Addressing algorithmic fairness through metrics and explanations

IDAI 2021 Summer School
Course 5

Catuscia Palamidessi, Miguel Couceiro

Speakers

Catuscia Palamidessi (catuscia@lix.polytechnique.fr)

Research Director at **Inria Saclay**, head of the **Comète** team. Catuscia's research interests include Machine Learning, Privacy, **Fairness**, Secure Information Flow, and Concurrency.

She is a recipient of an ERC advanced grant on privacy, and she is PI in various other projects including a French-German cooperation on cybersecurity. She is in the Editorial board of various journals, including the IEEE Transactions on Dependable and Secure Computing and the Journal of Computer Security. She is member of the advising committee of the French National Information Systems Security Agency (ANSSI).

Miguel Couceiro (miguel.couceiro@loria.fr)

Professor at the **University of Lorraine** (UL), head of the **ORPAILLEUR** team at **LORIA**. Miguel's research interests include Knowledge Discovery, Multicriteria Decision Making, and **Fair and Explainable Models**.

He is a member of the H2020 TAILOR project and of the Inria project Lab HyAIAI. He is the local coordinator of the European Erasmus Mundus Masters program LCT and the head of the 2nd year of the NLP Master's program at the University of Lorraine.

Fairness as non-discrimination

Fair model: that protects salient groups against discrimination

Discrimination: "unjust or prejudicial treatment of different categories of people, especially, on the grounds of race, age, or sex"

Example: Decision Making process...

- Human: Objective & Subjective reasoning
- Machine: Only objective **but** ...

Motivation: unfair algorithmic decisions

Algorithmic decisions: are objective **but** they can be unfair

Common "sources": Data Collection & Model choice/design

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

https://www.bbc.com/news/business-50365609

Motivation: unfair algorithmic decisions

https://www.bbc.com/news/technology-35902104

https://www.businessinsider.com/how-algorithms-can-be-racist-2016-4?IR=T

Other Critical applications of algorithmic decisions: loan requests, job applications, Stop & Frisk, etc.

Need of fairness: Unfair outcomes not only affect human rights, but they undermine public trust in ML & AI.

Guidelines/Rules: GDPR (in the EU), CCPA (in the US), etc.

Defining "fairness" in ML & AI

Based on **decision outcomes**, fairness can be assessed based on:

- **Fairness metrics**: individual & group fairness, equal opportunity, demographic parity, equal accuracy, etc.
- **Process fairness**: model's reliance on "sensitive features" (e.g., salient features such as race, age, or sex,. . .)

Two main approaches to dealing with ML unfairness:

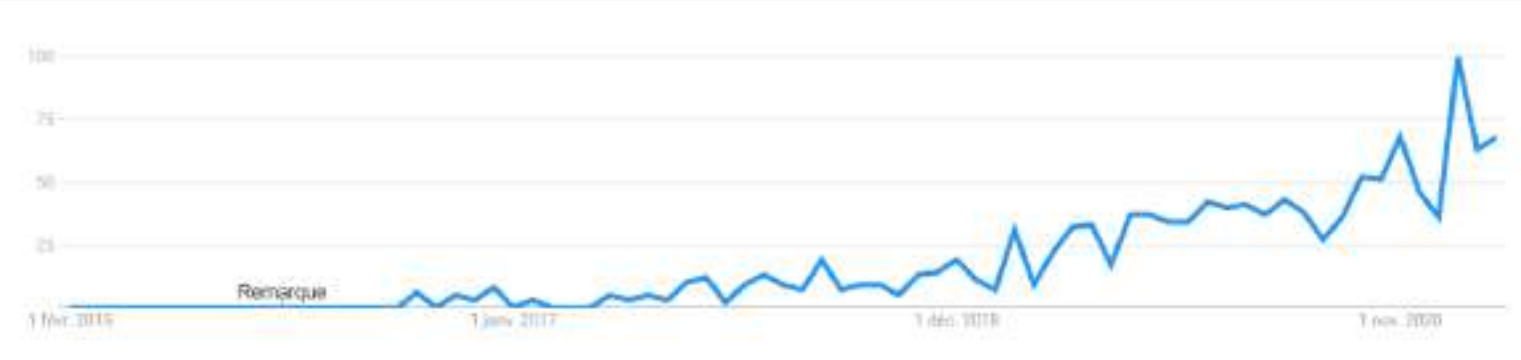
1. Enforce fairness constraints while learning
2. Forget sensitive/salient features

NB: No free lunch!

Fair and Explainable AI

This has propelled the interest in the design of fair and transparent AI systems

Explainable AI



Fairness AI

Source: `trends.google.fr`

Venues on Explainable, Fair & Trustworthy AI

Fairness and Bias in AI/ML:

- **FAccT**: ACM Conference on Fairness, Accountability, and Transparency
- **AIES**: AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society
- **FORC**: Symposium on Foundations of Responsible Computing
- **BIAS**: Int. Workshop on Algorithmic Bias in Search and Recommendation

Explainable AI/ML:

- **XAI**: eXplainable AI at AAAI
- **XKDD**: eXplainable Knowledge Discovery in Data Mining at ECML-PKDD
- **AIMLAI**: Advances in Interpretable Machine Learning and AI at ECML-PKDD

Seminars:

- **TrustML**: Bi-weekly Seminar Series of The Trustworthy ML Initiative

Other venues on Explainable, Fair & Trustworthy AI

Projects:

- **TAILOR** ICT-48 project: Foundations of Trustworthy AI integrating Learning, Optimisation and Reasoning
- **IPL HyAIAI**: Hybrid Approaches for Interpretable AI
- **NoBIAS**: A Marie Skłodowska-Curie Innovative Training Network

Tools:

- **FAT Forensics**: Python toolkit for evaluating Fairness, Accountability and Transparency of AI systems
- **AI Fairness 360**: IBM Open source toolkit for examining, reporting, and mitigating discrimination and bias in ML models
- **FixOut**: Python toolkit for rendering ML models fairer using explanations, feature dropout and ensembles

Course 5: Addressing algorithmic fairness through...

1st part (10h - 11h)

- Introduction, presentation and motivation
- Fairness notions and metrics
- Relation between fairness notions
- Relation between fairness, accuracy and privacy

2nd part (11h30 - 13h)

- Tackling data biases: balancing datasets
- Using explanations for assessing process fairness
- Addressing unfairness through unawareness: feature dropout and aggregation
- Some use cases and available resources
- Discussions