Using Computer Science to tackle unwanted bias in AI

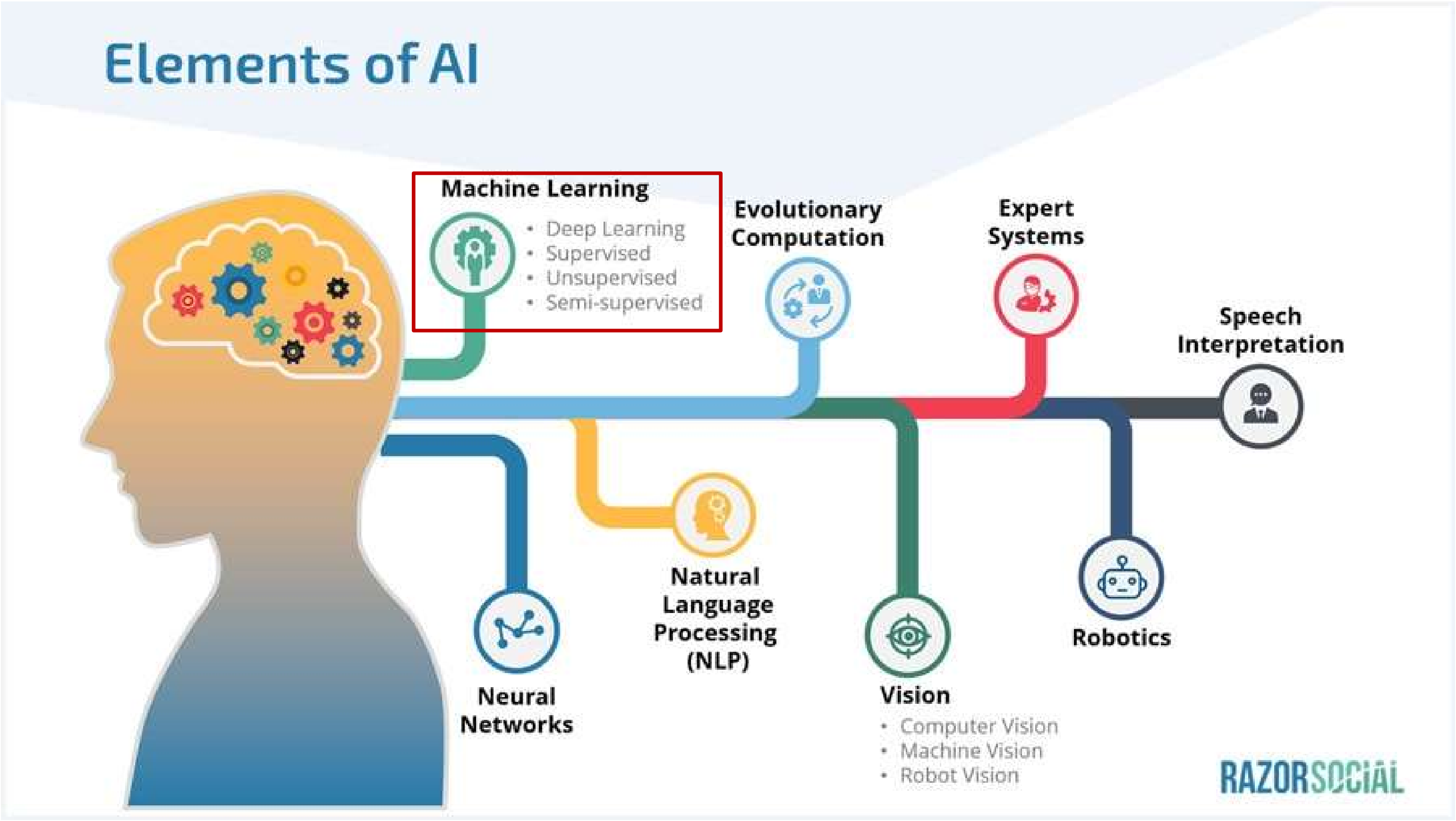
Daphne Lenders

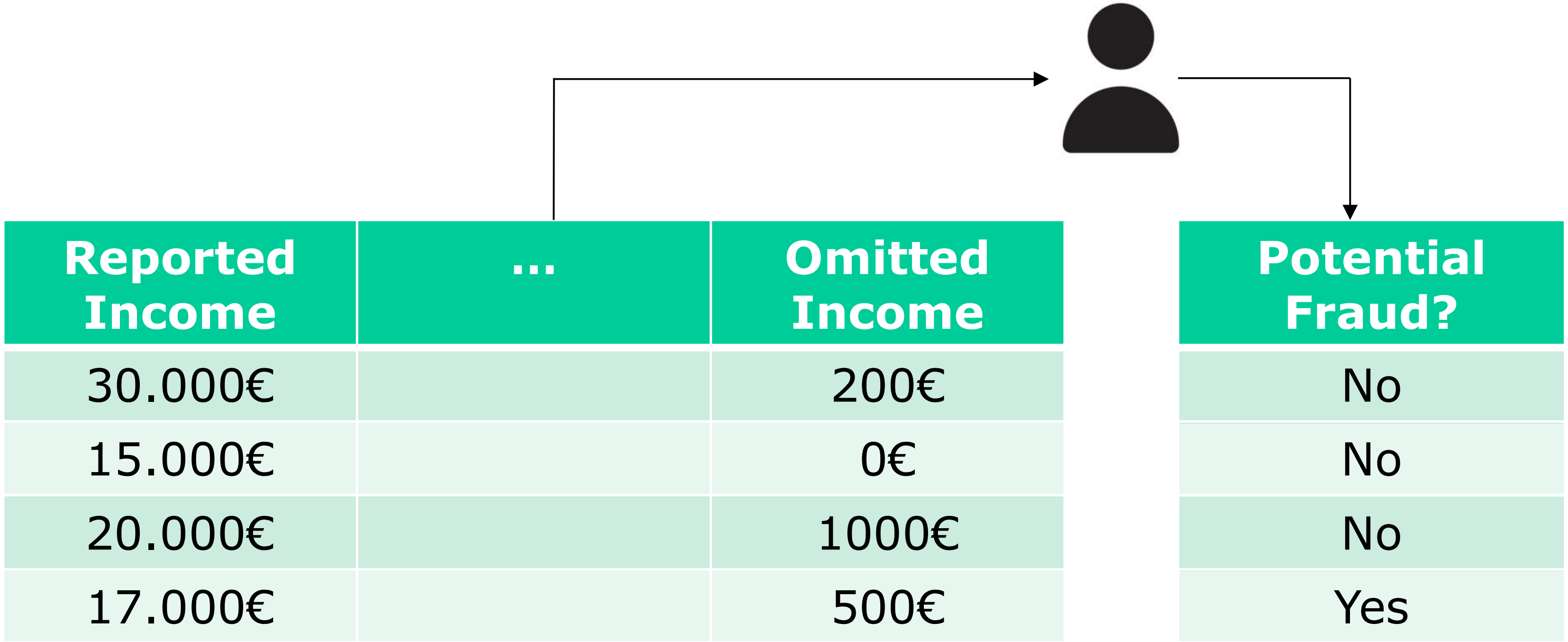
✉ daphne.lenders@uantwerpen.be

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

What is AI?


Elements of AI
Machine Learning
• Deep Learning
• Supervised
• Unsupervised
• Semi-supervised
Evolutionary Computation
Expert Systems
Speech Interpretation
Neural Networks
Natural Language Processing (NLP)
Vision
• Computer Vision
• Machine Vision
• Robot Vision
Robotics
RAZORSOCIAL



Reported Income	…	Omitted Income
30.000€		200€
15.000€		0€
20.000€		1000€
17.000€		500€

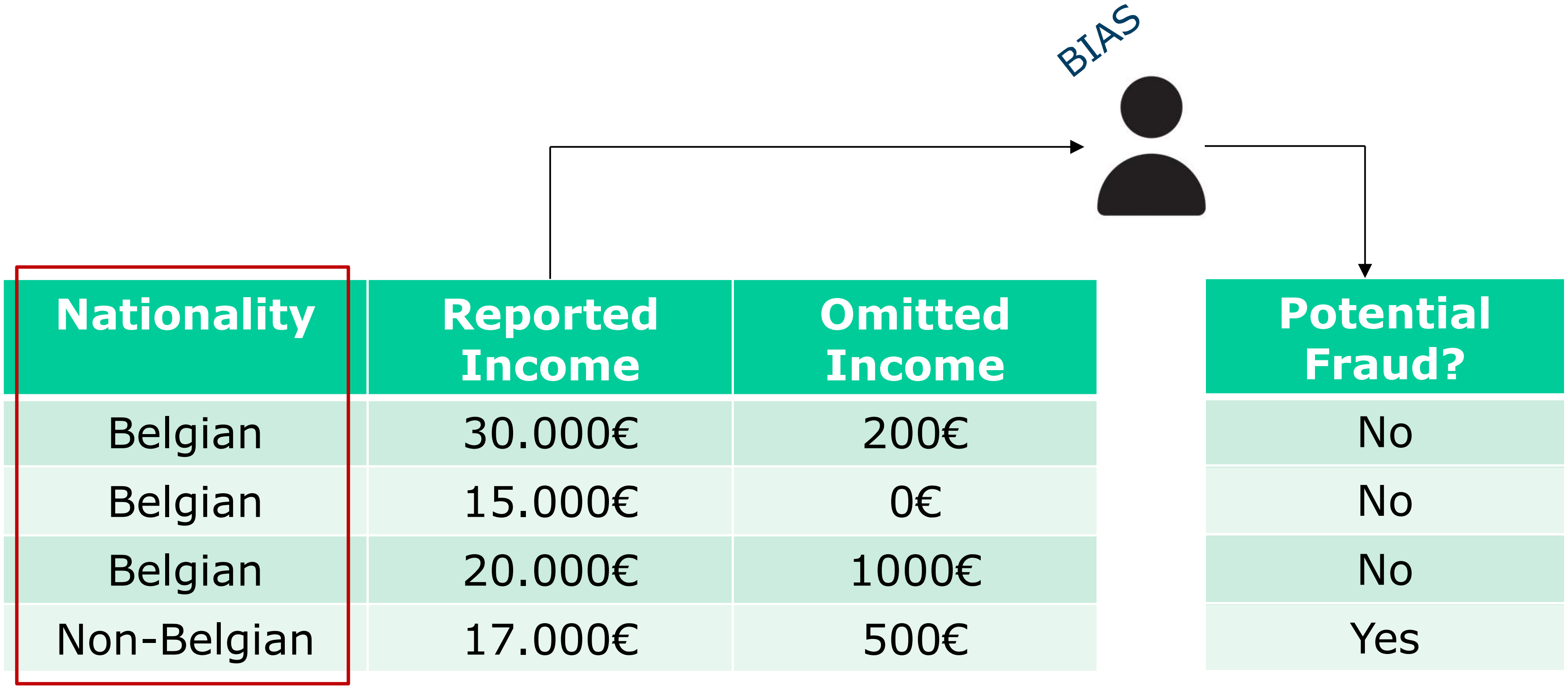
Potential Fraud?
No
No
No
Yes

Reported Income	…	Potential Fraud?
30.000€		No
15.000€		No
20.000€		No
17.000€		Yes

Fraud Prediction

Save time of tax administrators

Potential for higher accuracy



Nationality	Reported Income	Omitted Income
Belgian	30.000€	200€
Belgian	15.000€	0€
Belgian	20.000€	1000€
Non-Belgian	17.000€	500€

Potential Fraud?
No
No
No
Yes

Nationality	…	Potential Fraud?
Belgian		No
Belgian		No
Belgian		No
Non-Belgian		Yes

AI algorithms are…

- Taking over existing biases

AI algorithms are…

- Taking over existing biases

- Amplifying existing biases

Nationality	…	Potential Fraud?	AI Fraud Prediction
Belgian		No	No
…	….	…	…
Non-Belgian		Yes	Yes

AI algorithms are…

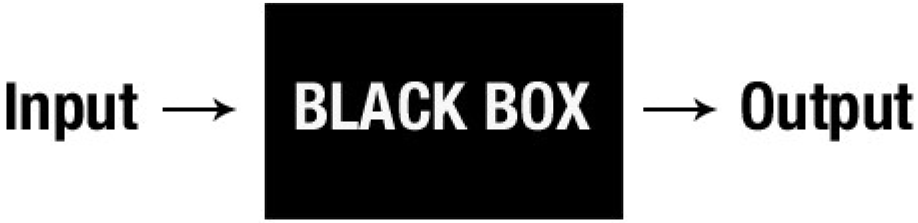
- Taking over existing biases

- Amplifying existing biases

Nationality	…	Potential Fraud?	AI Fraud Prediction
Belgian		No	No
…	….	…	…
Non-Belgian		Yes	Yes
Non-Belgian		No	Yes

AI algorithms are…

- Taking over existing biases
- Amplifying existing biases
- Black Boxes



Input → BLACK BOX → Output

So how can we measure bias?

- Demographic Parity
- Equal Opportunity
- Individual Fairness

So how can we measure bias?

- Demographic Parity
  - Comparing base rates of AI predictions
    - 10% of Belgian people are flagged
    - 30% of Non-Belgian people are flagged

    → 20% difference; thus unfair!

So how can we measure bias?

- Demographic Parity
  - Comparing base rates of AI predictions
    - 10% of Belgian people are flagged
    - 30% of Non-Belgian people are flagged

    → 20% difference; thus unfair!

What if Non-Belgian people commit more fraud?

So how can we measure bias?

- Equal Opportunity
  - Comparing how “correct” the classifier is across groups

Equal Opportunity

Belgian

Potential Fraud?	AI Fraud Prediction
No	No
No	No
No	No
Yes	Yes

Of all people who were not flagged by human, **100%** was not flagged by AI

Non-Belgian

Potential Fraud?	AI Fraud Prediction
No	No
No	Yes
Yes	Yes
Yes	Yes

Of all people who were not flagged by human, **50%** was not flagged by AI

Equal Opportunity

Belgian

Potential Fraud?	AI Fraud Prediction
No	No
No	No
No	No
Yes	Yes

Non-Belgian

Potential Fraud?	AI Fraud Prediction
No	No
No	Yes
Yes	Yes
Yes	Yes

This algorithm does not satisfy equal opportunity!

Equal Opportunity

Belgian

Potential Fraud?	AI Fraud Prediction
No	No
No	No
No	No
Yes	Yes

Non-Belgian

Potential Fraud?	AI Fraud Prediction
No	No
No	No
Yes	Yes
Yes	Yes

This algorithm does satisfy equal opportunity

But what about human bias?

So how can we measure bias?

- Equal Opportunity
  - Comparing how “correct” the classifier is across groups

  - Makes sure that existing bias is not increased
  - Does not solve the problem of bias in human labels

So how can we measure bias?

- Individual Fairness
  - Determine for one individual at a time whether s/he got discriminated

But how to measure this?

Measuring bias – No Silver Bullet

- There’s no such thing as a perfect definition of fairness

- How we measure bias may depend on the problem
  - Do we want to fundamentally change a decision process?
  - Are we okay with making mistakes?
  - How high is the risk of preserving existing biases?

- Legal Definitions?

How to tackle bias?

- Many approaches
- Again no silver bullet
- One example: Situation Testing

Situation Testing

- Approach taken from Social Sciences
- Trying to achieve Individual Fairness
- Idea: similar people should be treated alike

Nationality: Belgian

Reported Income: 20000€

Omitted Income: 500€

NO FRAUD

Nationality: Non-Belgian

Reported Income: 20000€

Omitted Income: 500€

FRAUD

Situation Testing

- Can’t always find “equal” people (that only differ on sensitive attribute)
- Look at “similar” people instead
- How to define similarity?
  - Computer Science can help here!

Nationality	Reported Income	Omitted Income	Potential Fraud?
Belgian	30.000€	200€	No
Belgian	20.000€	1000€	No
Non-Belgian	17.000€	500€	Yes

Situation Testing

- Still more questions
    - How many people do we have to compare?
    - How similar is similar enough?

Conclusion

- Tackling bias in AI has gained interest but …
  - No silver bullet
  - Still many open questions
  - Lack of a clear legal framework

It's Time to Combine Forces