Bias in AI Systems in Financial Services

Ángel Pavón Pérez

Supervisors: Harith Alani, Hasan Al-Madfai, Gregoire Burel and Miriam Fernandez

The Open University

Knowledge Media Institute

AI and Society

Current problem

Forbes

ENTERPRISE TECH

Artificial Intelligence Opens Up The World Of Financial Services

Financial Systems might be trained in data coming from our society and therefore, learn historical biases.

- **Biased data**

Credit amount	Purpose	Housing	Gender	Class
10000	car	renting	female	Bad risk
3000	repairs	own house	male	Good risk
50000	business	renting	female	Bad risk
5000	repairs	own house	male	Good risk
100000	house	own house	male	Bad risk

1st Problem

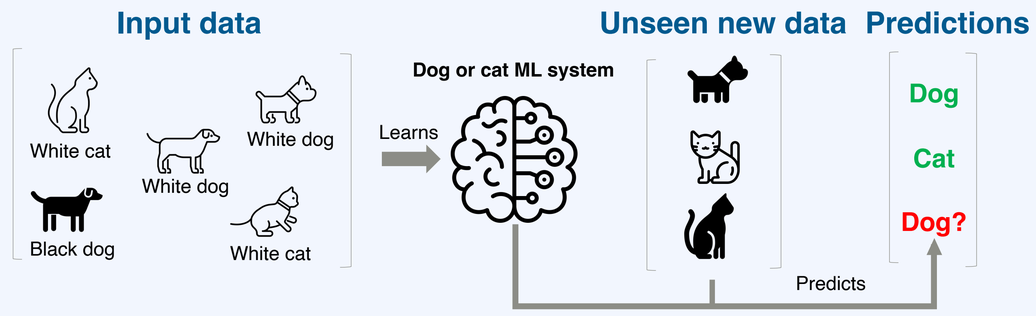
- All females in the sample have bad risk.
- A model can learn that *female → bad risk*

Remove gender?

2nd Problem

- All females in the sample have a rented house so removing gender and leaving housing will not solve the problem.
- A model can learn that *renting → bad risk*

How Machine Learning works



Current legislation doesn't allow to store sensitive information in financial systems which makes it difficult to analyse systems bias

Methods to deal with bias

- **Pre-training**: Modify data
- **In-Training**: Modify how the model learns
- **Post-Training**: Modify the model predictions

Some methods can be considered positive discrimination (Which might be illegal in some countries like the UK)

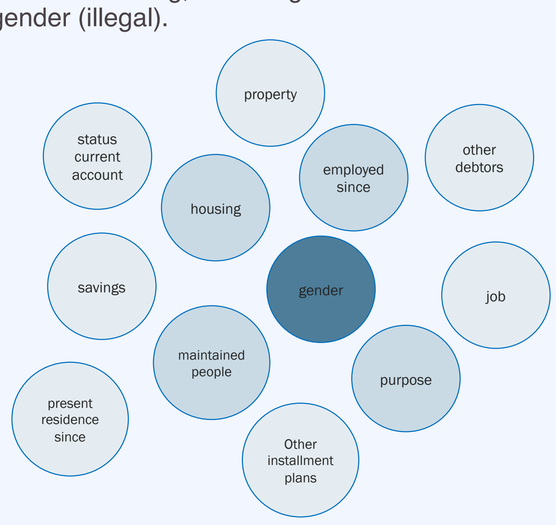
Experiments paper

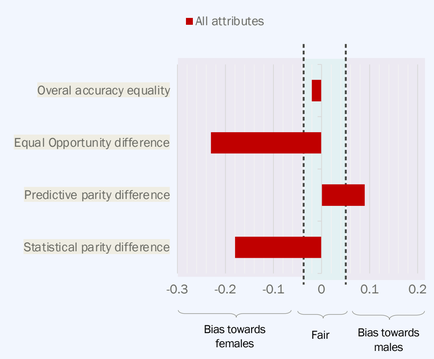
Experiments: How gender indirect representation can bias ML

We train three Machine Learning (ML) models for predicting the credit risk.

Model with all attributes

Model in which all available attributes have been used for training, including the sensitive attribute, gender (illegal).

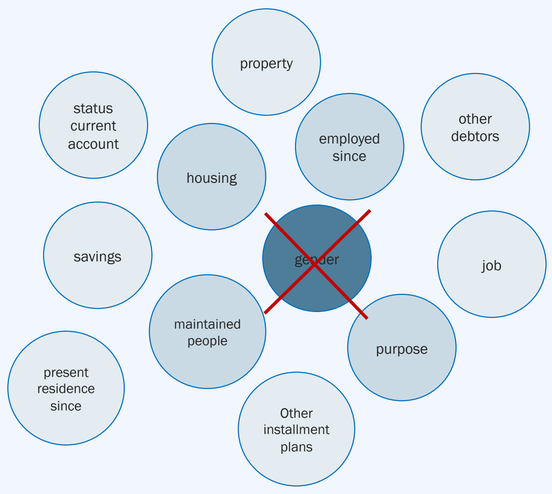




~20% more males than females get a good credit score

Model without gender

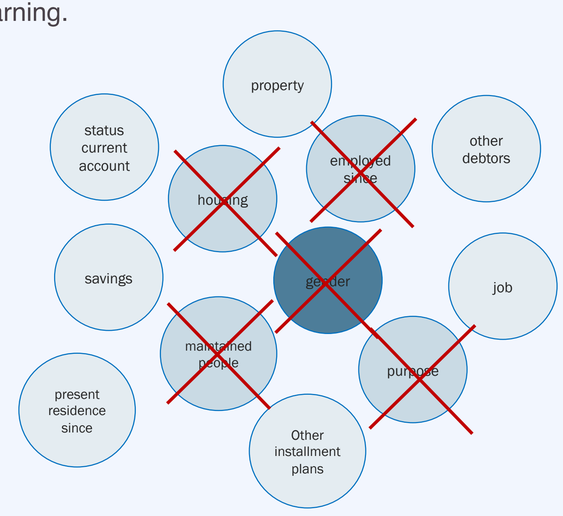
Model trained with all attributes except for our sensitive attribute, gender (current law).

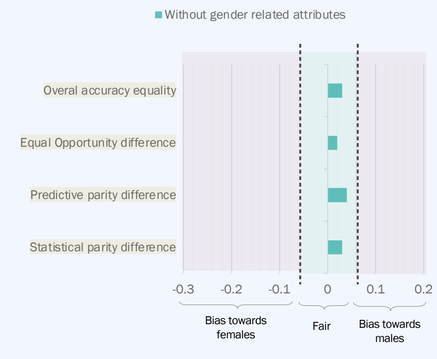


Indirect representation. Removing just gender is not enough.

Model without gender related attributes

Model trained using only the attributes considered gender-independent by statistical tests and machine learning.





Removing the identified related attributes reduces bias.

Current legislation doesn't consider gender indirect representation (just clear proxies like familiar status)

Contact

Ángel Pavón Pérez
Knowledge Media Institute, The Open University, Milton Keynes, UK
angel.pavon-perez@open.ac.uk