PRATER

Prior Offenses

2 armed robberies, 1
attempted armed
robbery

Subsequent Offenses

1 grand theft

LOW RISK

3

BRISHA BORDEN

Prior Offenses

4 juvenile
misdemeanors

Subsequent Offenses

None

HIGH RISK

8



Comment développer des IAs sans biais ?



Comment développer des IAs sans biais ?

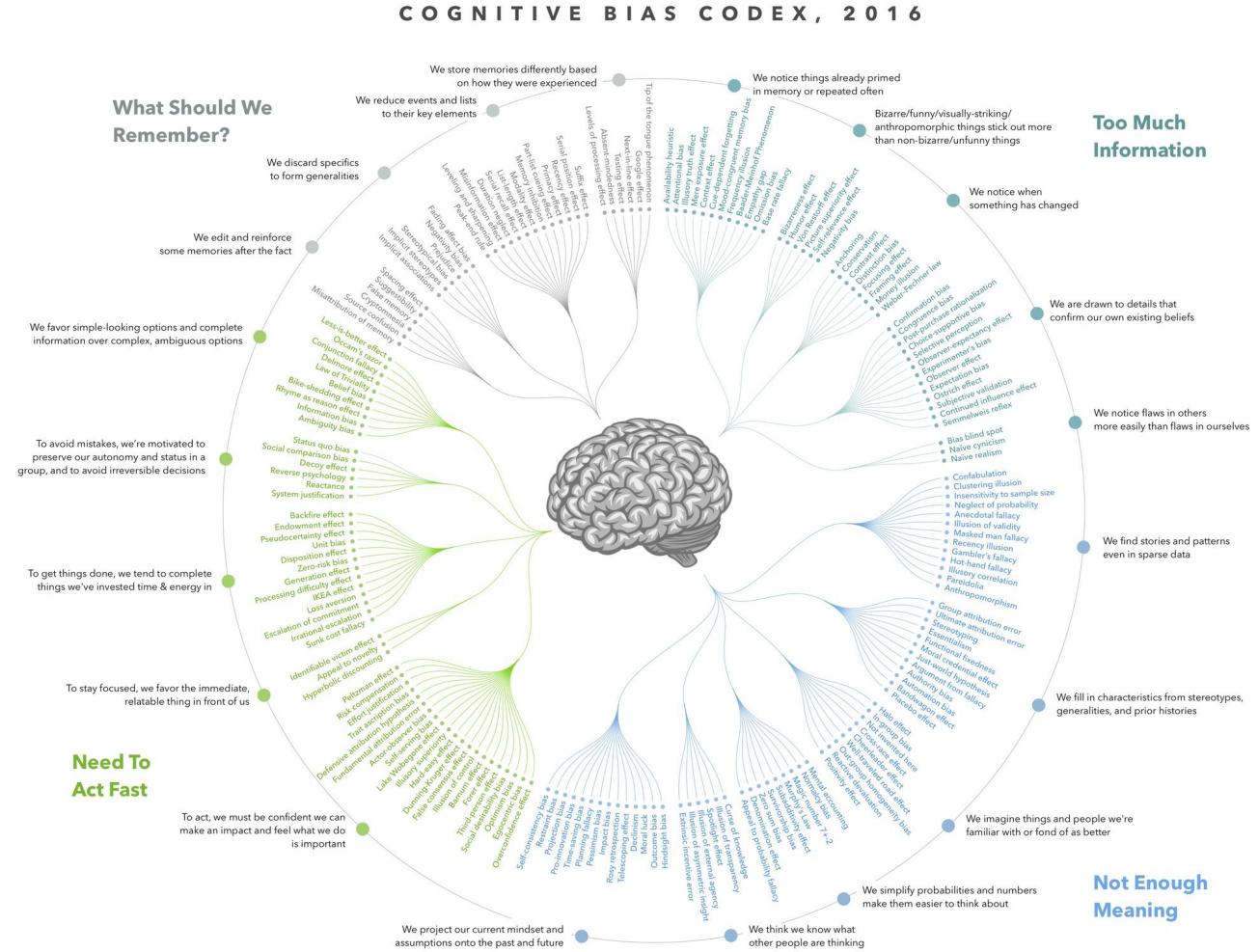
Le facteur humain

C'est quoi le biais humain ?

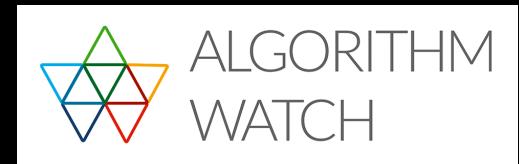
= Une distorsion que subit une information en entrant dans le système cognitif ou en sortant

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



PARTNERSHIP ON AI



DATA FOR GOOD



ALGORITHMIC JUSTICE LEAGUE

La reconnaissance faciale : Gender Shades, l'étude



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%

La reconnaissance faciale : Gender Shades, l'impact



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%

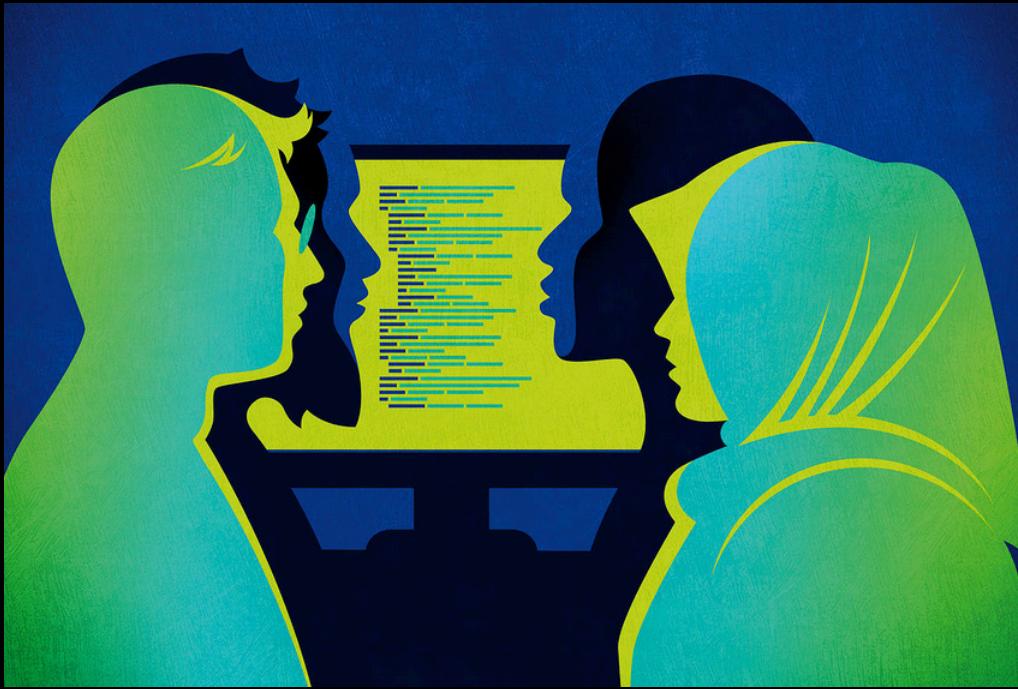
Débiaiser les humains : des équipes soucieuses du sujet

"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

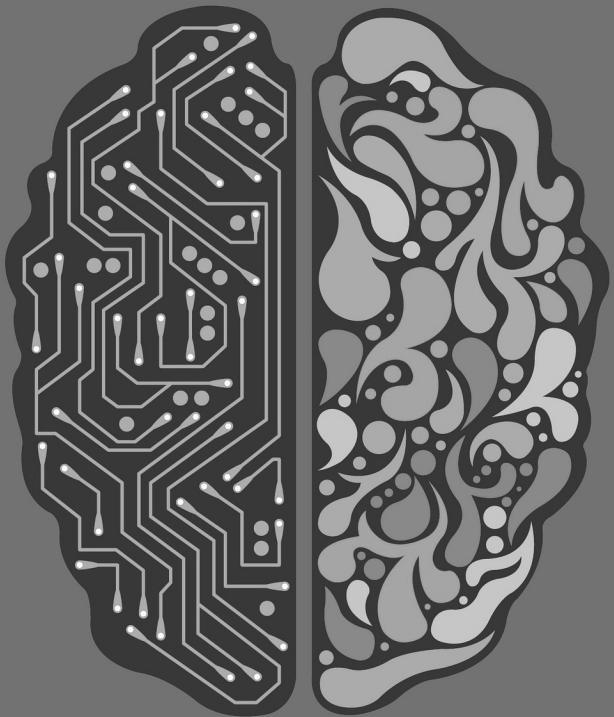


<https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865>



Comment développer des IAs sans biais ?

Les clés techniques



C'est quoi le biais algorithme ?

= 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

...



STOP-DISCRIMINATION.gouv.fr

MINISTÈRE DE LA JUSTICE

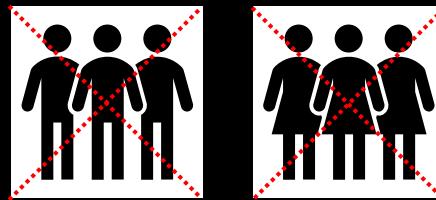
RGPD : des critères protégés en plus

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*)

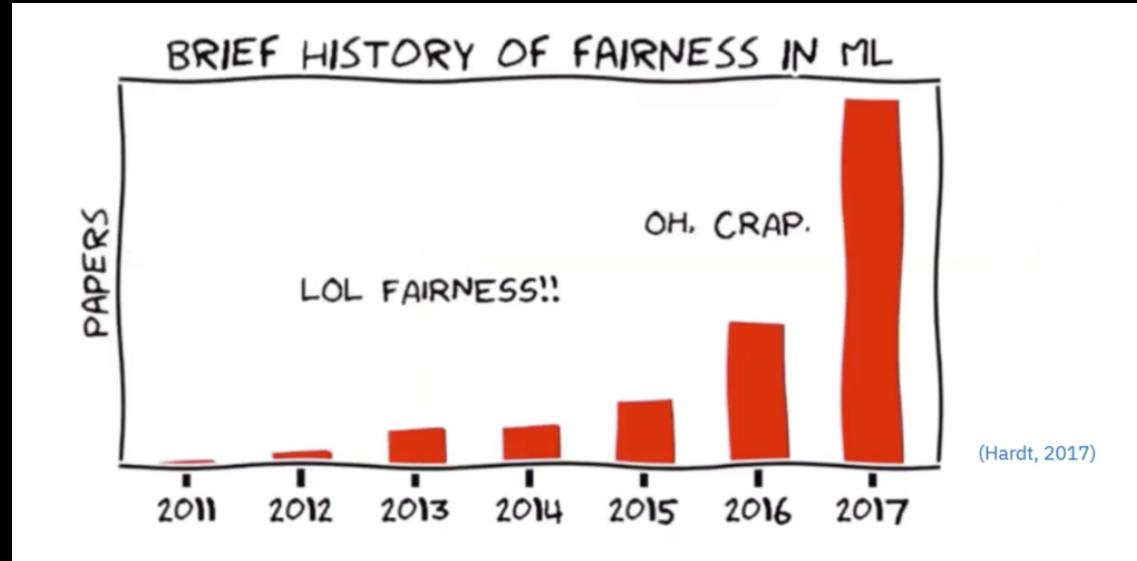
<https://www.cnil.fr/fr/reglement-europeen-protection-donnees/>

Equité de la procédure : l'approche ignorante



Ignorance
(Unawareness)

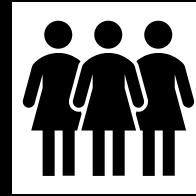
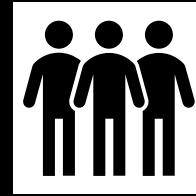
L'équité et les métriques d'équité : un sujet complexe



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 : $TP + TN \approx$	Intervention non-équitable sur les prédictions incorrectes

Equité de groupe : quelques métriques

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

Equité de groupe : quelques métriques

Noms	Définition	Critiques
Equal accuracy = Equalized accuracy	Autant de prédictions correctes pour chaque groupe : $TP + TN \approx$	Intervention non-équitable sur les prédictions incorrectes
Demographic parity = Statistical parity = Equal parity	Autant de prédictions favorables pour chaque groupe : - Favorable = positif : $TP + FP \approx$ - Favorable = négatif : $TN + FN \approx$	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 : $TP \approx$ (et $FN \approx$)	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 : $TP + TN \approx$	Intervention non-équitable sur les prédictions incorrectes
Demographic parity = Statistical parity = Equal parity	Autant de prédictions favorables pour chaque groupe : - Favorable = positif : $TP + FP \approx$ - Favorable = négatif : $TN + FN \approx$	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 : $TP \approx$ (et $FN \approx$)	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 : $FP \approx$ (et $TN \approx$)	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 Research Trusted AI

Home Demo Resources Events Community

These are ten state-of-the-art bias mitigation algorithms that can address bias throughout AI systems. Add more!

Optimized Pre-processing Use to mitigate bias in training data. Modifies training data features and labels. →	Reweighting Use to mitigate bias in training data. Modifies the weights of different training examples. →	Adversarial Debiasing Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions. →	Reject Option Classification Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer. →	Disparate Impact Remover Use to mitigate bias in training data. Edits feature values to improve group fairness. →	Learning Fair Representations Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes. →	Prejudice Remover Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective. →	Calibrated Equalized Odds Post-processing Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels. →
Equalized Odds Post-processing Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer. →	Meta Fair Classifier Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric. →						

Are individuals treated similarly? Are privileged and unprivileged groups treated similarly? Find out by using metrics like these that measure individual and group fairness.

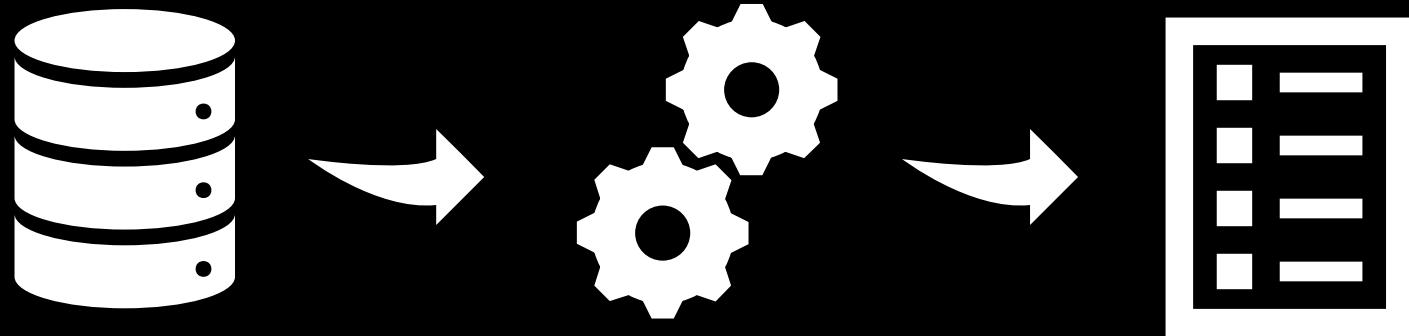
Statistical Parity Difference The difference of the rate of favorable outcomes received by the unprivileged group to that of the privileged group. →	Equal Opportunity Difference The difference of true positive rates between the unprivileged and the privileged groups. →	Average Odds Difference The average difference of false positive rate (false positives/negatives) and true positive rate (true positives/positives) between unprivileged and privileged groups. →	Disparate Impact The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group. →	Theil Index Measures the inequality in benefit allocation for individuals. →	Euclidean Distance The average Euclidean distance between the samples from the two datasets. →	Mahalanobis Distance The average Mahalanobis distance between the samples from the two datasets. →	Manhattan Distance The average Manhattan distance between the samples from the two datasets. →
---	---	--	--	---	---	---	---

There are more than 70 metrics in the GitHub repository already. Add new metrics to the repository and use the Slack channel to let the community know about them.

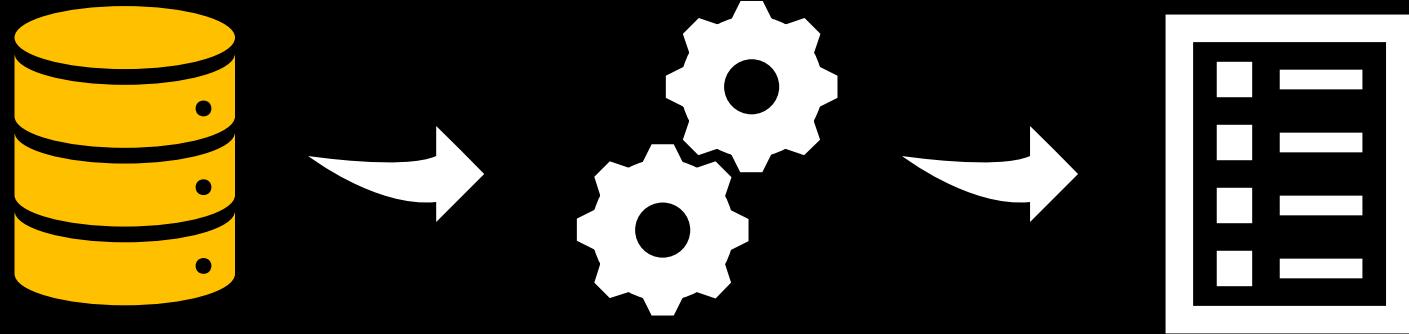
[IBM AI Fairness 360 Open Source Toolkit](#)

[API Docs](#) ↗ [Get Code](#) ↗

Débiaiser les IAs

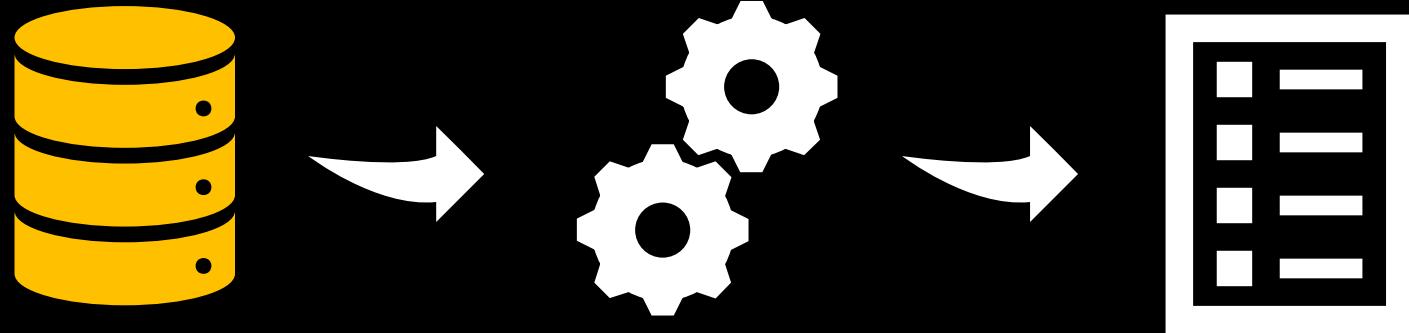


Débiaiser les IAs : agir sur les données



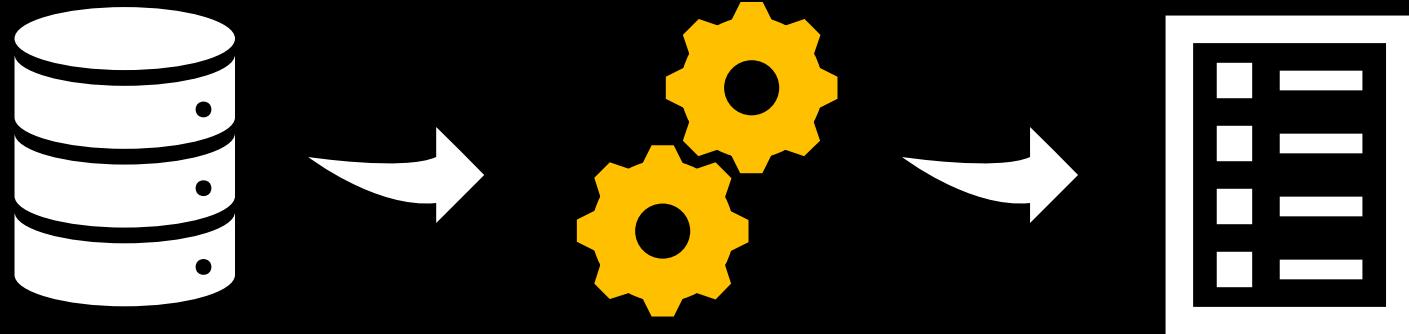
Optimized Pre-Processing

Débiaiser les IAs : agir sur les données



**Optimized Pre-Processing
Reweighting**

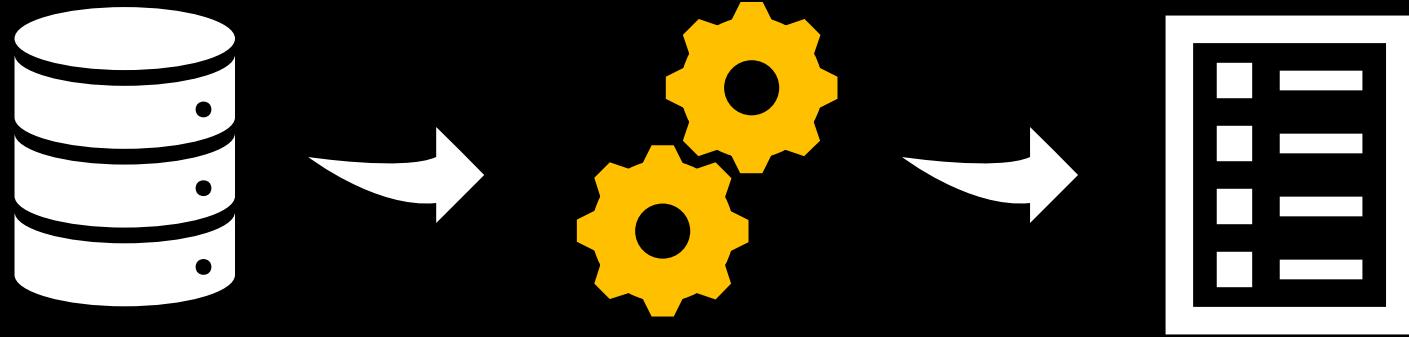
Débiaiser les IAs : agir sur le modèle



**Optimized Pre-Processing
Reweighting**

Prejudice Remover

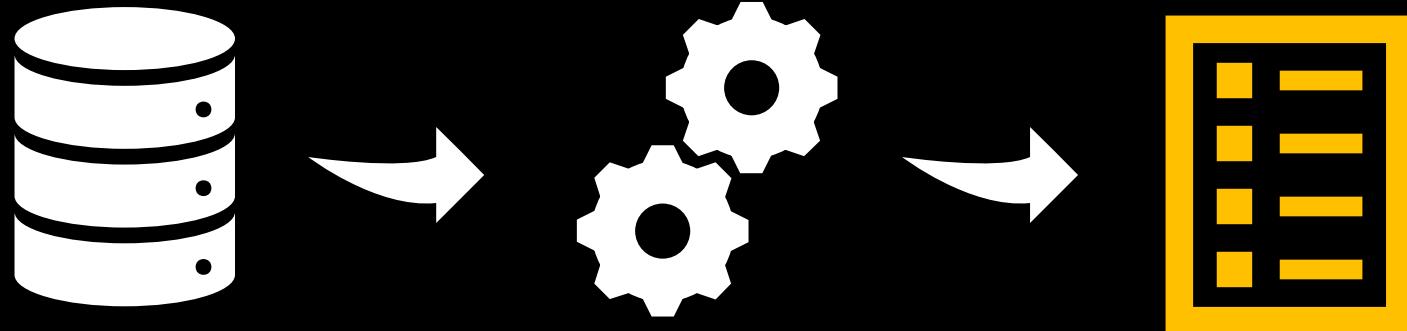
Débiaiser les IAs : agir sur le modèle



**Optimized Pre-Processing
Reweighting**

**Prejudice Remover
Adversarial Debiasing**

Débiaiser les IAs : agir sur les prédictions

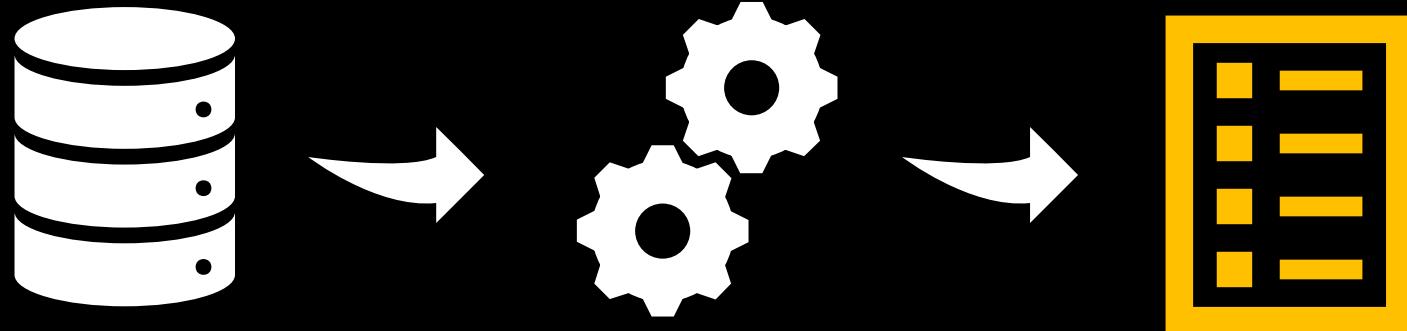


**Optimized Pre-Processing
Reweighting**

**Prejudice Remover
Adversarial Debiasing**

Equalized Odds Post-Processing

Débiaiser les IAs : agir sur les prédictions



**Optimized Pre-Processing
Reweighting**

**Prejudice Remover
Adversarial Debiasing**

**Equalized Odds Post-Processing
Reject Option Classification**

Quelques outils & frameworks actuels

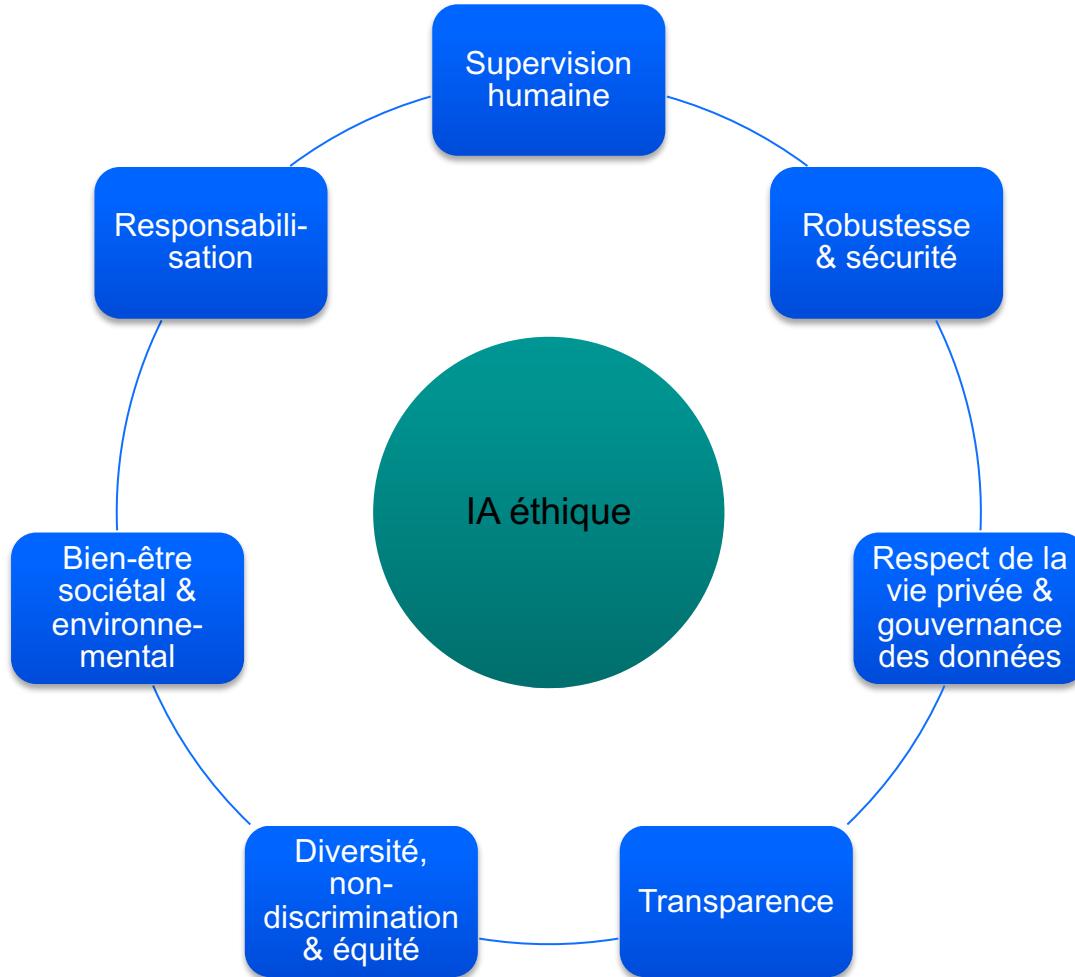
Nom	Date de lancement	Outil/API	Disponibilité	Explicabilité	Estimation Equité	Atténuation de biais
<u>Audit AI</u>  pymetrics	05/2018	API Python Notebooks Jupyter	<u>Open source</u>	Non	~10 métriques	Non
<u>What-If Tool</u> 	09/2018	Outil graphique Intégration TensorBoard & notebooks Démo web Notebooks Jupyter	<u>Open source</u>	Oui : des données Non : de décisions	5 métriques	Variateur de seuil
<u>AI Fairness 360</u> 	09/2018	API Python Démo web Notebooks Jupyter	<u>Open source</u>	Non	70+ métriques	10+ algorithmes
<u>AI Openscale</u> 	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
<u>Teach & Test AI</u> -> accenture	02/2018	?	Offre payante	Oui ?	Oui ?	?

Démo

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

Conclusion

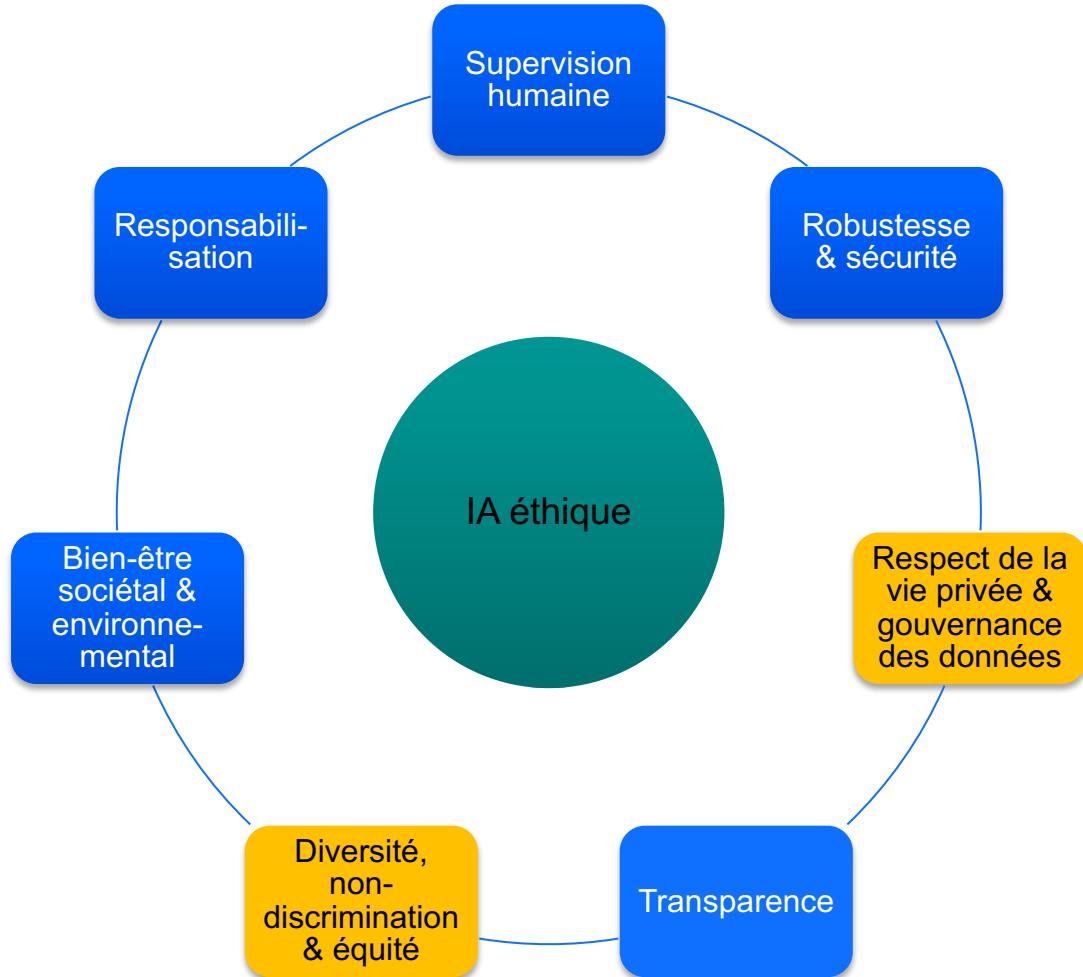
Pour une IA éthique



7

Principes
(selon la commission
européenne)

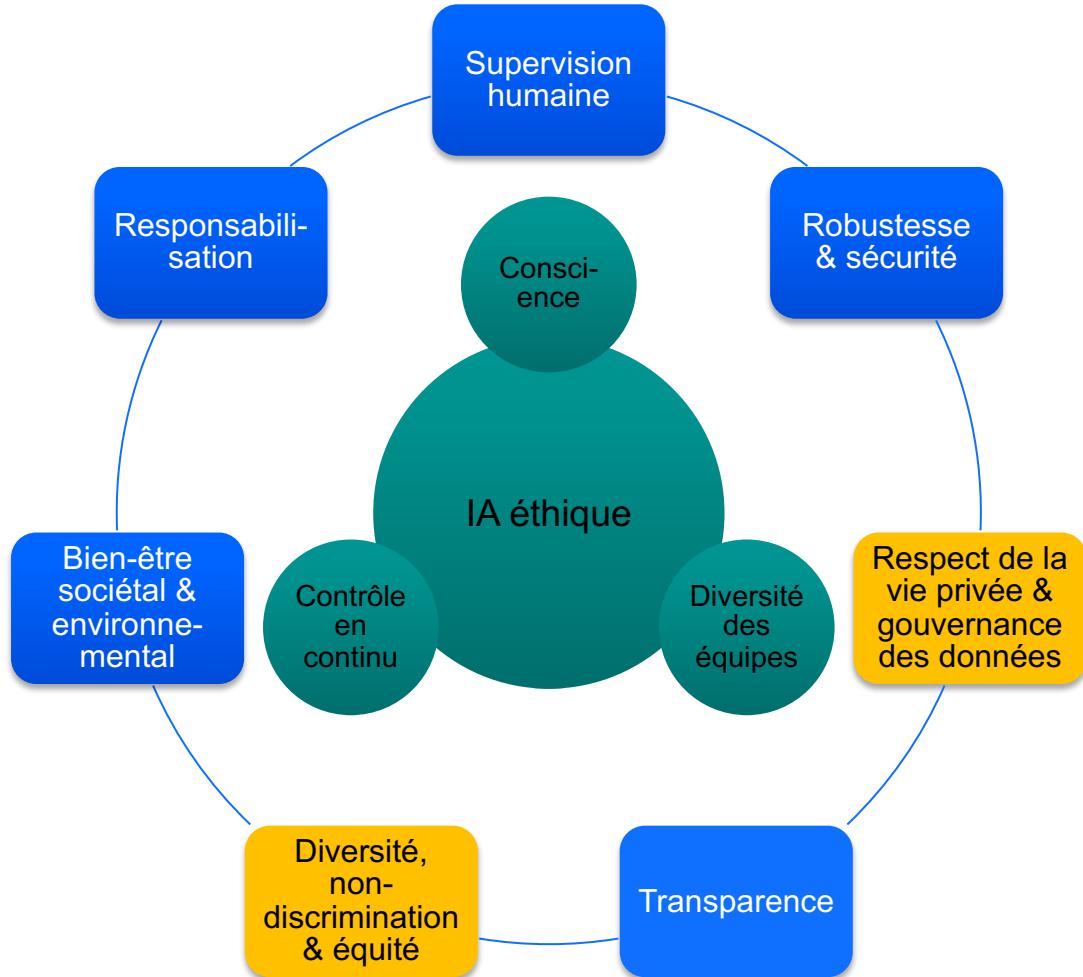
Pour une IA éthique



7

Principes
(selon la commission
européenne)

Pour une IA éthique

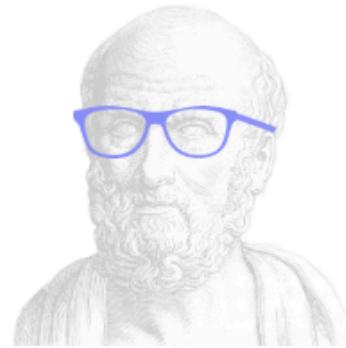


7
Principes
(selon la commission
européenne)

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>