Bias and Fairness in AI

Somaieh Nikpoor
Research Advisor AI/ML- Government of Canada
Iasa- BILT Conference – July 2021

BLOG POST
RESEARCH

20 JUL 2016

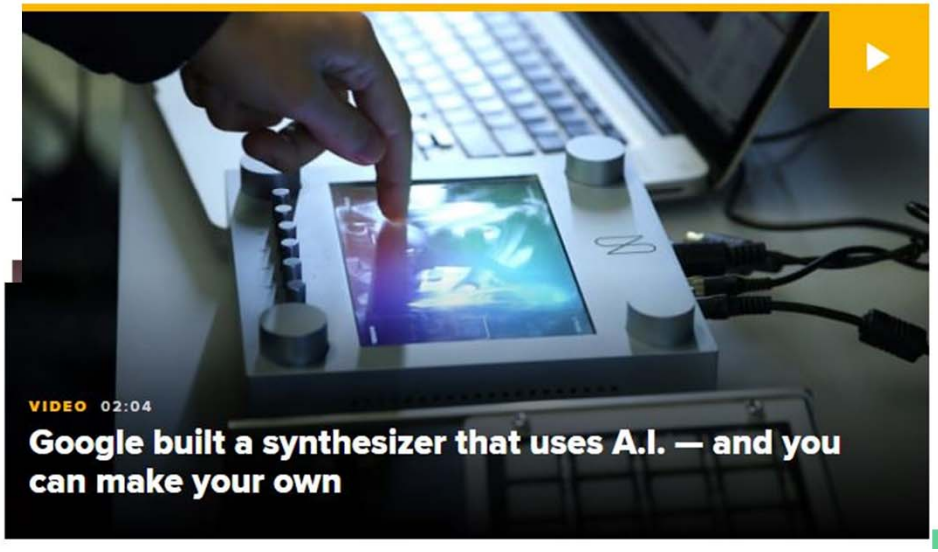
DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

3 Advances Changing the Future of Artificial Intelligence in Manufacturing

BY PETER DORFMAN · MANUFACTURING · JAN 3 2016 · 5 MIN READ

This Machine Learning-Powered Software Teaches Kids To Be Better Writers

Quill.org offers corrections to kids' sentence construction and composition, giving teachers more time to focus on students.



INSIDER FEATURE

How AI can help you stay ahead of cybersecurity threats

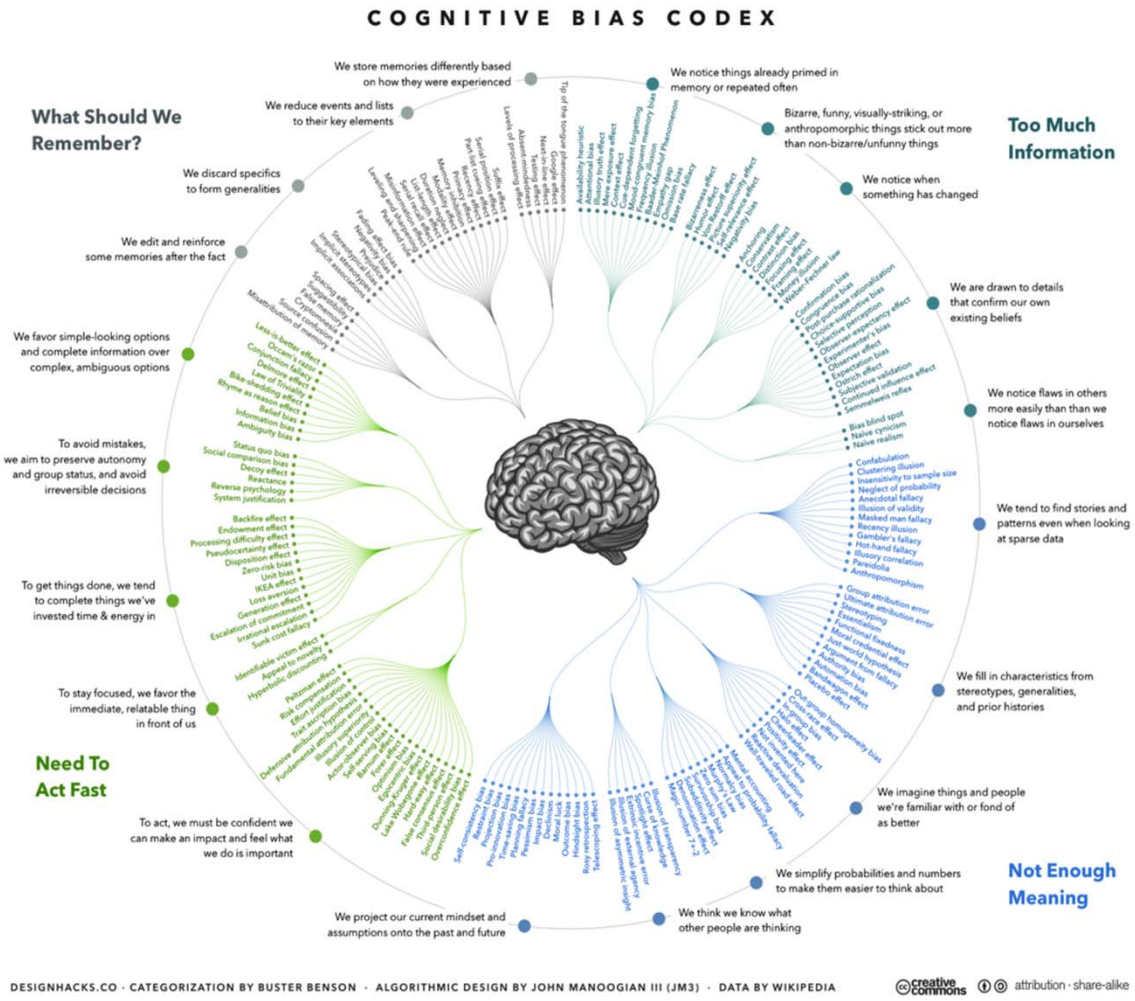
Artificial intelligence and machine learning can be force multipliers for under-staffed security teams needing to respond faster and more effectively to cyber threats.

- Humans are prone to hundreds of proven biases

- Humane biases categories:
  - Too much information
  - Not enough meaning
  - What should we remember
  - Need to act fast

Design by JM3 and John Manoogian

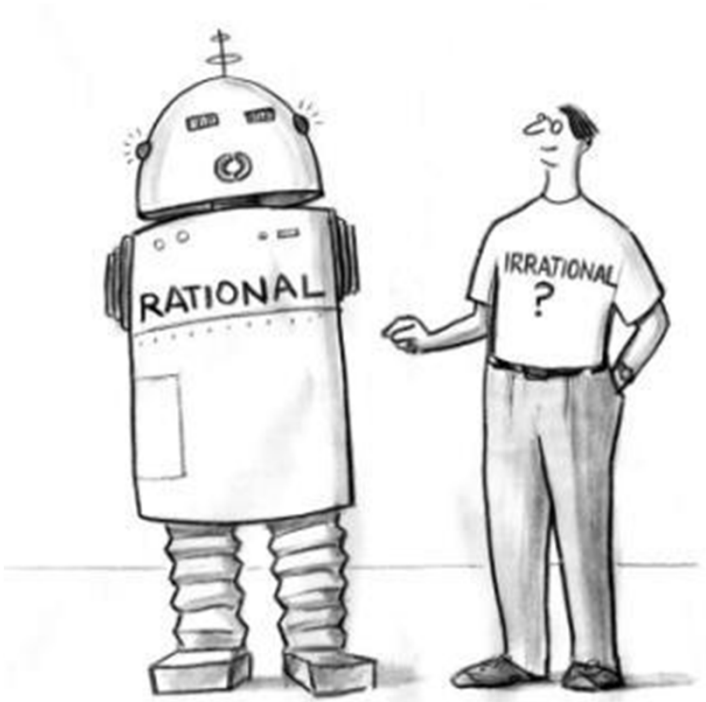
Concept + categorization by Buster Benson



Why it Matters

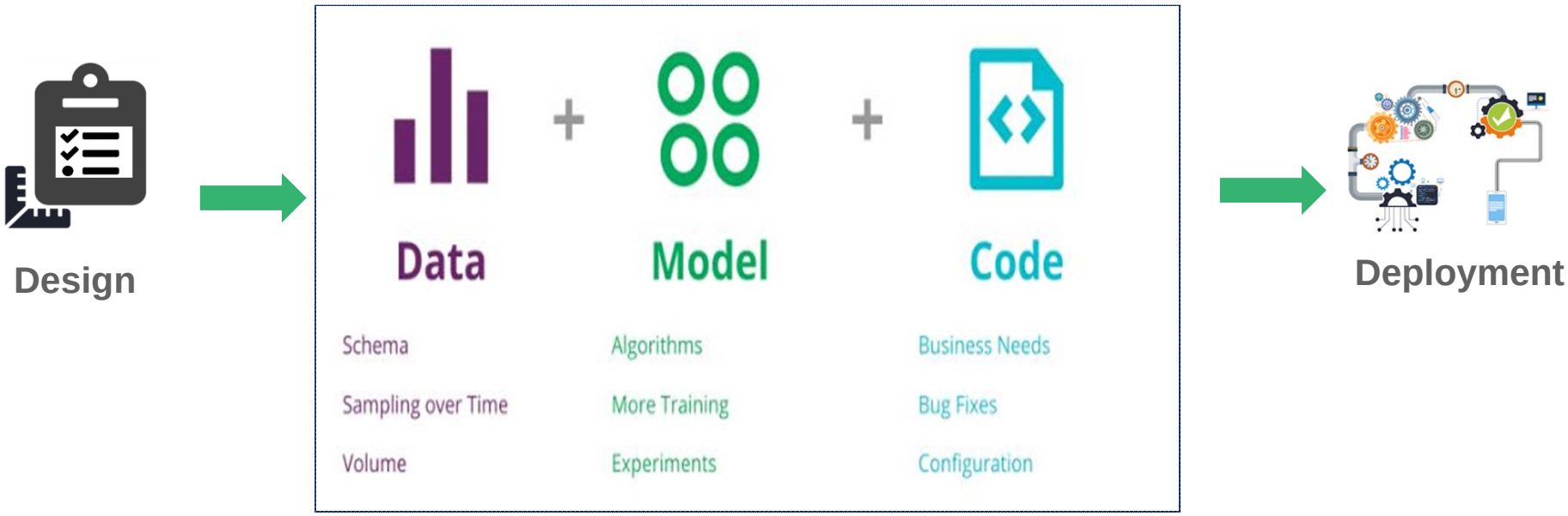
According to Rachel Thomas (2018):

- Algorithm and human decision making are used differently
- Algorithmic systems used at scale/ relatively cheap
- Create feedback loop



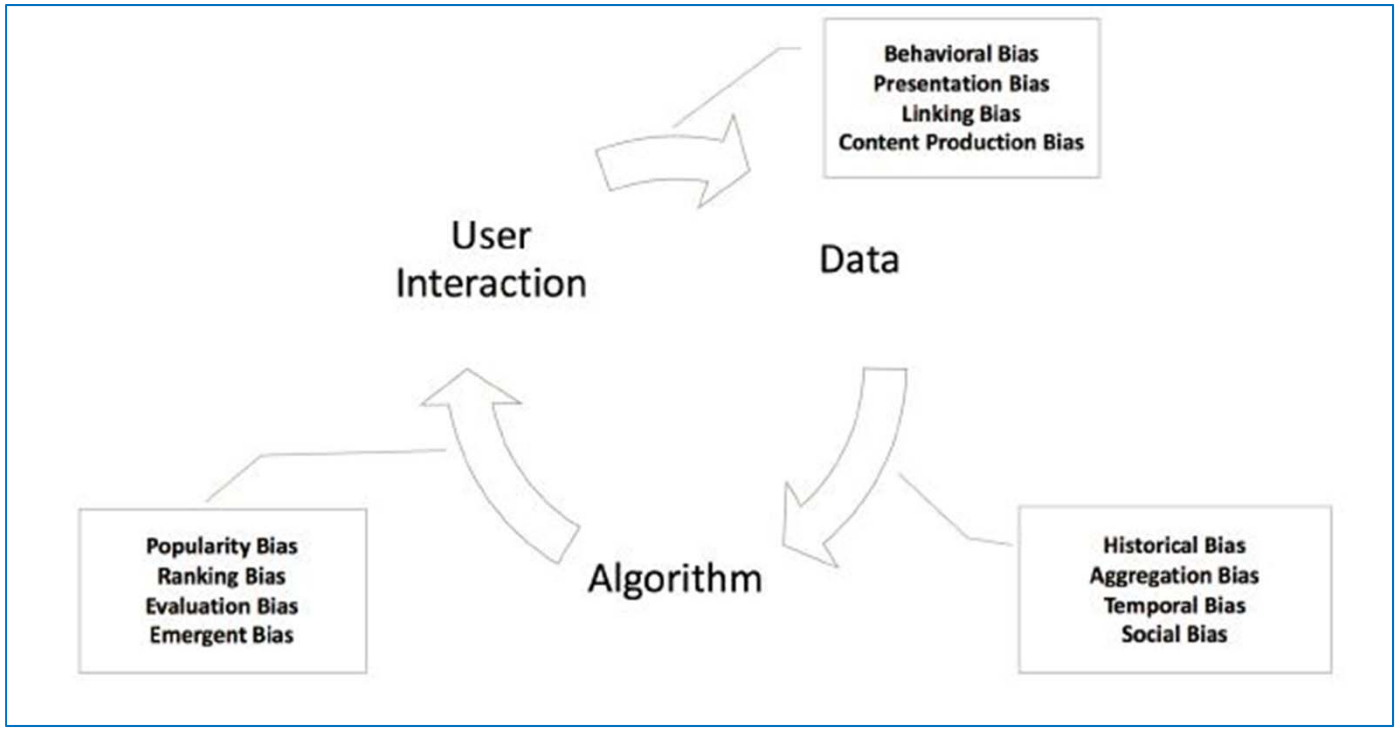
Economind

Bias can occur in every step of the ML pipeline



@Sato, Wider and Windheuser, 2019

Types of Bias in Data



Source: Mehrabi et al, 2019

Data Bias Measures

According to Olteanu etal (2018):

1. Population Biases: Differences in demographics or other user characteristics between a user population represented in a dataset or platform and a target population (Hargitta , 2007)

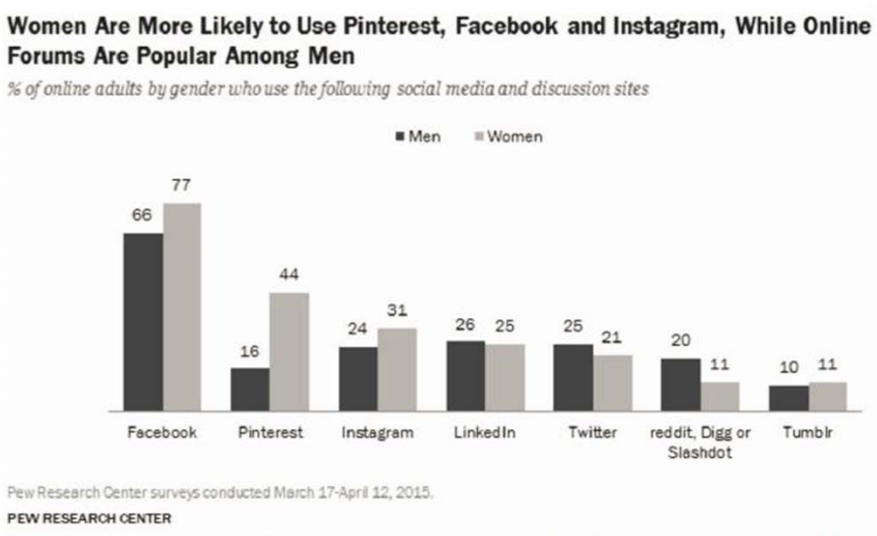
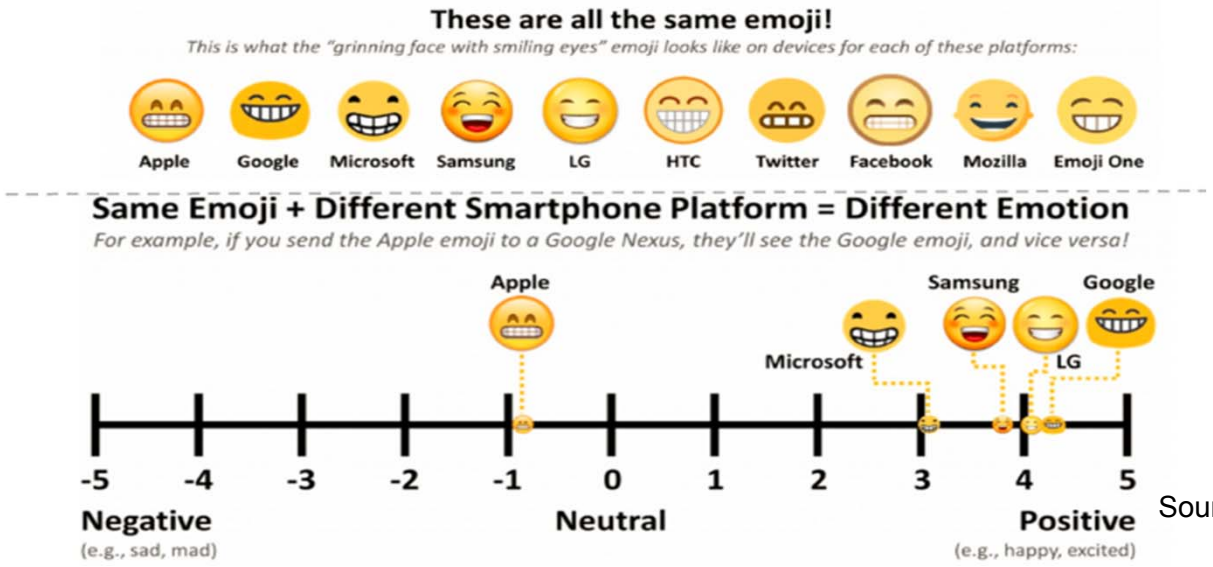


Figure from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

2. Behavioral Biases

- Differences in user behavior across platforms or contexts, or across users represented in different datasets (Miller et al. ICWSM'16)

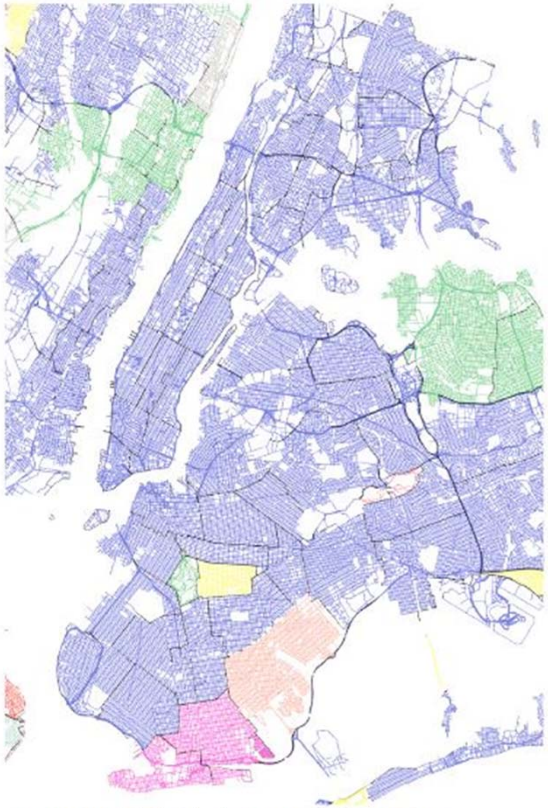


Source: group ens

3. Content Production Biases

- Lexical, syntactic, semantic, and structural differences in the contents generated by users (Rao etal, 2010, Mocanu et al. PlosOne 2013)

The second language by district or municipality (in the case of New Jersey state) is shown. Blue - Spanish, Light Green - Korean, Fuchsia - Russian, Red - Portuguese, Yellow - Japanese, Pink - Dutch, Grey - Danish, Coral - Indonesian.
Figure source [Mocanu et al. PlosOne 2013]



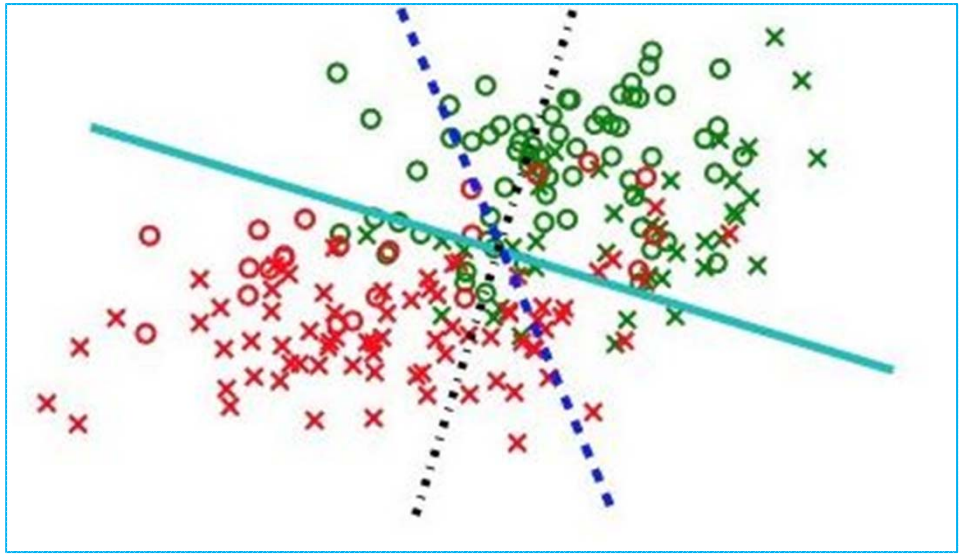
Language polarization in New York City, NY, USA.

4. Linking Biases: Differences in the attributes of networks obtained from user connections, interactions, or activity (Gilbert et al,2009)

5. Temporal Biases: Differences in populations and behaviors over time (Hannak et al, 2017)

- In order to minimize bias, how do we define and measure fairness?

- What is means for a decision to be fair?



@Zafar et al. AISTATS2017

How to Define Fairness?

- So many definitions of fairness
- An interesting tutorial by **Arvind Narayanan: 21 fairness definitions and their politics**
- An article by Verma, Sahil, and Julia Rubin. **"Fairness Definitions Explained.“**
- Another interesting tutorial by **Jon Kleinberg: Inherent Trade-Offs in Algorithmic Fairness**

Definition	Citation #
Group fairness or statistical parity	208
Conditional statistical parity	29
Predictive parity	57
False positive error rate balance	57
False negative error rate balance	57
Equalised odds	106
Conditional use accuracy equality	18
Overall accuracy equality	18
Treatment equality	18
Test-fairness or calibration	57
Well calibration	81
Balance for positive class	81
Balance for negative class	81
Causal discrimination	1
Fairness through unawareness	14
Fairness through awareness	208
Counterfactual fairness	14
No unresolved discrimination	14
No proxy discrimination	14
Fair inference	6

@Verma, etal, 2018

Definition of Fairness

- **Group Fairness:** equitable outcomes across demographic groups.

- **Individual Fairness**: was first proposed in *Fairness Through Awareness* by Cynthia Dwork et al: Similar individuals (based on a metric that defines how similar two given individuals are in the context of a decision-making task) should be treated similarly.

Individual Fairness Group Fairness

Equality Equity

Types of Group Fairness

- It is very important to select the right type of fairness;
- A wrong metric can lead to harmful decisions

Two most common definition: (Barocas and Hardt, 2017)

- Demographic Parity
- Equal Opportunity

Demographic Parity

- The proportion of each segment of a protected class (e.g. gender) should receive the positive outcome at equal rates (positive outcome : “getting to university”, “getting a loan” or “being shown an ad” .

When to use Demographic Parity

- Improve the state of our current world
- Historical biases may have affected the quality of our data
- Support the unprivileged group and to prevent the reinforcement of historical biases

Equal Opportunity

- States that each group should get the positive outcome at equal rates, assuming that people in this group qualify for it.
- Equal Opportunity requires the positive outcome to be independent of the protected class.

When to use Equal Opportunity

- Predicting the positive outcome correctly
- False Positives are not costly to the user nor the company

(-

The AI Ethics Landscape

- European Commission legal framework for AI
- Ethical Guidelines for trustworthy AI’ - the European Union
- GDPR
- OECD Principles on AI - OECD
- Asilomar AI Principles - Future of Life Institute

Takeaways

- **Bias** can occur at **Every Step** of the AI-Building Process
- It is fundamental to **define the fairness goals** and definition when scoping an AI project
- **Interdisciplinary** collaboration among researchers is required to ensure ethical development of AI systems
- **Improving** fairness would be an **ever evolving process** as the interactions between people and machines is changing every day.

Questions?