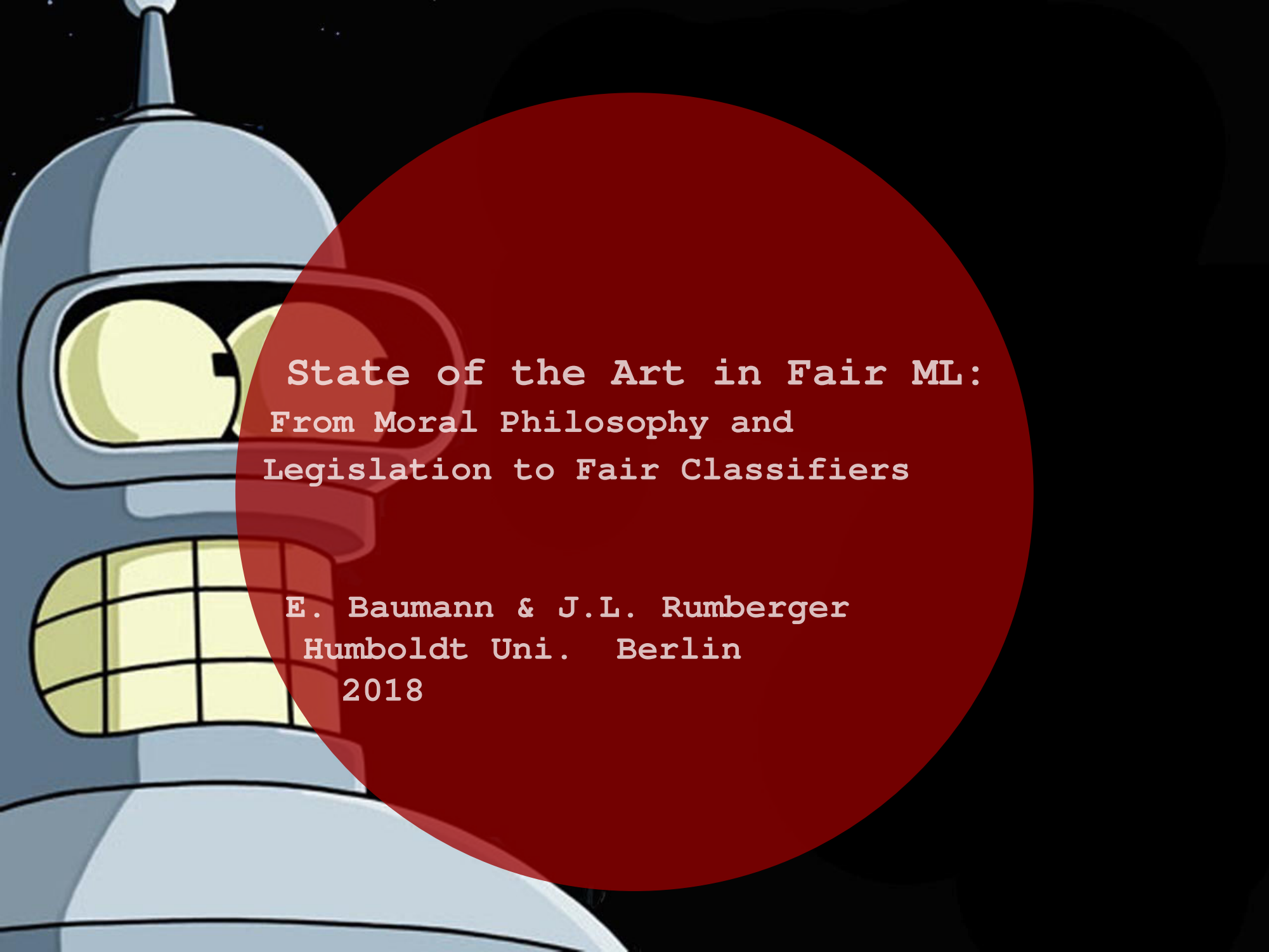


# State of the Art in Fair ML:
## From Moral Philosophy and Legislation to Fair Classifiers

E. Baumann & J.L. Rumberger
Humboldt Uni. Berlin
2018

# PROBLEMS OF DEEP LEARNING

M.R. Minar and J. Naher. "Recent Advances in Deep Learning: An Overview."; 2018

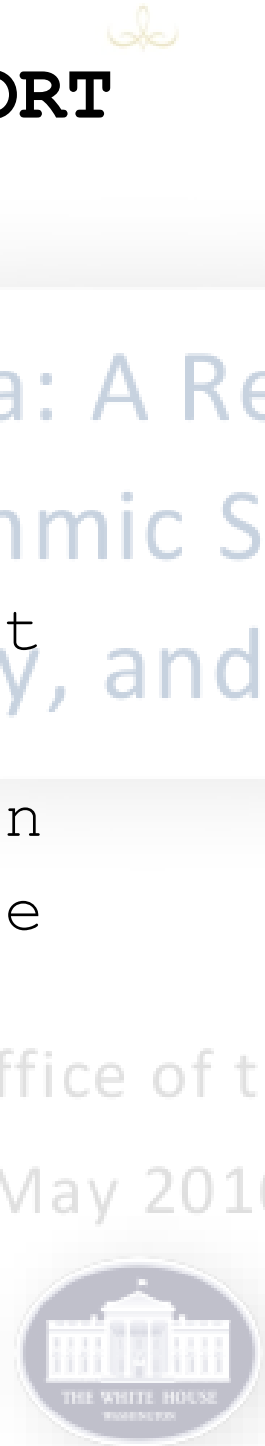
- Requiring a lot of data.
- Limited capacity.
- Inability to deal with hierarchical structures.
- Struggling with open-ended interfaces.
- Inability to distinguish causation from correlation.

# PROBLEMS OF DEEP LEARNING

M.R. Minar and J. Naher. "Recent Advances in Deep Learning: An Overview."; 2018

- Requiring a lot of data.
- Limited capacity.
- Inability to deal with hierarchical structures.
- Struggling with open-ended interfaces.
- Inability to distinguish causation from correlation.
- Not transparent.

# WHITE HOUSE REPORT




- Access to Credit
- Employment
- Higher Education
- Criminal Justice

# GENERAL DATA PROTECTION REGULATION (GDPR)

adopted by EU in 2016 and implemented since May 2018

- Objective: Protection of individuals in relation to processing and movement of personal data.
- Protected attributes:
  - Race
  - Gender
  - Religion and Belief
  - Disability, chronic and mental illness
  - Age
  - Sexual Orientation

# WHAT IS THE GOAL?

-> Equality of Opportunity (EoO)

- Formal EoO
- Substantive EoO

# NOTATION

$X$ : features

$Y$ : label

$\hat{Y}$ : predicted label

$A$ : protected features

# DISCRIMINATORY BEHAVIOR OF ALGORITHMS

If protected characteristics are in feature set $(A \in X)$.

# DISCRIMINATORY BEHAVIOR OF ALGORITHMS

If protected characteristics are in feature set ($A \in X$). Simply remove $A$.

-> Usually does not make it more fair, but can even worsen.

# DISCRIMINATORY BEHAVIOR OF ALGORITHMS

If protected characteristics are in feature set ($A \in X$). Simply remove $A$.

CS.Barocas et al.: Fairness and Machine Learning. NIPS Tutorial (2017)

-> Usually does not make it more fair, but can even worsen.

C.R.Sugimoto et al.: Big data is not a monolith. MIT Press (2016)

Everything might reveal everything else.

-> Protected characteristics can be predicted.

# DISCRIMINATORY BEHAVIOR OF ALGORITHMS

- Danford and Reece [47] apply machine learning methods on instagram photos to predict the probability of an individual to suffer from depression. The resulting model beats average human practitioners in unassisted diagnosis accuracy. Suffering from mental illness is a protected variable under german legislation.

- Narayanan and Shmatikov [42] present several statistical methods to de-anonymize large sparse datasets. They apply these methods to de-anonymize individuals from the Netflix Prize dataset based on information about users from the comments and ratings in the publicly available Internet Movie Database (IMDB). The former dataset was made public by netflix in order to organize a data science competition. After successful de-anonymization a lawsuit against netflix was filed and the data science competition was discontinued due to privacy concerns. The authors found that besides real

# DISCRIMINATORY BEHAVIOR OF ALGORITHMS

Where does it come from?

- Selection and Confirmation Bias
- Limited Features
- Sample Size Disparity

# HOW TO DETECT DISCRIMINATION

Architecture:

- Transparency
- Post-hoc interpretability

Investigation:

- Observational Approach
- Causal Reasoning

# HOW TO DETECT DISCRIMINATION

## Architecture

- Interpretability of ML models has yet to be defined.

  B.Goodman and S.Flaxman; arXiv:1606.08813. (2016)

  - Understand-able and articulate-able.
- GDPR, *right to explanation*: not defined what exactly explanation entails.

# HOW TO DETECT DISCRIMINATION

## Architecture: Transparency

- Explain the workings of an algorithm.
- Visible neural networks (VNNS): visualize hidden representations.

# HOW TO DETECT DISCRIMINATION

## Architecture: Post-hoc interpretability

- Obtain information of already trained models to understand decisions.
- Interpretable models:

  Christoph Molnar: Interpretable Machine Learning. github (2018)

  - linear regr. and log. regr.
  - decision trees
  - model-agnostic meta-learning (MAML)

# HOW TO DETECT DISCRIMINATION

## Architecture

Issues:

- Requires ML models to never surpass human ability.
- Evaluation could hide discrimination.

# HOW TO DETECT DISCRIMINATION

## Investigation: Observational Approach

- Uses easily observable characteristics $(X, A, Y, \hat{Y})$.
- Analyzes conditional probability of $Y$ and $\hat{Y}$ given $X$ and $A$.
- Applicable on any classifier

# MITIGATION OF DISCRIMINATION

## Observational Fairness Criteria

- Group/Demographic Parity
- Individual
  - Equal Odds
  - Equal Opportunity
- Calibration

# MITIGATION OF DISCRIMINATION

## Observational Fairness Criteria

- Group/Demographic Parity:
  Positive classification $\hat{Y}$ statistically independent from A.
- Individual
  - Equal Odds
  - Equal Opportunity
- Calibration

# MITIGATION OF DISCRIMINATION

## Observational Fairness Criteria

- Group/Demographic Parity:
  Positive classification $\hat{Y}$ statistically independent from A.
- Individual:
  Similar individuals get similar classification.
  - (strong) Equal Odds:
    classific. & misclassific.
  - (weak) Equal Opportunity:
    only pos. classifications
- Calibration

# MITIGATION OF DISCRIMINATION

## Observational Fairness Criteria

- Group/Demographic Parity:
  Positive classification $\hat{Y}$ statistically independent from A.
- Individual:
  Similar individuals get similar classificationn.
  - (strong) Equal Odds:
    classific. & misclassific.
  - (weak) Equal Opportunity:
    only pos. classifications
    --> used by US Equal Employment Opport. Commission (EEOC)
- Calibration

# MITIGATION OF DISCRIMINATION

## Observational Fairness Criteria

Issues:

- No two criteria can be applied simultaneously.
- The need to choose for a criteria can lead to additional discrimination.
- Individual fairness criteria make use of A.
- Individual fairness criteria allow dependencies between A and classifier.
- Disregard of long-term impact of decisions.

# HOW TO DETECT DISCRIMINATION

## Investigation: Causal Reasoning

- Information on causation instead of correlation.
- Analyze via causality graphs how variables are generation and interconnected.

# MITIGATION OF DISCRIMINATION

## Causal Fairness Criteria

- Unresolved discrimination
- Proxy discrimination
- Multi-world approach
- Counter-factual fairness

# MITIGATION OF DISCRIMINATION

## Causal Fairness Criteria

Issues:

- Require domain knowledge.

# DISCUSSION

- Main problem is the data-set.
- Second main problem is lack of awareness.
- Best approach is to combine observable and causal mitigation procedures together with an transparent and interpretable model.


QUESTIONS ?
THE
GENDER
BENDER