# Can I trust my algorithm ?

J-M. Loubes

Institut de Mathématiques de Toulouse
&
Artificial and Natural Intelligence Toulouse Institute

Chair: M. Serrurier, B. Laurent, C. Benesse, L. Bethune, L. De Lara, W. Todo, A. G Sanz,
Collaborations : L. Risser, F. Gamboa, N. Asher, J. Eynard and T. Boissin, A. Picard,

**[ Bias ]**

An unfair/irrelevant information that
influences a decision

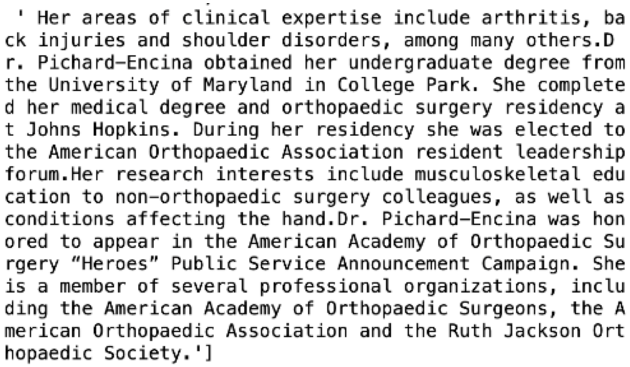
1

# PRINCIPLE OF MACHINE LEARNING

*Biased decisions that are 'accurate but unfair' ? (Bias in Bios Dataset)*

## SUCCESS OF NEURAL-NETWORKS TO TREAT COMPLEX DATA

Example of the « Bios dataset », which was made public by linkedin/microsoft, to predict the job occupation using neural networks.



```
 ' Her areas of clinical expertise include arthritis, ba
ck injuries and shoulder disorders, among many others.D
r. Pichard-Encina obtained her undergraduate degree from
the University of Maryland in College Park. She complete
d her medical degree and orthopaedic surgery residency a
t Johns Hopkins. During her residency she was elected to
the American Orthopaedic Association resident leadership
forum.Her research interests include musculoskeletal edu
cation to non-orthopaedic surgery colleagues, as well as
conditions affecting the hand.Dr. Pichard-Encina was hon
ored to appear in the American Academy of Orthopaedic Su
rgery "Heroes" Public Service Announcement Campaign. She
is a member of several professional organizations, inclu
ding the American Academy of Orthopaedic Surgeons, the A
merican Orthopaedic Association and the Ruth Jackson Ort
hopaedic Society.']
```

**Input data**

(A biography on linkedin)

**Data preparation**
(generally by using a generic pre-trained neural-network)

**Optimal data representation**

(embedding)

**Prediction**
(using a specific neural-network)

"Surgeon"

**Job recommendation**

(out of a list of known jobs)

Mimics the recommendations made in a reference **training set**
(here more than 400K recommendations)

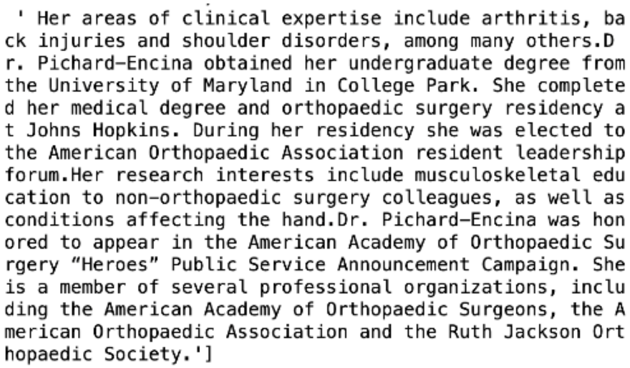
1

# PRINCIPLE OF MACHINE LEARNING

*Biased decisions that are ‘accurate but unfair’ ? (Bias in Bios Dataset)*

## SUCCESS OF NEURAL-NETWORKS TO TREAT COMPLEX DATA

Example of the « Bios dataset », which was made public by linkedin/microsoft, to predict the job occupation using neural networks.



```
 ' Her areas of clinical expertise include arthritis, ba
ck injuries and shoulder disorders, among many others.D
r. Pichard-Encina obtained her undergraduate degree from
the University of Maryland in College Park. She complete
d her medical degree and orthopaedic surgery residency a
t Johns Hopkins. During her residency she was elected to
the American Orthopaedic Association resident leadership
forum.Her research interests include musculoskeletal edu
cation to non-orthopaedic surgery colleagues, as well as
conditions affecting the hand.Dr. Pichard-Encina was hon
ored to appear in the American Academy of Orthopaedic Su
rgery "Heroes" Public Service Announcement Campaign. She
is a member of several professional organizations, inclu
ding the American Academy of Orthopaedic Surgeons, the A
merican Orthopaedic Association and the Ruth Jackson Ort
hopaedic Society.']
```

**Input data**

(A biography on linkedin)

**Data preparation** (generally by using a generic pre-trained neural-network)

**Optimal data representation**

(embedding)

**Prediction** (using a specific neural-network)

“Surgeon”

**Job recommendation**

(out of a list of known jobs)

He -> She ; His -> Her

Mimics the recommendations made in a reference **training set** (here more than 400K recommendations)

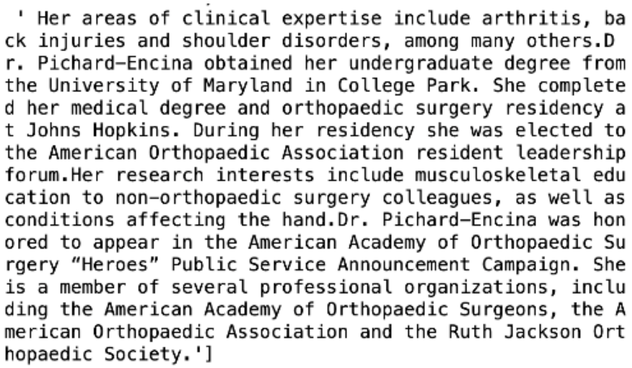
1

# PRINCIPLE OF MACHINE LEARNING

*Biased decisions that are 'accurate but unfair' ? (Bias in Bios Dataset)*

## SUCCESS OF NEURAL-NETWORKS TO TREAT COMPLEX DATA

Example of the « Bios dataset », which was made public by linkedin/microsoft, to predict the job occupation using neural networks.

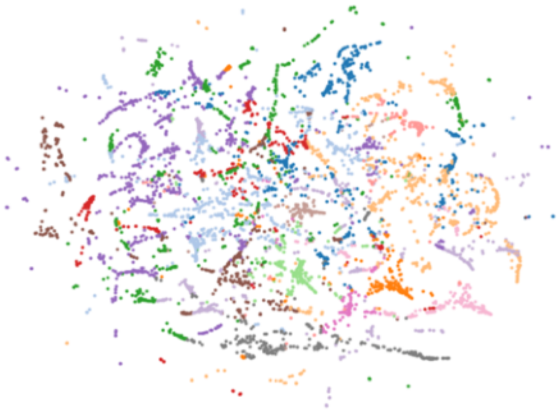


```
 ' Her areas of clinical expertise include arthritis, ba
ck injuries and shoulder disorders, among many others.D
r. Pichard–Encina obtained her undergraduate degree from
the University of Maryland in College Park. She complete
d her medical degree and orthopaedic surgery residency a
t Johns Hopkins. During her residency she was elected to
the American Orthopaedic Association resident leadership
forum.Her research interests include musculoskeletal edu
cation to non–orthopaedic surgery colleagues, as well as
conditions affecting the hand.Dr. Pichard–Encina was hon
ored to appear in the American Academy of Orthopaedic Su
rgery "Heroes" Public Service Announcement Campaign. She
is a member of several professional organizations, inclu
ding the American Academy of Orthopaedic Surgeons, the A
merican Orthopaedic Association and the Ruth Jackson Ort
hopaedic Society.']
```

**Input data**

(A biography on linkedin)

**Data preparation** (generally by using a generic pre-trained neural-network)



**Optimal data representation**

(embedding)

**Prediction** (using a specific neural-network)

**Job recommendation**

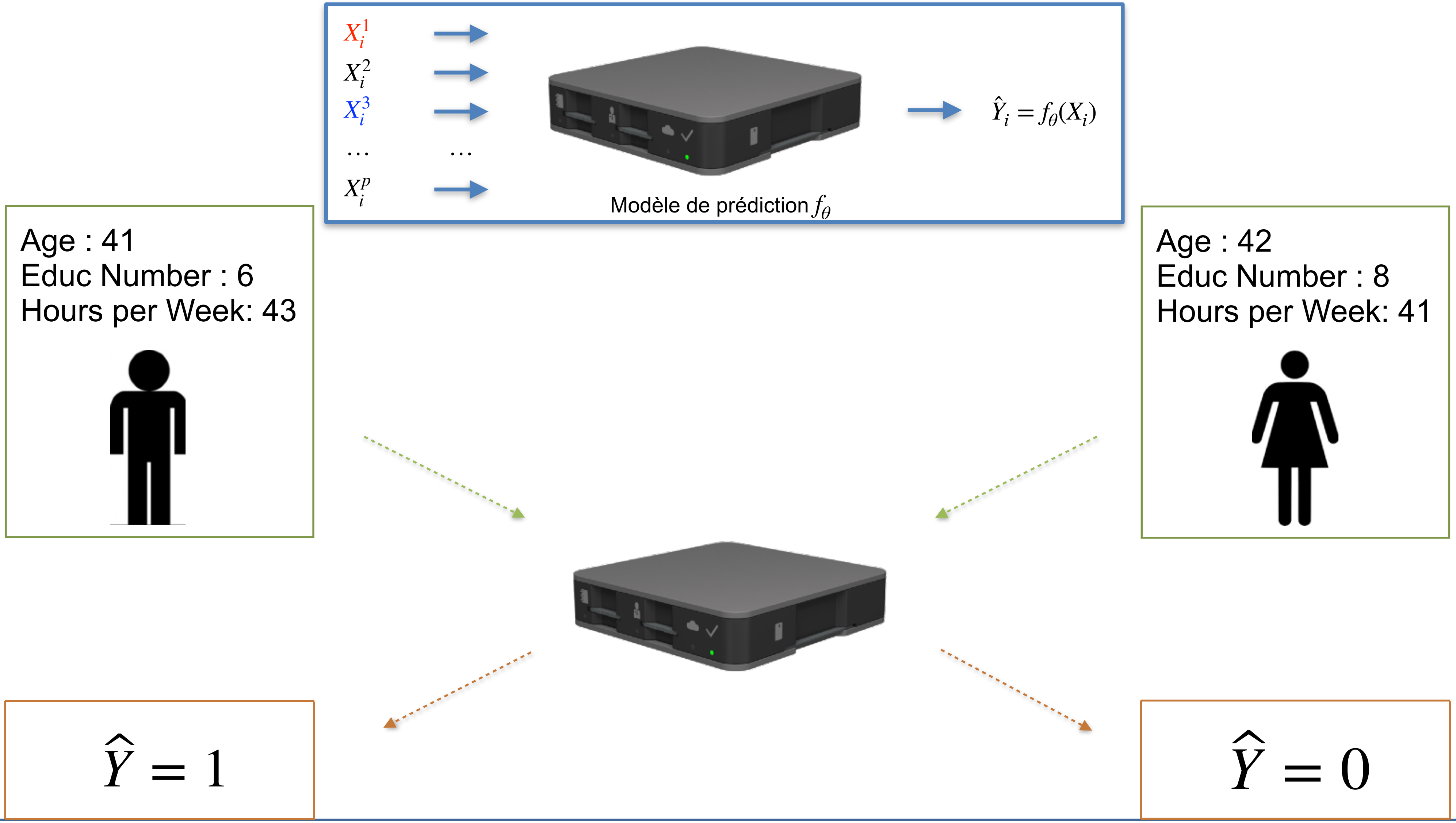
(out of a list of known jobs)

He -> She ; His -> Her

Mimics the recommendations made in a reference **training set** (here more than 400K recommendations)

# PRINCIPLE OF MACHINE LEARNING

*Biased decisions that are ‘accurate but unfair’ ? (Adult Dataset)*



$X_i^1$

$X_i^2$

$X_i^3$

$\dots$ $\dots$

$X_i^p$

$\hat{Y}_i = f_\theta(X_i)$

Modèle de prédiction $f_\theta$

Age : 41
Educ Number : 6
Hours per Week: 43

Age : 42
Educ Number : 8
Hours per Week: 41

$\widehat{Y} = 1$

$\widehat{Y} = 0$

1

# DECISION NON EXPLICABLE ET DÉLOYALE

*Illustration : Prédire un Risque de Crédit*

La variable « Genre » semble liée à la décision algorithmique mais d’un point de vue légal, c’est un **délit**

**Article 225-1 du code pénal** prévoit 3 ans de prison et 45.000 € d’amende

Discrimination directe → Traitement défavorable d’une personne fondé sur un critère prohibé 👩🏾 🏳️‍🌈 🤰 ☭

« Mais ce n’est pas de ma faute c’est celle de l’algorithme »

1

# BIAS IN MACHINE LEARNING

*Is it a problem we tackle ? From moral to legal point of view*

- An A.I. algorithm suffers from **unfairness** if its outcomes $Y$ (decisions) are fully or partly based on a variable A that *should* not play a decisive role in the decision making process.

- A **chosen** Variable A is denoted by **sensitive attribute.**
  → It divides the observations into subgroups (e.g.: Males/Females).
  → The prediction algorithm should not show a different behavior over these subsets.
  → The variable A is chosen by the practitioner. Its choice is driven by legal, ethic or technical concerns.

**Artificial Intelligence Act (April 2021) by European Commission**

- Definitions of High Risk domains of a applications (health, finance, public services, transports …)
- Performance matters but not only : notions of equity, transparency and robustness
- Need to **definitions of norms** to measures bias (AFNOR, IEEE, .[1]..)

1

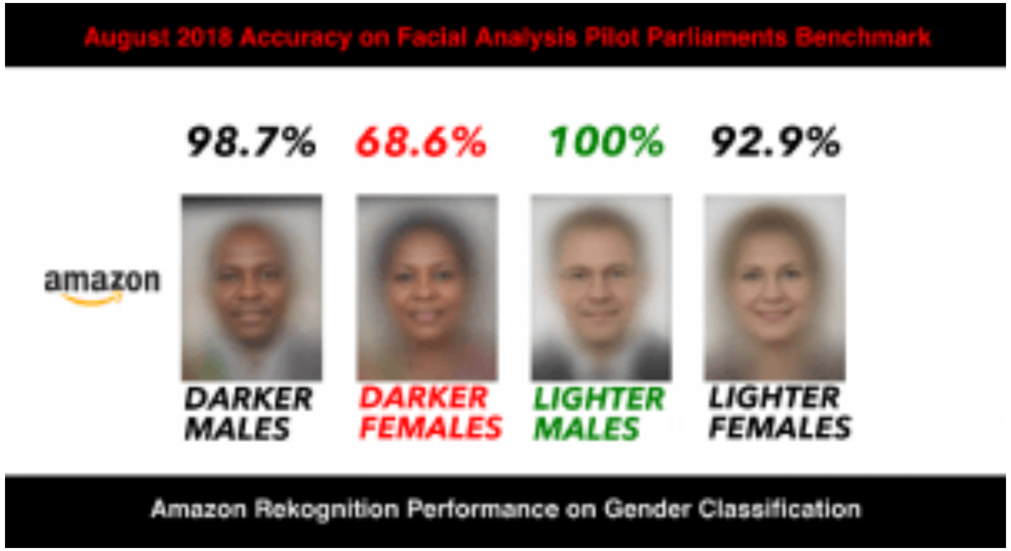
# NECESSITE DE SE PRÉMUNIR DES RISQUES DE DÉLOYAUTÉ

*Questions clés dans l'optique de mettre sur le marché des algorithmes d'IA*

- Les algorithmes peuvent-ils produire des décisions discriminatoires ?
- Pourquoi
- Comment définir, quantifier, et détecter les biais ?
- Peut-on auditer les algorithmes ?
- Peut-on corriger les algorithmes biaisés ?

# FAIRNESS IN MACHINE LEARNING

*Principle : independence w.r.t to the protected attribute*



August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

| | 98.7% | 68.6% | 100% | 92.9% |
|---|---|---|---|---|
| amazon | DARKER MALES | DARKER FEMALES | LIGHTER MALES | LIGHTER FEMALES |

Amazon Rekognition Performance on Gender Classification

**Statistical Parity**: a variable (gender) influences the outcome while it should not play any role

→ *Same decision for all groups A*

$$\hat{Y} = g(X) \perp A$$

**Equality of Odds:** the performance of the algorithm is degraded for given subgroups

→ *Same performance for all groups A*

$$\hat{Y} = (g(X) \perp A) \,|\, Y$$

# FAIRNESS IN MACHINE LEARNING

## *In the Jungle of Fairness in the Literature*

- Input observations are $(X, A)$
- Output observations (available in the learning sample) are $Y$
- Decision rules to predict $Y$ are $\widehat{Y} = f(X, A)$

**Different measures of (group) fairness in classification case that may be incompatible**

1. Disparate Treatment $P(\hat{Y} = 1 \mid A = 0) \,/\, P(\hat{Y} = 1 \mid A = 1)$

2. Avoiding Disparate Treatment : $P(\hat{Y} = i \mid A = 0, Y = j) \; - \; P(\hat{Y} \mid A = 1, Y = j)$

3. Predictive Parity $P(Y = i \mid \hat{Y} = j, A = 0) \; - \; P(Y = i \mid \hat{Y} = j, A = 1)$

If the decision is a function of a **score S,** previous definitions can be extended to the score
Or use the notion of **score balance** $E(S \mid (Y, A)) = E(S \mid Y)$

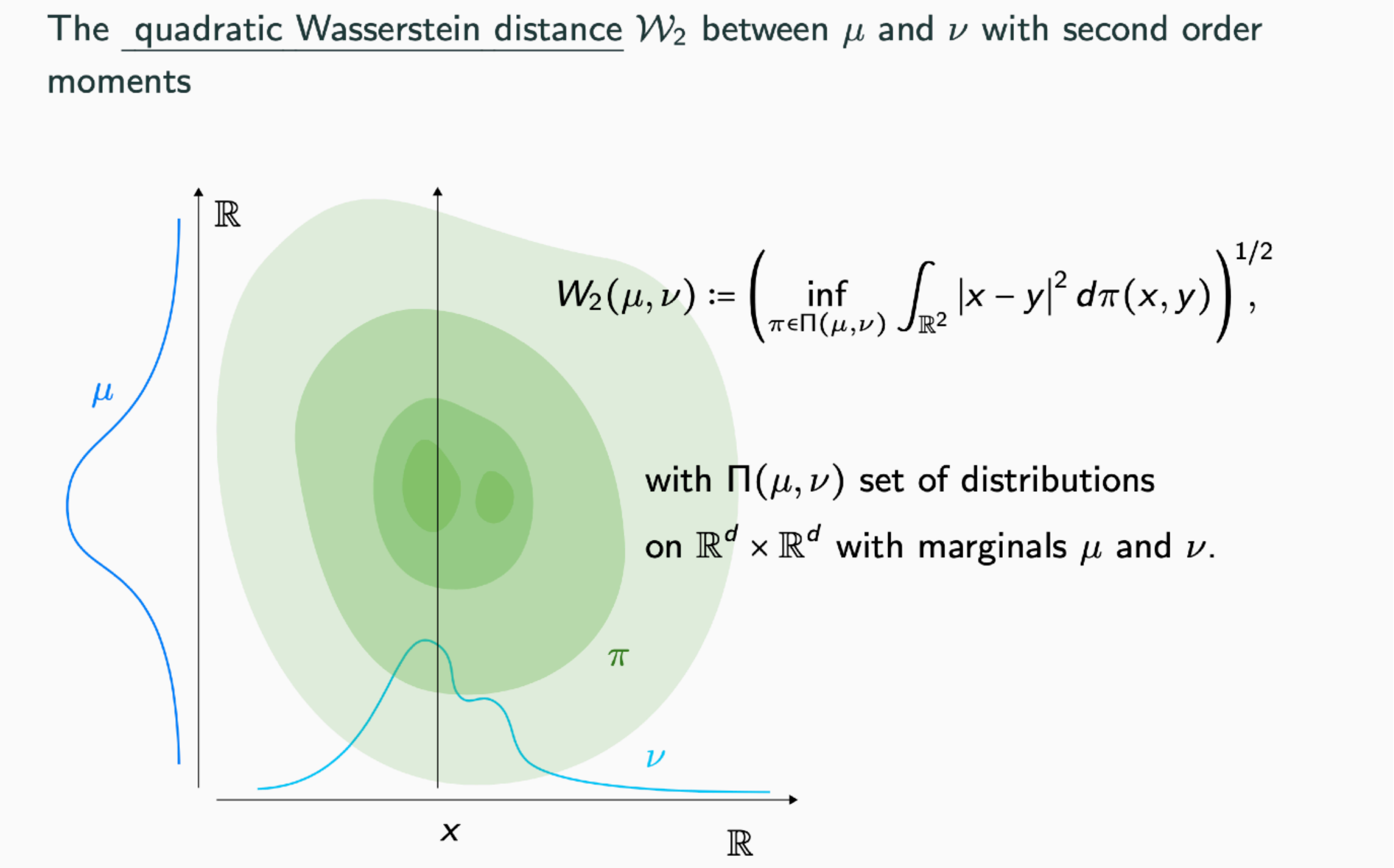
Extensions to the regression case and other applications (ranking, recommendations …)

2

# FROM OT THEORY TO FAIRNESS

*Definition of OT-based measure*

$$W_2^2(\mu,\nu) := \min_{X\sim\mu,Y\sim\nu} E\|X-Y\|^2$$



The quadratic Wasserstein distance $\mathcal{W}_2$ between $\mu$ and $\nu$ with second order moments

$$W_2(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 \, d\pi(x,y)\right)^{1/2},$$

with $\Pi(\mu,\nu)$ set of distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mu$ and $\nu$.

$$\eta_a(g) := \mathscr{L}(g(X,A)\,|\,A=a)$$

**Fairness Measure :**

$$\arg\min_{\nu} \int W_2^2(\mu_A(g),\nu)\,d\mathbb{P}(A)$$

2

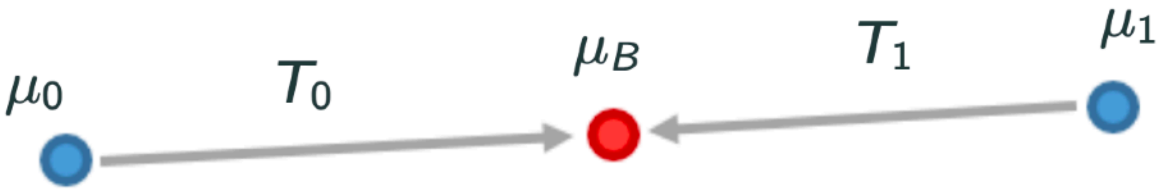
# OT FAIRNESS MEASURE IS OPTIMAL

*Wasserstein’s barycenter as an optimal fair regressor’s distribution*

* **Statistical Parity of g implies that** $\nu_a(g) = \nu(g)$ **for all A=a**

equality, achieved for the Wasserstein Barycenter of $P_{\mu_A} = \sum_{a=1}^{k} \pi_a \delta_{\mu_a}$

$$\mathscr{E}_{\text{Fair}}(\mathscr{F}) = \inf_{g \in \mathscr{F}} \mathbb{E}_A W_2^2(\mu_A, \nu(g)).$$



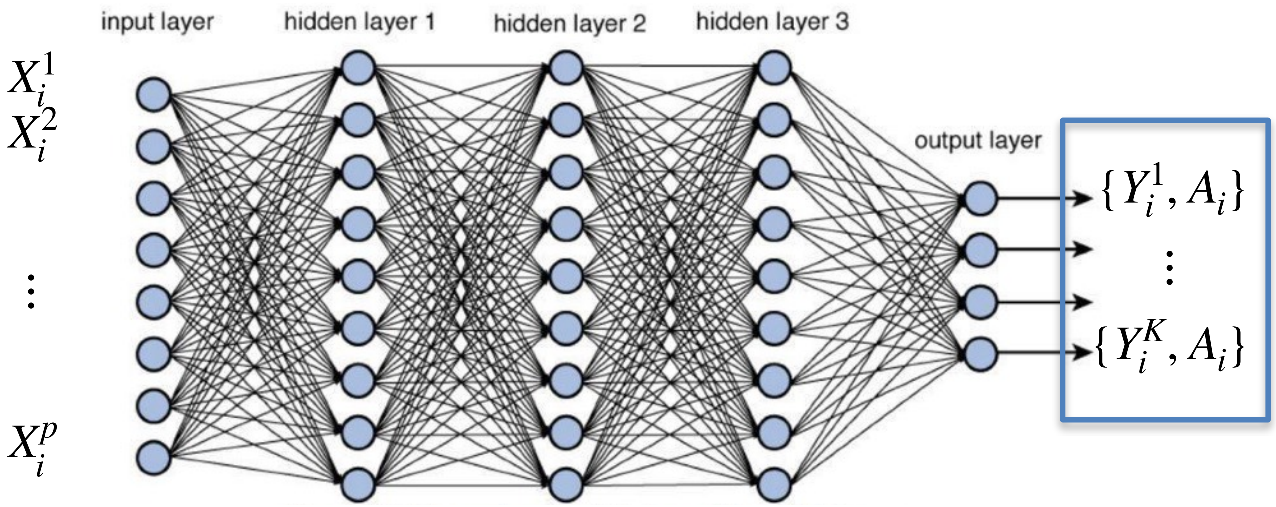
**Outcome** : New feasible Tests based on the asymptotic distribution of OT-cost (or Sinkhorn costs)

3

# BIAS MITIGATION

*Pre - In - or Post -  processing the data*

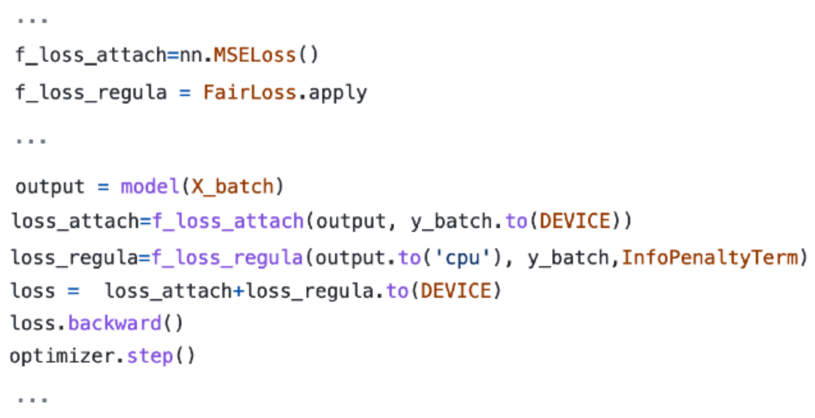
1/  **Pre-processing** the learning sample (**Fair Representations)**
2/  **Controlling** the Optimization step
3/ **Post-processing** the output of the algorithm



Parameters $\theta$

$$\hat{\theta} = \arg\min_{\theta} R(\theta) + \lambda W_2^2(\mu_{\theta,0}^n, \mu_{\theta,1}^n)$$

**Scalable regularisation term for PyTorch**

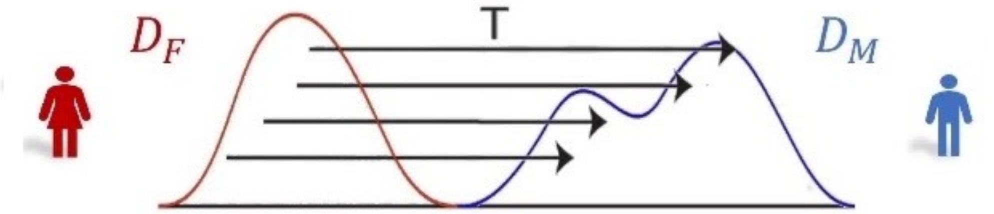


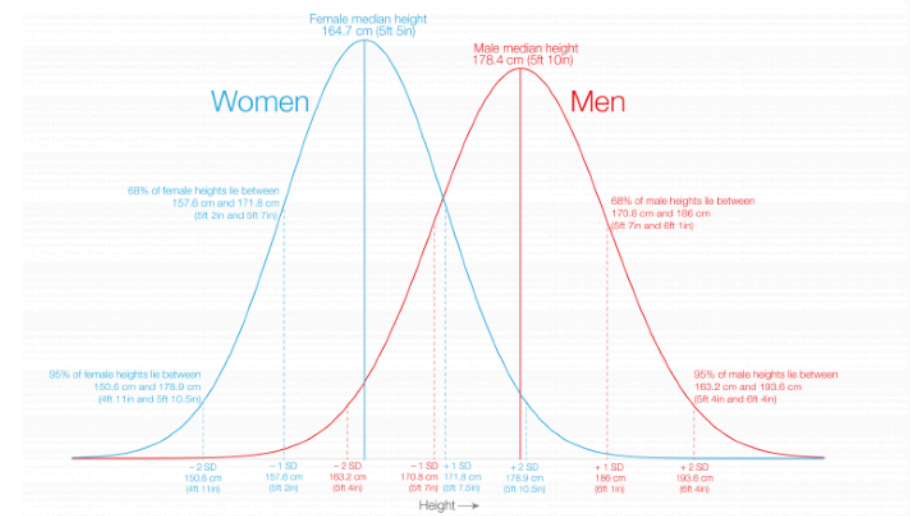
```
...
f_loss_attach=nn.MSELoss()
f_loss_regula = FairLoss.apply
...
output = model(X_batch)
loss_attach=f_loss_attach(output, y_batch.to(DEVICE))
loss_regula=f_loss_regula(output.to('cpu'), y_batch,InfoPenaltyTerm)
loss =  loss_attach+loss_regula.to(DEVICE)
loss.backward()
optimizer.step()
...
```

3

# BIAS MITIGATION OF LOCAL BIAS

## *Modeling using Counterfactual*





Note: this distribution of heights is not globally representative since it does not include all world regions due to data availability.
Data source: Jelenkovic et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts.
This is a visualization from OurWorldinData.org, where you find data and research on how the world is changing. Licensed under CC-BY by the author Cameron Appel.

**Figure 2:** Distribution of female and male height

What if I were a man ?
How would my other characteristics change accordingly ?
Counterfactuals that are plausible

Bob is 1m86 tall if Bob was Alice, he would be ?? tall

**Idea** : the counterfactual operation switching S from s to s' can be seen as a mass transportation plan pushing $\mu_{A=1}$ towards $\mu_{A=0}$

Transport-based Counterfactual Model and training with an individual biased penalty

4

# WHAT’S NEXT

*Trial : Trust & Reliability in Machine Learning*

- Understanding the **sources of bias** in the Algorithm beyond bias in the data

- From local to global : Discovering where the bias lies in the data , i.e **zones of unfairness**
Which reveals **hidden bias** (uncoded variables or intersectional)

- **Auditing** algorithms in a black box setting
**(**with a limited exploration budget as a constraint …)

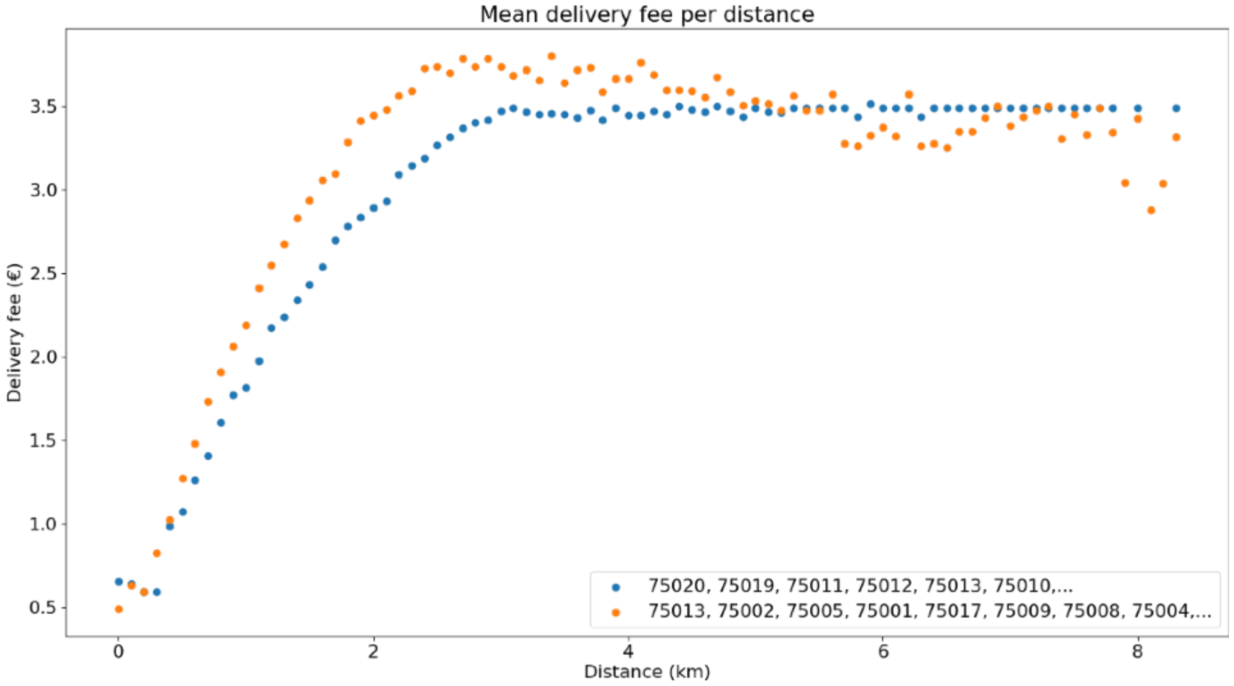
- Towards Auditing Generative Models (Chat-GPT, Stable Diffusion…. and more to come)

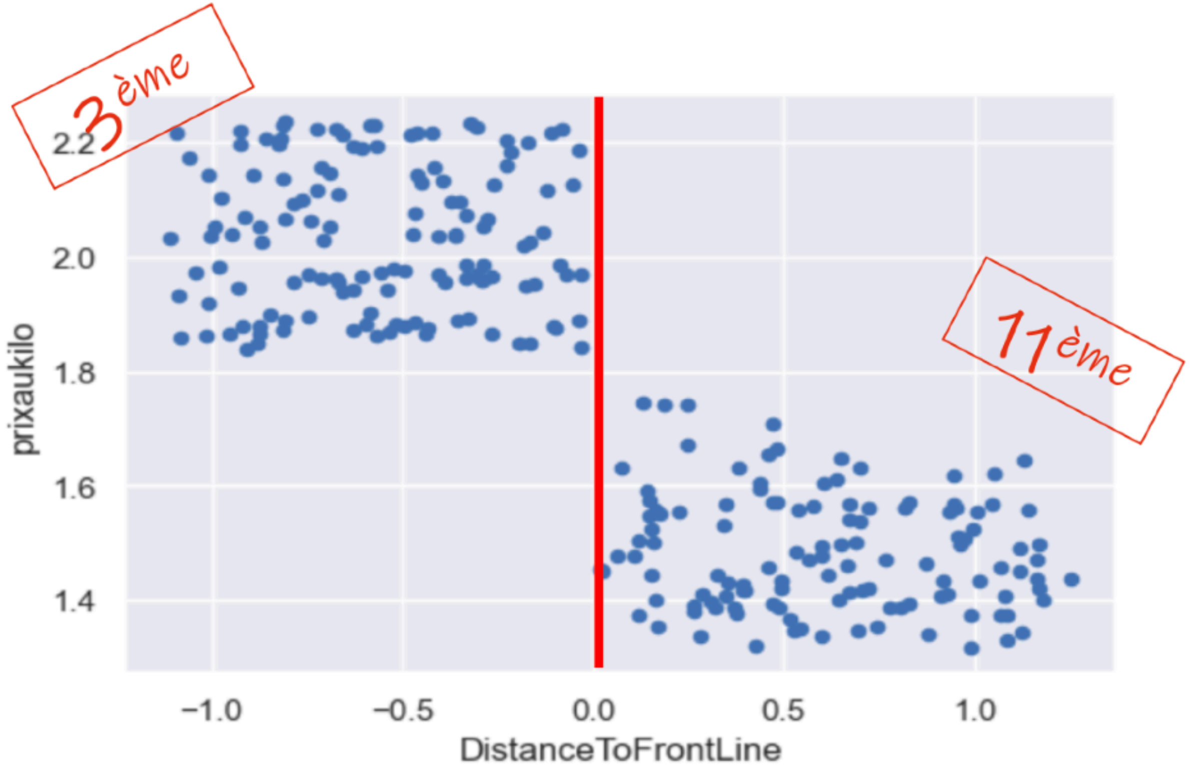
4

# WHAT’S NEXT

*Trial : Trust & Reliability in Machine Learning*

- From bias to disloyalty : finding the « true » loss function





**Disloyalty : the price depends on the location**

Merci !

# WE LC O ME

## TO THE ANITI DAYS 2023

#ANITIDAYS

𝕏 @ANITI_Toulouse

in ANITI Toulouse

# ANITI: Brief Research Overview

16th november 2023

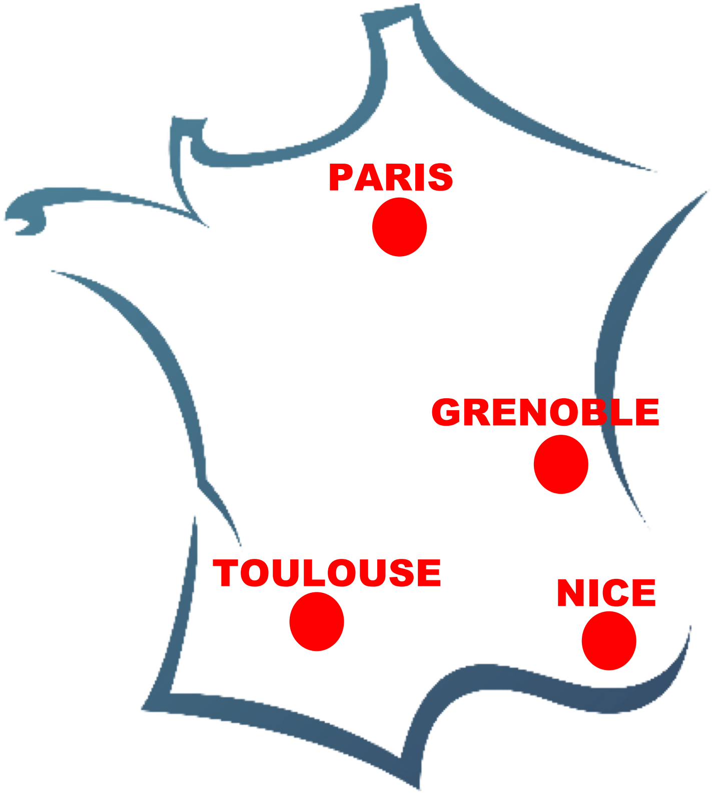
ANITI

Université de Toulouse

# What is ANITI ?

# 3iA: Interdisciplinary Institutes for AI

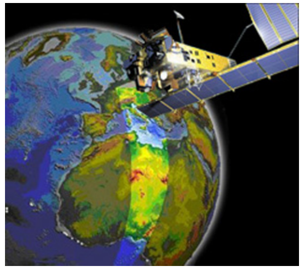
- Networked centers for research, education and economic development, with high international visibility
- Decision: April 26, 2019
- 4-year duration, renewable



ANITI

# Our original Ambition

Make possible the **sustainable** use and development of AI in **human critical applicative sectors** (transport…) **and in industry 4.0**



*Acceptability*

*Fairness*

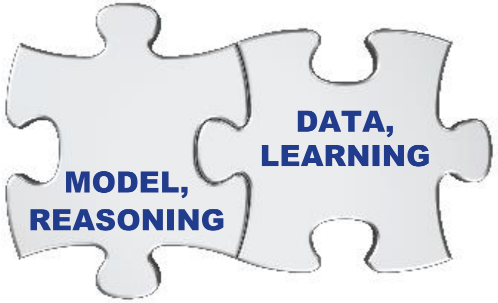
*Explainability*

*Robustness*

*Scalability*

*Adaptability*

**Hybrid AI:** efficient combination of **Model-based** & **Data-based AI**



MODEL, REASONING

DATA, LEARNING

# ANITI main topics

- Hybrid AI. Exploiting the virtues of model-based and data driven AI.
- Injecting knowledge in data driven methods
- Using analyses/models of natural intelligence to help AI and vice versa
- Multiple disciplines (statistics, automated reasoning, physical models), different, fruitful perspectives on ML, with implications for predictive maintenance, language, certifiability, robotics

# A collaborative institute

Coordinated by

Université de Toulouse

16 Research institutes and labs

7 Universities and engineering schools

31 companies

and economic clusters

121

Researchers

115

PhDs and Pods

49

Engineers

15+

Libraries

685+

(180+ in A/A* conf)

Publications

3800

(+90% vs 2018-2019)

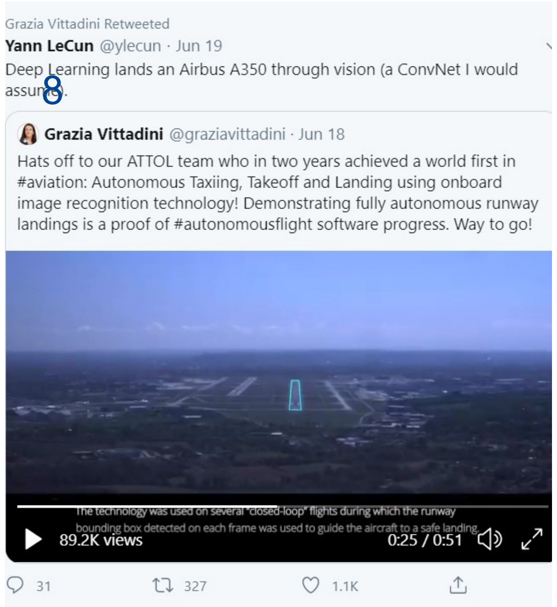
AI students in 2021-2022

# SIGNIFICANT CONTRIBUTIONS TO THE OPERATIONS OF EXISTING COMPANIES

AIRBUS

## “ANITI pave the way to go from demonstrators to certified products”

2024



Grazia Vittadini Retweeted

**Yann LeCun** @ylecun · Jun 19

Deep Learning lands an Airbus A350 through vision (a ConvNet I would assume).

**Grazia Vittadini** @graziavittadini · Jun 18

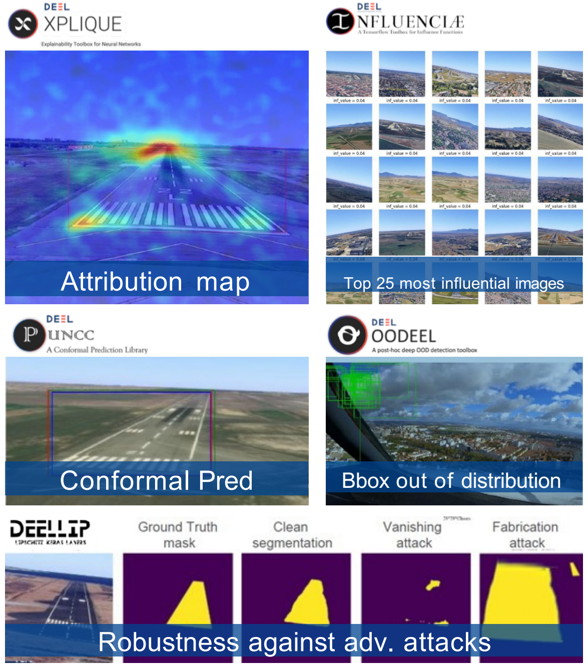
Hats off to our ATTOL team who in two years achieved a world first in #aviation: Autonomous Taxiing, Takeoff and Landing using onboard image recognition technology! Demonstrating fully autonomous runway landings is a proof of #autonomousflight software progress. Way to go!

The technology was used on several "closed-loop" flights during which the runway bounding box detected on each frame was used to guide the aircraft to a safe landing.

89.2K views 0:25 / 0:51

31 327 1.1K

2030


DEEL
XPLIQUE
Explainability Toolbox for Neural Networks
Attribution map
DEEL
INFLUENCIÆ
A Tensorflow Toolbox for Influence Functions
inf_value = 0.04
Top 25 most influential images
DEEL
PUNCC
A Conformal Prediction Library
Conformal Pred
DEEL
OODEEL
A post-hoc deep OOD detection toolbox
Bbox out of distribution
DEELLIP
LIPSCHITZ KERAS LAYERS
Ground Truth mask
Clean segmentation
Vanishing attack
Fabrication attack
Robustness against adv. attacks

NXP

## “ANITI contributes to preparing the future of automotive Advanced Drivers Assistant Systems”

| Patents | 3IA ANITI CHAIRS |
|---|---|
| Integrating neural network IP with safety | Towards certification of ML based systems |
| Collective perception with V2X and AI | AI for ATM and large scale urban mobility |
| Radar Interference mitigation with V2X and AI | |
| Multi Target extraction for automotive radar with AI | Deep learning with semantic cognitive and biological constraints |
| ML for analog circuits simulation | Fusion based inference from heterogeneous data |

Achievements :

- Significant technical findings for AI-based automotive ADAS features, for a greener & safer transportation
- 10 technical papers published (+ speakers at conferences for half of them)

# SIGNIFICANT CONTRIBUTIONS TO THE OPERATIONS OF EXISTING COMPANIES

vitesco TECHNOLOGIES

**“ANITI improves our operational excellence and our overall”**

> RUL4TEST — ANITI PhD student → VT hiring

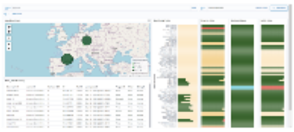
> PIKE — ANITI PhD student → VT hiring

> RUL4ROB — ANITI PhD student → ...

Quality Improvement

Process mining & predicted flow

Predictive maintenance



VITESCO TECHNOLOGIES

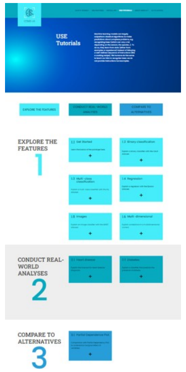
LA TRIBUNE

**IA : Vitesco et l'institut Aniti s'associent pour une industrie 4.0 plus poussée**

Renault Air Liquide Quantmetry Part of Capgemini Invent AIRBUS sopra steria

**“We are providing more explainable AI capabilities for our Engineers and Customers”**

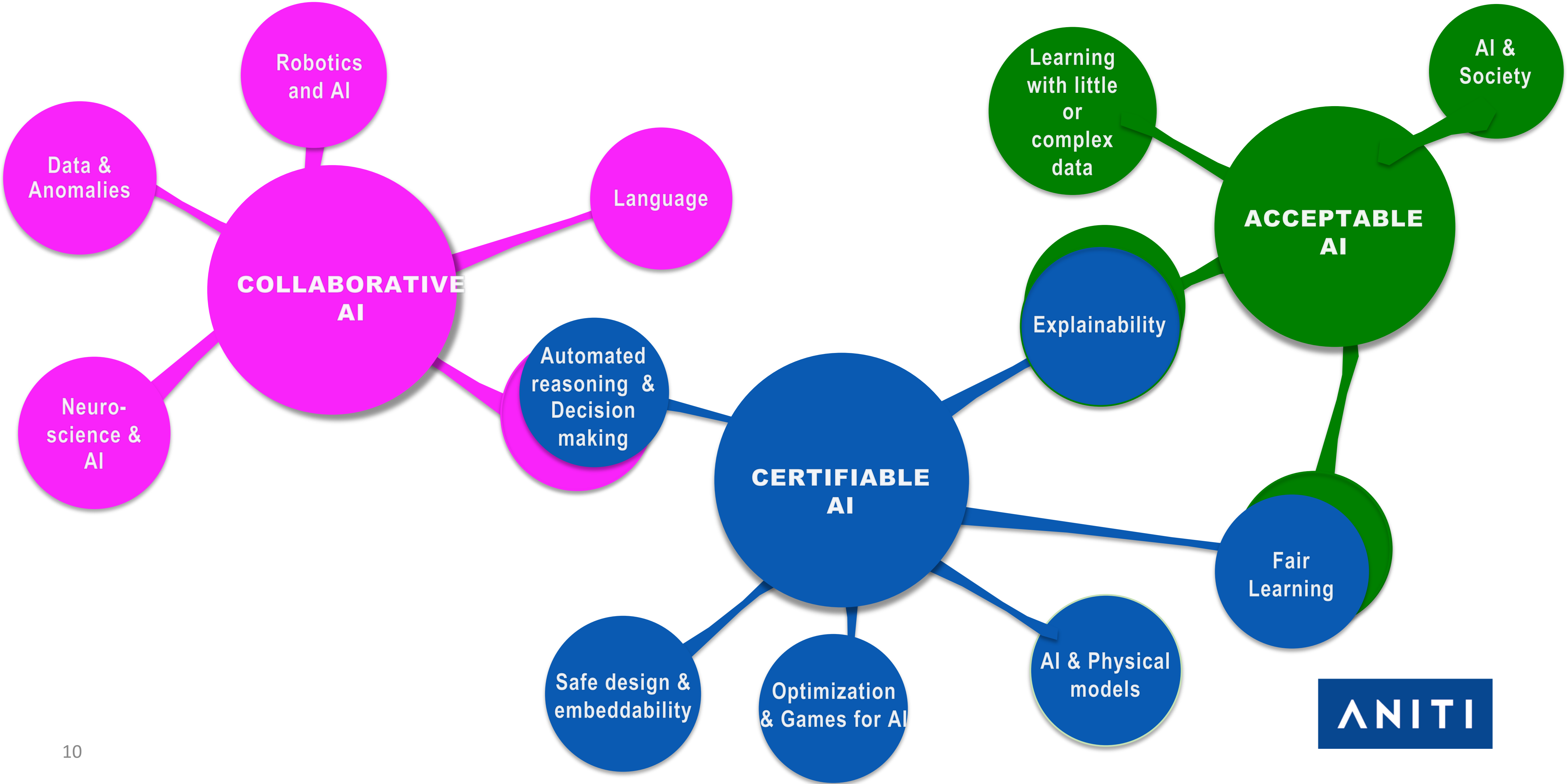
GEMS-AI
Globally Explaining Models under Stress
cnrs
ANITI
INSTITUT de MATHÉMATIQUES de TOULOUSE



machine Learning, Optimal Transport, Wasserstein Barycenter, Transfert Learning, Adversarial Learning, Robustness

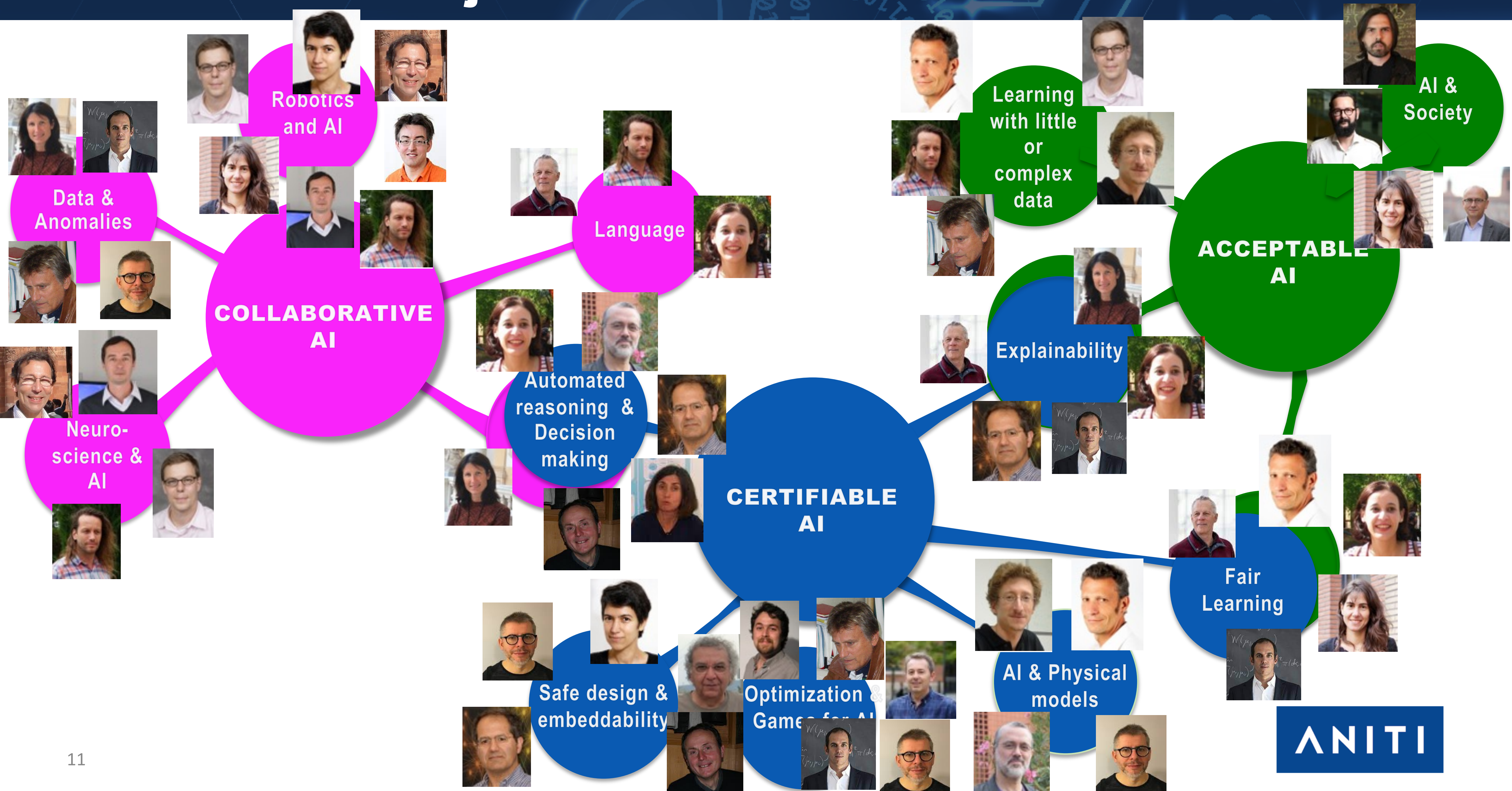
CNRS innovation prize (project Ethik-IA)

# Integrative Progams and Themes



ANITI

# Chair Pis and Projects



COLLABORATIVE AI

Robotics and AI

Data & Anomalies

Language

Neuro-science & AI

Automated reasoning & Decision making

CERTIFIABLE AI

Safe design & embeddability

Optimization & Games for AI

AI & Physical models

Fair Learning

Explainability

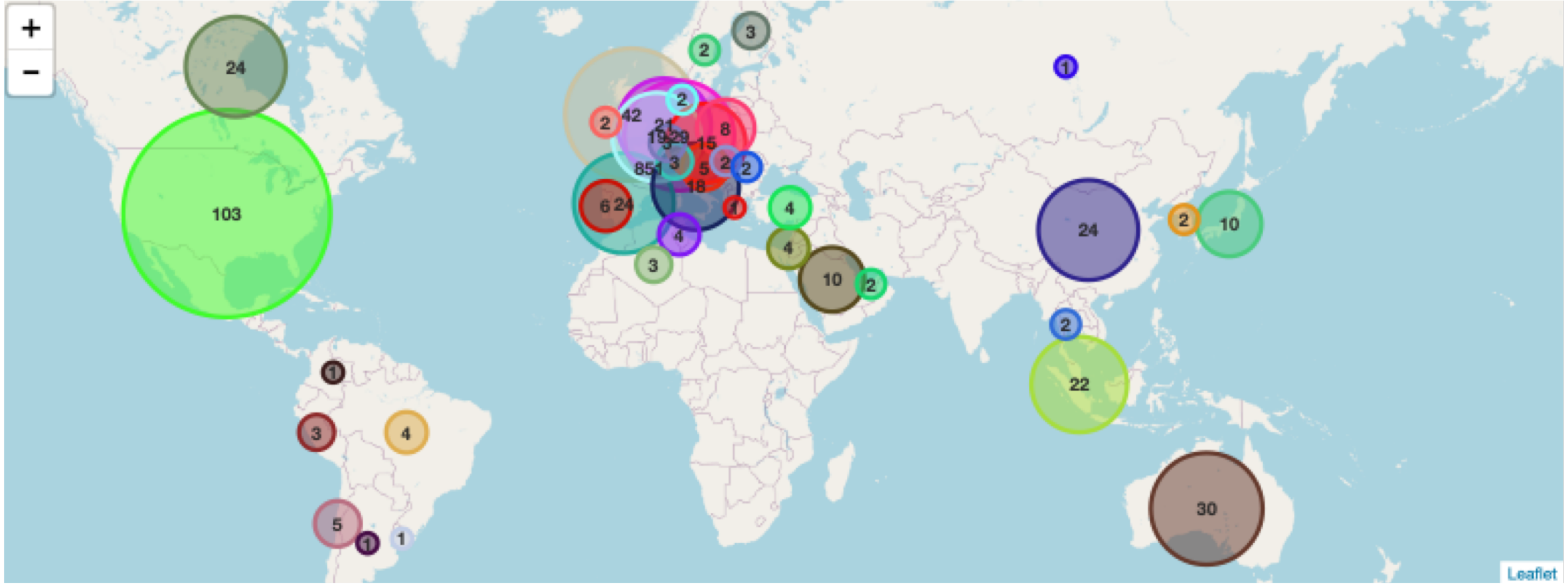
ACCEPTABLE AI

Learning with little or complex data

AI & Society

# Our assets (II) : International and national collaborations

at the international level (H2020 projects, Singapore, DEEL with Canada, collaborations with India, Germany, Japan and the United States)

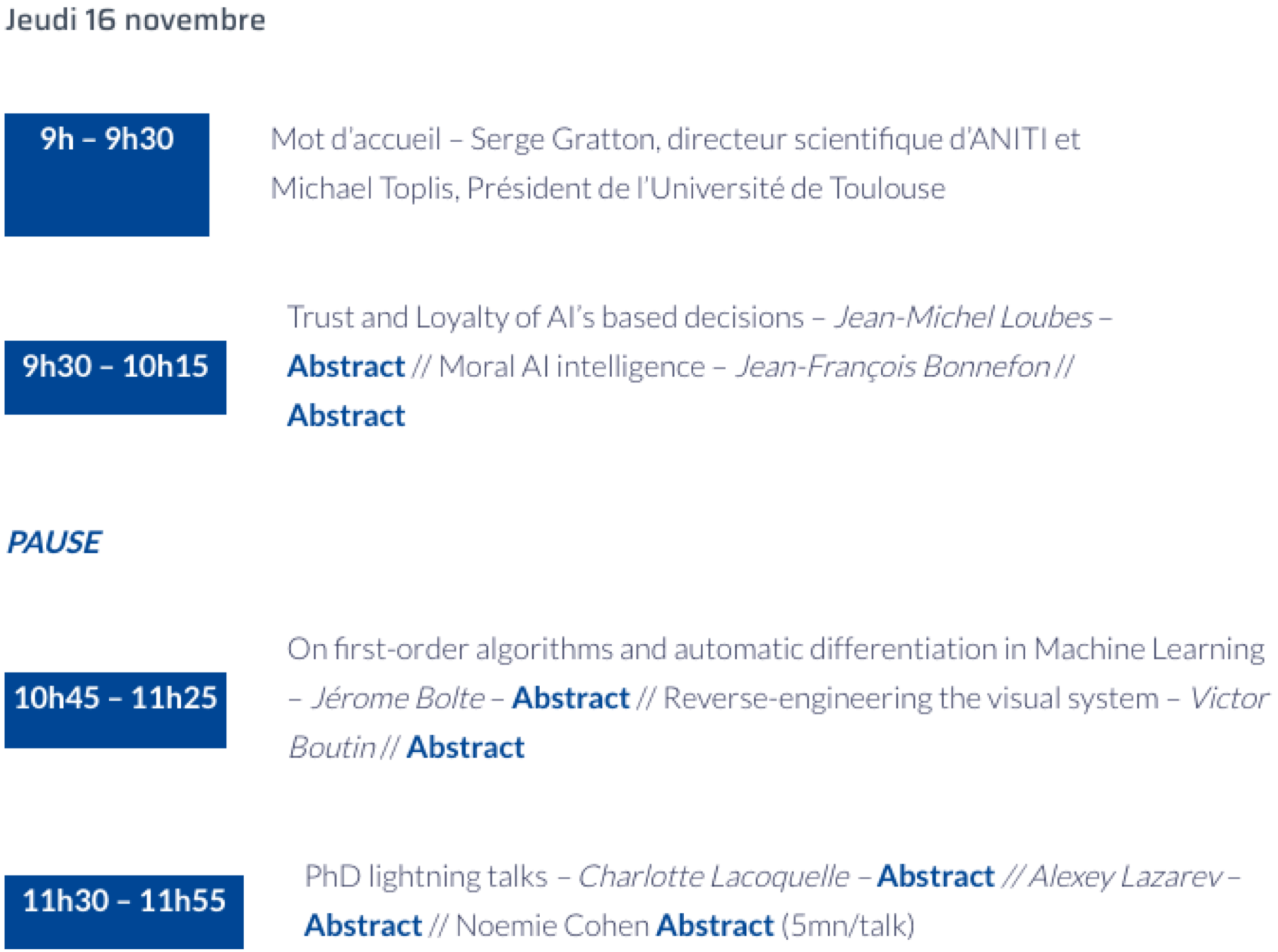


at the national level, with the **Confiance.AI** program and **3IA network**

# Coming next

- Call for clusters answered September 28th, 2023: **Sustainable ecosystem** enabling wide development of Efficient, Frugal and Trustworthy AI:
  - 400 researchers, +3000 students trained to AI, a comprehensive set of dispositive in Education, Research and Transfer
- Audition at ANR: October 30th, 2023. Answer to 8 questions on ML expertise, innovation in education, computational resources, capacity to attract talents, transfer to industry
- Informal feedback calls for reduction in amplitude of the cluster. Excellence at international level. Keep the scientific focus and consortium
- Next document to be produced by December 7th based on a formal answer not available yet

# DAY 1



Jeudi 16 novembre

**9h – 9h30** Mot d'accueil – Serge Gratton, directeur scientifique d'ANITI et Michael Toplis, Président de l'Université de Toulouse

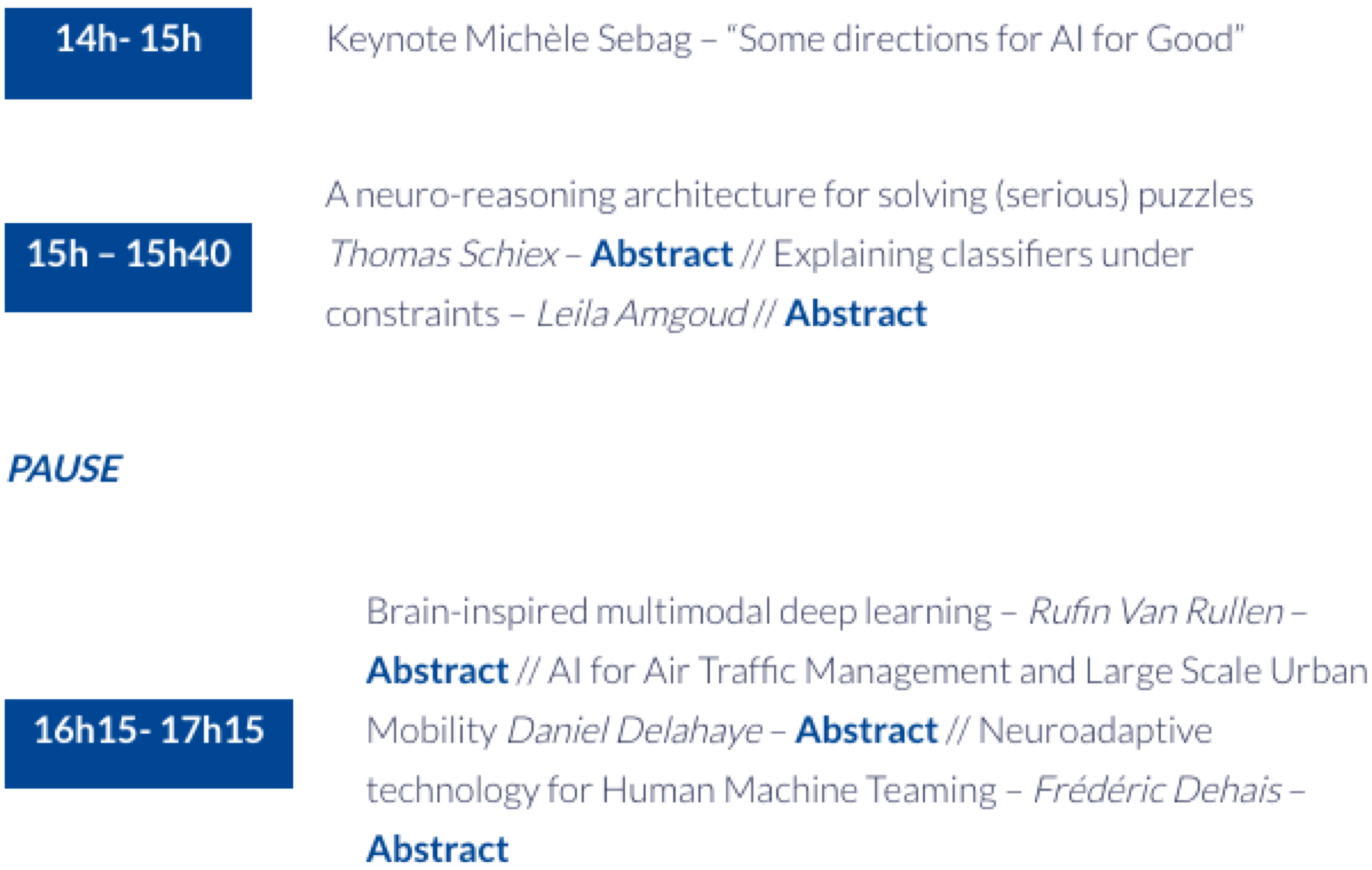
**9h30 – 10h15** Trust and Loyalty of AI's based decisions – *Jean-Michel Loubes* – **Abstract** // Moral AI intelligence – *Jean-François Bonnefon //* **Abstract**

***PAUSE***

**10h45 – 11h25** On first-order algorithms and automatic differentiation in Machine Learning – *Jérome Bolte* – **Abstract** // Reverse-engineering the visual system – *Victor Boutin //* **Abstract**

**11h30 – 11h55** PhD lightning talks – *Charlotte Lacoquelle* – **Abstract** *// Alexey Lazarev* – **Abstract** // Noemie Cohen **Abstract** (5mn/talk)

# DAY 1



| | |
|---|---|
| 14h- 15h | Keynote Michèle Sebag – "Some directions for AI for Good" |
| 15h – 15h40 | A neuro-reasoning architecture for solving (serious) puzzles *Thomas Schiex* – **Abstract** // Explaining classifiers under constraints – *Leila Amgoud* // **Abstract** |
| ***PAUSE*** | |
| 16h15- 17h15 | Brain-inspired multimodal deep learning – *Rufin Van Rullen* – **Abstract** // AI for Air Traffic Management and Large Scale Urban Mobility *Daniel Delahaye* – **Abstract** // Neuroadaptive technology for Human Machine Teaming – *Frédéric Dehais* – **Abstract** |

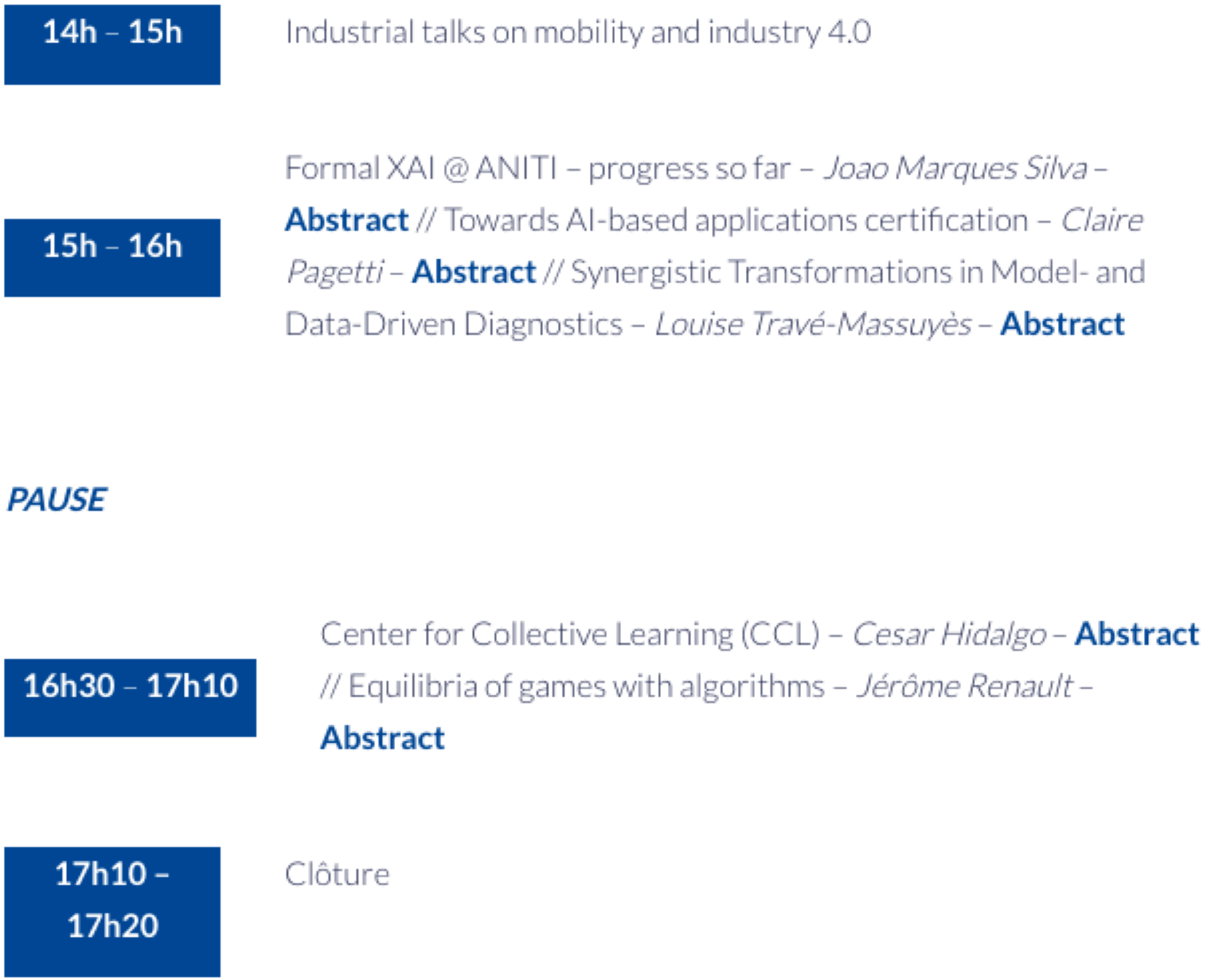
# DAY 2

**Vendredi 17 novembre**

| | |
|---|---|
| **9h30** | Welcome coffee |
| **10h – 11h** | Cognitive and interative robotics – *Rachid Alami* – **Abstract** // Artificial and Natural Movement – *Nicolas Mansard* – **Abstract** // Solving scheduling problems with Constraint Programming and Graph Neural Networks – *Florent Teichteil-Koenigsbuch & Hélène Fargier* – **Abstract** |
| **11h – 11h50** | AI for physical models with geometric tools – *Reda Chhaibi & Serge Gratton* – **Abstract** // Generative models for satellite image analysis – *Mathieu Fauvel* – **Abstract** |
| **11h50 – 12h15** | PhD Lightning talks – *Anthony Favier – Reverdi Justin – Iryna De Albuquerque* (5mn/talk) |

***PAUSE DÉJEUNER***

# DAY 2



| | |
|---|---|
| **14h – 15h** | Industrial talks on mobility and industry 4.0 |
| **15h – 16h** | Formal XAI @ ANITI – progress so far – *Joao Marques Silva* – **Abstract** // Towards AI-based applications certification – *Claire Pagetti* – **Abstract** // Synergistic Transformations in Model- and Data-Driven Diagnostics – *Louise Travé-Massuyès* – **Abstract** |
| ***PAUSE*** | |
| **16h30 – 17h10** | Center for Collective Learning (CCL) – *Cesar Hidalgo* – **Abstract** // Equilibria of games with algorithms – *Jérôme Renault* – **Abstract** |
| **17h10 – 17h20** | Clôture |