Fairness, Accountability, Confidentiality, and Transparency (**F**ACT) in AI

9 April 2020

Ana Lucic
University of Amsterdam

FACT-AI

Fairness, Accountability, Confidentiality and Transparency in AI

> *"Algorithmic systems are being adopted in a growing number of contexts, fueled by big data. These systems filter, sort, score, recommend, personalize, and otherwise shape human experience, increasingly making or informing decisions with major impact on access to, e.g., credit, insurance, healthcare, parole, social security, and immigration. Although these systems may bring myriad benefits, they also contain inherent risks, such as codifying and entrenching biases; reducing accountability, and hindering due process; they also increase the information asymmetry between individuals whose data feed into these systems and big players capable of inferring potentially relevant information."*

https://facctconference.org

FACT-AI

Fairness

- AI system should avoid discrimination across people and communities.

Accountability

- AI system should be able to justify its recommendations or actions to users and other stakeholders, and be reliable at all times.

Confidentiality

- The output or actions of the AI system should not reveal secrets.

Transparency

- The AI system should be able to explain to users and other interested stakeholders why and how its results were obtained.

FACT-AI

Fairness

- **AI system should avoid discrimination across people and communities.**

Accountability

- AI system should be able to justify its recommendations or actions to users and other stakeholders, and be reliable at all times.

Confidentiality

- The output or actions of the AI system should not reveal secrets.

Transparency

- The AI system should be able to explain to users and other interested stakeholders why and how its results were obtained.

FACT-AI

Fairness, Accountability, Confidentiality and Transparency in AI

“Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people’s behavior. As a result, algorithms can reinforce human prejudices.”

https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html

Overview of Talk

- Motivating Examples
- Types of Bias
- Sources of Unfairness
- Mitigating Unfairness
  - Identifying protected characteristics
  - Fairness objectives
  - Fairness interventions
    - Pre-processing data
    - Modifying algorithm
    - Post-processing outputs
  - Evaluating and monitoring fairness interventions
- Conclusion

Motivating Examples

PREDICTIVE POLICING: USING MACHINE LEARNING TO DETECT PATTERNS OF CRIME

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers lack, like making conversation and reading social cues.

But people have biases and predilections. They make hiring decisions, often unconsciously, based on similarities that have nothing to do with the job requirements — like whether an applicant has a friend in common, went to the same school or likes the same sports.

That is one reason researchers say traditional job searches are broken. The question is how to make them better.

A new wave of start-ups — including Gild, Entelo, Textio, Doxa

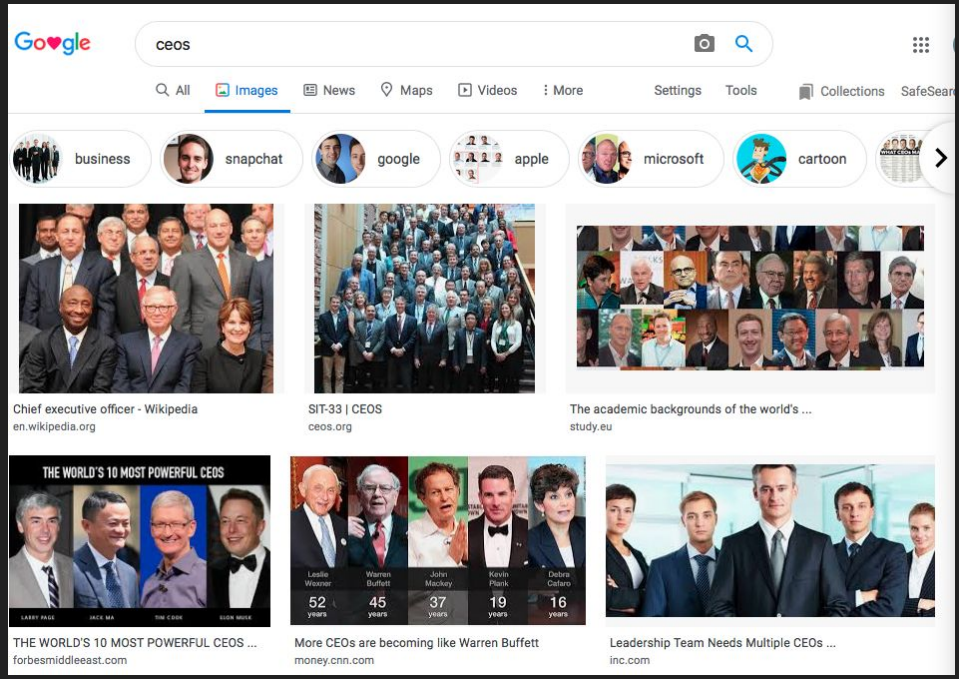
Motivating Examples

Beauty contest judged by AI and the robots discriminate against dark skin

3 days ago | Published by : Avinash Nandakumar

Septo: The first international beauty contest judged by "machines" was supposed to use objective factors such as facial symmetry and wrinkles to identify the most attractive contestants. After Beauty.AI launched this year, roughly 6,000 people from more than 100 countries submitted photos in the hopes that artificial intelligence, supported by complex algorithms, would determine that their faces most closely resembled "human beauty".

But when the results came in, the creators were dismayed to see that there was a glaring factor linking the winners: the robots did not like people with dark skin.



Motivating Example: COMPAS Recidivism

- When a defendant is convicted of a crime, a judge needs to decide how long the sentence will be.

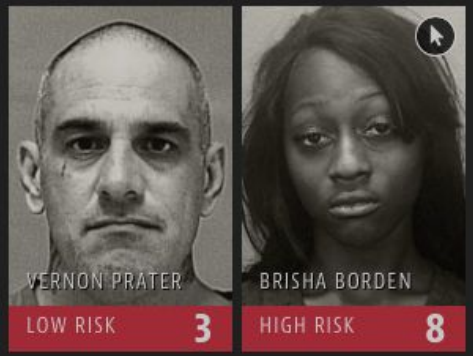
- In the US, an algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is sometimes used to help with this.
  - Predicts whether someone is “high risk” or “low risk” of committing another crime.
  - “High risk” implies longer prison time.

Machine Bias: https://v2-www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Motivating Example: COMPAS Recidivism

- Although COMPAS **does not use race explicitly**, it systematically makes different types of mistakes for black people compared to white people.
  - Black people are more likely to be **incorrectly predicted as "high risk".**
  - White people are more likely to be **incorrectly predicted as "low risk".**

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

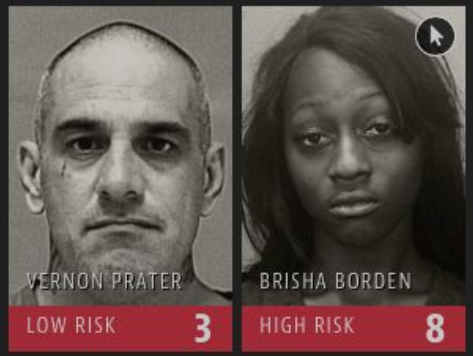
Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Motivating Example: COMPAS Recidivism

- How is this possible?
  - Model includes **proxy variables** such as financial problems, social environment, and residential instability, which can be correlated with race.

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

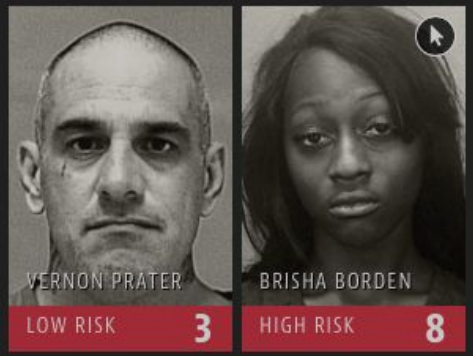
Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Motivating Example: COMPAS Recidivism

- Result: Model **accidentally learns race** and discriminates against people based on it, even though race is **not explicitly shown** to the model!
- Goal: avoid such “accidental” learning of protected characteristics.

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Algorithmic Fairness

- In the context of algorithmic decision-making, fairness is:

  *The absence of any prejudice or favoritism towards an individual or group based on **protected characteristics such as race, gender, etc.***

- An algorithm is *unfair* if treats different groups differently based on protected characteristics.
  - e.g., An algorithm makes mistakes about males more often than about females.
- Unfairness is usually the result of some sort of *bias.*

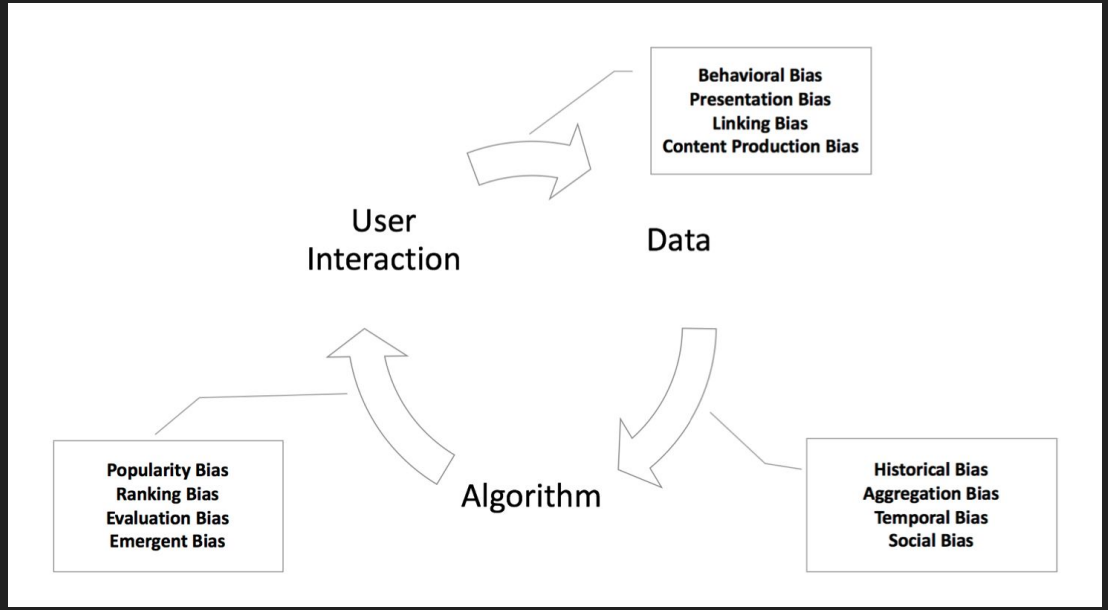
Types of Bias

Mehrabi et al. mention **23 types** of bias:

- Historical bias
- Representational bias
- Measurement bias
- …
- Observer bias
- Funding bias

Feedback loop mechanism

- ML model makes decisions, its outcomes affect future data collected for subsequent training



https://arxiv.org/abs/1908.09635

Sources of Unfairness

- The world is unfair:
  - Model trained on data obtained from this world simply reflects this.
- Data might encode existing **biases:**
  - e.g. Historical bias – lack of female CEOs
  - e.g. Labels are not “Committed a crime?” but the proxy “Was arrested.”
- Data collection feedback loops
  - e.g. Only observe “Paid back loan?” if the loan was granted, counterfactual case is unavailable.
- Different populations have different properties
  - e.g. “SAT score” might correlate with label differently in populations employ SAT tutors more.

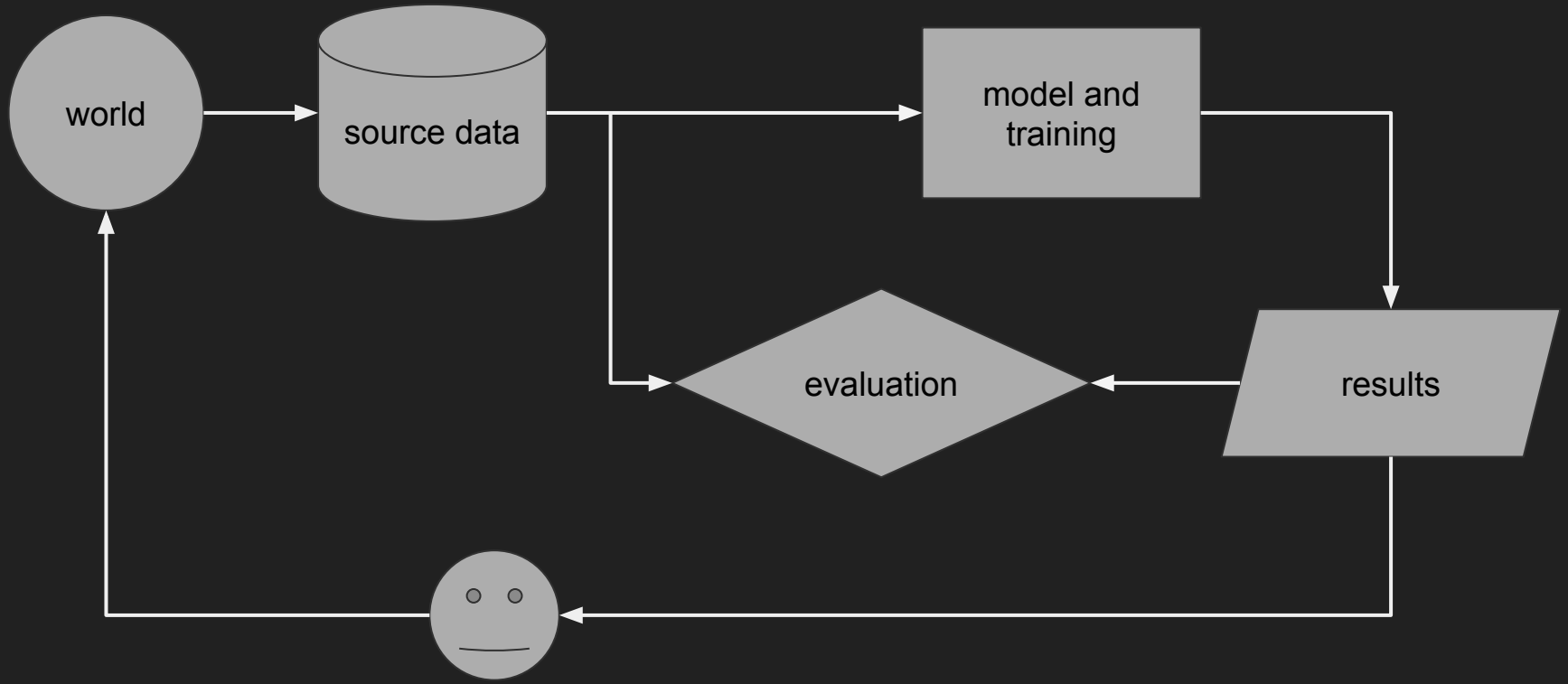
Sources of Unfairness

- Less data about minority populations.
  - This is by definition.
  - Results in lack of representation of minority population.

- The burden of exploration might disproportionately fall on certain populations.
  - This can result in worse model performance on these populations.

- Model includes features that are proxies to protected attributes.
  - Postal code can be a proxy for race if there exists a national history of racial segregation.

Sources of Unfairness



Mitigating Unfairness

1. Identify protected characteristic.
2. Define fairness objective.
3. Choose fairness intervention.
4. Evaluate and monitor fairness intervention.

Mitigating Unfairness

1. **Identify protected characteristic.**
2. Define fairness objective.
3. Choose fairness intervention.
4. Evaluate and monitor fairness intervention.

Examples of Protected Attributes

Mitigating Unfairness

1. Identify protected characteristic.
2. **Define fairness objective.**
3. Choose fairness intervention.
4. Evaluate and monitor fairness intervention.

Types of Fairness Objectives

Individual fairness says similar individuals should be treated similarly

- e.g., Two applicants with the same ability to repay a loan should receive the same decision

Group fairness says each salient group of people should be treated comparably

- e.g., Female job applicants should not be denied more often than male.
- Often concerned with a protected class or sensitive characteristic (e.g., age, gender, race).

Types of Group Unfairness

Disparate treatment: members of different groups are *treated* differently

- Applying different standards to people of different ethnicities

Disparate impact: members of different groups obtain different *outcomes*

- Men pass the employment test at a higher rate than other genders

Disparate mistreatment: members of different groups have different *error* rates

- A risk assessment tool is more likely to misclassify a black defendant as high risk

Examples of Fairness Objectives

Given an algorithm for job candidates that predicts hired/not hired, and a binary protected attribute (e.g., gender):

(1) **Equalized Odds:**

- The algorithm makes similar mistakes on both males and females.
- It would be *unfair* if the classifier systematically overestimates men's worthiness of being hired while systematically underestimating women's worthiness of being hired.
- Both men and women should have similar rates for true positives and false positives.

Hardt et. al. Equality of Opportunity in Supervised Learning: https://arxiv.org/abs/1610.02413

Examples of Fairness Objectives

Given an algorithm for job candidates that predicts hired/not hired, and a binary protected attribute (e.g., gender):

(2) Equality of Opportunity:

- Both men and women have equal opportunities of being hired.
- The proportion of men that are correctly predicted as being hired is the same as the proportion of women correctly predicted as being hired.

Hardt et. al. Equality of Opportunity in Supervised Learning: https://arxiv.org/abs/1610.02413

Examples of Fairness Objectives

Given an algorithm for job candidates that predicts hired/not hired, and a binary protected attribute (e.g., gender):

(3) Demographic Parity:

- Both men and women have equal probabilities of being hired, regardless of the group's correct probability of getting hired
- The proportion of men that are predicted as being hired is the same as the proportion of women predicted as being hired.

Dwork et al. Fairness through Awareness https://arxiv.org/abs/1104.3913

Examples of Fairness Objectives

For more definitions, see:

- Arvind Narayanan, 21 Fairness Definitions and their Politics (FAT* 2018):
  - Tutorial video: https://www.youtube.com/watch?v=jIXIuYdnyyk
  - Notes: https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

- Ninareh Mehrabi et al., A Survey on Bias and Fairness in Machine Learning
  - https://arxiv.org/abs/1908.09635

Mitigating Unfairness

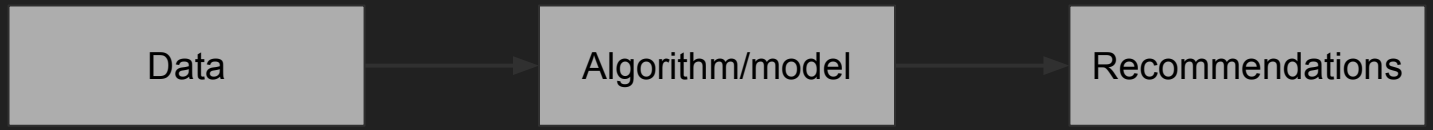
1. Identify protected characteristic.
2. Define fairness objective.
3. **Choose fairness intervention.**
4. Evaluate and monitor fairness intervention.

Types of Fairness Interventions

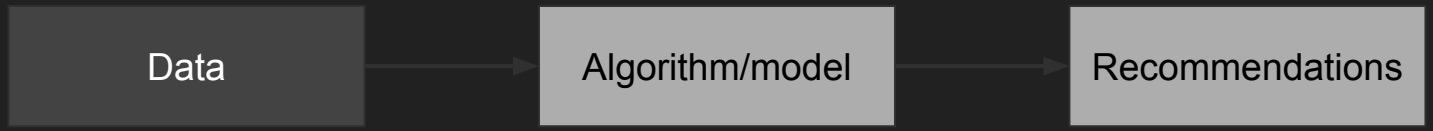
Pre-processing the data

Modifying the algorithm

Post-processing outputs of algorithm



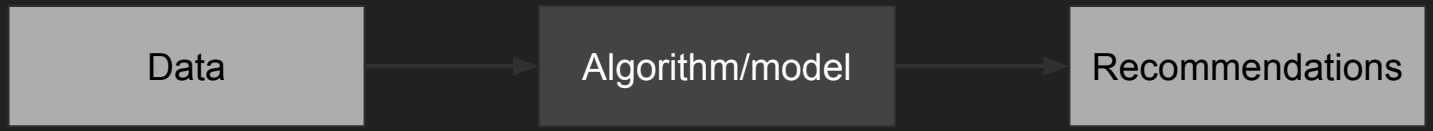
Pre-processing the data



If bias is present in the data, we can de-bias before building a model

- Data relabeling/repair
  - Feldman et al. Certifying and removing disparate impact: https://arxiv.org/abs/1412.3756
  - Salimi et al. Data Management for Causal Algorithmic Fairness: https://arxiv.org/abs/1908.07924
- Data augmentation (Luizos & Welling 2016)
  - Luizos & Welling. The Variational Fair Autoencoder: https://arxiv.org/abs/1511.00830

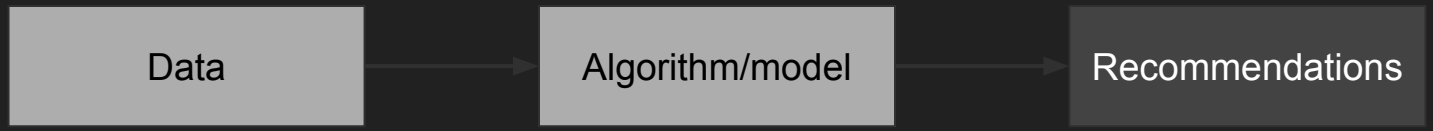
Modifying the algorithm



Alter the objective of the algorithm to emphasize fairness

- Through regularization
  - Bechavod & Ligett. Penalizing Unfairness in Binary Classification: https://arxiv.org/abs/1707.00044
  - Bordia & Bowman. Identifying and Reducing Gender Bias in Word-Level Language Models: https://arxiv.org/abs/1904.03035
- Localizing fairness
  - Nasim Sonboli. Localized Fairness in Recommender Systems: https://dl.acm.org/doi/10.1145/3314183.3323845

Post-processing algorithm outcomes



Re-ranking results from algorithm to produce fairer recommendations

- Greedy approach
  - Zehlike et al. FA*IR: A Fair Top-K Ranking Algorithm: https://arxiv.org/abs/1706.06368
- Constraint Satisfaction
  - Joachims et al. Policy Learning for Fairness in Ranking: https://arxiv.org/pdf/1902.04056.pdf
  - Geyik et al. Fairness-Aware Ranking in Search and Recommendation Systems with Application to LinkedIn Talent Search: https://arxiv.org/abs/1905.01989

Mitigating Unfairness

1. Identify protected characteristic.
2. Define fairness objective.
3. Choose fairness intervention.
4. **Evaluate and monitor fairness intervention.**

Evaluate and Monitor Fairness Interventions

- Majority of existing methods for fairness interventions are applied in a static setting, but the world is not static.

- In a hiring pipeline, the results of the algorithm are fed back into society by deciding who gets/doesn’t get hired, changing the population of people who are fed into the next iteration of training the algorithm.

- Although applying fairness interventions is meant to reduce discriminatory algorithmic behaviour, it is possible to induce harm in the long-term.

Tools for Evaluation and Monitoring

- For more information about dynamic impacts of fair ML, see Liu et al. Delayed Impacts of Fair Machine Learning:
  - Paper: https://arxiv.org/abs/1803.04383
  - Blog: https://bair.berkeley.edu/blog/2018/05/17/delayed-impact/

- ML Fairness Gym is a tool for exploring long-term impacts of fairness interventions in ML systems from Google:
  - Code: https://github.com/google/ml-fairness-gym/
  - Blog: https://ai.googleblog.com/2020/02/ml-fairness-gym-tool-for-exploring-long.html
  - Follow up paper: https://arxiv.org/abs/1911.05489

Conclusion

- Fairness is a complex idea with many different (and often competing!) definitions.
  - Need to decide which definition is best suited for a particular problem.

- Existing intervention methods are deployed at three different stages of the pipeline:
  - Preprocessing the data.
  - Modifying the algorithm.
  - Post-processing outputs.

- Evaluating and monitoring should ideally be done in a dynamic fashion.

Thanks!

Contact info:

Email: a.lucic@uva.nl

GitHub: https://github.com/a-lucic

Twitter: https://twitter.com/__alucic