Thomas Souverain

PhD Student on AI Ethics

AI Phi - ENS

2023-12-07

Applied AI Ethics

1) AI Ethics Principles

2) My PhD work

pôle emploi

ENS

Institut | Nicod

Thomas Souverain

PhD Student on AI Ethics

AI Phi - ENS

2023-12-07

Applied AI Ethics

1) AI Ethics Principles

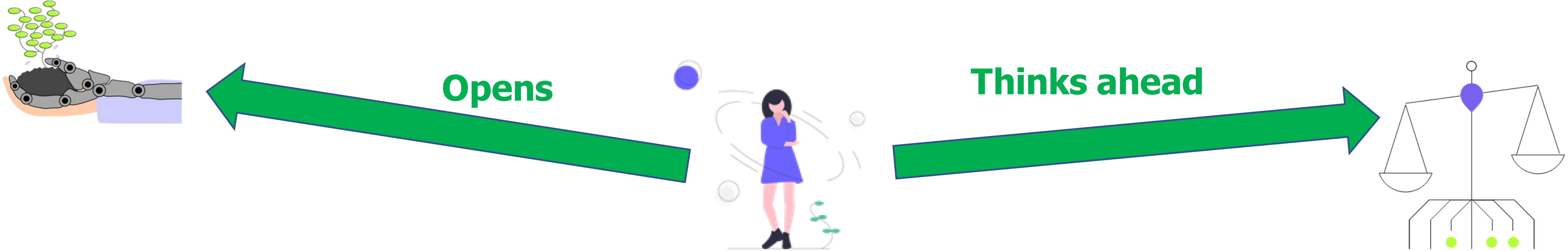
2) My PhD work

1) Why introducing ethics… For AI?

- ❑ **Automatism** : training and parameters setting lacks of transparency
- ❑ **Big data**

=> Machine Learning models may have a broad, fast and unexpected impact on society

But what is AI Ethics?



Opens

Thinks ahead

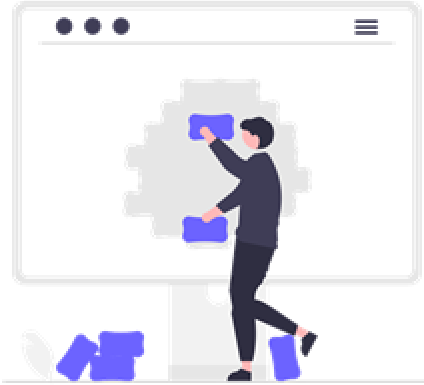
Common Values

Law

For the designer

- ✓ Professional best practices: be aware and evaluate **risks** and impacts
- ✓ Incorporate the expectations of impacted citizens => **trust**

1) AI Ethics Principles



I know and forestall the **risks** of my model

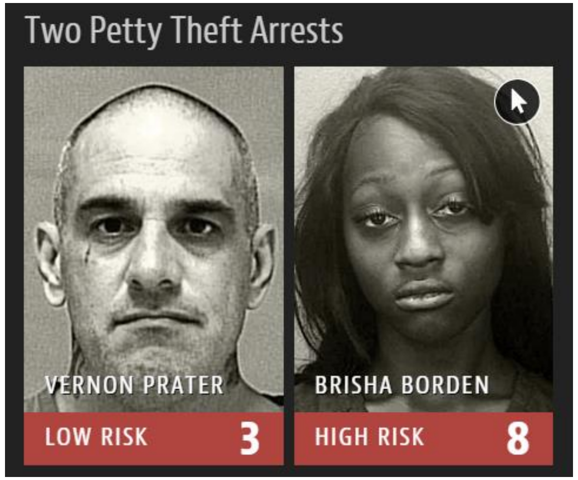
TECHNICAL SAFETY

How will AI work on roads or **unexpected situations** during training?

RESPECT OF PRIVACY

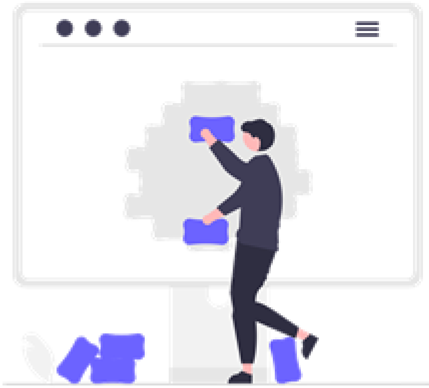
Granting I don't use data **without consent**

FAIRNESS



Focusing on **stat performance** may hide **discriminations**

1) AI Ethics Principles



I am **not blind** on my tool
I am aware and concerned by its impacts

EXPLAINABILITY

"Why am I rejected for a loan?"

RESPONSIBILITY

UIA
Unemployment Insurance Agency

Anticipate the consequences once deployed
Ensuring that fraudsters (34 000) are real ones...

1) AI Ethics Principles

Now, how to **implement** them?

TECHNICAL SAFETY

RESPECT OF PRIVACY

FAIRNESS

EXPLAINABILITY

RESPONSIBILITY

AUTONOMY

LONG-TERM IMPACT

Example – loan lending

Now, how to **implement** them?

TECHNICAL SAFETY

Maturity level: 4/5

Check for finer secure paths between banks

RESPECT OF PRIVACY

Maturity level: 4/5

GDPR respected
Process for non-EU data?

FAIRNESS

Maturity level: 3/5

Except legal requirements (gender), lack of investigation of banking impact by group

EXPLAINABILITY

Maturity level: 2/5

Personal features influence
Not enough logic in method

RESPONSIBILITY

Maturity level: 5/5

Clear accountability of bank officers / managers by case

AUTONOMY

Maturity level: 4/5

Bank officer / manager always validating

LONG-TERM IMPACT

Maturity level: 1/5

No clear position on AI replacing / assisting for credits

Thomas Souverain

PhD Student on AI Ethics

AI Phi - ENS

2023-12-07

Applied AI Ethics

1) AI Ethics Principles

2) My PhD work

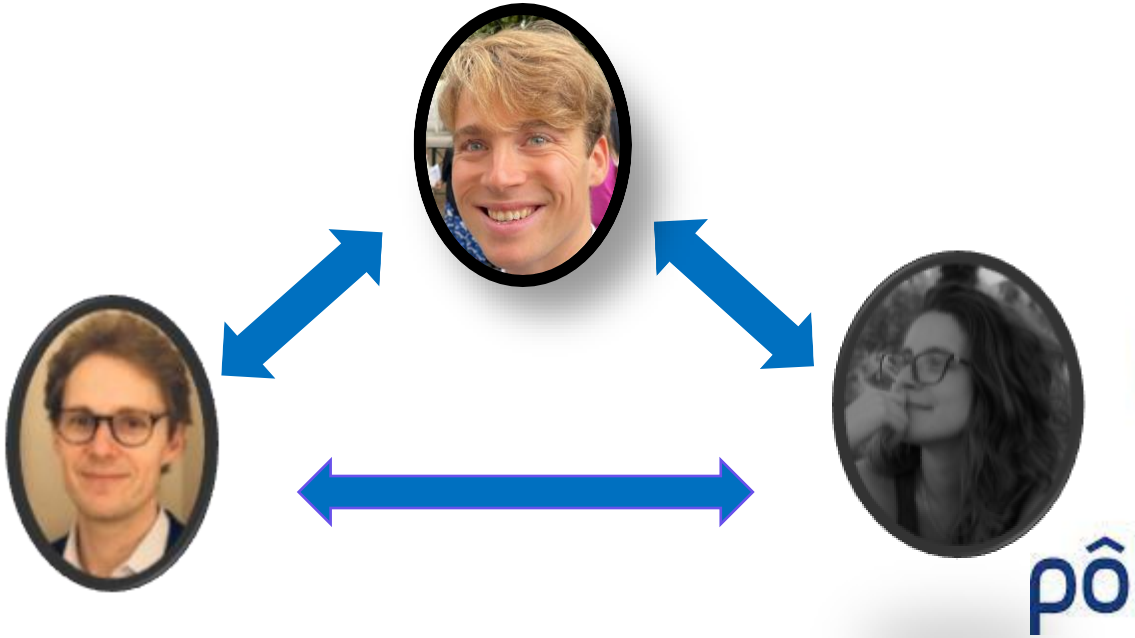
Prelude - My PhD in Philosophy of AI

"Is it possible to explain AI?

Technical solutions, ethical issues in algorithmic loan lending and job offering"

Since 2020, I dig into specific use cases with data-scientist teams

Programming, guiding, and analysing their ethical impact



ENS

ENS Ulm (Paul Egré)
Expert in Logics and Epistemology

pôle emploi

dreamquark

Example – fairness in income prediction

As a data-scientist, I coded a package to **detect** and **mitigate** group Inequalities, which fits the preferences of the end user (e.g. banker)

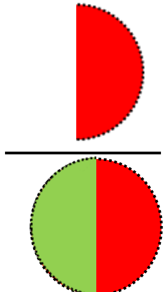
Paves the way for applied analysis of **fairness metrics** in AI
Translate a vision of the "fairest as possible" repartition of **resources**

Statistical Parity

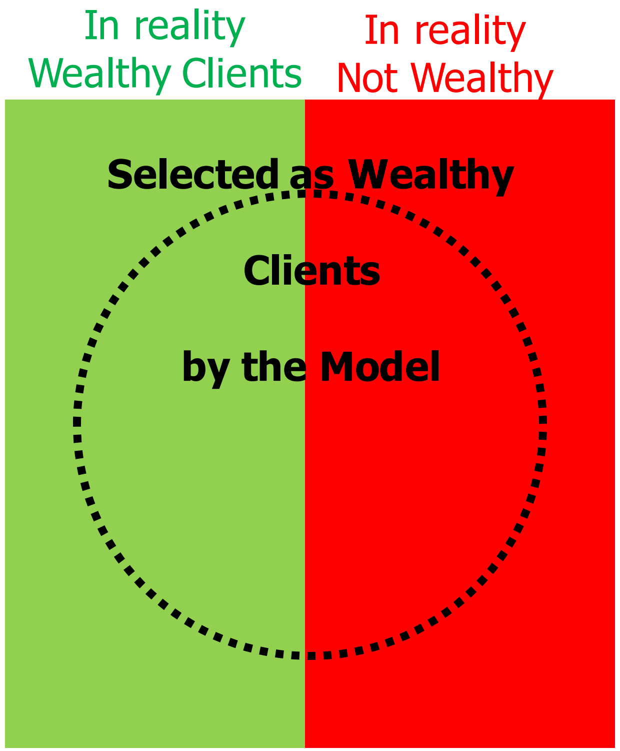
"Reformist"

$$\textbf{Accuracy} = \frac{|y_{pred} = y_{true}|}{nb\_clients}$$

False Positive Rate
How many 'not wealthy' clients are wrongly selected?



$$= \frac{|true\ positive|}{|true\ positive| + |false\ positive|}$$

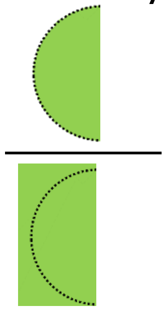


Equalized Odds

"Prudent"

$$\textbf{ROC AUC} = \frac{\textbf{True Positive Rate}}{\textbf{False Positive Rate}}$$

True Positive Rate (Recall)
How many real 'wealthy' clients are detected?



$$= \frac{|true\ positive|}{|true\ positive| + |false\ negative|}$$

In progress – AI, Explanation and Trust by Pôle Emploi advisors

Neural networks assist the advisor to select only LEGal job Offers (**LEGO**)

Why are there **only 40% of job offers** on which LEGO **alerts** advisors which are fully corrected before being published?

- Diverse hypotheses : time consuming, unclear explainability, some rules are controversial or not deeply understood, lack of trust in AI...

- I lead a **survey** to address this designer - user gap

Link technical solutions (AI, explainability) - **trust**

Thomas Souverain

PhD Student on AI Ethics

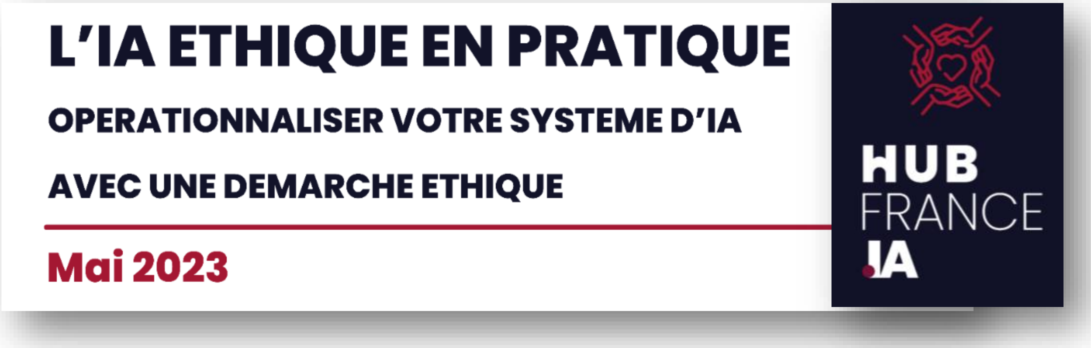
IJM Workshop

2023-11-17

Thanks for your attention !

To know more on AI Ethics into practices

https://www.hub-franceia.fr/groupe-ethique/



L'IA ETHIQUE EN PRATIQUE

OPERATIONNALISER VOTRE SYSTEME D'IA

AVEC UNE DEMARCHE ETHIQUE

Mai 2023

HUB FRANCE IA

pôle emploi

ENS

Institut | Nicod

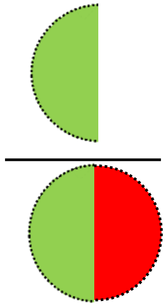
« Statistical Performance » Measures

$$\textbf{PR AUC} = \frac{\textbf{Precision}}{\textbf{Recall}}$$

$$\textbf{ROC AUC} = \frac{\textbf{True Positive Rate}}{\textbf{False Positive Rate}}$$

Precision

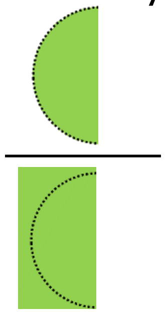
How many selected clients are really 'wealthy'?



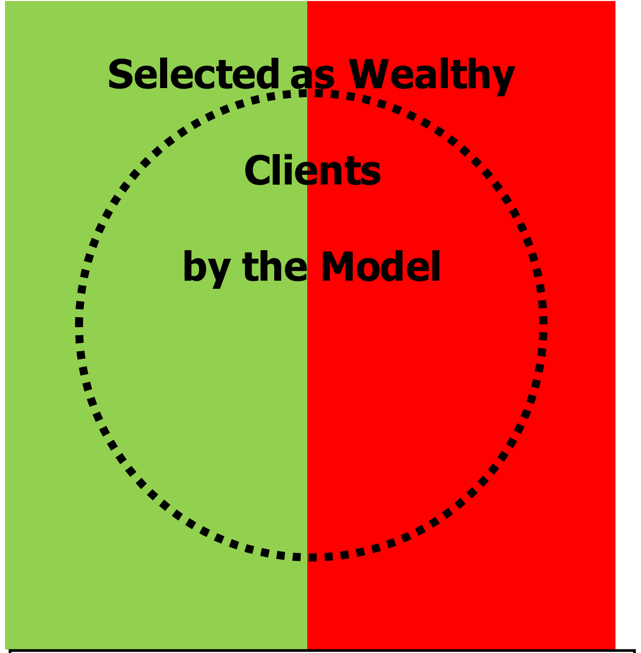
$$= \frac{|true\ positive|}{|true\ positive| + |false\ positive|}$$

True Positive Rate (Recall)

How many real 'wealthy' clients are detected?



$$= \frac{|true\ positive|}{|true\ positive| + |false\ negative|}$$



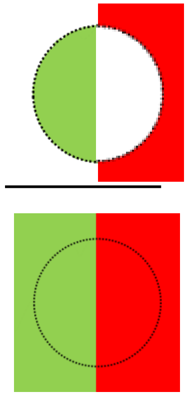
In reality Wealthy Clients

In reality Not Wealthy

Selected as Wealthy Clients by the Model

$$\textbf{Accuracy} = \frac{|y_{pred} = y_{true}|}{nb\_clients}$$

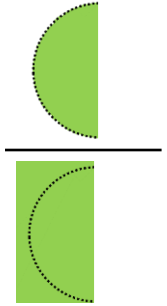
How many clients are well detected?



$$= \frac{|true\ positive| + |true\ negative|}{nb\_clients}$$

True Positive Rate (Recall)

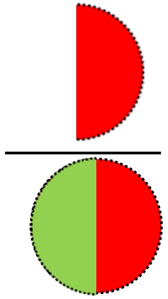
How many real 'wealthy' clients are detected?



$$= \frac{|true\ positive|}{|true\ positive| + |false\ negative|}$$

False Positive Rate

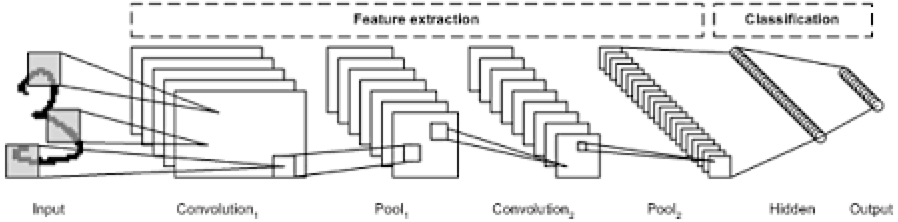
How many 'not wealthy' clients are wrongly selected?

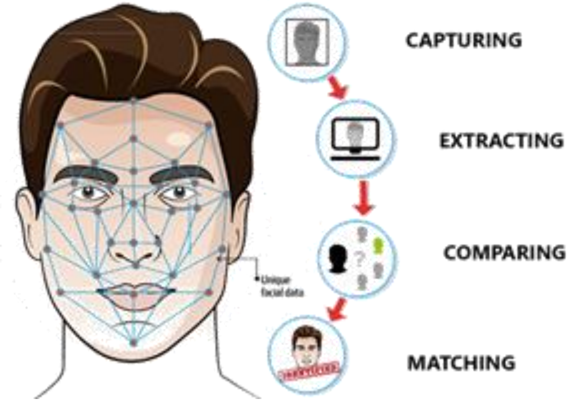


$$= \frac{|true\ positive|}{|true\ positive| + |false\ positive|}$$

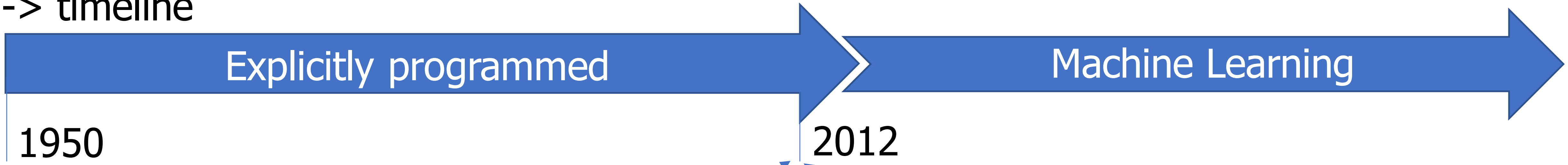
What we mean by « AI » : machine learning led to its new summer

-> artificial intelligence (AI) : system which **spontaneously** performs **tasks** which were commonly thought to be exclusive to **human intelligence**





-> timeline

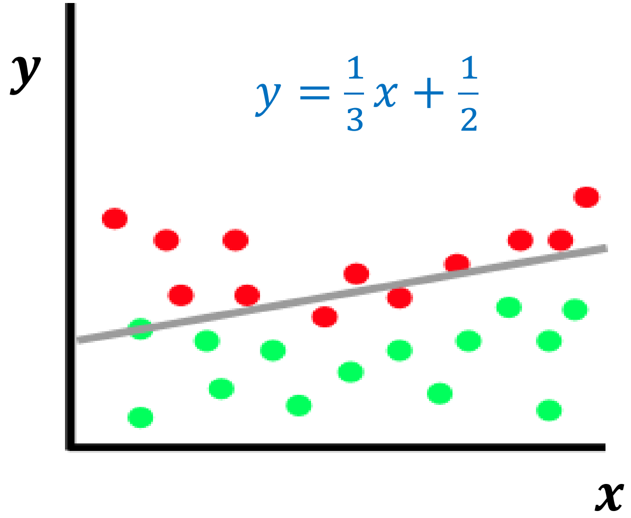


GPU

Big data

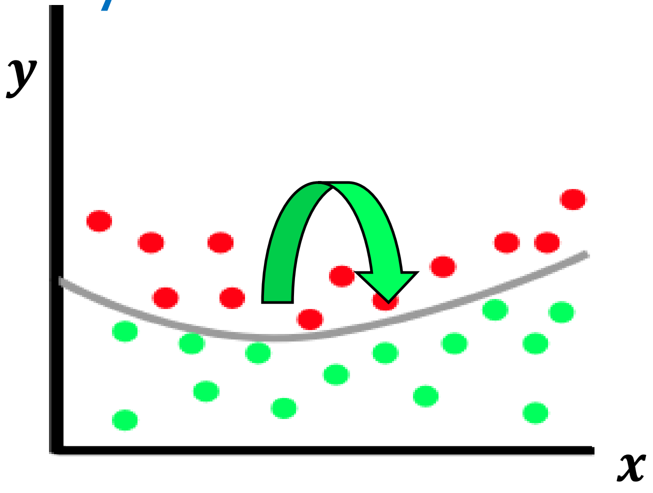
Problem: The Current AI Models rely on Opacity

-> AI performance



Rule only implemented by humans

Not explicitly programmed



Parameters self-adjusting to fit the data

=> in architecture and training : « **black box** » of AI models

-> philosophy : : « **why** » does this AI model handle data that way, do these **operations** have a **meaning for humans**?

(1) Logics

2. Justice Categories