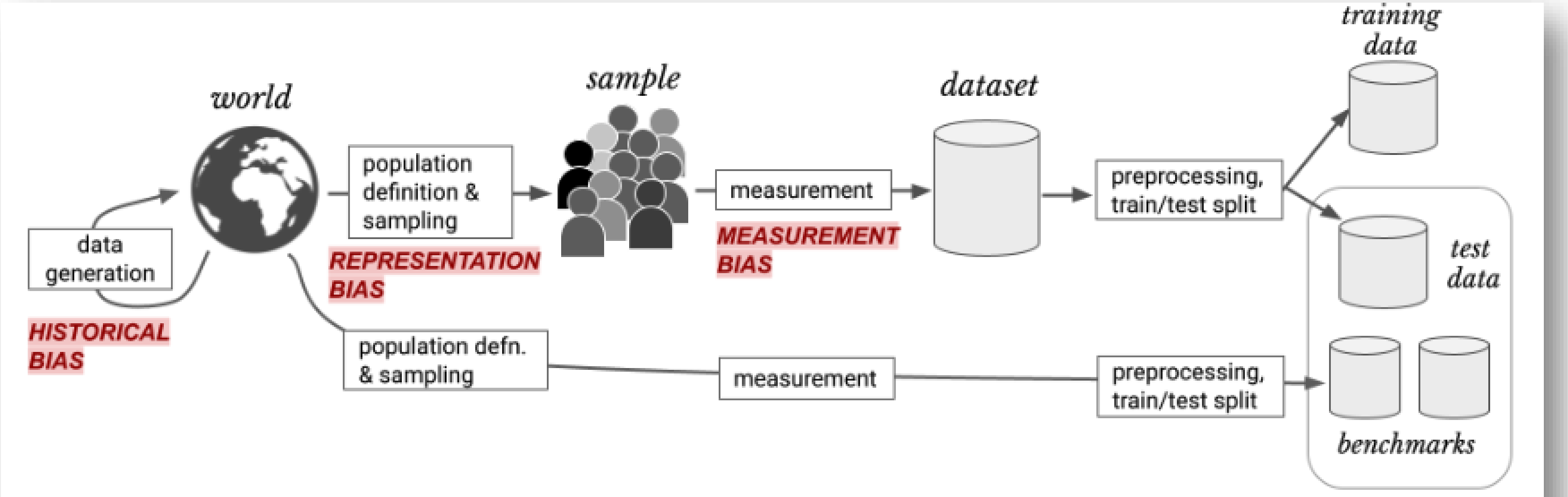
Justice, Fairness, Bias (Part 2)

The Big Three

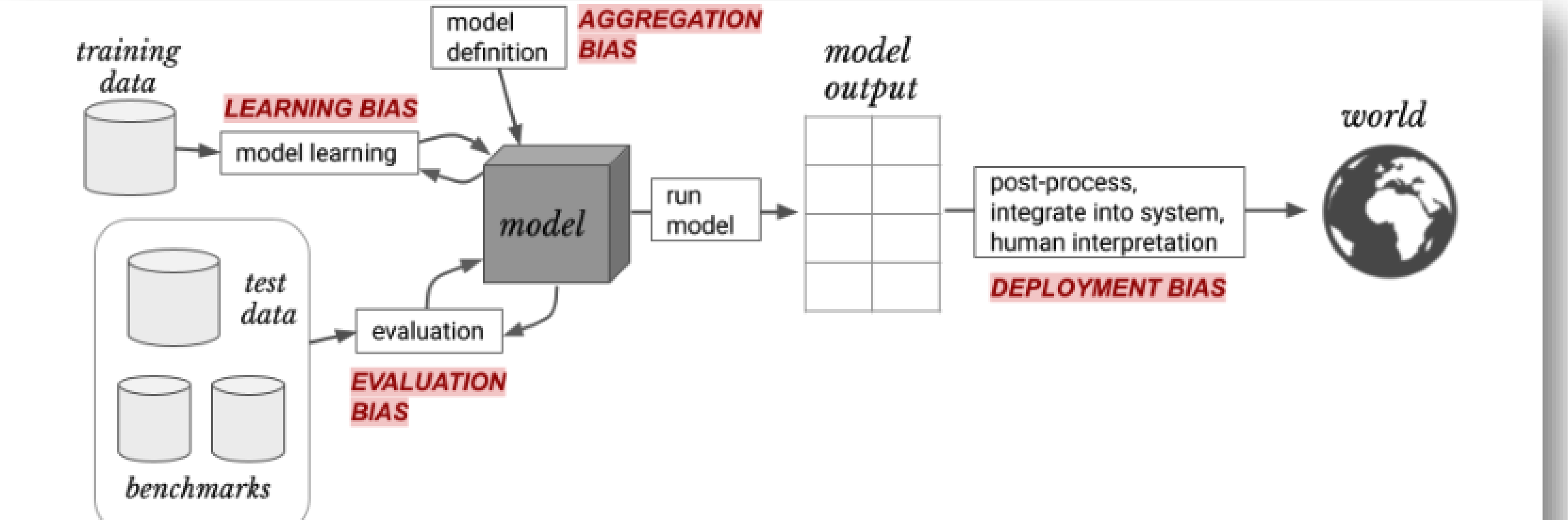
Bias -- Recap



(a) Data Generation

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 17, 1–9.

Bias -- Recap

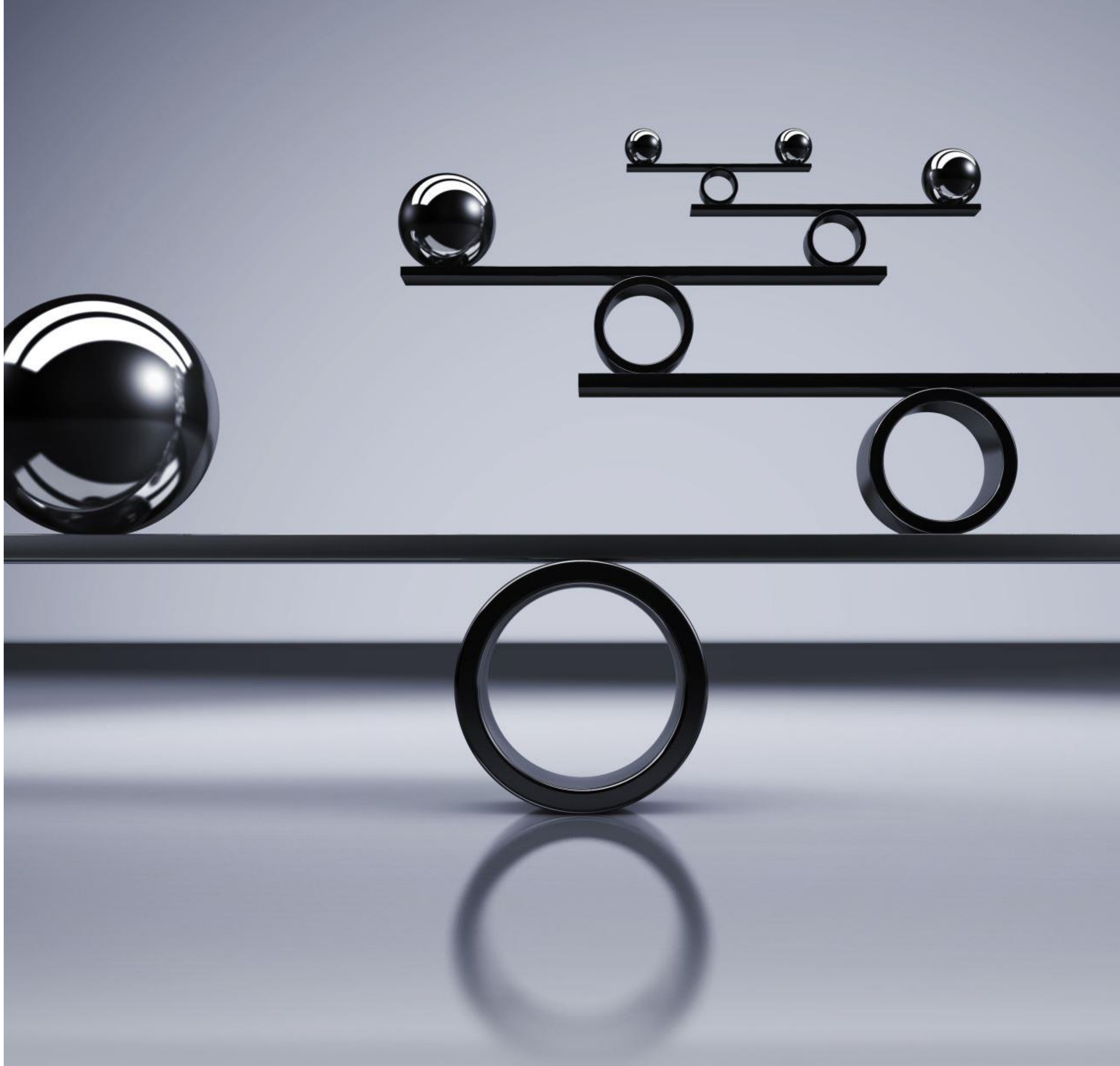


(b) Model Building and Implementation

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 17, 1–9.

Outline

- Algorithmic **Fairness**
  - Group Fairness
  - Individual Fairness
- Data **Justice**
- Watchdogs



De-biasing Algorithms

- Increasing awareness about different types of bias is essential.
- We will now have a closer look at how to design an AI system that would not discriminate.

Fairness Through Awareness

Cynthia Dwork[*] Moritz Hardt[†] Toniann Pitassi[‡] Omer Reingold[§]
Richard Zemel[¶]

November 30, 2011

Abstract

We study *fairness in classification*, where individuals are classified, e.g., admitted to a university, and the goal is to prevent discrimination against individuals based on their membership in some group, while maintaining utility for the classifier (the university). The main conceptual contribution of this paper is a framework for fair classification comprising (1) a (hypothetical) task-specific metric for determining the degree to which individuals are similar with respect to the

2018 ACM/IEEE International Workshop on Software Fairness

Fairness Definitions Explained

Sahil Verma
Indian Institute of Technology Kanpur, India
vsahil@iitk.ac.in

Julia Rubin
University of British Columbia, Canada
mjulia@ece.ubc.ca

ABSTRACT

Algorithm fairness has started to attract the attention of researchers in AI, Software Engineering and Law communities, with more than twenty different notions of fairness proposed in the last few years. Yet, there is no clear agreement on which definition to apply in each situation. Moreover, the detailed differences between multiple definitions are difficult to grasp. To address this issue, this paper

training data containing observations whose categories are known. We collect and clarify most prominent fairness definitions for classification used in the literature, illustrating them on a common, unifying example – the German Credit Dataset [18]. This dataset is commonly used in fairness literature. It contains information about 1000 loan applicants and includes 20 attributes describing each applicant, e.g., credit history, purpose of the loan, loan amount

Algorithmic Fairness

- We can talk about fairness when people are not discriminated against based on their membership to a specific group.
- Fairness definition? The most famous discussion about fairness definitions come from Arvind Narayanan.
- There are two main categories: group fairness (statistical fairness) and individual fairness.

21 Definitions of Fairness -- https://www.youtube.com/watch?v=jIXIuYdnyyk

Fairness through Blindness

- We can ignore all irrelevant or protected attributes in our dataset.

Some Statistical Measures

- Predicted outcomes
- Predicted and actual outcomes
- Predicted probabilities and actual outcomes

Y

O

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Predicted Outcomes -- Statistical Parity

- We aim to equalize two groups S (e.g., protected set) and T (e.g., complement of S) at the level of predicted outcomes.

$$P[O=1|S] = P[O=1|T]$$

- Conditional statistical parity extends this one by allowing conditioning on a set of factors.

$$P[O=1|X, S] = P[O=1|X, T]$$

Statistical Parity -- Problems

- Self-fulfilling Prophecy

'A self-fulfilling prophecy is the psychological phenomenon of someone "predicting" or expecting something, and this "prediction" or expectation coming true simply because the person believes or anticipates it will and the person's resulting behaviors align to fulfill the belief. This suggests that people's beliefs influence their actions.'

Example: Give loans to people in S who are least credit-worth

Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.

Statistical Parity -- Problems

- Reverse Tokenism

Example: Pick a token from T, who is more qualified than any member of S, and deny their loan. Then, you have an excuse to deny a loan for a member of S.

Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.

Predicted and Actual Outcomes

- COMPAS, Gender Shades examples fall within this category.
- Error rate balance suggests that FNR and FPR should be equal across different groups.

Equalized Odds

$P[O=1|Y=i, S] = P[O=1|Y=i, T]$

$P[O=1|Y=0, S] = P[O=1|Y=0, T]$ FP Error Rate (Predictive Equality)

$P[O=0|Y=1, S] = P[O=0|Y=1, T]$ FN Error Rate (Equal Opportunity)

Predicted and Actual Outcomes

- COMPAS, Gender Shades examples fall within this category.
- Predictive Parity (PPV) : The probability of a subject with positive predictive value to truly belong to the positive class.

$$P[Y=1|O=1, S] = P[Y=1|O=1, T]$$

Outcome Test

Gender Shades

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77-91. PMLR, 2018.

Predicted Probabilities and Actual Outcomes

- Calibration is one of the well-known definitions in this category.
- Calibration focuses on the fraction of correct positive predictions.
- For any given predicted probability score *r* in [0,1], the probability of having actually a good outcome should be equal for S, T:

$$P[Y=1|R=r, S] = P[Y=1|R=r, T]$$

Calibration Example

s	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(Y=1 \mid S=s, G=m)$	1.0	1.0	0.3	0.3	0.4	0.6	0.6	0.7	0.8	0.8	1.0
$P(Y=1 \mid S=s, G=f)$	0.5	0.3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

S. Verma and J. Rubin, "Fairness Definitions Explained," *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 2018, pp. 1-7.

Individual Fairness

- Treat similar individuals similarly.
- Fairness is task-specific, similarity measure should be defined for the purpose of the task.
- We should aim for a similar distribution over outcomes.
- Problem: Which factors to consider to represent individuals? How to define a distance metric?

Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.

Data Justice

Original Research Article

Big Data & Society
July–December 2017: 1–14

Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053951717736335
journals.sagepub.com/home/bds
SAGE

What is data justice? The case for connecting digital rights and freedoms globally

Linnet Taylor

Abstract

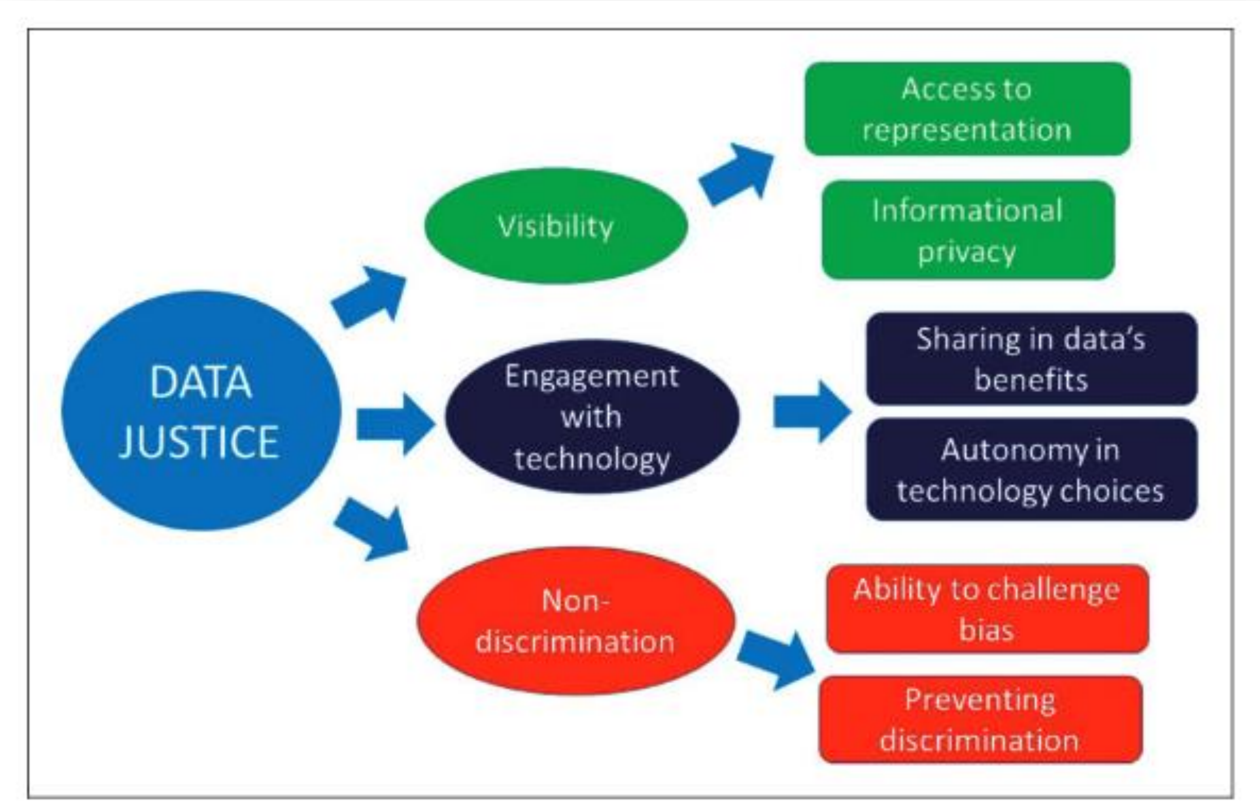
The increasing availability of digital data reflecting economic and human development, and in particular the availability of data emitted as a by-product of people's use of technological devices and services, has both political and practical implications for the way people are seen and treated by the state and by the private sector. Yet the data revolution is so far primarily a technical one: the power of data to sort, categorise and intervene has not yet been explicitly connected to a social justice agenda by the agencies and authorities involved. Meanwhile, although data-driven discrimination is advancing at a similar pace to data processing technologies, awareness and mechanisms for combating it are not. This paper posits that just as an idea of justice is needed in order to establish the rule of law, an idea of *data justice* – fairness in the way people are made visible, represented and treated as a result of their production of digital data – is necessary to determine ethical paths through a datafying world. Bringing together the emerging scholarly perspectives on this topic, I propose three pillars as the basis of a notion of international data justice: (in)visibility, (dis)engagement with technology and antidiscrimination. These pillars integrate positive with negative rights and freedoms, and by doing so challenge both the basis of current data protection regulations and the growing assumption that being visible through the data we emit is part of the contemporary social contract.

Keywords

Privacy, ethics, development, discrimination, representation, surveillance

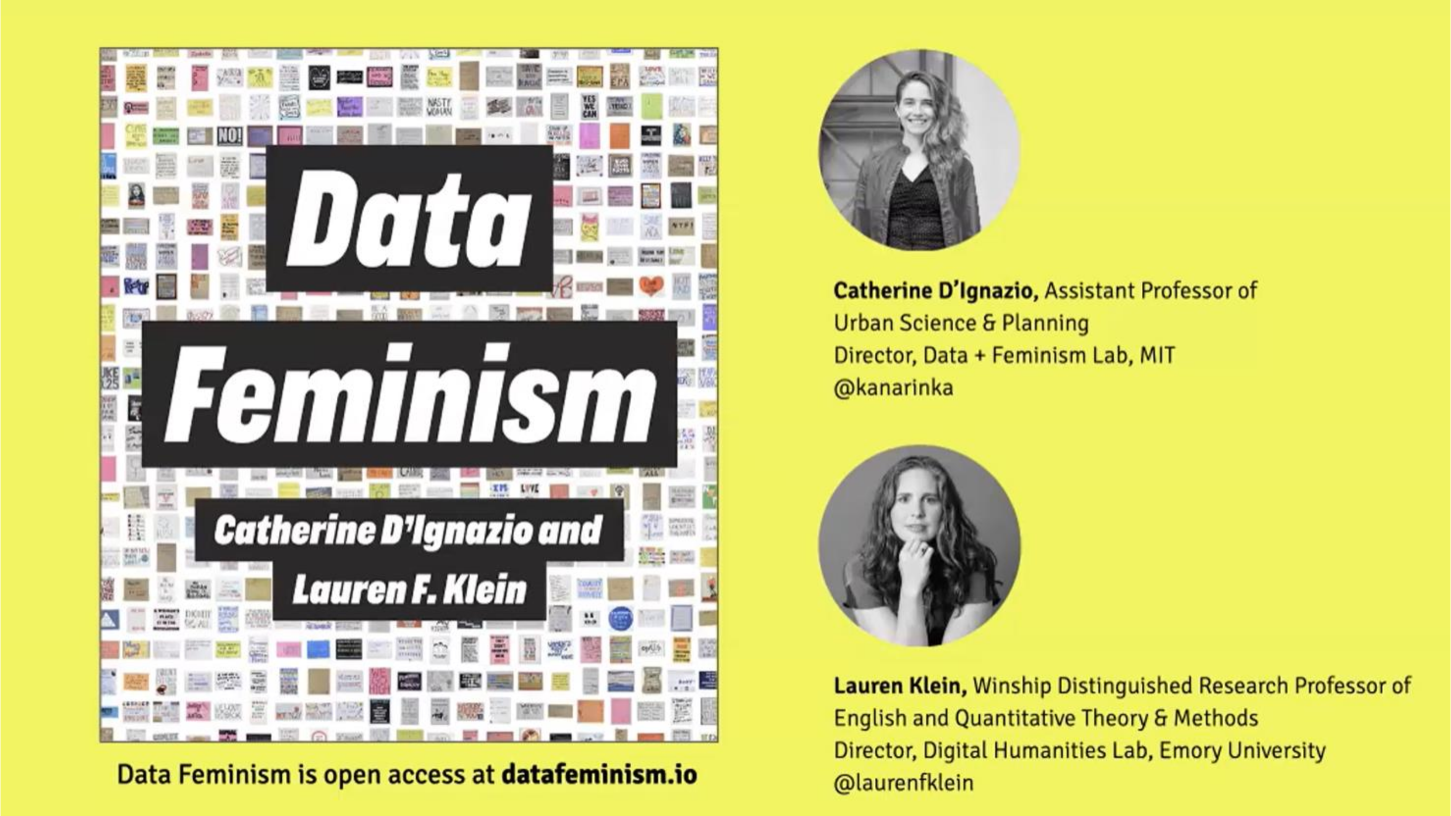
Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. Big Data & Society.

Data Justice



Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. Big Data & Society.

Data Justice: Power Asymmetries



Watchdogs

For Data Justice

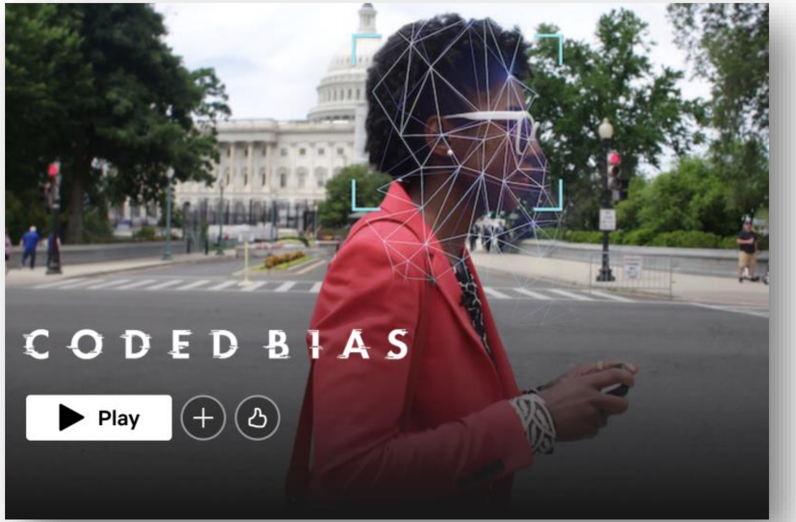
Algorithmic Justice League – AJL (USA)

- The Algorithmic Justice League is an organization that combines art, research, policy guidance and media advocacy to illuminate the social implications and harms of AI.
- AJL is a cultural movement towards
  - Equitable AI (agency and control, affirmative consent, centering justice)
  - Accountable AI (transparency, continuous oversight, redress harms)
- AJL recognizes the limitations of Ethical AI, which does not create any mandatory requirements or ban certain uses of AI. They focus on creating action.

https://www.ajl.org/

Algorithmic Justice League – AJL

- They lead projects, workshops.
- They provide algorithmic audits.
- You can join AJL to act now, donate, expose AI harms and biases, spread the word and so on.



https://www.ajl.org/spotlight-documentary-coded-bias

Ada Lovelace Institute (UK)

- An independent research institute
- They have a mission to ensure data and AI work for people and society
- They represent people to fight against power asymmetries
- Core values: research, policy and practice

https://www.adalovelaceinstitute.org/

Algorithm Watch (Germany)

- Algorithm Watch is a non-profit research and advocacy organization.
- They analyze automated decision-making systems to measure their impact on society.
- Algorithm Watch maintains AI Ethics Guidelines Global Inventory that includes 173 guidelines (April 2020).
- They have many projects to investigate how algorithms work in practice.

https://inventory.algorithmwatch.org/

Algorithm Watch

ALGORITHM WATCH

- An initial evaluation done in 2019 shows that AI ethics guidelines lack enforcement mechanisms (10 out of 160 mention this).
- Policies mostly include voluntary commitments/general recommendations.
- Other Issues: Guidelines come from wealthy countries.
- "*The question arises whether guidelines that can neither be applied nor enforced are not more harmful than having no ethical guidelines at all. Ethics guidelines should be more than a PR tool for companies and governments.*"

https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/
https://algorithmwatch.org/en/ethical-ai-guidelines-binding-commitment-or-simply-window-dressing/

Algorithm Watch – Example Case

- A professional association (the Institute of Electrical and Electronics Engineers – IEEE) publishes "Ethically Aligned Design" in 2016.
- The report includes general principles about transparency, human rights, accountability and many others.
- Algorithm Watch approaches Facebook, Google and Twitter to challenge them about how they implement the IEEE principles.

https://algorithmwatch.org/en/ieee-ethically-aligned-design-guidelines-fail-to-gain-traction/

Summary

- Algorithmic **Fairness**
  - Group Fairness
  - Individual Fairness
- Data **Justice**
- Watchdogs