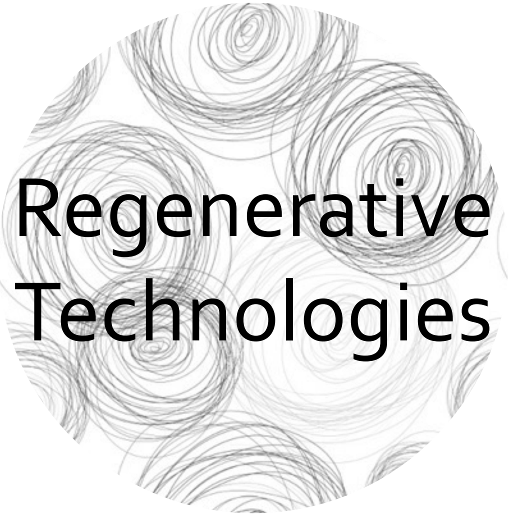
INFO901
16.03.2022

Unbiased data? Fair AI? Forget it!



Regenerative Technologies

SUSTAINABILITY LAB

SMART | Sustainable Market Actors for Responsible Trade (2016 – 2020)

Futuring Nordics | Futuring Sustainable Nordic Business Models (2019 – 2023)

Circular Energy | Circular Energy for a Sustainable Circular Economy (2021 – 2023)

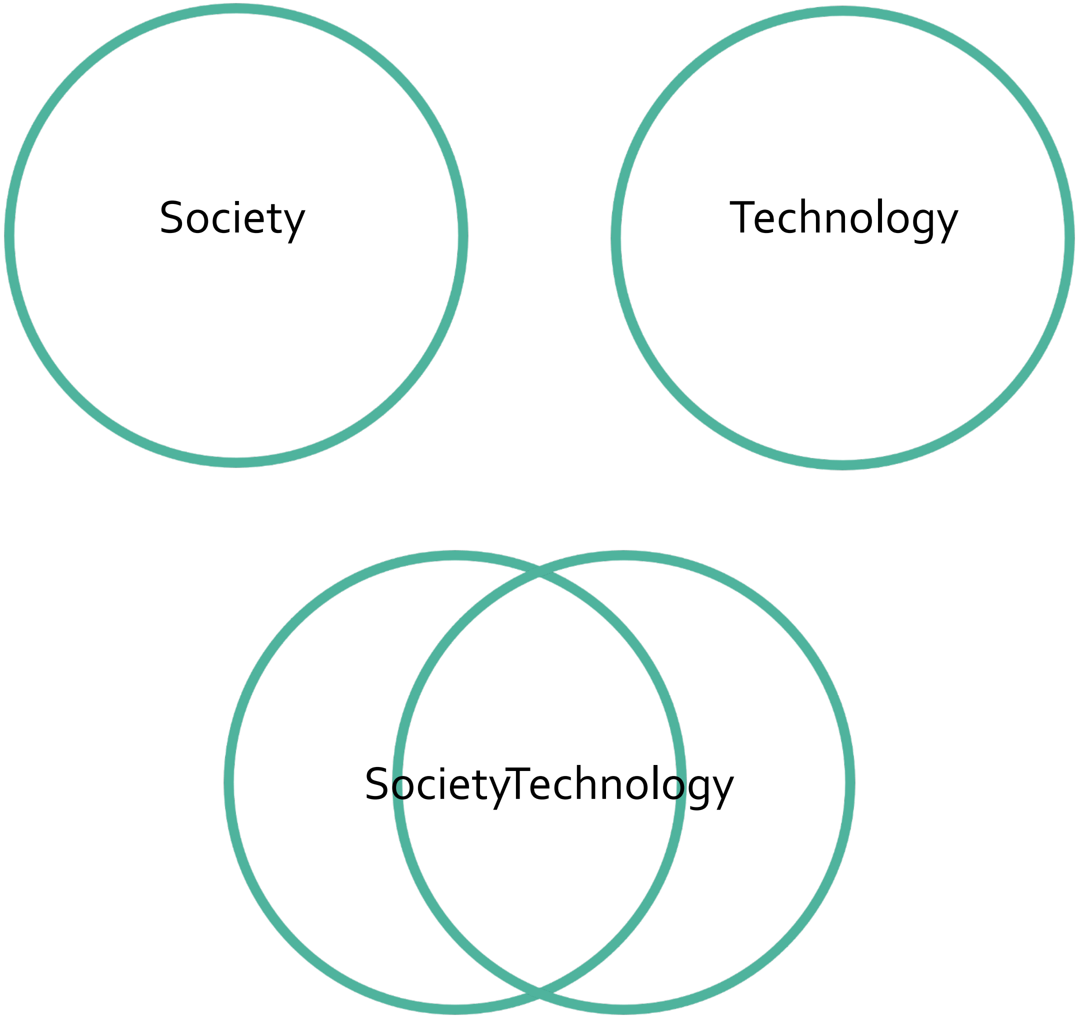
More a proposal than an abstract …

While we all are affected, or will be, by algorithms, some of us are more vulnerable than others to biased data and unfair AI. Is a focus on unbiased data and fair AI the solution? Is there a universal understanding of fairness? Are there sources of neutral data or can we make existing data sets unbiased? If we answer ‘yes’ on these questions, does it mean that AI can be neutral? In this lecture we will engage with the understanding that technology is not neutral and explore what this means for working towards unbiased data and fair AI.

[The] concept of ***imagined objectivity*** emphasizes the role that cultural assumptions and personal preconceptions play in upholding this false belief: one imagines (wrongly) that datasets and algorithms are less partial and less discriminatory than people and thus more "objective."

(Ruha Benjamin in D'Ignazio & Klein, 2020)

Relationship technology - society

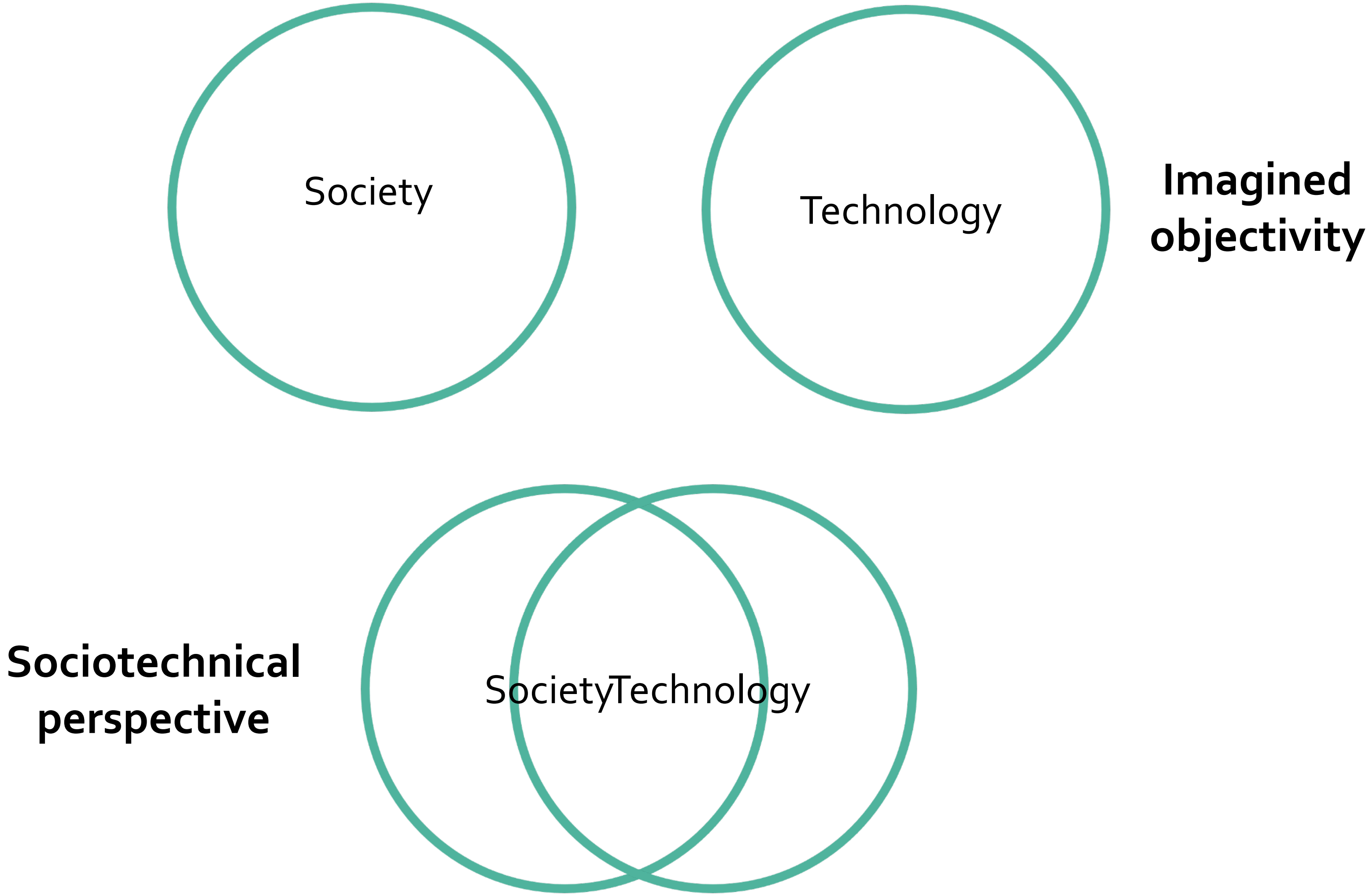


Society

Technology

Society Technology

Sociotechnical perspective



Society

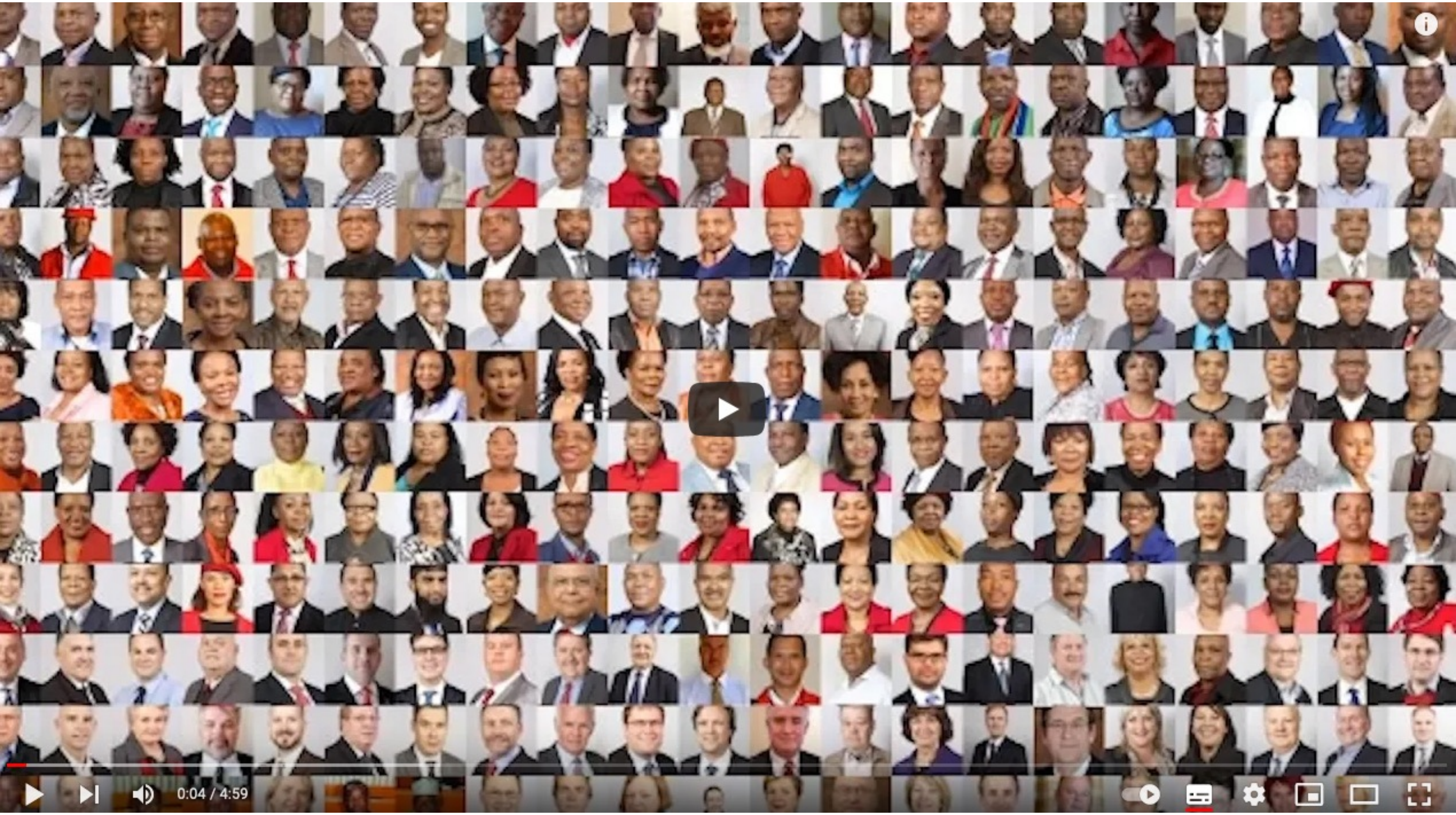
Technology

Imagined objectivity

Sociotechnical perspective

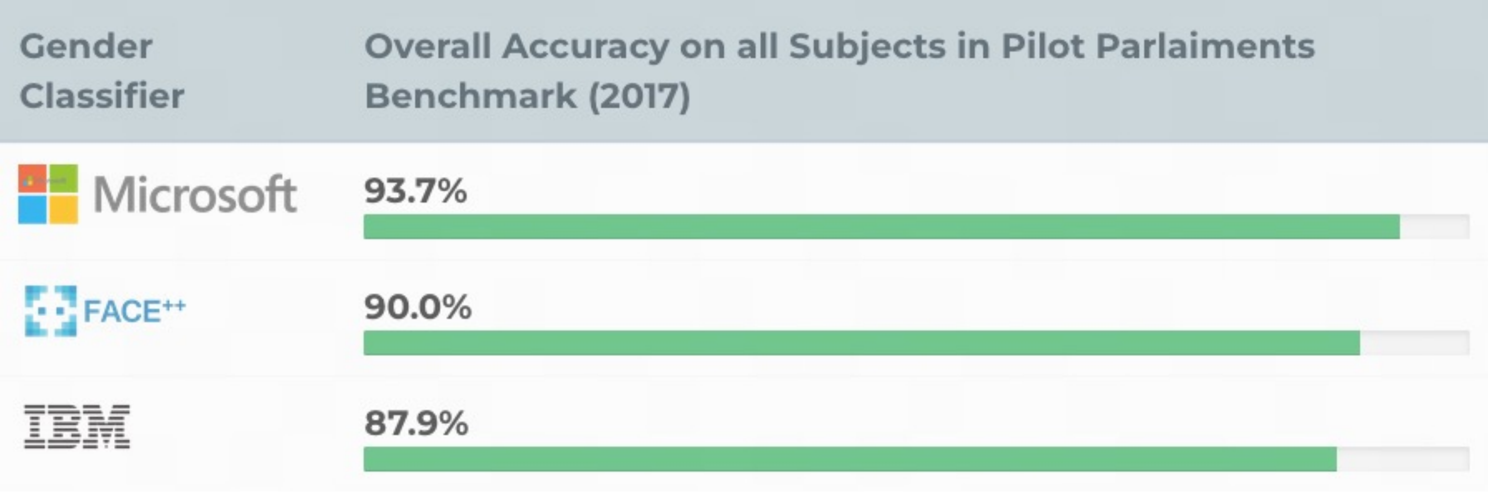
SocietyTechnology

Gender Shades



Source: https://youtu.be/TWWsW1w-BVo

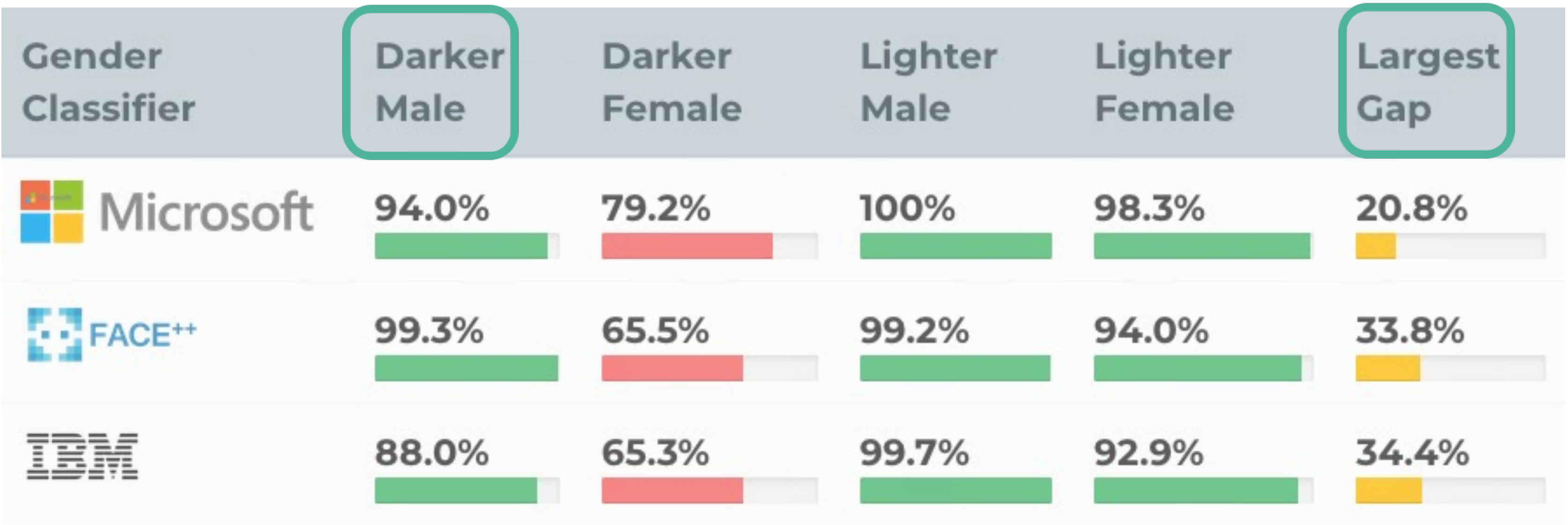
Aggregation bias



Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

Taken at face value, gender classification accuracies ranging from 87.9% to 93.7% on the PPB dataset, suggest that these classifiers can be used for all populations represented by the benchmark. A company might justify the market readiness of a classifier by presenting performance results in **aggregate** (Buolamwini & Gebru, 2018).

Intersectional bias



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Yet a gender and phenotypic breakdown of the results shows that performance differs substantially for distinct sub-groups. Classification is 8.1% – 20.6% worse on female than male subjects and 11.8% – 19.2% worse on darker than lighter subjects (Buolamwini & Gebry, 2018).

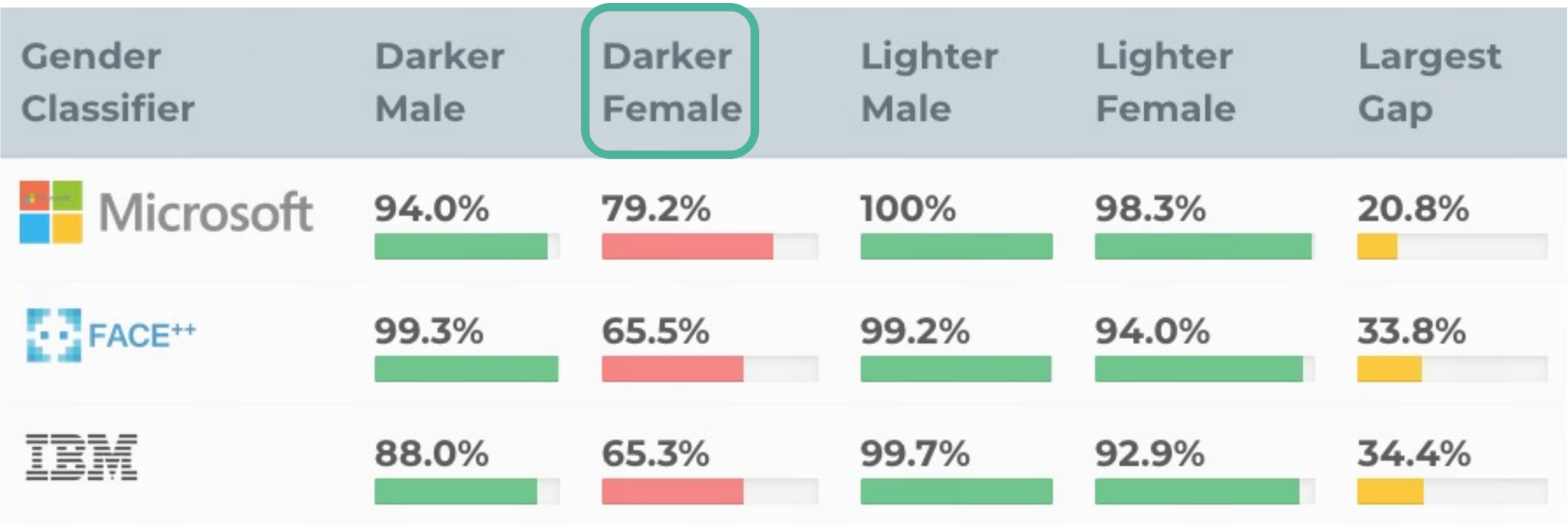
Intersectionality

Kimberlé Crenshaw, law professor at Columbia and UCLA coined the term intersectionality 30 years ago to describe the way people's social identities can overlap:

"It's basically a lens, a prism, for seeing the way in which various forms of inequality often operate together and exacerbate each other. We tend to talk about race inequality as separate from inequality based on gender, class, sexuality or immigrant status. What's often missing is how some people are subject to all of these, and **the experience is not just the sum of its parts**."

https://time.com/5786710/kimberle-crenshaw-intersectionality/

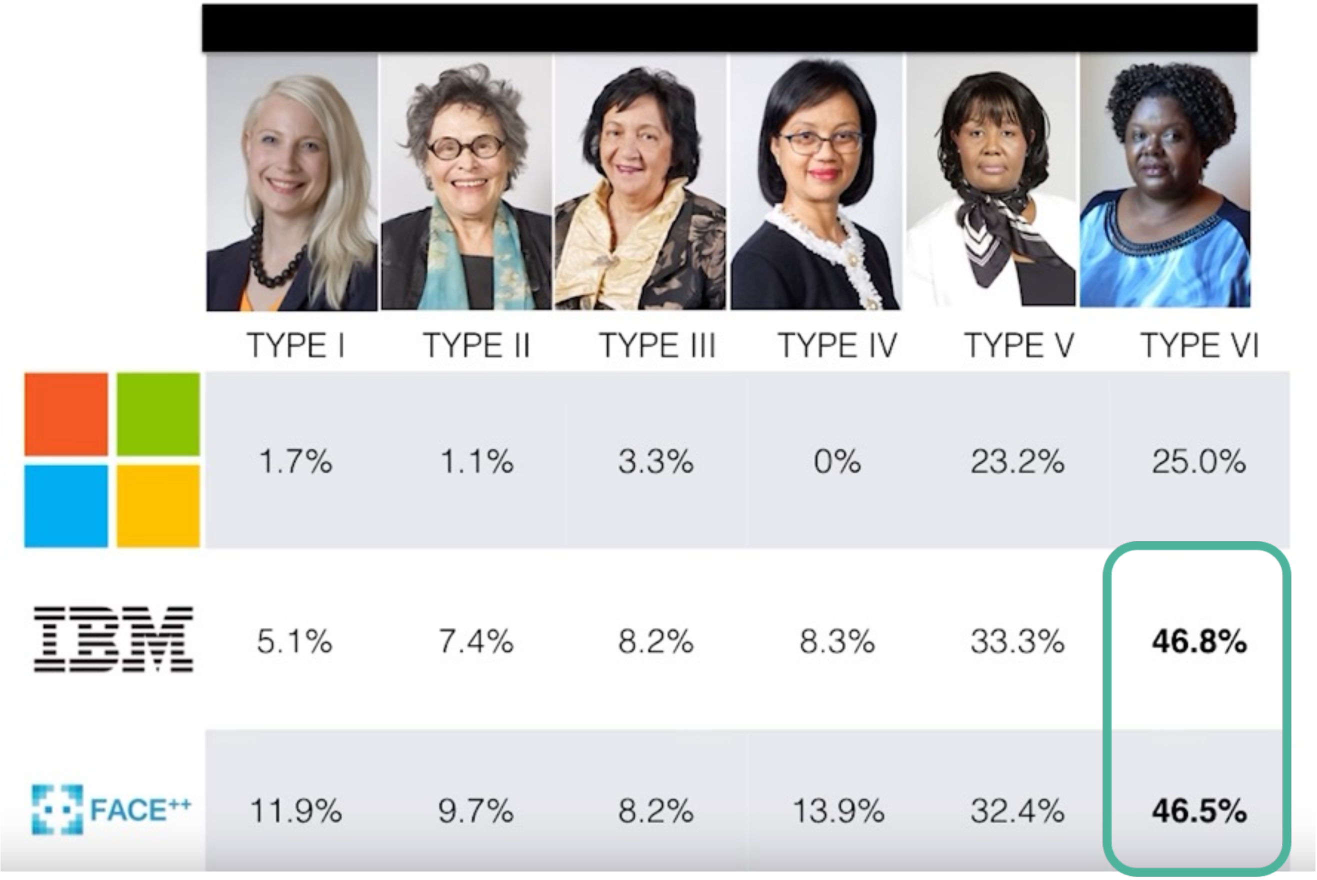
Intersectionality



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

The intersectional error analysis that targets gender classification performance on darker female, lighter female, darker male, and lighter male subgroups provides more answers. Darker females have the highest error rates for all gender classifiers ranging from 20.8% – 34.7% (Buolamwini & Gebru, 2018).

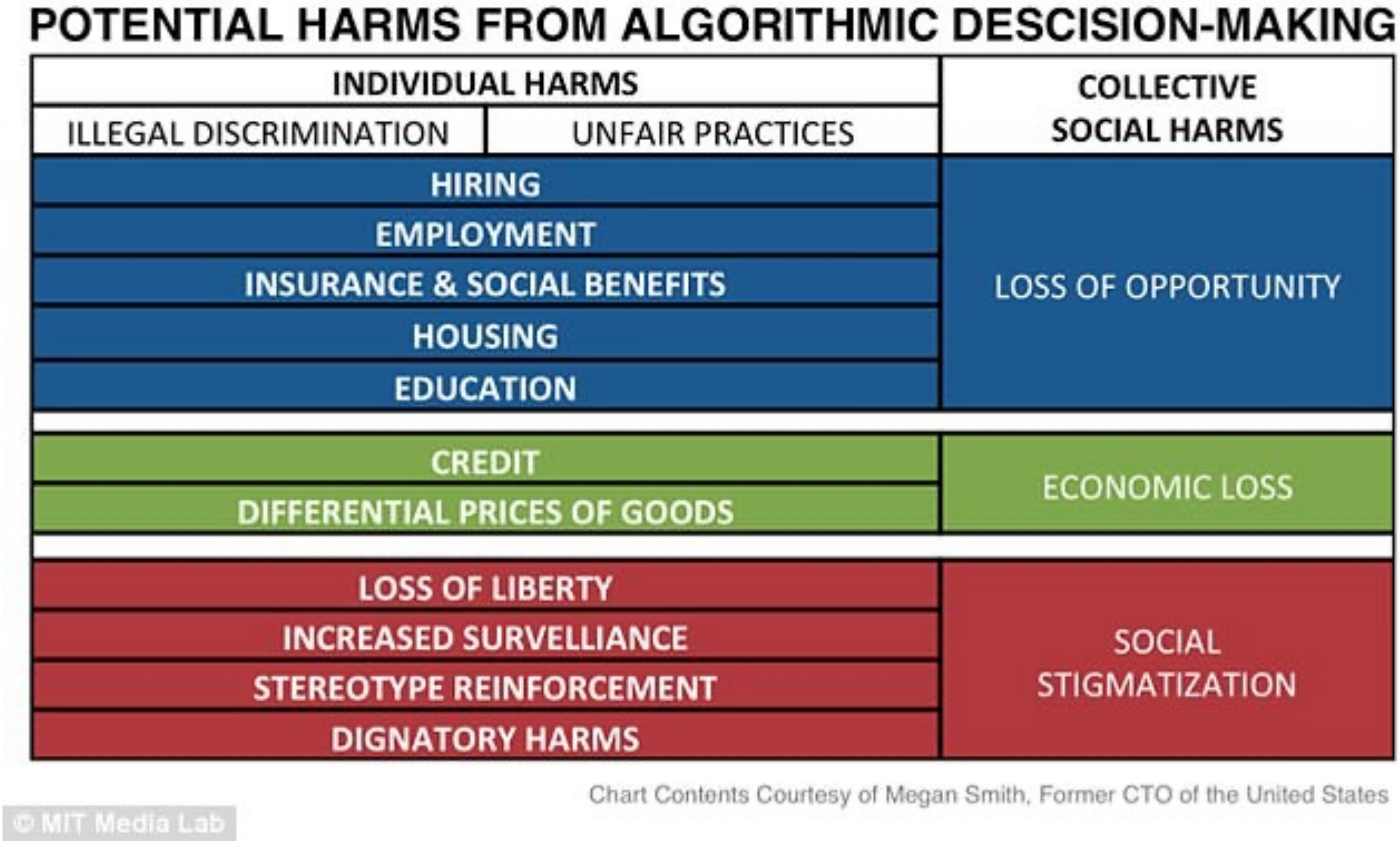
Intersectionality



	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
	1.7%	1.1%	3.3%	0%	23.2%	25.0%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	**46.8%**
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	**46.5%**

"In fact, as we tested women with darker and darker skin, the chances of being correctly gendered came close to a coin toss". Bulamwini in Gender Shades (https://youtu.be/TWWsW1w-BVo)

Unfair or flawed AI



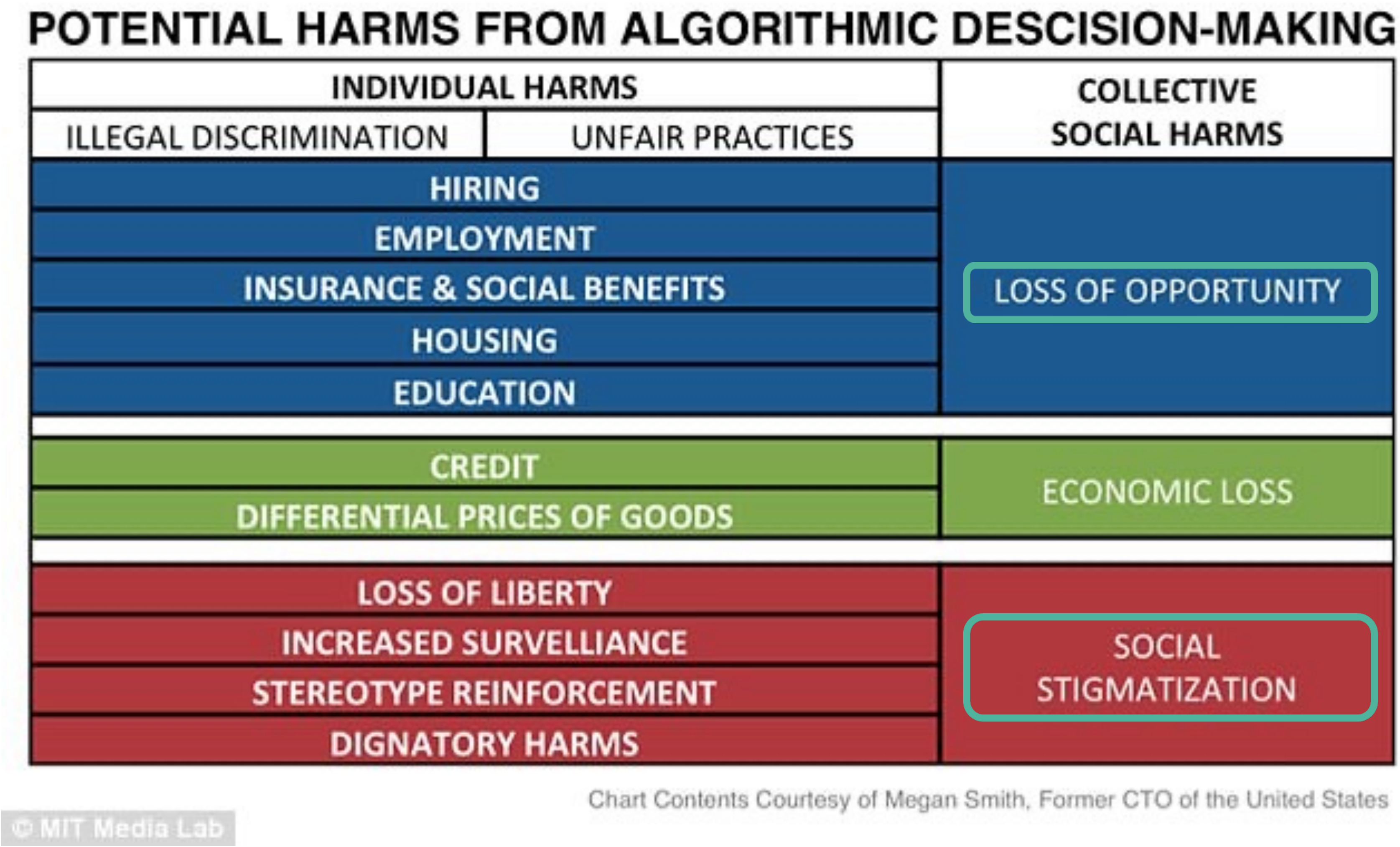
POTENTIAL HARMS FROM ALGORITHMIC DESCISION-MAKING

INDIVIDUAL HARMS		COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES	
HIRING		LOSS OF OPPORTUNITY
EMPLOYMENT		
INSURANCE & SOCIAL BENEFITS		
HOUSING		
EDUCATION		
CREDIT		ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS		
LOSS OF LIBERTY		SOCIAL STIGMATIZATION
INCREASED SURVELLIANCE		
STEREOTYPE REINFORCEMENT		
DIGNATORY HARMS		

Chart Contents Courtesy of Megan Smith, Former CTO of the United States

Buolamwini & Gebru, 2018)

Unfair or flawed AI



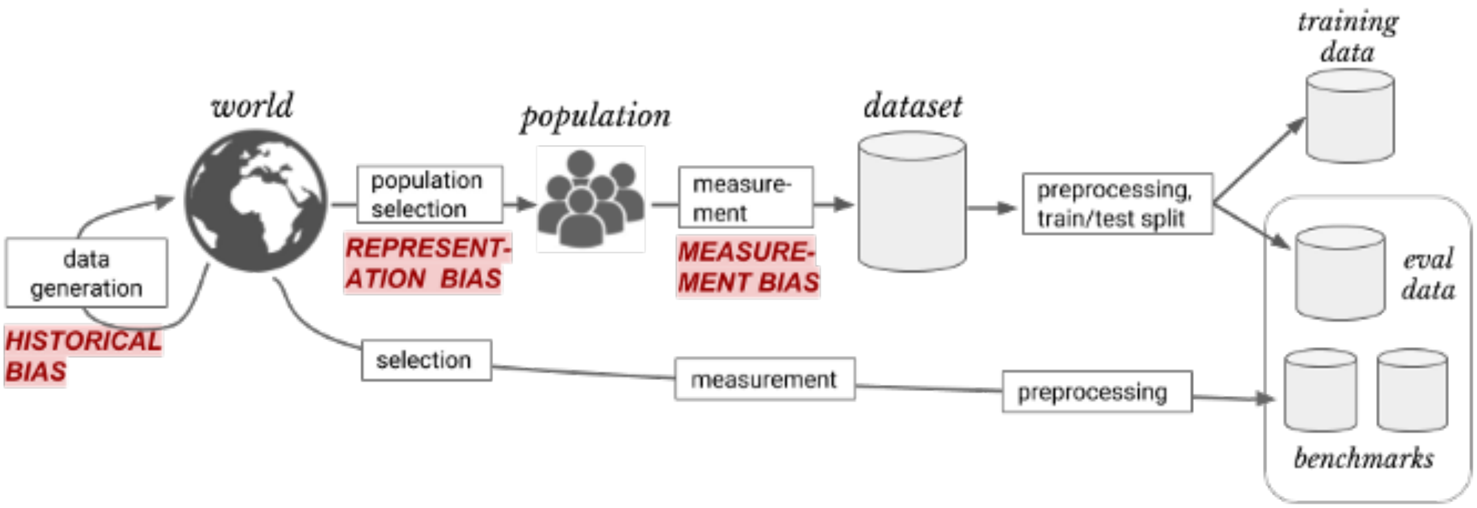
POTENTIAL HARMS FROM ALGORITHMIC DESCISION-MAKING

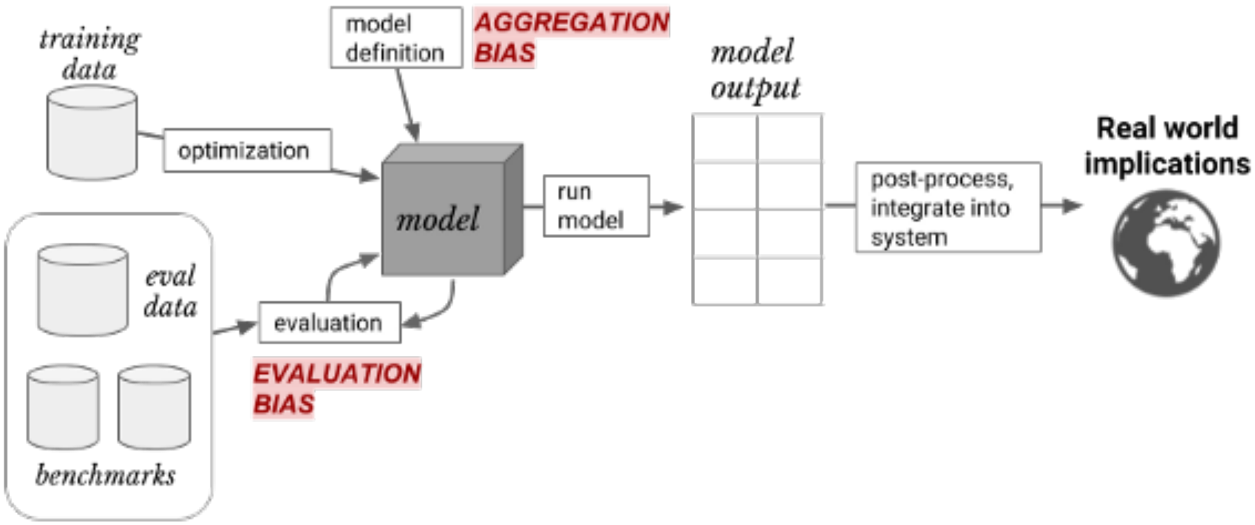
INDIVIDUAL HARMS		COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES	
HIRING		LOSS OF OPPORTUNITY
EMPLOYMENT		
INSURANCE & SOCIAL BENEFITS		
HOUSING		
EDUCATION		
CREDIT		ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS		
LOSS OF LIBERTY		SOCIAL STIGMATIZATION
INCREASED SURVELLIANCE		
STEREOTYPE REINFORCEMENT		
DIGNATORY HARMS		

Chart Contents Courtesy of Megan Smith, Former CTO of the United States

Buolamwini & Gebru, 2018)

Biased Data

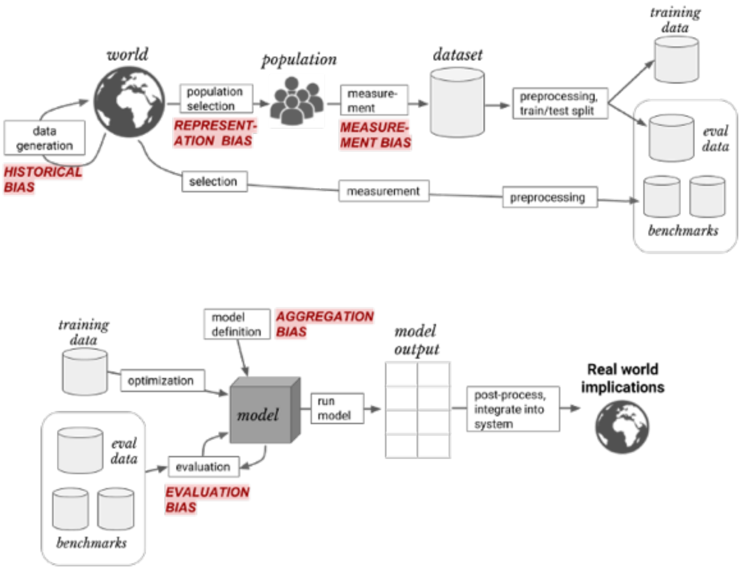




(Suresh, 2019; Suresh & Guttag, 2021)

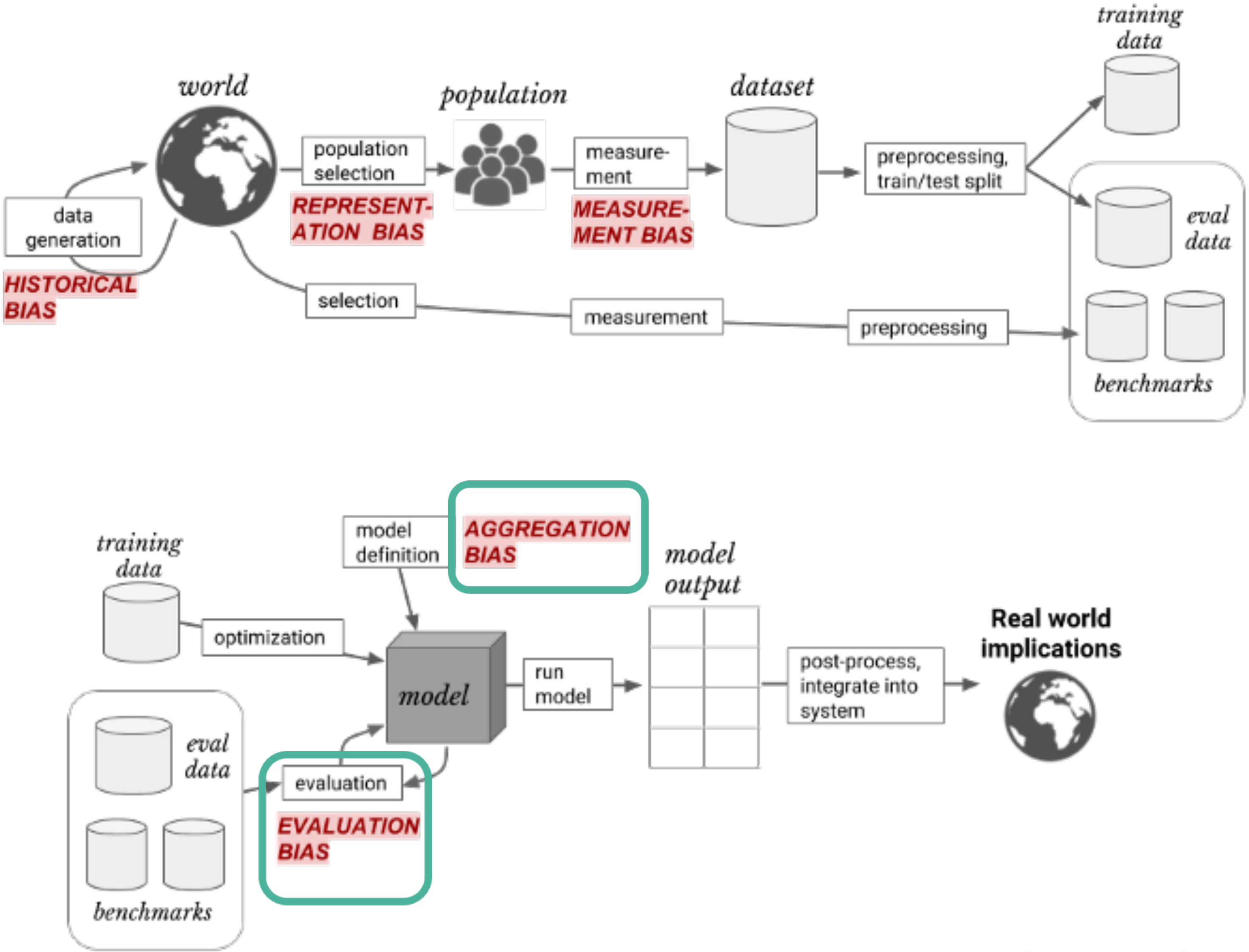
Biased Data

- **Historical bias** arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.
- **Representation bias** arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently causes worse performance, for some part of the final population.
- **Measurement bias** arises when choosing and measuring the particular features and labels of interest. Features considered to be relevant to the outcome are chosen, but these can be incomplete or contain group- or input-dependent noise. In many cases, the choice of a single label to create a classification task may be an oversimplification that more accurately measures the true outcome of interest for certain groups.
- **Evaluation bias** occurs during model iteration and evaluation, when the testing or external benchmark populations do not equally represent the various parts of the final population. Evaluation bias can also arise from the use of performance metrics that are not granular or comprehensive enough.
- **Aggregation bias** arises when flawed assumptions about the population affect model definition. In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.



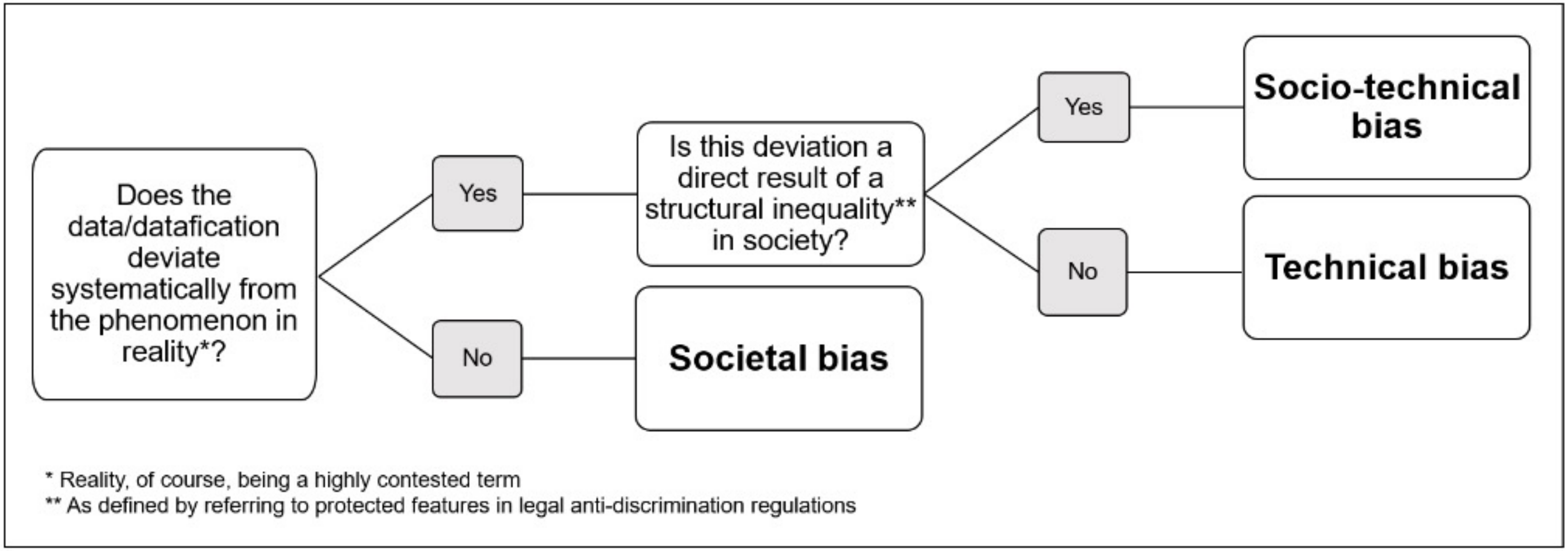
(Suresh, 2019; Suresh & Guttag, 2021)

Biased Data



(Suresh, 2019; Suresh & Guttag, 2021)

Socio-technical typology of bias


Does the data/datafication deviate systematically from the phenomenon in reality*?
Yes
No
Is this deviation a direct result of a structural inequality** in society?
Societal bias
Yes
No
Socio-technical bias
Technical bias

* Reality, of course, being a highly contested term

** As defined by referring to protected features in legal anti-discrimination regulations

FIGURE 1: Bias scheme

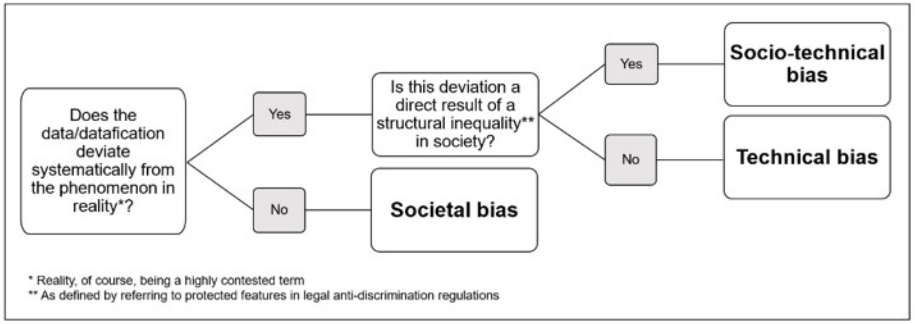
Lopez, Paola (2021). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems

Socio-technical typology of bias

- **Technical bias**: Any kind of technical or conceptual mis-measurement and misconception

- **Socio-technical bias**: A discrepancy between what is to be represented and what is being represented, and this discrepancy is a direct result of structural inequalities.

- **Societal bias**: Arise when structural inequalities are reflected in the respective data, albeit correctly.


Does the data/datafication deviate systematically from the phenomenon in reality*?
Yes
No
Is this deviation a direct result of a structural inequality** in society?
Societal bias
Yes
No
Socio-technical bias
Technical bias

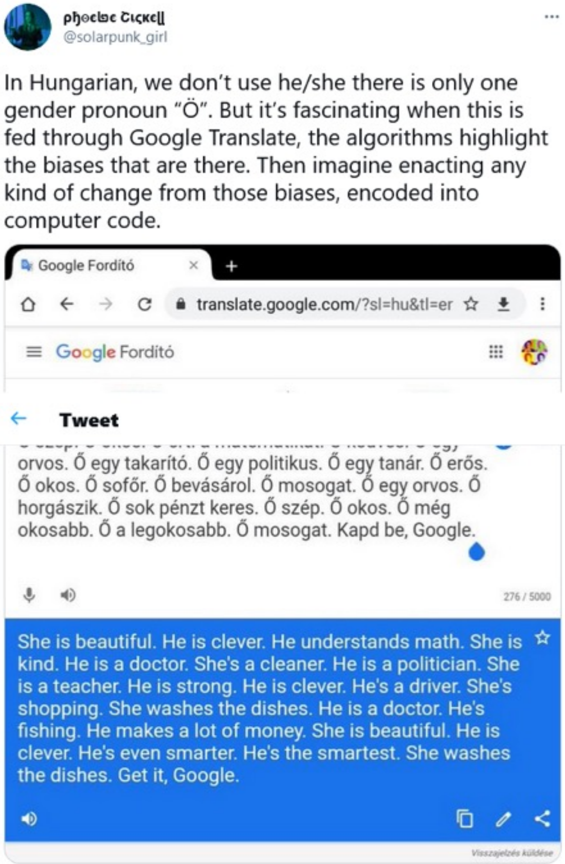
* Reality, of course, being a highly contested term
** As defined by referring to protected features in legal anti-discrimination regulations

FIGURE 1: Bias scheme

Lopez, 2021

Technical bias

- Any kind of technical or conceptual mis-measurement and misconception

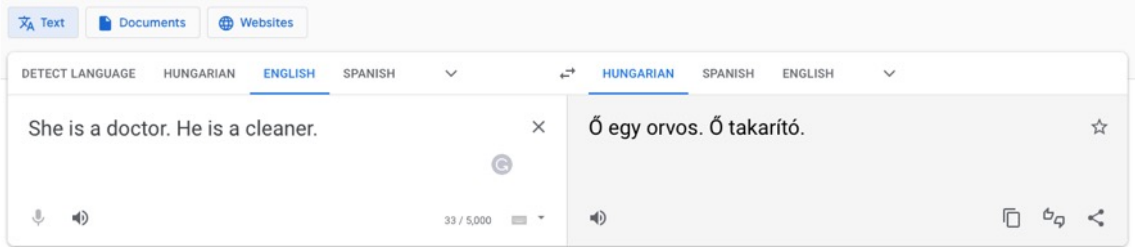


phoebe tickell
@solarpunk_girl

In Hungarian, we don't use he/she there is only one gender pronoun "Ö". But it's fascinating when this is fed through Google Translate, the algorithms highlight the biases that are there. Then imagine enacting any kind of change from those biases, encoded into computer code.

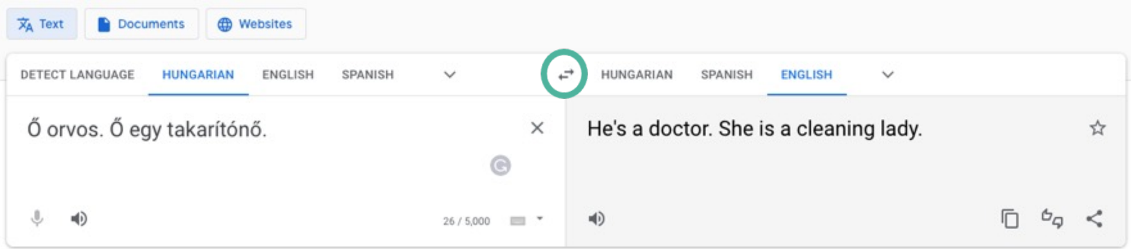
orvos. Ő egy takarító. Ő egy politikus. Ő egy tanár. Ő erős. Ő okos. Ő sofőr. Ő bevásárol. Ő mosogat. Ő egy orvos. Ő horgászik. Ő sok pénzt keres. Ő szép. Ő okos. Ő még okosabb. Ő a legokosabb. Ő mosogat. Kapd be, Google.

She is beautiful. He is clever. He understands math. She is kind. He is a doctor. She's a cleaner. He is a politician. She is a teacher. He is strong. He is clever. He's a driver. She's shopping. She washes the dishes. He is a doctor. He's fishing. He makes a lot of money. She is beautiful. He is clever. He's even smarter. He's the smartest. She washes the dishes. Get it, Google.



She is a doctor. He is a cleaner.

Ő egy orvos. Ő takarító.



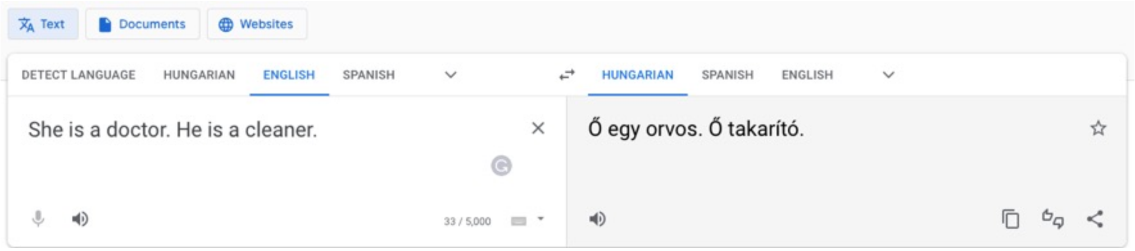
Ő orvos. Ő egy takarítónő.

He's a doctor. She is a cleaning lady.

Lopez, 2021

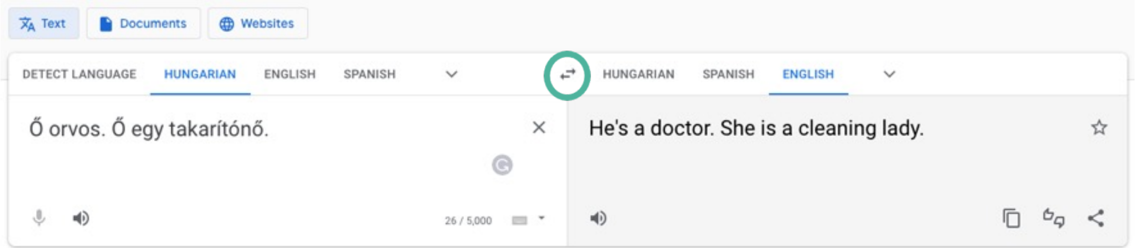
Technical bias

- Any kind of technical or conceptual mis-measurement and misconception



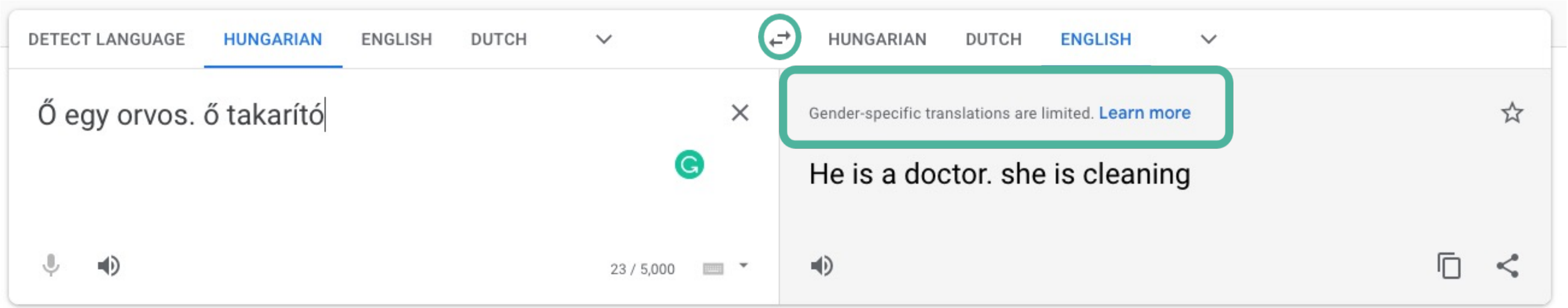
She is a doctor. He is a cleaner.

Ő egy orvos. Ő takarító.



Ő orvos. Ő egy takarítónő.

He's a doctor. She is a cleaning lady.



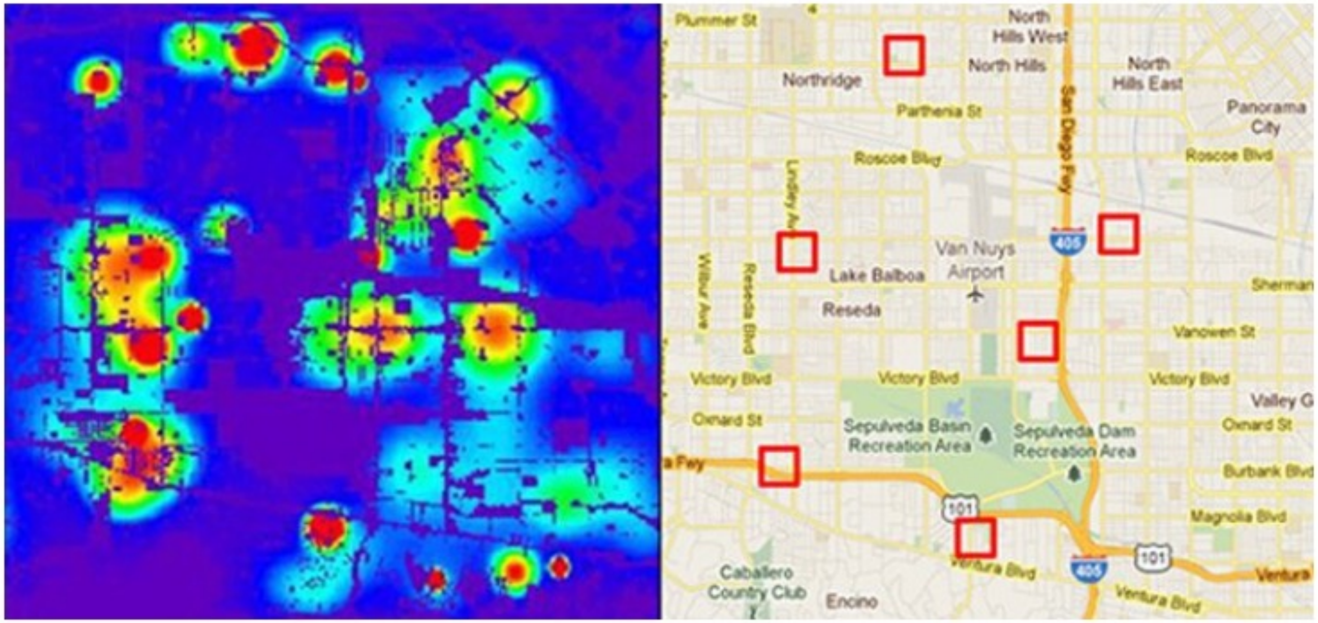
Ő egy orvos. ő takarító

Gender-specific translations are limited. Learn more

He is a doctor. she is cleaning

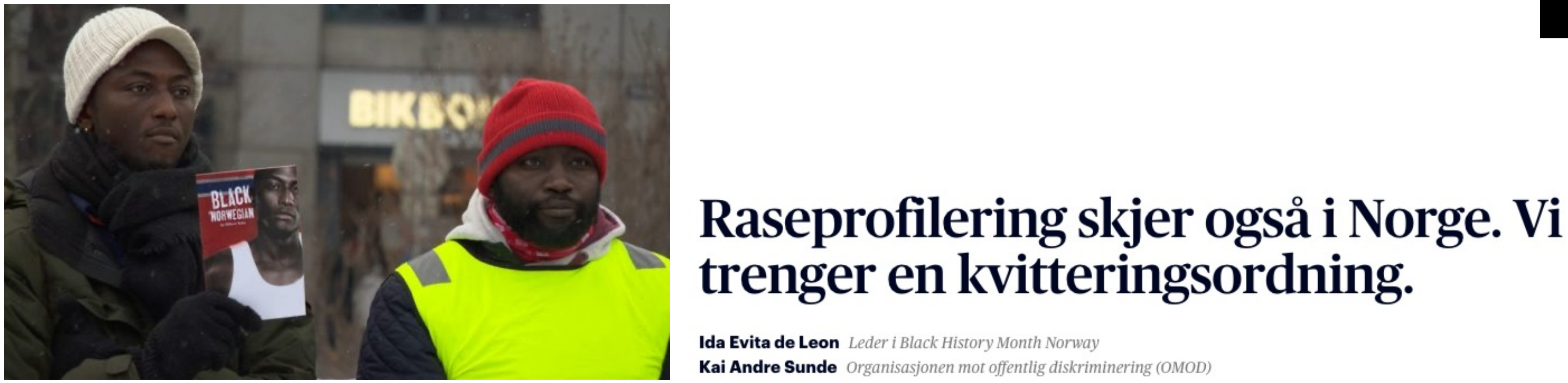
Socio-technical bias

- ○ A discrepancy between what is to be represented and what is being represented, and this discrepancy is a direct result of structural inequalities.



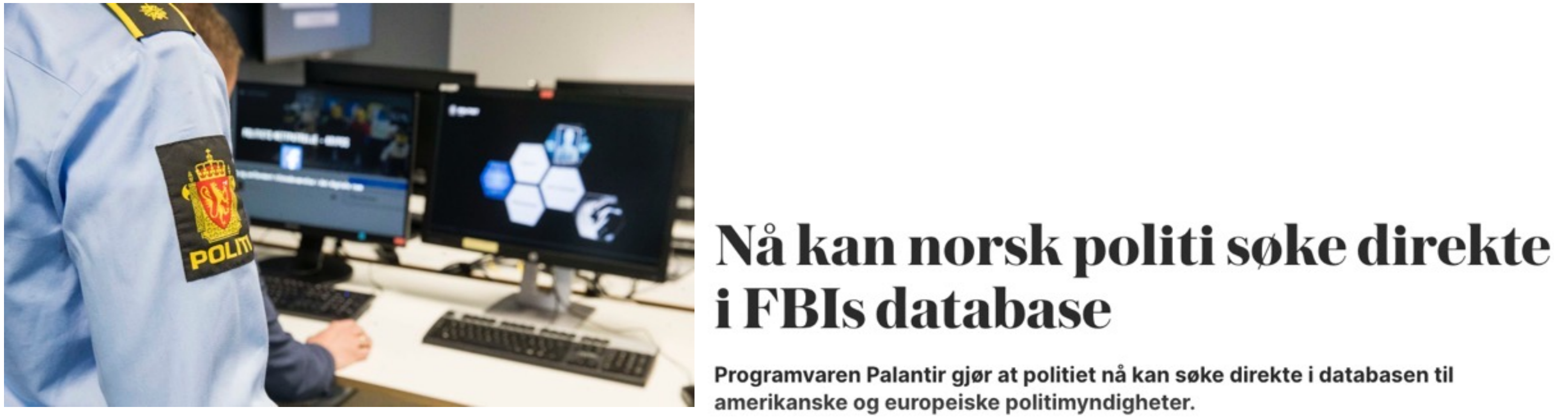
Predictive policing is built around algorithms that identify potential crime hotspots.. PredPol

Lopez, 2021



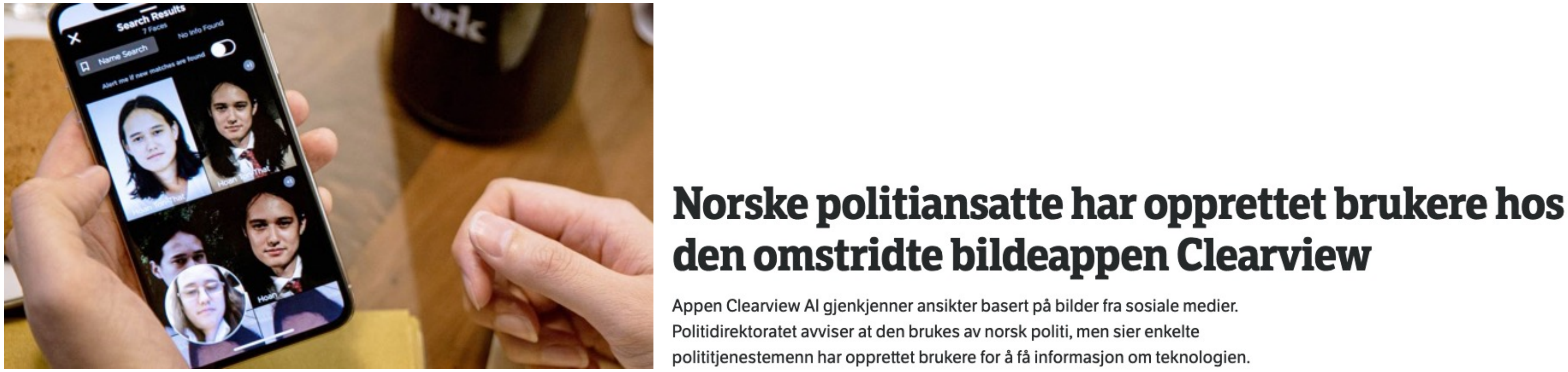
Raseprofilering skjer også i Norge. Vi trenger en kvitteringsordning.

Ida Evita de Leon *Leder i Black History Month Norway*
Kai Andre Sunde *Organisasjonen mot offentlig diskriminering (OMOD)*



Nå kan norsk politi søke direkte i FBIs database

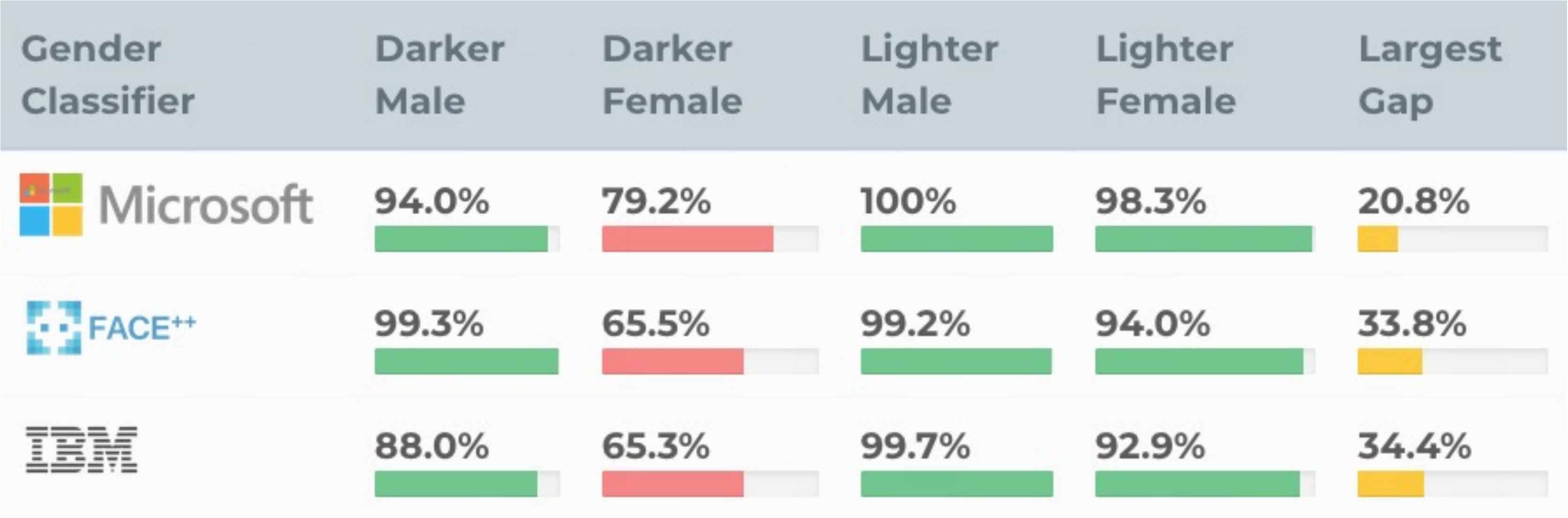
Programvaren Palantir gjør at politiet nå kan søke direkte i databasen til amerikanske og europeiske politimyndigheter.



Norske politiansatte har opprettet brukere hos den omstridte bildeappen Clearview

Appen Clearview AI gjenkjenner ansikter basert på bilder fra sosiale medier. Politidirektoratet avviser at den brukes av norsk politi, men sier enkelte polititjenestemenn har opprettet brukere for å få informasjon om teknologien.

Socio-technical bias



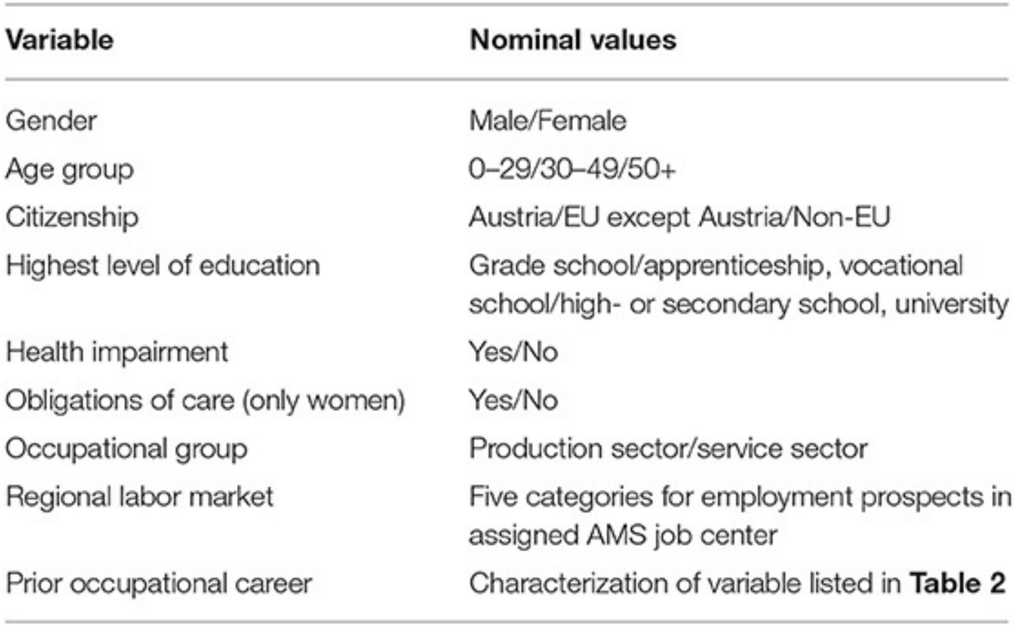
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini & Gebru, 2018)

Socio-technical bias

Societal bias

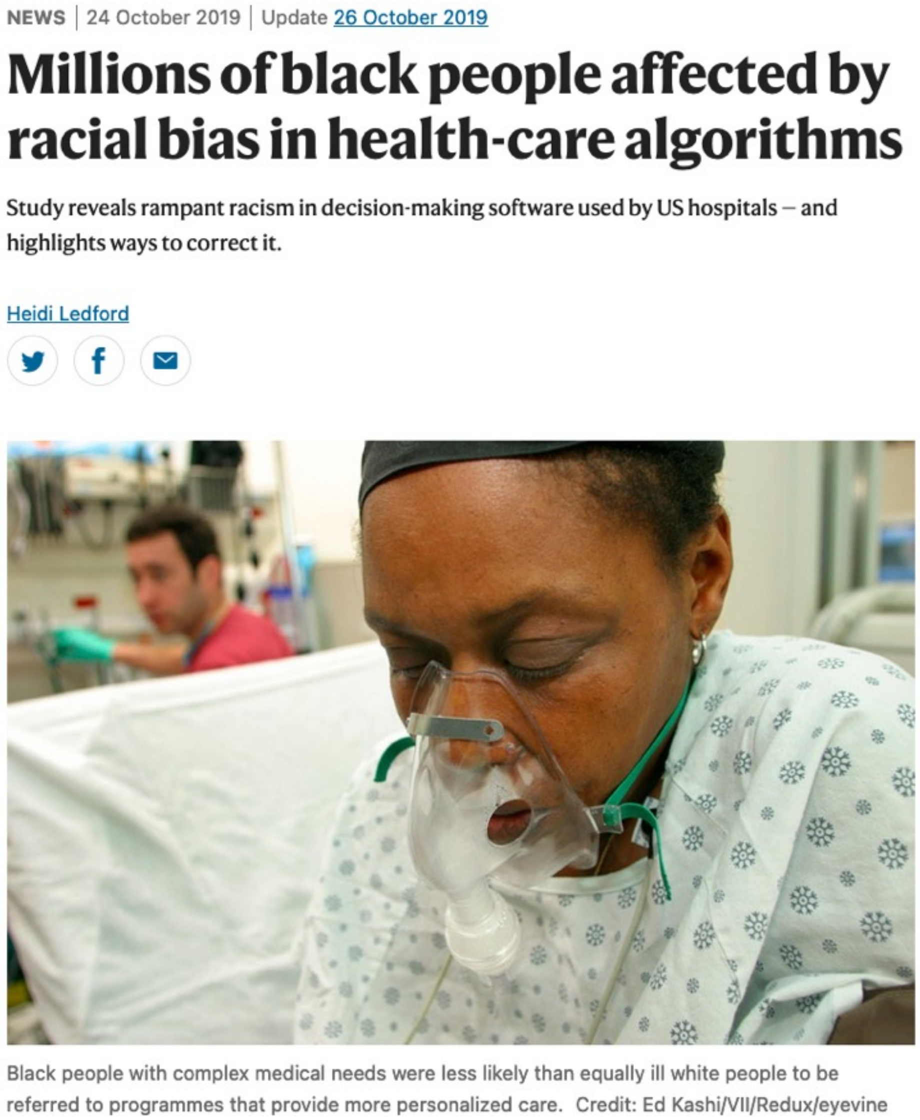
- **Societal bias**: When structural inequalities are reflected in the respective data, albeit correctly.



Variable	Nominal values
Gender	Male/Female
Age group	0–29/30–49/50+
Citizenship	Austria/EU except Austria/Non-EU
Highest level of education	Grade school/apprenticeship, vocational school/high- or secondary school, university
Health impairment	Yes/No
Obligations of care (only women)	Yes/No
Occupational group	Production sector/service sector
Regional labor market	Five categories for employment prospects in assigned AMS job center
Prior occupational career	Characterization of variable listed in **Table 2**

Lopez, 2021

Can biases be fixed?



NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Heidi Ledford

Black people with complex medical needs were less likely than equally ill white people to be referred to programmes that provide more personalized care. Credit: Ed Kashi/VII/Redux/eyevine

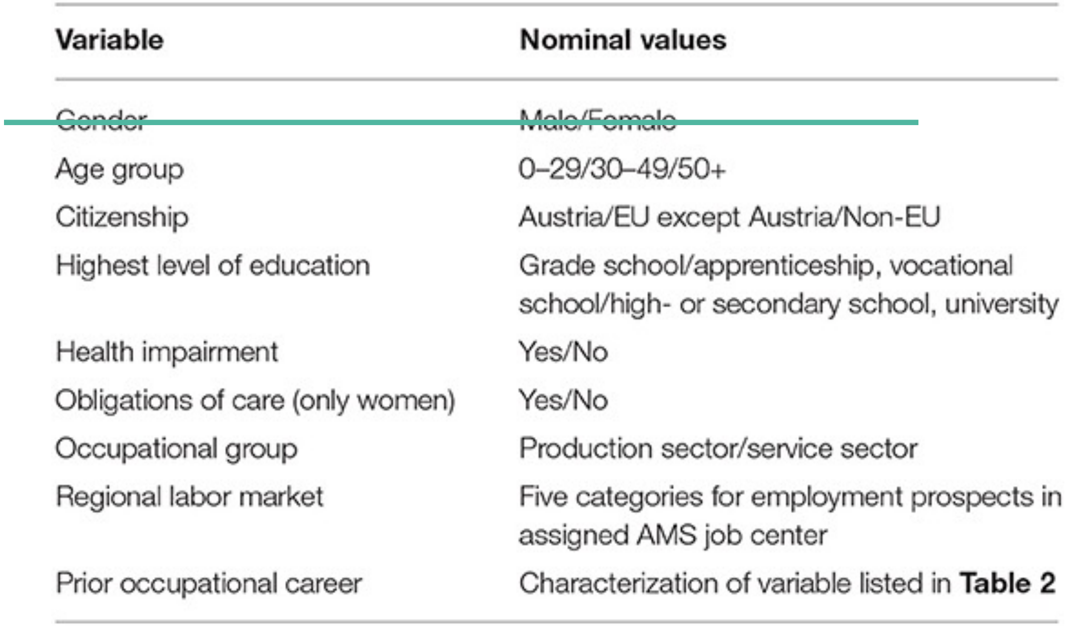
An algorithmic system built on a perfect datafication can reinforce inequalities – **depending on its context of use**.

- Healthcare
- Preventive policing
- Austrian Social Services

(Lopez, 2021)

Societal bias

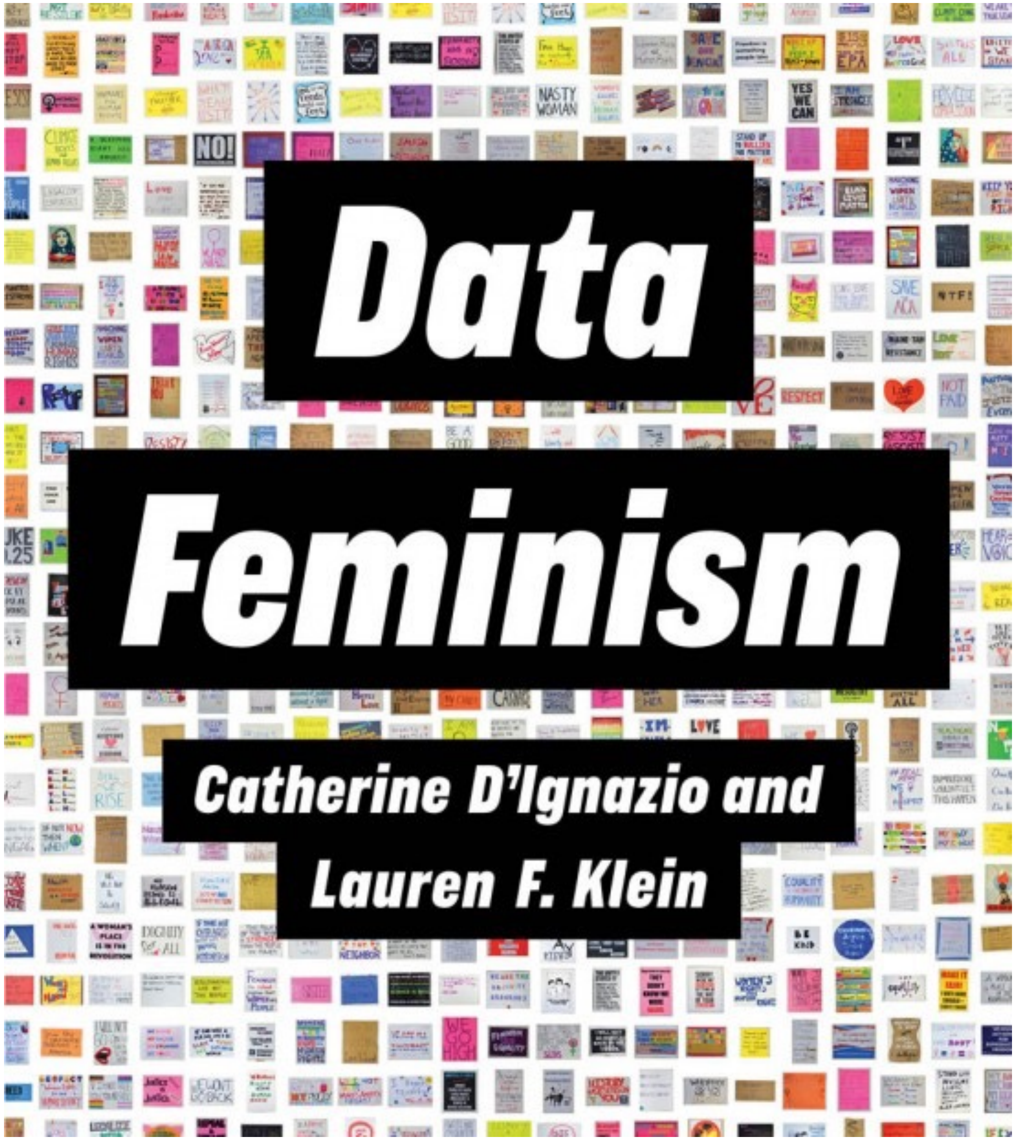
- **Societal bias**: When structural inequalities are reflected in the respective data, albeit correctly.



Variable	Nominal values
~~Gender~~	~~Male/Female~~
Age group	0–29/30–49/50+
Citizenship	Austria/EU except Austria/Non-EU
Highest level of education	Grade school/apprenticeship, vocational school/high- or secondary school, university
Health impairment	Yes/No
Obligations of care (only women)	Yes/No
Occupational group	Production sector/service sector
Regional labor market	Five categories for employment prospects in assigned AMS job center
Prior occupational career	Characterization of variable listed in **Table 2**

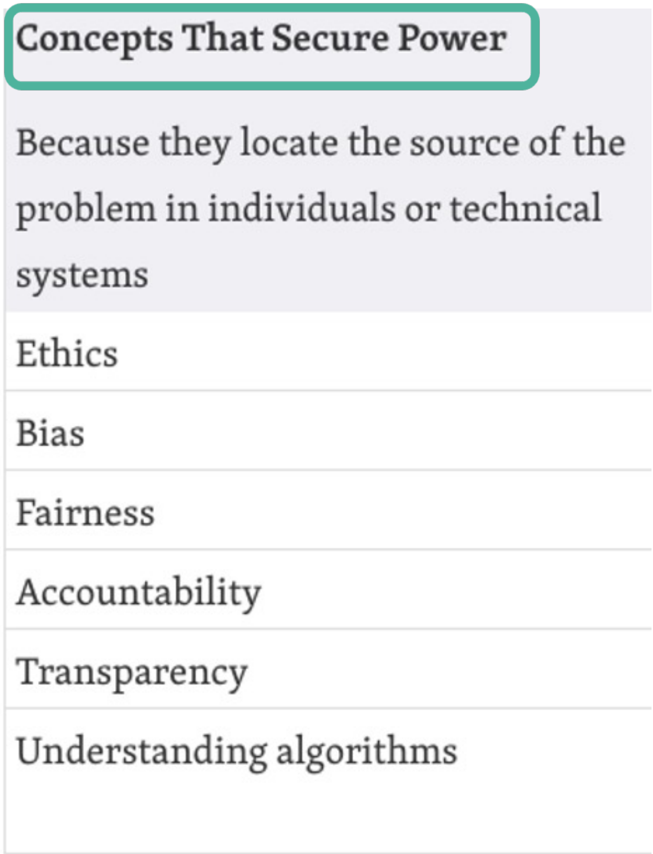
Lopez, 2021

Context of use …



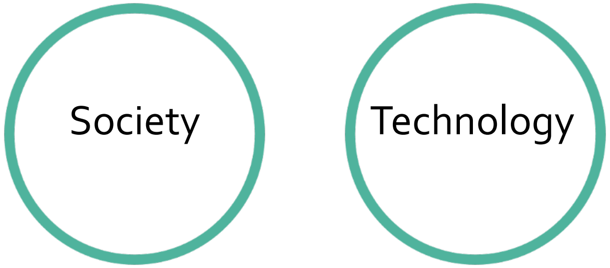
D'Ignazio & Klein, 2020

Data ethics



Concepts That Secure Power
Because they locate the source of the problem in individuals or technical systems
Ethics
Bias
Fairness
Accountability
Transparency
Understanding algorithms

"Addressing bias in a dataset is a tiny technological Band-Aid for a much larger problem. Even the values mentioned here, which seek to address instances of bias in data-driven systems, are themselves non-neutral, as they locate the source of the bias in individual people and specific design decisions. So how might we develop a practice that results in **data-driven systems that challenge power at its source**?
(D'Ignazio & Klein, 2020)



Society

Technology

From data ethics to data justice

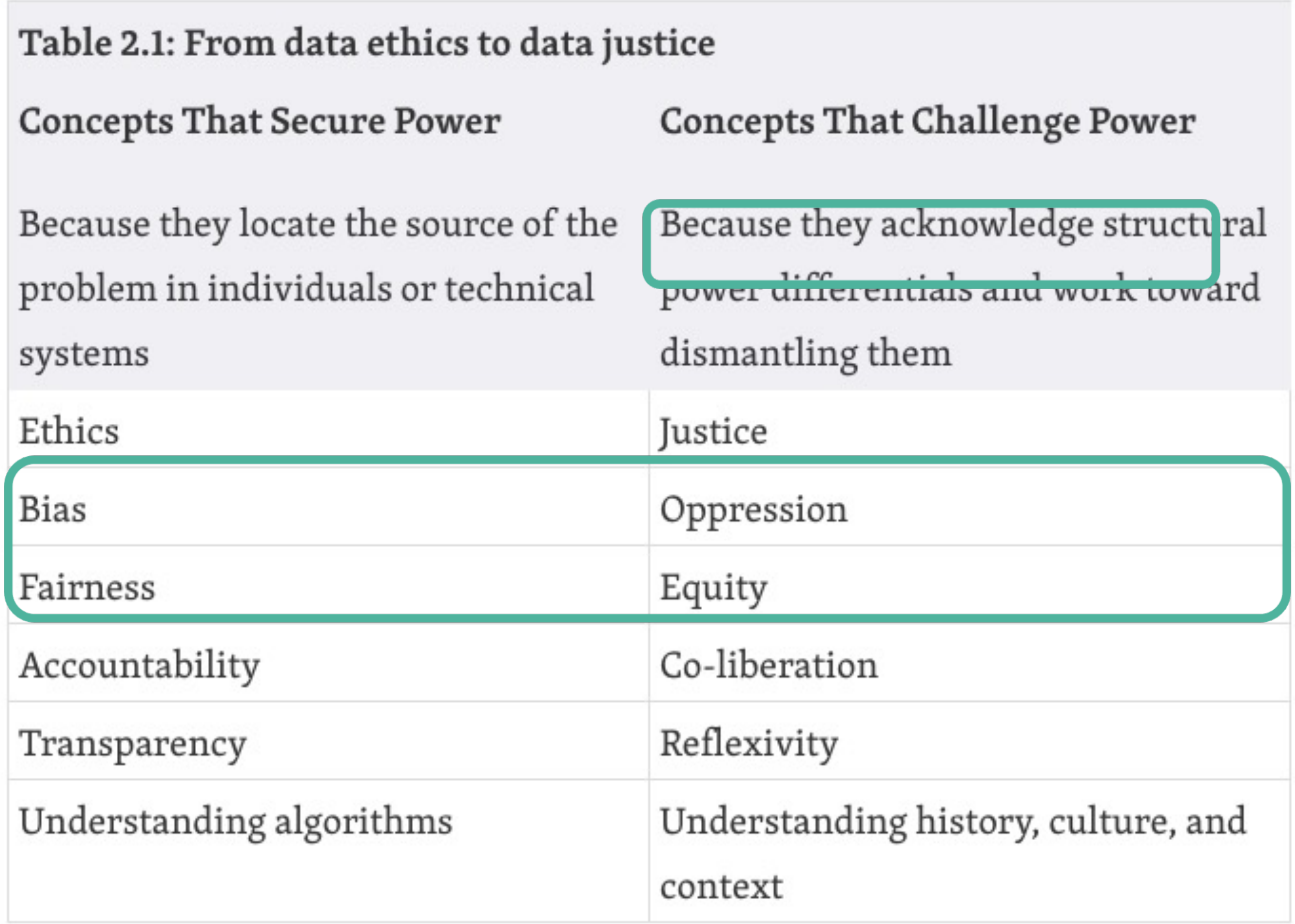


Table 2.1: From data ethics to data justice

Concepts That Secure Power	**Concepts That Challenge Power**
Because they locate the source of the problem in individuals or technical systems	Because they acknowledge structural power differentials and work toward dismantling them
Ethics	Justice
Bias	Oppression
Fairness	Equity
Accountability	Co-liberation
Transparency	Reflexivity
Understanding algorithms	Understanding history, culture, and context

SocietyTechnology

D'Ignazio & Klein, 2020

Unbiased data? Fair AI?

- **Bias – Oppression**

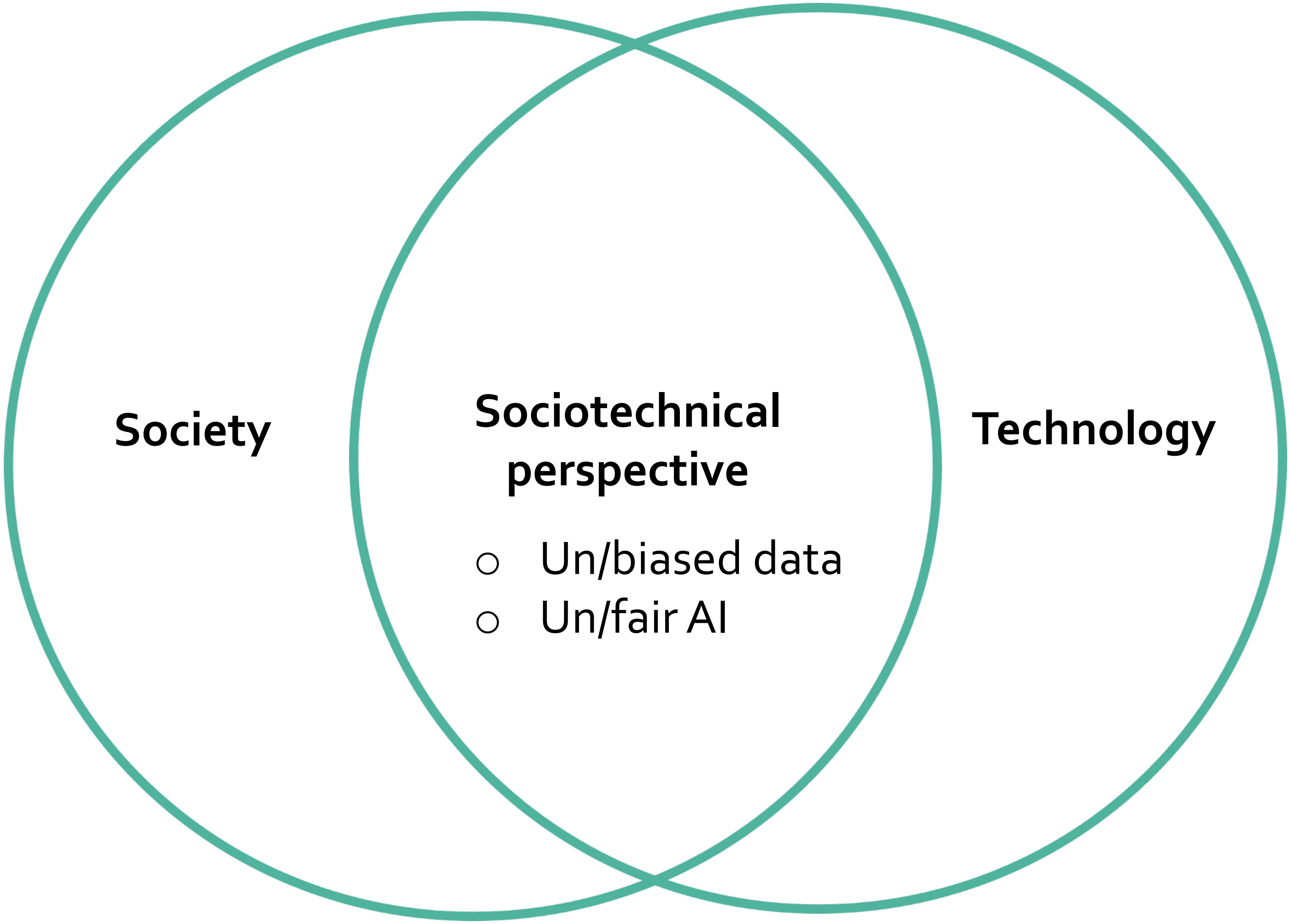
While bias remains a serious problem, it should not be viewed as something that can be fixed after the fact. Instead, we must look to understand and design systems that address the source of the bias: structural oppression. Starting from the assumption that oppression is the problem, not bias, leads to fundamentally different decisions about what to work on, who to work with, and when to stand up and say that a problem cannot and should not be solved by data and technology.

- **Fairness – Equity**

Working toward a world in which everyone is treated equitably, not equally, means taking into account these present power differentials and distributing (or redistributing) resources accordingly. Equity is much harder to model computationally than equality—as it needs to take time, history, and differential power into account—but it is not impossible.

D'Ignazio & Klein, 2020

Sociotechnical perspective



Society

Sociotechnical perspective

- ○ Un/biased data
- ○ Un/fair AI

Technology

7 Principles of Data Feminism

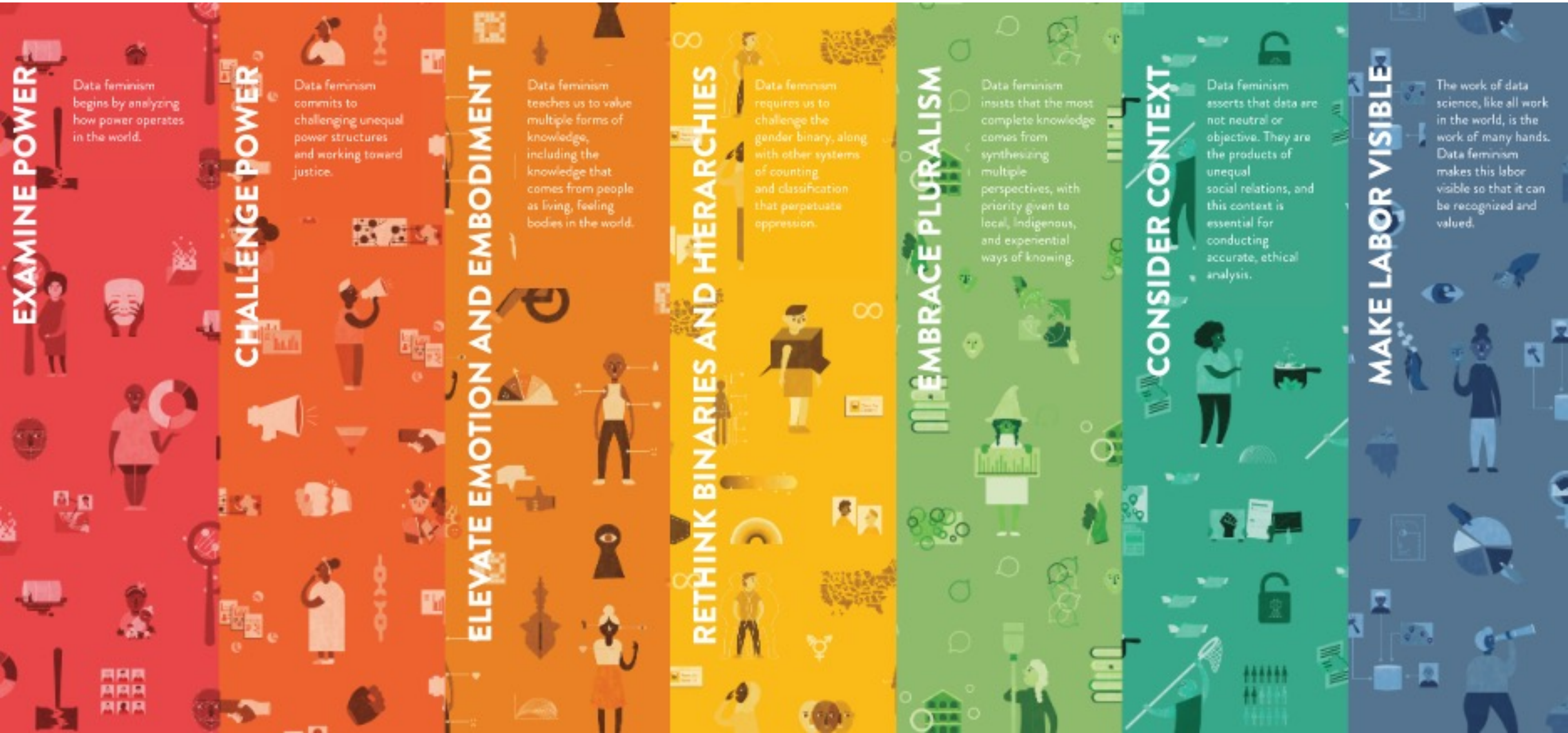
- Examine power: Data feminism begins by analyzing how power operates in the world
- Challenge power: Data feminism commits to challenging unequal power structures and working towards justice
- Rethink binaries and hierarchies: Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world
- Elevate emotion and embodiment: Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression
- Embrace pluralism: Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experimental ways of knowing
- Consider context: Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis
- Make labor visible: The work of data science, like all the work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued

D'Ignazio & Klein, 2020

Thank you!

Good luck with your research projects!

majava@ifi.uio.no



EXAMINE POWER

Data feminism begins by analyzing how power operates in the world.

CHALLENGE POWER

Data feminism commits to challenging unequal power structures and working toward justice.

ELEVATE EMOTION AND EMBODIMENT

Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.

RETHINK BINARIES AND HIERARCHIES

Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.

EMBRACE PLURALISM

Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.

CONSIDER CONTEXT

Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.

MAKE LABOR VISIBLE

The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued.

References and additional reading I

1. Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley. https://www.ruhabenjamin.com/race-after-technology
2. Birhane, A., & Guest, O. (2020). Towards decolonising computational sciences. *ArXiv:2009.14258 [Cs]*. http://arxiv.org/abs/2009.14258
3. Brayne, S. (2020). *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press. https://doi.org/10.1093/oso/9780190684099.001.0001
4. Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press. https://mitpress.mit.edu/books/artificial-unintelligence
5. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html
6. Crawford, K. (2021). *Atlas of AI*. Yale University Press. https://www.katecrawford.net/
7. Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*. https://doi.org/10.21428/96c8d426
8. Crenshaw, K. (1990). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1300.

References and additional reading II

9. D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press Open. https://mitpressonpubpub.mitpress.mit.edu/data-feminism
10. Hagendorff, T. (2021). Blind spots in AI ethics. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00122-8
11. Lopez, P. (2021). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, *10*(4). https://doi.org/10.14763/2021.4.1598
12. Suresh, H. (2021, June 25). The Problem with "Biased Data." *Medium*. https://harinisuresh.medium.com/the-problem-with-biased-data-5700005e514c
13. Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305
14. Timcke, S. (2020). *Algorithms and the Critical Theory of Technology* (SSRN Scholarly Paper ID 3551467). Social Science Research Network. https://doi.org/10.2139/ssrn.3551467
15. Tkacz, N., Henrique da Mata Martins, M., Porto de Albuquerque, J., Horita, F., & Dolif Neto, G. (2021). Data diaries: A situated approach to the study of data. *Big Data & Society*, *8*(1), 205395172199603. https://doi.org/10.1177/2053951721996036
16. Wilson, C., & van der Velden, M. (2022). Sustainable AI: An integrated model to guide public sector decision-making. *Technology in Society*, *68*, 101926. https://doi.org/10.1016/j.techsoc.2022.101926