ethix

Algorithmic Bias: *Values and Trust*

Lea Strohm – Innovation Scientist & Lab Manager

AMLD2020 Trust&AI Track, 28.01.20

WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

The Apple Card Didn't 'S That's the Problem

The way its algorithm determines credit lines make

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

TECH ARTIFICIAL INTELLIGENCE

AI is worse at identifying household items from lower-income countries

An example of AI bias reflecting global inequalities

By James Vincent | Jun 11, 2019, 1:20pm EDT

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Bias – what are we talking about

- Tendency
- Systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others
- An inclination of temperament or outlook
  - especially **:** a personal and sometimes unreasoned judgment **:** prejudice
- An instance of such prejudice

Bias – what are we talking about

Statistical Bias: Biased Model

- *Underfitting:* High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting) => Incorrect predictions
- *Overfitting:* *the algorithm* makes predictions based on ‘wrong’ characteristics in training data => poor performance on test data/real world

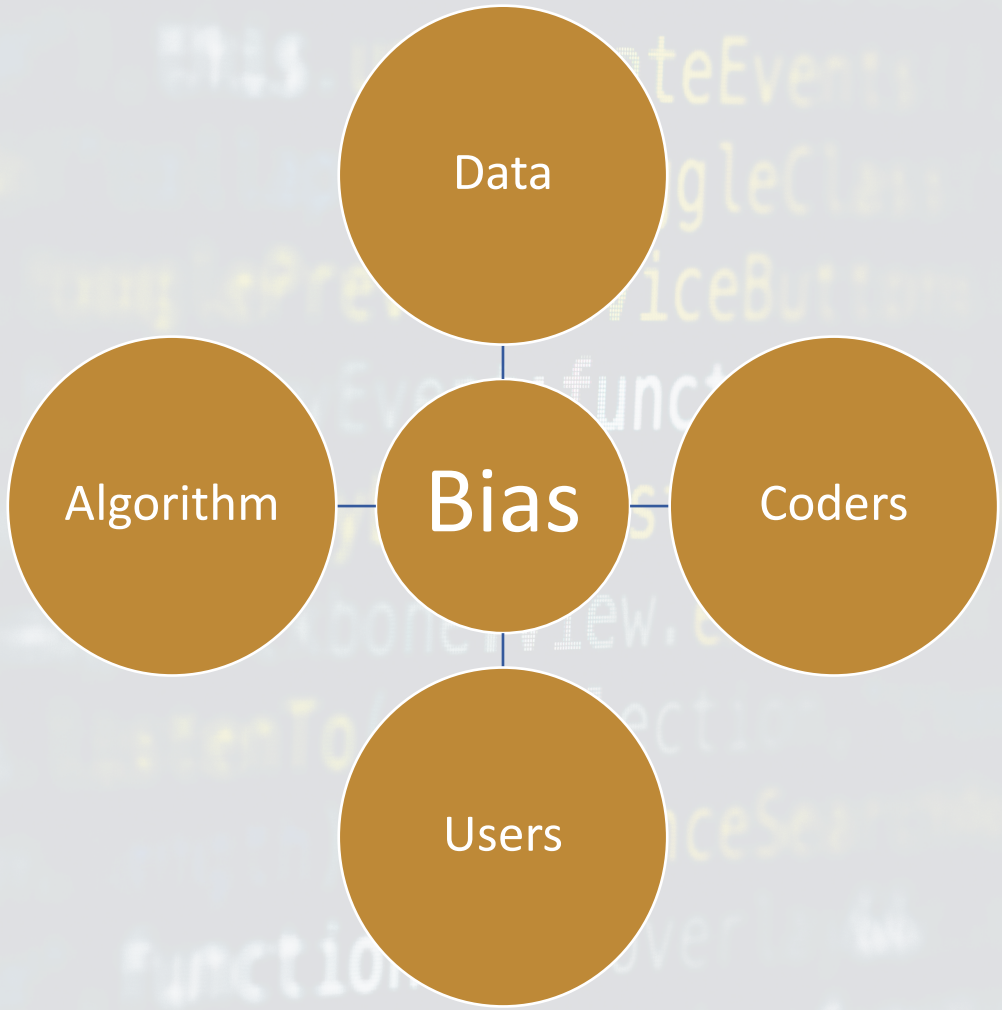
Social Bias: Biased Outcome

- Biased Outcome: Groups of people get (illegally or immorally) discriminated, based on certain characteristics (e.g. gender, race, religion, address etc.)

Bias – what are we talking about

Sources of bias:

1. The Data
2. The algorithm that reinforces the bias in the data
3. The humans that program the algorithms
4. The humans that deploy the algorithms.



Consequences of Bias

- Unjustified Discrimination:
  - Fairness – a core ethical principle – at risk.

- Social Bias:
  - Unintentional & unjustified discrimination *People do not trust an application that is perceived or experienced as unfair.*

- Statistical Bias:
  - Badly performing algorithms: *People do not trust an application that misperforms too often*.

EU Principles for Trustworthy AI

- 4 Core Ethical Principles

Respect for human autonomy	Prevention of harm
Fairness	Explicability

- 7 Key Requirements
  - Human agency & oversight
  - Technical robustness and safety
  - Privacy & data governance
  - Transparency
  - Diversity, non-discrimination & fairness
  - Societal & environmental wellbeing
  - Accountability

ethix Trust Framework

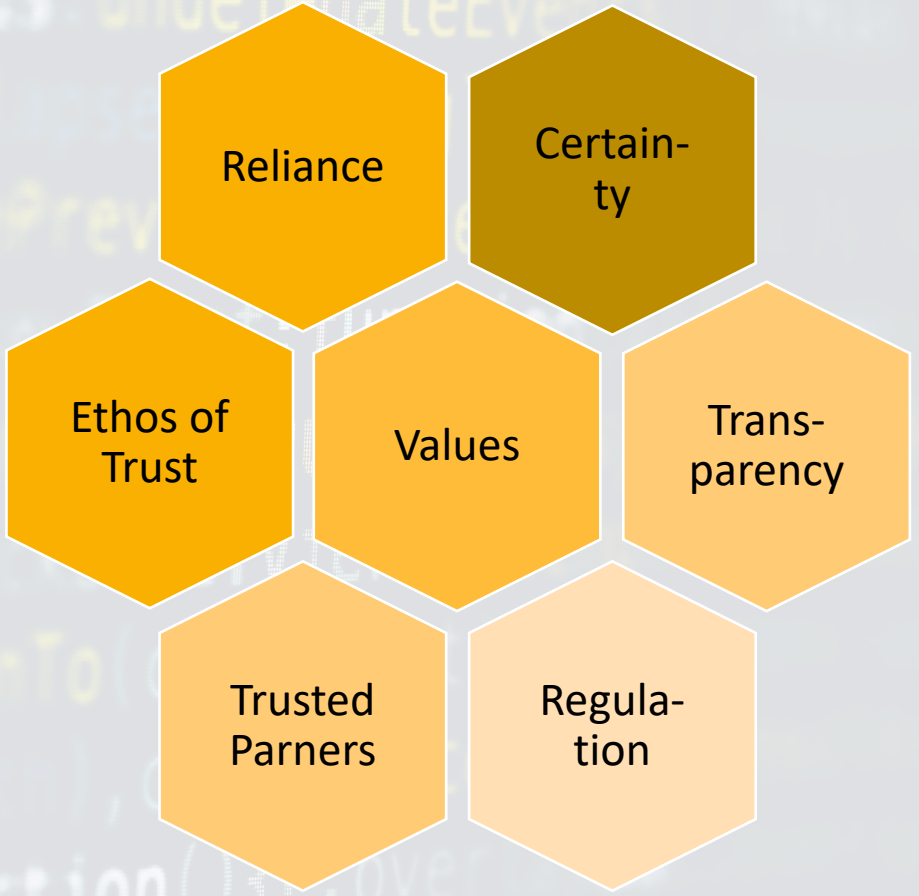
- Whitepaper: Trust in Innovation
- Definition of Trust
  - *Trust is a future-oriented, interpersonal relationship between at least two individuals. It involves a prediction: Person A makes a prediction about person B's goodwill or intentions in the future. There is a perception of shared values and beliefs with person B and vice versa.*
- *goodwill or intentions in the future = trustworthiness*

ethix Trust Framework

- Elements of Trustworthiness: What makes a trustworthy person?
  - Competence
  - Honesty
  - Vulnerability
- Context-specific trust vs. generalized trust?
  - I trust someone *to do X* vs I trust someone *in general*.
  - Crucial for the context of technology

Building Blocks of Trust

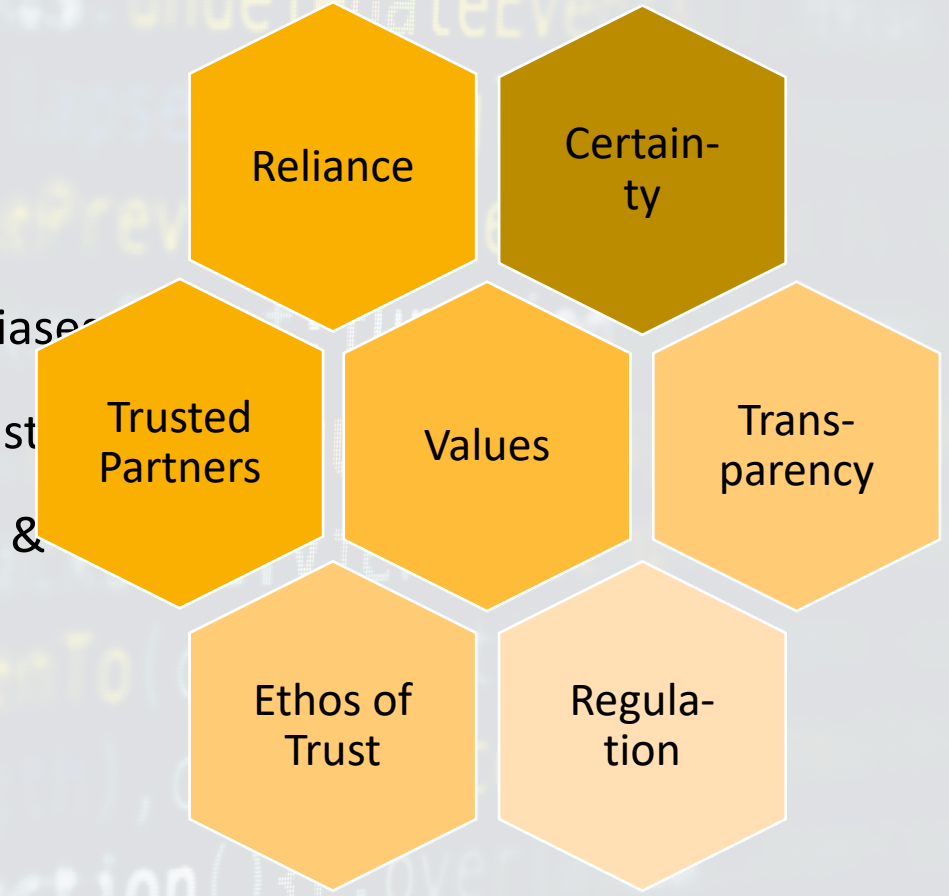
- Seven Building Blocks
  - Reliance: Predictability
  - Certainty: minimizing the unknown
  - Transparency: receiving information
  - Third party Regulation (e.g. guidelines)
  - Information from Trusted Partners
  - Ethos of trust: climate of generalized trust
  - Values: values of innovators are aligned with mine



What to do?

Sources of bias:

- The data: Assure the dataset is representative of the population the algorithm will be used on
- The algorithm: Test & Audit the algorithm for potential biases
- The programmers: Provide ethics training for data scientist
- The users: Actively inform on intended target population & limitations



Thank you

Lea Strohm

@Lea_st1
strohm@ethix.ch

ethix.ch
@EthixInnovation