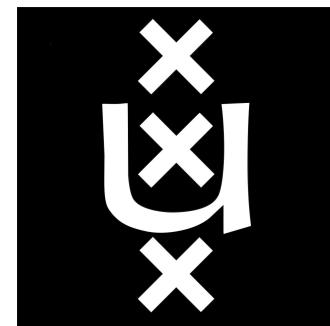


AI for Society

BSc AI 2020/21



Who we are

Tobias Blanke, university professor AI & Humanities



Giovanni Colavizza*, assist. prof. Digital Humanities



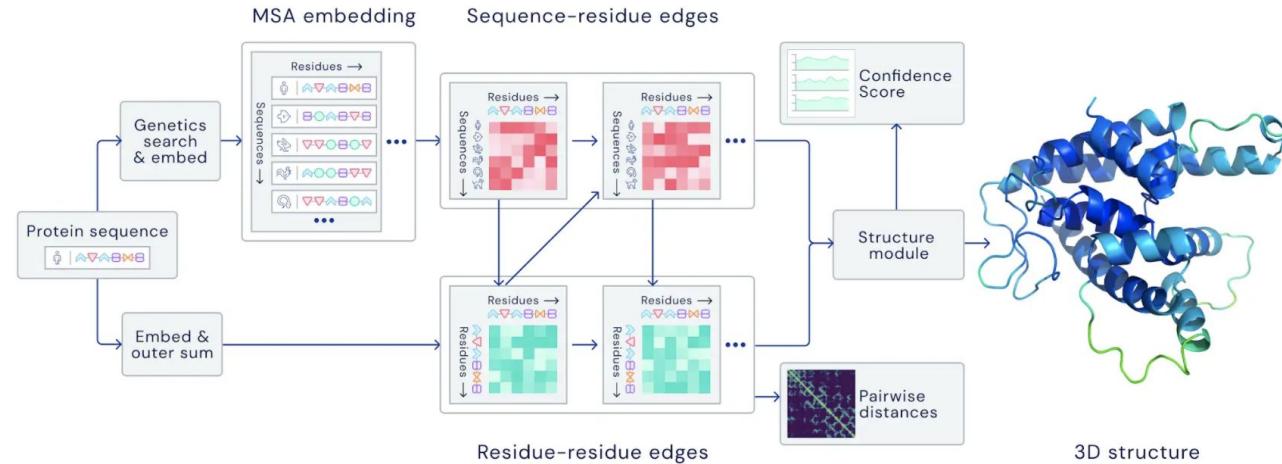
Zarah van Hout, research assistant



* Main point of contact, via Canvas (preferably) or g.colavizza@uva.nl

What is AI for Society about?

AI is increasingly being used to great benefit; notorious examples include **DeepMind's Alpha Fold**, which recently solved protein folding..



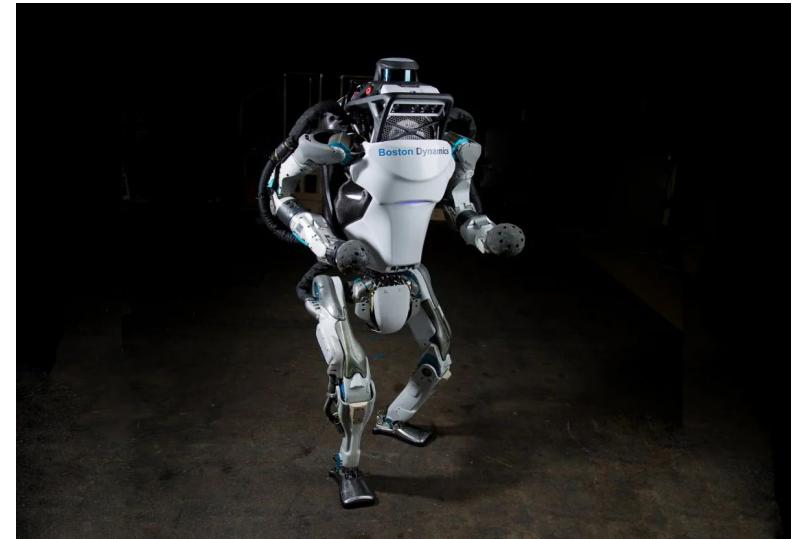
AN OVERVIEW OF THE MAIN NEURAL NETWORK MODEL ARCHITECTURE. THE MODEL OPERATES OVER EVOLUTIONARILY RELATED PROTEIN SEQUENCES AS WELL AS AMINO ACID RESIDUE PAIRS, ITERATIVELY PASSING INFORMATION BETWEEN BOTH REPRESENTATIONS TO GENERATE A STRUCTURE.

What is AI for Society about?

AI is increasingly being used to great benefit;
SpaceX's landing rockets..



Boston Dynamics' robots..



What is AI for Society about?

Many states are seeing AI as of strategic value..

HOLONIQ, GLOBAL INTELLIGENCE

Global AI Strategy Landscape

50 National Artificial Intelligence Policies as at February 2020.

Argentina December 2018, the "National Plan of Artificial Intelligence" falls under the Innovative Argentina 2030 Plan and the 2030 Digital Agenda.	Australia March 2019, "AI Roadmap focused on specialization in health, infrastructure and natural resources. Planning for an additional 10,000 AI specialists by 2030."	Belgium March 2019, "AI & Belgium" launched and includes seven major objectives.
Canada 2017 federal budget announced five-year \$250m plan. Led by CIFAR. Research and talent focus. First National AI Strategy.	Chile December 2018, Ministry of Science, Technology, Knowledge, and Innovation created a committee of 10 experts to develop.	China July 2017, China launched the most comprehensive AI strategy globally with 2030 targets for a \$1T RMB AI industry.
Denmark March 2019, Denmark announced the National Strategy for Artificial Intelligence with four key objectives.	Estonia – Krotts Strategy May 2019, Estonia AI experts, led by government CIO produced a road map, later adopted by the Estonian National AI Strategy in July 2019.	Colombia November 2019, first draft issued for National Policy for Digital Transformation. Medellin to become an AI & Robotics Centre of Excellence.
Hungary October 2019, Hungary announced an AI Action Plan, the first pillar of a national AI strategy, expected in 2020.	Finland June 2019, "Leading the Way into the Age of Artificial Intelligence" identified 11 key areas following May 2017 Steering Group announcements.	France May 2019, €15 billion plan announced in 2018 influenced by the "Vitale Report" to transform France into a global leader in AI.
India March 2018, AGID released a White Paper called "AI at the service of citizens," which was edited by the AI Task Force.	Indonesia June 2018, working paper on using AI to ensure social growth, inclusion and positioning the country as a leader in AI.	Iceland Iceland Artificial Intelligence Society (IASI) inaugurated under Smart Iceland Act in October 2018. National Strategy expected in 2020.
Japan March 2017, Japan's AI policy, the "Artificial Intelligence Technology Strategy," was announced second only to China with "Society 5.0."	Kenya February 2018, task force to create a five-year strategy on national use of emerging technologies.	Lithuania April 2019, government announced "Artificial Intelligence Strategy" to modernise and expand the current AI ecosystem and ensure that the nation is ready."
Malaysia April 2018, Malaysia revealed a National Artificial Intelligence Framework expanding the National Big Data Analytics Framework.	Mexico October 2019, "A Strategy and Vision for Artificial Intelligence in Mexico: Homaging the AI Revolution," serves as a foundation for building full AI strategy.	Luxembourg May 2019, launched "Artificial Intelligence" a strategic vision for Luxembourg."
Norway January 2020, Norway issued its National Strategy for Artificial Intelligence.	Pakistan President's Interim Task Force for Artificial Intelligence launched December 2018, focused on training beginners in AI and advanced technology.	New Zealand November 2018, AI NZ published a roadmap for developing a full national strategy.
Qatar October 2019, National AI Strategy as a blueprint produced by Qatar Computing Research Institute (QCRI).	Philippines November 2019, Ateneo School of Management, Technology and Entrepreneurship (ASME) appointed to craft an AI roadmap.	Poland November 2019, "Assumptions for the AI strategy in Poland" an action plan towards developing an AI strategy.
South Korea May 2018, five-year AI development plan launched with \$1958 budget.	Saudi Arabia September 2019, Royal decree to establish an AI center, to align with the Kingdom's Vision 2030 programs.	Singapore May 2017, AI Singapore is a five-year, \$150 million national program launched to enhance Singapore's capabilities in AI.
Spain March 2019, the Spanish Ministry of Science, Innovation and Universities launched the R&D Strategy in Artificial Intelligence.	Sweden National Approach for Artificial Intelligence launched in May 2018.	Portugal February 2019, AI Portugal 2030, seeks to promote economic growth, scientific excellence, and human development using with AI.
Turkey AI Task Force and Steering Committee to develop a National AI Strategy. The strategy was scheduled to be published in the first quarter of 2019.	United Arab Emirates October 2019, national strategy. First country to create a Ministry of AI and first in the Middle East to launch an AI strategy.	South Africa Intabto Future Production Technologies Initiative launched in 2018 with aim to advancing manufacturing sector.
	United Kingdom April 2018, "Sector Deal" announced £124 billion as part of the UK's larger industrial strategy.	Thailand Thailand's Digital Economy and Society (DESI) Ministry has drafted the country's first artificial intelligence (AI) ethics guidelines.
	United States of America February 2019, Executive Order to promote and protect AI technology. AI.gov launched Mar 2019. Followed by the National Artificial Intelligence Research and Development Strategic Plan.	Vietnam Ministry of Information and Communications developing a broad AI strategy.

Holon IQ

<https://www.holoniq.com/notes/50-national-ai-strategies-the-2020-ai-strategy-landscape>

Source: HoloniQ and source government strategy and policy papers.

www.holoniq.com

What is AI for Society about - AI in 2020?

A screenshot of a web browser displaying a blog post from deeplearning.ai. The URL in the address bar is blog.deeplearning.ai/blog/the-batch-biggest-ai-stories-of-2020-covid-triage-fun-with-gans-disinfo-whack-a-mole.... The page title is "COPING WITH COVID". Below the title, there is a section titled "The Batch" with a sub-section "COPING WITH COVID". A red button labeled "View all Archives" is visible.

A screenshot of the "THE BATCH" newsletter header. It features the "DeepLearning.AI" logo, the title "THE BATCH" in large blue letters, and a small graphic of a Christmas tree. Below the title, it says "December 23, 2020 Essential news for deep learners". There are links for "Subscribe" and "Tips".

Dear friends,

Every year for the past decade, I flew to Singapore or Hong Kong to celebrate my mother's birthday with her on December 22. This year, for the first time, we did it via Zoom. Despite the distance, I was warmed that my family could gather from the U.S., Singapore, Honk Kong, and New Zealand and sing a poorly synchronized "Happy Birthday To You."



Coping with Covid

This Snowman Does Not Exist

Representing the Underrepresented

Algorithms Against Disinformation

What is AI for Society about?

We face major transparency issues in advertising

The screenshot shows a GitHub repository page for 'awful-ai'. The top navigation bar includes icons for back, forward, search, and refresh, followed by the URL 'github.com/daviddao/awful-ai'. To the right are star, fork, clone, and other repository management buttons, along with an 'Update' button.

The repository header shows 'master' (selected), 2 branches, 0 tags, a 'Go to file' button, and a 'Code' dropdown. The main content area displays three commits:

- cd6c483 4 days ago (47 commits) - Adding AI-based machine gun incident to list by [daviddao](#)
- 14 months ago - Create FUNDING.yml by [.github](#)
- 4 days ago - Adding AI-based machine gun incident to list by [README.md](#)

A large callout box highlights the first commit: 'Adding AI-based machine gun incident to list' by [daviddao](#). Below this, the README.md file is shown with its content:

Awful AI

Awful AI is a curated list to track *current* scary usages of AI - hoping to raise awareness to its misuses in society

Artificial intelligence in its current state is [unfair](#), [easily susceptible to attacks](#) and [notoriously difficult to control](#). Often, AI systems and predictions [amplify existing systematic biases](#) even when the data is balanced. Nevertheless, more and more concerning the uses of AI technology are appearing in the wild. This list aims to track *all of them*. We hope that *Awful AI* can be a platform to spur discussion for the development of possible preventive technology (to fight back!).

The right sidebar contains sections for 'About', 'Releases', and 'Sponsor this project'.

About: Awful AI is a curated list to track current scary usages of AI - hoping to raise awareness. [twitter.com/dwddao](#)

Releases: No releases published

Sponsor this project: [daviddao ...](#) [Sponsor](#)

[Learn more about GitHub Sponsors](#)

Algorithms are deciding who gets the first vaccines. Should we trust them?

Clashes over who gets vaccinated first could ramp up as supplies roll out nationwide. Transparency over how these mathematical formulas work is critical, their designers say.

← → C science.sciencemag.org/content/366/6464/447



SHARE



RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer^{1,2,*}, Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan^{5,*†}

* See all authors and affiliations

Science, 25 Oct 2019;
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Article

Figures & Data

Info & Metrics

eLetters

PDF

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions.

Obermeyer et al. find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.



Science

Vol 366, Issue 6464
25 October 2019

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

PDF

Help

ARTICLE TOOLS

- Email
- Print
- Alerts
- Share
- Download Powerpoint
- Request Permissions
- Citation tools

MY SAVED FOLDERS

- Save to my folders

STAY CONNECTED TO SCIENCE

- Facebook
- Twitter

Advertisement



'(...) the algorithm uses health costs as a proxy for health needs.'

What could be the problem with this assumption?

Why would algorithm designers do this?

What alternative proxies might there be?

Type into the chatbox ...

Weekly topics

1. Introduction + Fairness and Bias (**TODAY**)
2. Explainable and Interpretable AI
3. AI to Study Society and Culture
4. The Economics of AI

Course logistics

Each week we meet for a **lecture (Mondays, 11:00-13:00)** and a **laboratory (Wednesdays, 11:00-15:00)**. Yes, we will do a lunch break for labs.

Lectures are lectures on the topic at hand. We will pre-record each one and leave time for Q&A. At the end of each lecture, we will hand-out a **group assignment**.

Laboratories give you time to work on your group assignments, plus there will be some reporting back and discussion.

Assessment

60% for 4 group assignments (weekly deadlines)

40% for a final individual essay (take-home, after the course ends)

For more details, see Canvas

How to communicate and code of conduct

Please use Canvas to email Giovanni for any question on the course. Try to use the Discussion section first, to get support among your peers.

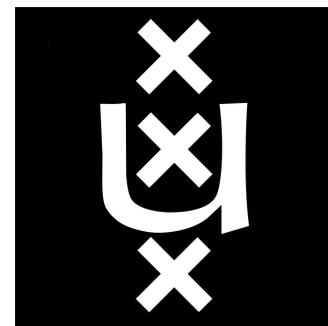
We will mainly use Announcements and class time to communicate with you.

Please take close note of the applicable **code of conduct** regarding group and individual work, including plagiarism. We will use anti-plagiarism software to check your submissions. We trust you will do your own work.

Q&A

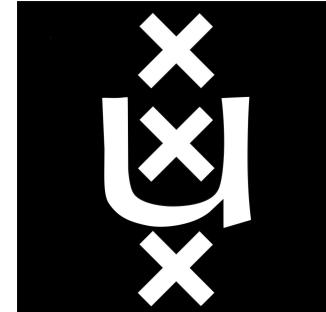
AI for Society

BSc AI 2020/21



Week 1: Fairness and Bias

Giovanni Colavizza

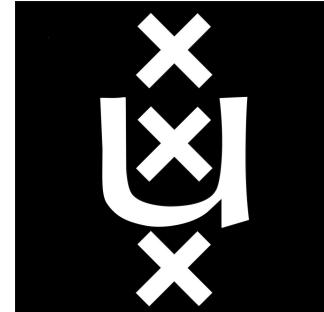


Topics

1. Fairness
2. Bias
3. Countermeasures
4. Assignment

PART 1: Fairness

Giovanni Colavizza



Facial recognition

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%



Recidivism detection



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

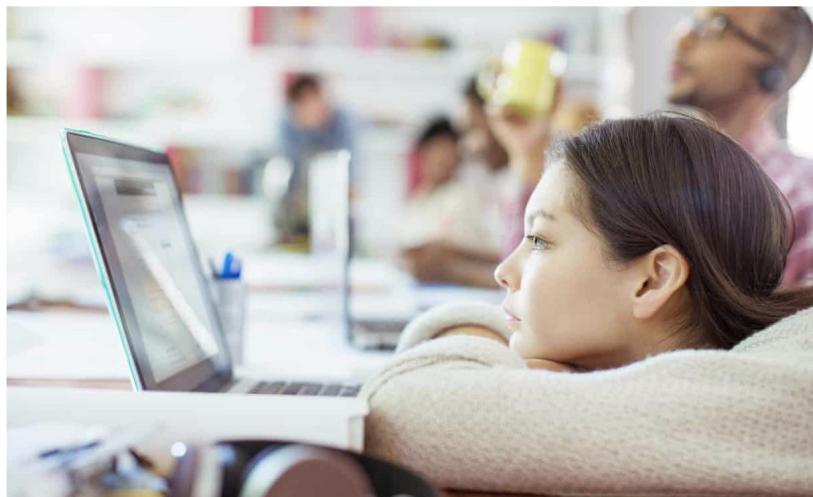
May 23, 2016

Targeted ads

• This article is more than 5 years old

Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



▲ One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

Questions

- If a classifier works with 100% accuracy, who is affected by it?
- What if it works 90% uniformly?
- What if it works 90% non-uniformly, e.g., 95% for one non-protected group and 60% for a protected group?
- Ethical questions:
 - What should be done?
 - Which requirements should this classifier abide to?
 - Should it be forbidden or regulated? If so, how?
 - Who is responsible?

By the way.. Advertising

Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

40 Pages • Posted: 15 Oct 2016 • Last revised: 12 Mar 2018

Anja Lambrecht

London Business School

Catherine E. Tucker

Massachusetts Institute of Technology (MIT) - Management Science (MS)

Date Written: March 9, 2018

Abstract

We explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields. This ad was explicitly intended to be gender-neutral in its delivery. Empirically, however, fewer women saw the ad than men. This happened because younger women are a prized demographic and are more expensive to show ads to. An algorithm which simply optimizes cost-effectiveness in ad delivery will deliver ads that were intended to be gender-neutral in an apparently discriminatory way, due to crowding out. We show that this empirical regularity extends to other major digital platforms.

What do we mean by ‘fairness’?

Fairness is an elusive concept.

Oftentimes, it refers to lack of bias in decisions. In a related way, fairness might entail the balanced treatment of sub-populations and individuals.

Fairness is elusive also because some people mainly care about **equality** of opportunity, while others emphasize **equity** in the outcomes.

> Learn 21 definitions of ‘fairness’ by Arvind Narayanan:

<https://www.youtube.com/watch?v=jIXluYdnyyk>.

What do we mean by ‘ethics’?

“Ethics is based on well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, **fairness**, or specific virtues.”*

Any ethical framework is contingent to time, place, culture. Some components are so widely shared that might have deep roots into our evolutionary history too (e.g., reciprocity).

The first question to ask when you hear ‘Ethical AI’ is: **whose ethics?** Often, that goes unstated. Please always ask yourself the question during this course (and afterwards).

* Velasquez et al. 2010. *What is Ethics?* <https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics>.

What do we mean by ‘fairness’?

Many **formal definitions of fairness** have been proposed.

Their goal is to formalize what it means for an AI (e.g., a classifier) to be ‘fair’, according to some notion of fairness. It is then up to someone (Governments? Individuals? Companies?) to decide which notion of fairness is appropriate.

Note that definitions of fairness are often (mathematically and morally) mutually exclusive.

We mainly use (you will find more in there):

- Suresh and Guttag. 2020. *A Framework* <https://arxiv.org/abs/1901.10002>
- Mitchell et al. 2021. *Algorithmic Fairness* <https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-042720-125902>
- Mehrabi et al. 2019. *A Survey* <https://arxiv.org/abs/1908.09635>

Families of definitions of fairness

- **Group-Independent Predictions** require that the decisions that are made are independent (or conditionally independent) of group membership. For example, the *demographic parity criterion* states that the proportion of each segment of a protected class (e.g., gender) should receive the positive outcome at equal rates.
- **Equal Metrics Across Groups** require equal prediction metrics of some sort (this could be accuracy, true positive rates, false positive rates, and so on) across groups. For example, the *equality of opportunity criterion* requires equal true positive/negative rates across groups.
- **Individual Fairness** requires that individuals who are similar with respect to the prediction task are treated similarly. There is an assumption that an ideal feature space exists in which to compute similarity, and that those features are recoverable in the available data. For example, *fairness through (un)awareness* tries to identify a task-specific similarity metric in which individuals who are close according to this metric are also close in outcome space.
- **Causal Fairness** definitions place some requirement on the causal graph that generated the data and outcome. For example, *counterfactual fairness* requires that there is not a causal pathway from a sensitive attribute to the outcome decision

How to formalize a definition

Notation (assume, w/o loss of generality, a binary classifier 1/0):

- D is your decision {0,1}
- Y is the ground truth {0,1}
- X are non-protected covariates (e.g., income and age)
- A are protected covariates (e.g., race or sex)

How to formalize a definition

- *Demographic parity* (also called statistical parity or group fairness) entails that outcomes need to be balanced across groups, irrespective of ground truth and covariates.
 - $P(D = 1|A = 1) = P(D = 1|A = 0)$ and possibly the same for $D = 0$.

How to formalize a definition

- *Demographic parity* (also called statistical parity or group fairness) entails that outcomes need to be balanced across groups, irrespective of ground truth and covariates.
 - $P(D = 1|A = 1) = P(D = 1|A = 0)$ and possibly the same for $D = 0$.
- *Equality of opportunity* entails that the TP and/or TN rates are balanced.
 - TP: $P(D = 1|Y = 1, A = 1) = P(D = 1|Y = 1, A = 0)$
 - TN: $P(D = 0|Y = 0, A = 1) = P(D = 0|Y = 0, A = 0)$

How to formalize a definition

- *Demographic parity* (also called statistical parity or group fairness) entails that outcomes need to be balanced across groups, irrespective of ground truth and covariates.
 - $P(D = 1|A = 1) = P(D = 1|A = 0)$ and possibly the same for $D = 0$.
- *Equality of opportunity* entails that the TP and/or TN rates are balanced.
 - TP: $P(D = 1|Y = 1, A = 1) = P(D = 1|Y = 1, A = 0)$
 - TN: $P(D = 0|Y = 0, A = 1) = P(D = 0|Y = 0, A = 0)$
- *Fairness through (un)awareness* entails that similar individuals will get similar results.
 - Unawareness: $P(D = 1|X_i) = P(D = 1|X_j)$ where X_i is very similar to X_j
 - Awareness: $P(D = 1|X_i, A = 1) = P(D = 1|X_j, A = 0)$ where X_i is very similar to X_j

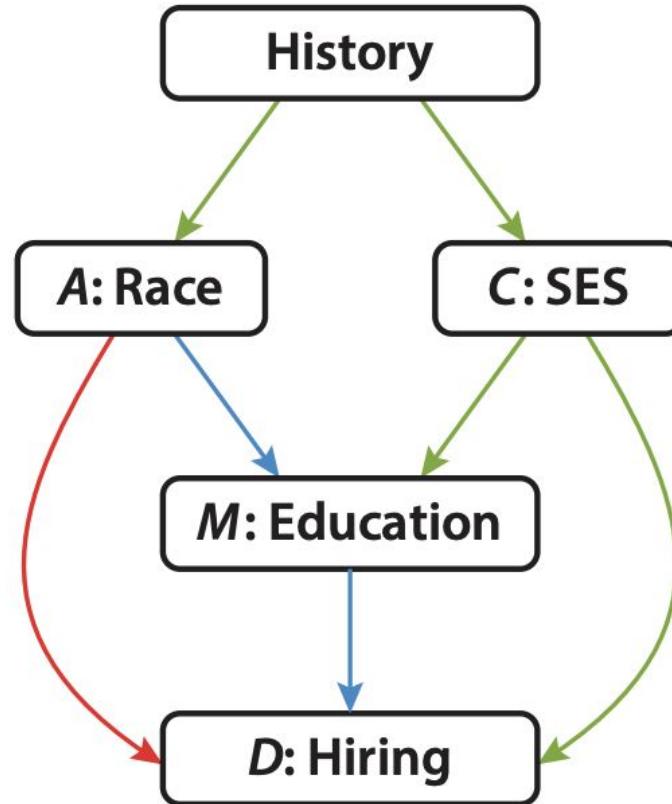
How to formalize a definition

	$Y = 1$	$Y = 0$	$P(Y=1 D)$	$P(Y=0 D)$
$D = 1$	True positive	False positive	$P(Y=1 D=1)$: Positive predictive value	$P(Y=0 D=0)$: False discovery rate
$D = 0$	False negative	True negative	$P(Y=1 D=0)$: False omission rate	$P(Y=0 D=0)$: Negative predictive value
$P(D=1 Y)$	$P(D=1 Y=1)$: True positive rate	$P(D=1 Y=0)$: False positive rate		
$P(D=0 Y)$	$P(D=0 Y=1)$: False negative rate	$P(D=0 Y=0)$: True negative rate		$P(D=Y)$: Accuracy

Counterfactual fairness

Red and (usually) blue would be forbidden; green would be (usually) allowed, yet this is also questionable.

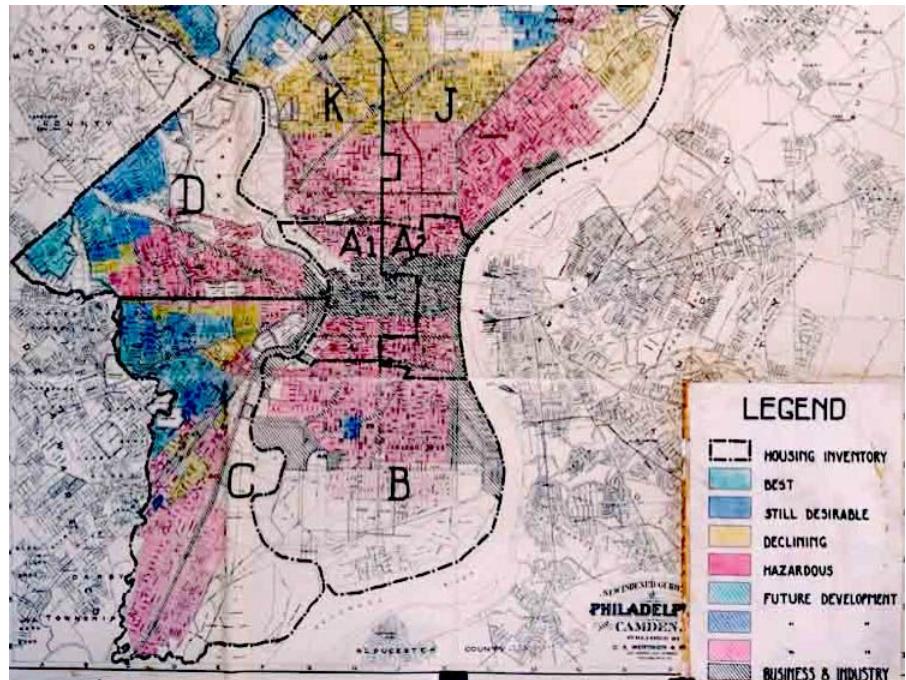
Green (back-door) paths bring back the discussion on “correlation vs causation” and the burden of proof.



An example: Redlining

“Redlining is the systematic denial of various services or goods by US federal government agencies, local governments, or the private sector either directly or through the selective raising of prices.”

<https://en.wikipedia.org/wiki/Redlining>



An example: Redlining

Notation:

- D is your decision {0,1}, e.g., to give credit or not.
- Y is the ground truth {0,1}
- X are non-protected covariates (e.g., zip code)
- A are protected covariates (e.g., race)

An example: Redlining

- *Demographic parity.*
 - $P(D = 1|A = 1) = P(D = 1|A = 0)$ and possibly the same for $D = 0$.
 - Would be fair (in the equity sense) to groups, but unfair (in the equality sense) to individuals. It would also be economically inefficient. Can you think why?

An example: Redlining

- *Demographic parity.*
 - $P(D = 1|A = 1) = P(D = 1|A = 0)$ and possibly the same for $D = 0$.
 - Would be fair (in the equity sense) to groups, but unfair (in the equality sense) to individuals. It would also be economically inefficient. Can you think why?
- *Equality of opportunity.*
 - TP: $P(D = 1|Y = 1, A = 1) = P(D = 1|Y = 1, A = 0)$
 - TN: $P(D = 0|Y = 0, A = 1) = P(D = 0|Y = 0, A = 0)$
 - Would be fair (in the equality sense) but *to work it would require to drop non-protected covariates which are highly correlated with protected ones*. Can you think why?

An example: Redlining

- *Demographic parity.*
 - $P(D = 1|A = 1) = P(D = 1|A = 0)$ and possibly the same for $D = 0$.
 - Would be fair (in the equity sense) to groups, but unfair (in the equality sense) to individuals. It would also be economically inefficient. Can you think why?
- *Equality of opportunity.*
 - TP: $P(D = 1|Y = 1, A = 1) = P(D = 1|Y = 1, A = 0)$
 - TN: $P(D = 0|Y = 0, A = 1) = P(D = 0|Y = 0, A = 0)$
 - Would be fair (in the equality sense) but *to work it would require to drop non-protected covariates which are highly correlated with protected ones*. Can you think why?
- *Fairness through (un)awareness* entails that similar individuals will get similar results.
 - Unawareness: $P(D = 1|X_i) = P(D = 1|X_j)$ where X_i is very similar to X_j
 - It would be unfair in both equity and equality senses. Can you think why?
 - *This is the argument that was used to justify redlining.*
 - Awareness: $P(D = 1|X_i, A = 1) = P(D = 1|X_j, A = 0)$ where X_i is very similar to X_j
 - Likely similar to equality of opportunity. Can you think why?
- What about *counterfactual fairness*? Would zip-codes be allowed there?

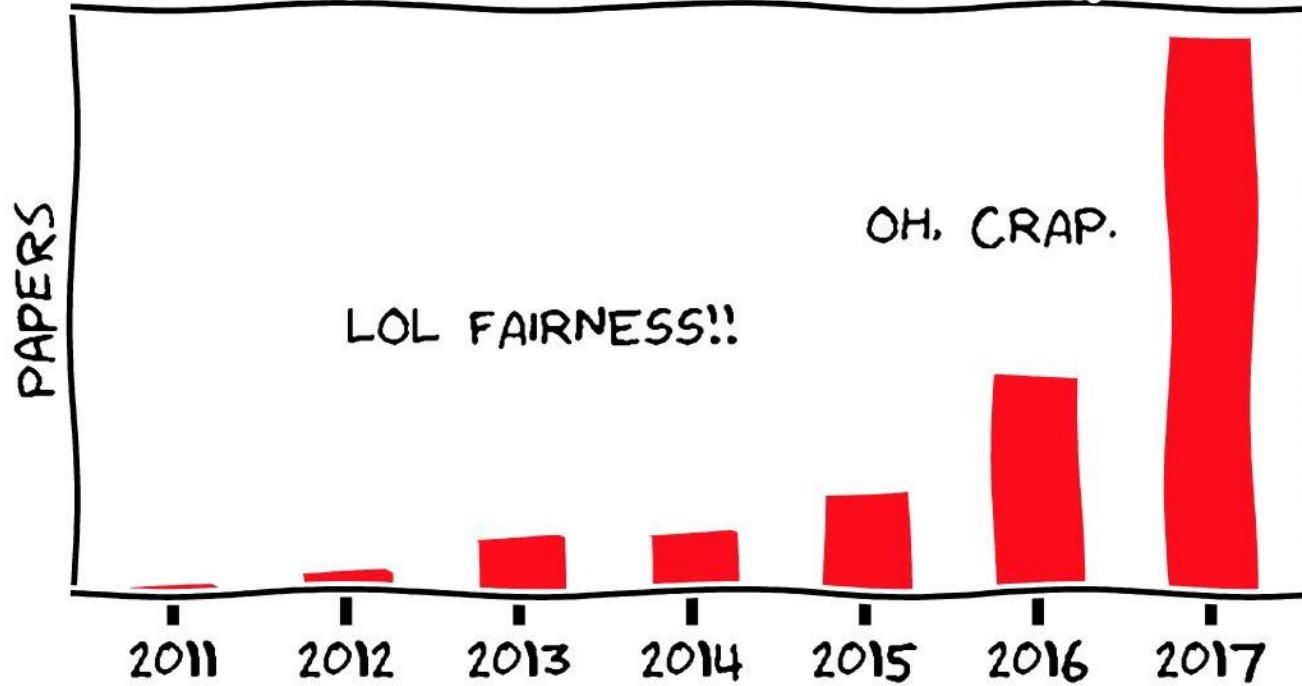
Why care about 'fairness'?

Disparate impact: errors do not affect everyone in the same way. For example, being flagged at high-risk of recidivism by error (FP) usually has worse consequences than being flagged at low-risk by error (FN).

Allocative vs representational harm: allocative harm refers to unfair allocation of resources (e.g., hiring or mortgage decisions), representational refers to the unfair depiction of individuals or groups (e.g., stereotyping).

> Watch Kate Crawford's lecture 'The trouble with bias': https://www.youtube.com/watch?v=fMym_BKWQzk.

BRIEF HISTORY OF FAIRNESS IN ML



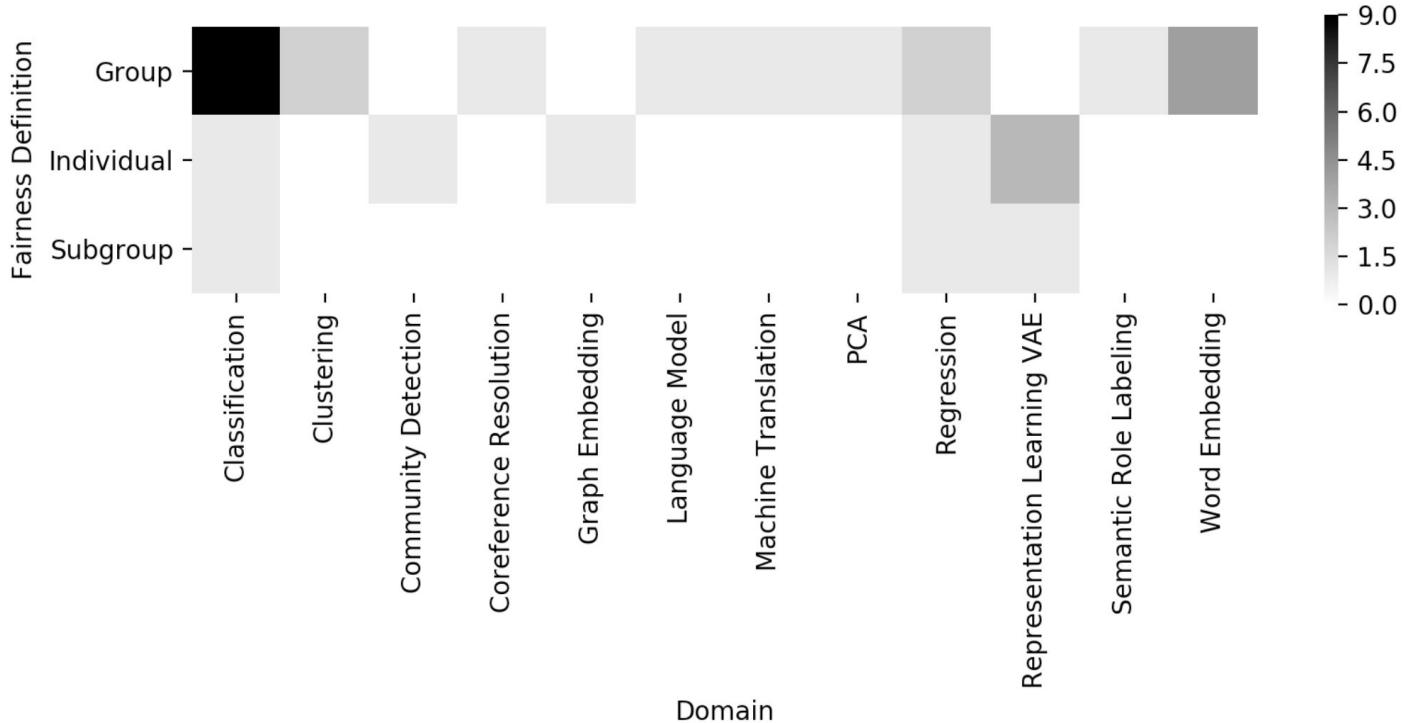
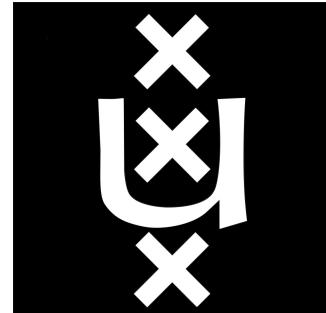


Fig. 7. Heatmap depicting distribution of previous work in fairness, grouped by domain and fairness definition.

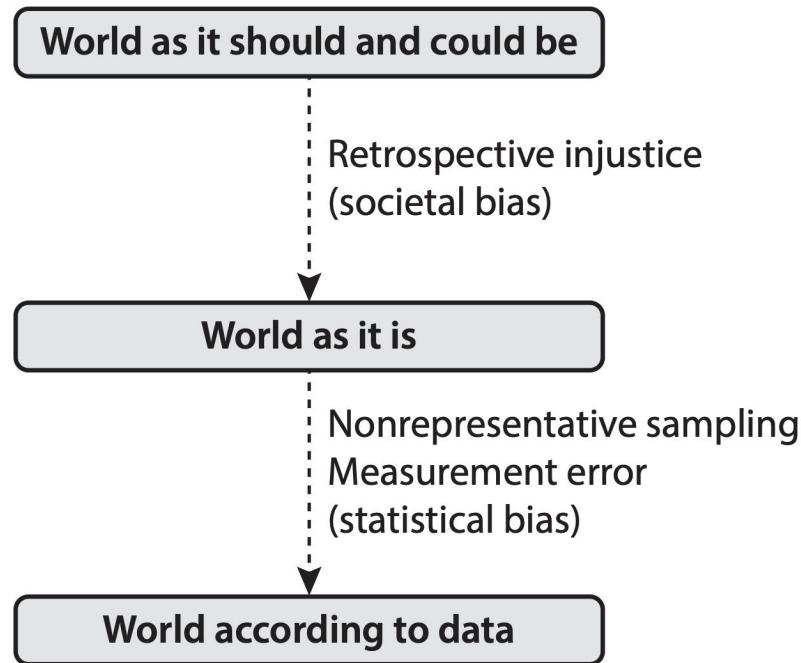
Q&A

PART 2: Bias

Giovanni Colavizza



What do we mean by ‘bias’?



Statistical bias

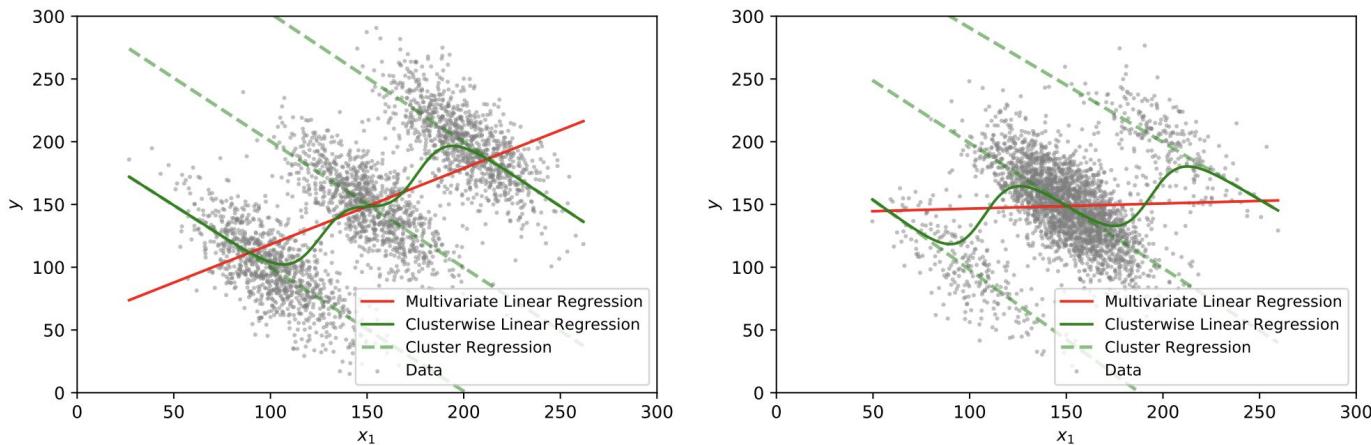
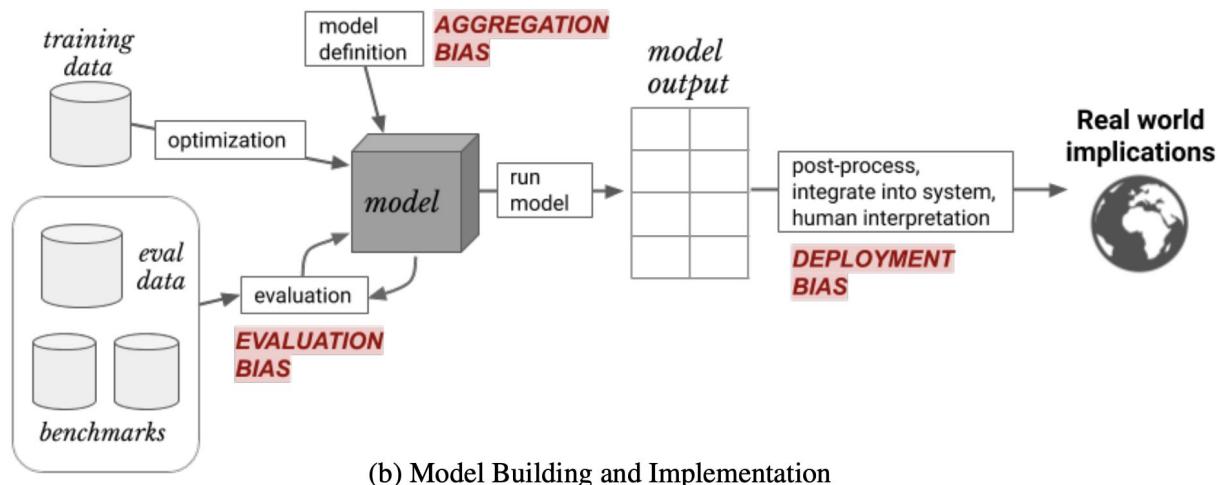
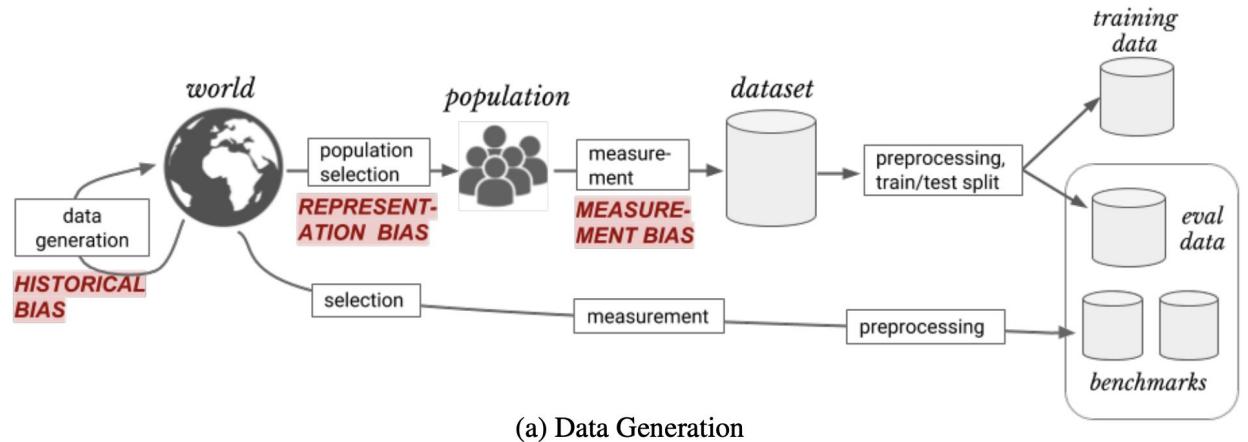


Fig. 1. Illustration of biases in data. Red line shows the regression (MLR) for the entire population, while dashed green lines are regressions for each subgroup, and the solid green line is the unbiased regression. (a) When all subgroups are of equal size, then MLR shows a positive relationship between the outcome and the independent variable. (b) Regression shows almost no relationship in less balanced data. The relationships between variables within each subgroup, however, remain the same. (Credit: Nazanin Alipourfard)

Bias in AI



Examples of bias in AI

- Historical bias
- Representation bias
- Measurement bias
- Evaluation bias
- Aggregation bias
- Deployment 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%

Probably yes
Maybe
Probably not



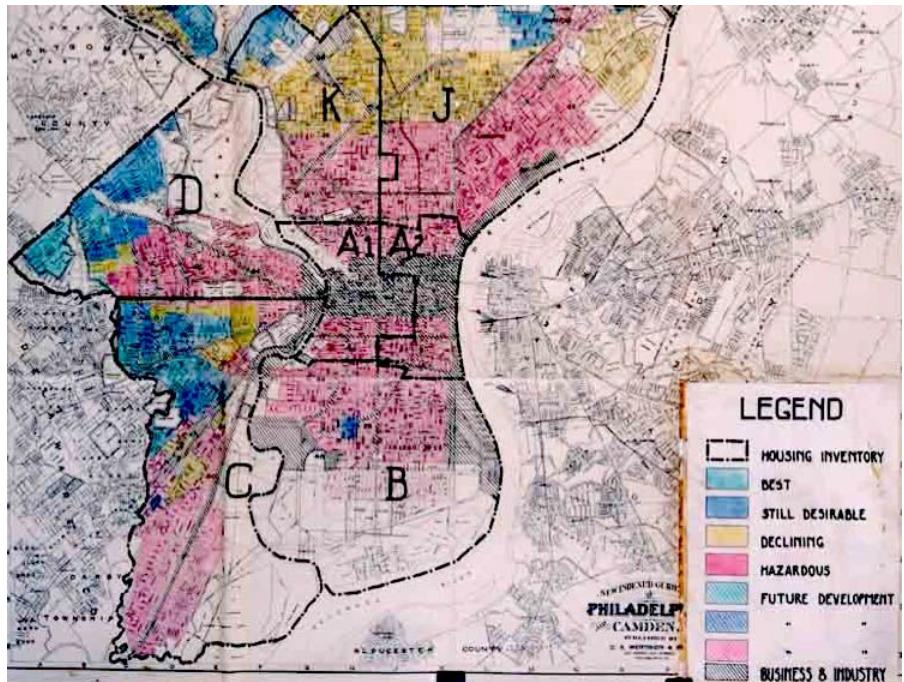
Examples of bias in AI

- Historical bias
- Representation bias
- Measurement bias
- Evaluation bias
- Aggregation bias
- Deployment bias

Probably yes

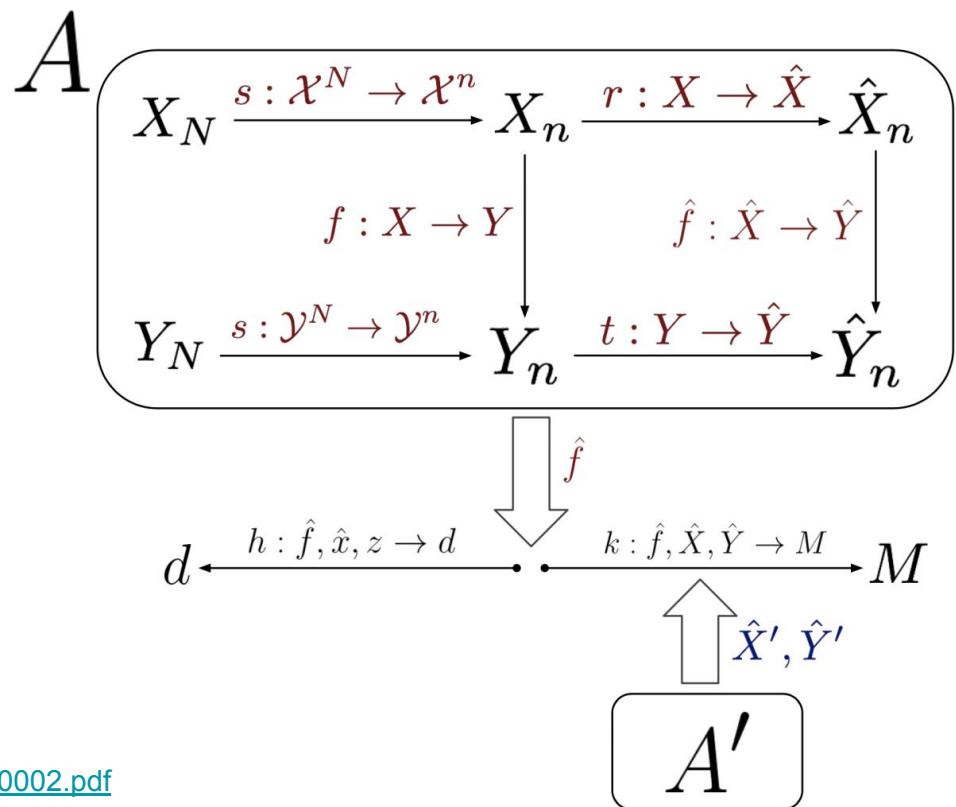
Maybe

Probably not



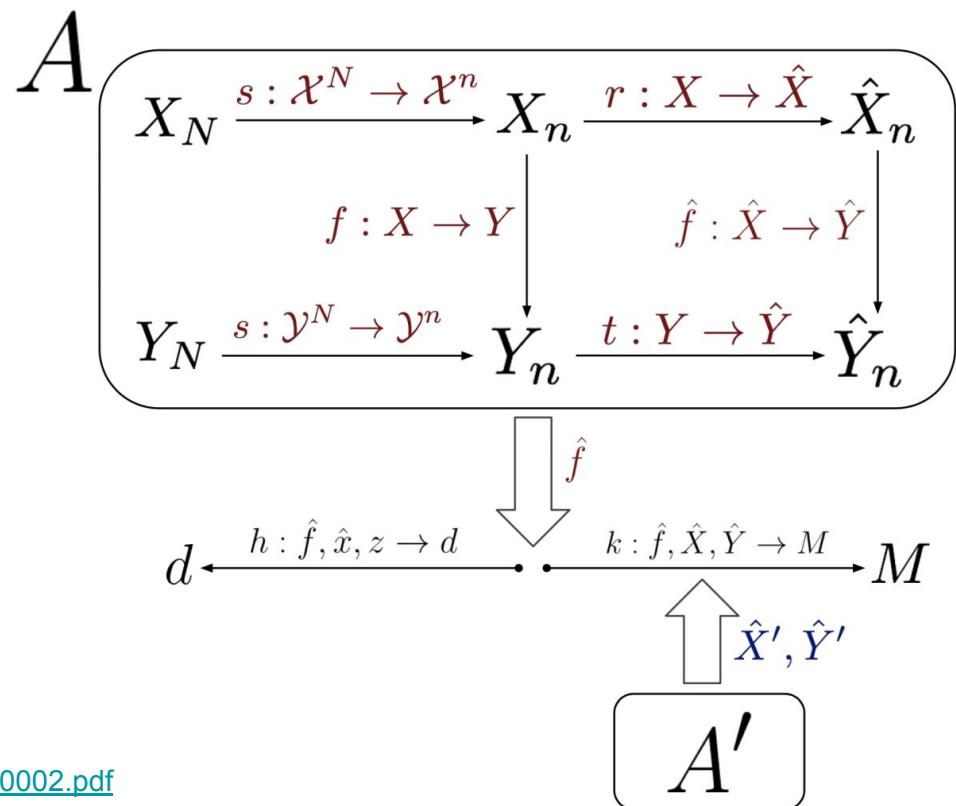
How to reason about AI bias formally?

Using **data transformations**
is one possible way



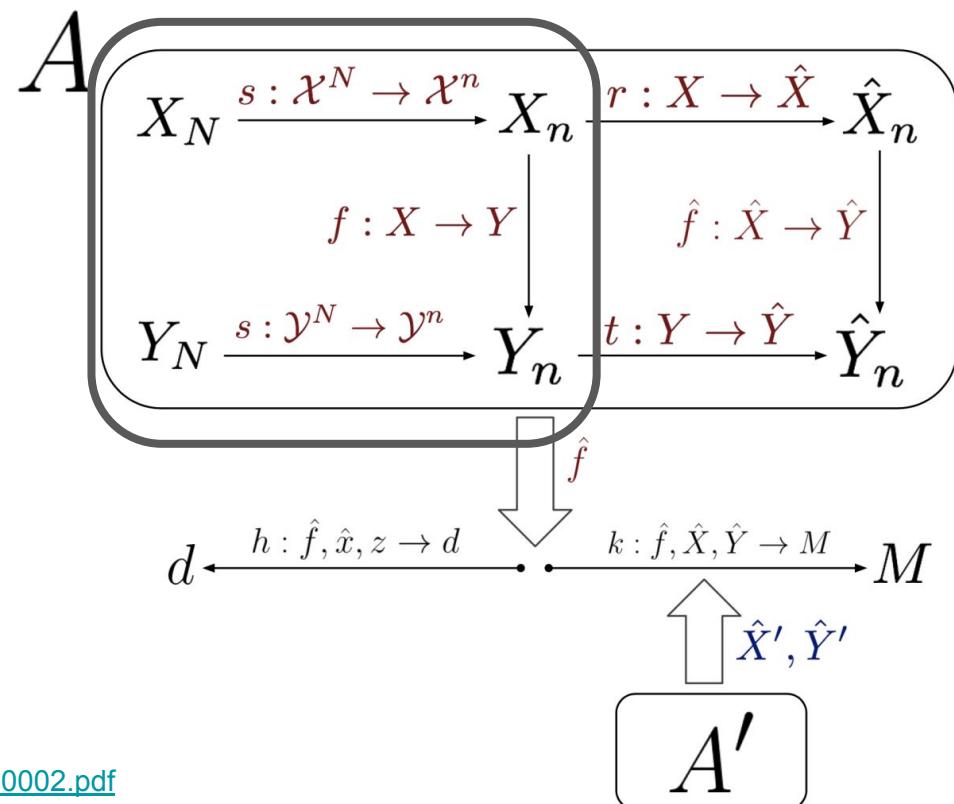
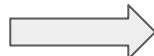
How to reason about AI bias formally?

\mathbf{X} are real features, \mathbf{Y} are outcomes. For example, \mathbf{X} might be the skills of candidates, \mathbf{Y} might be the hiring decisions.



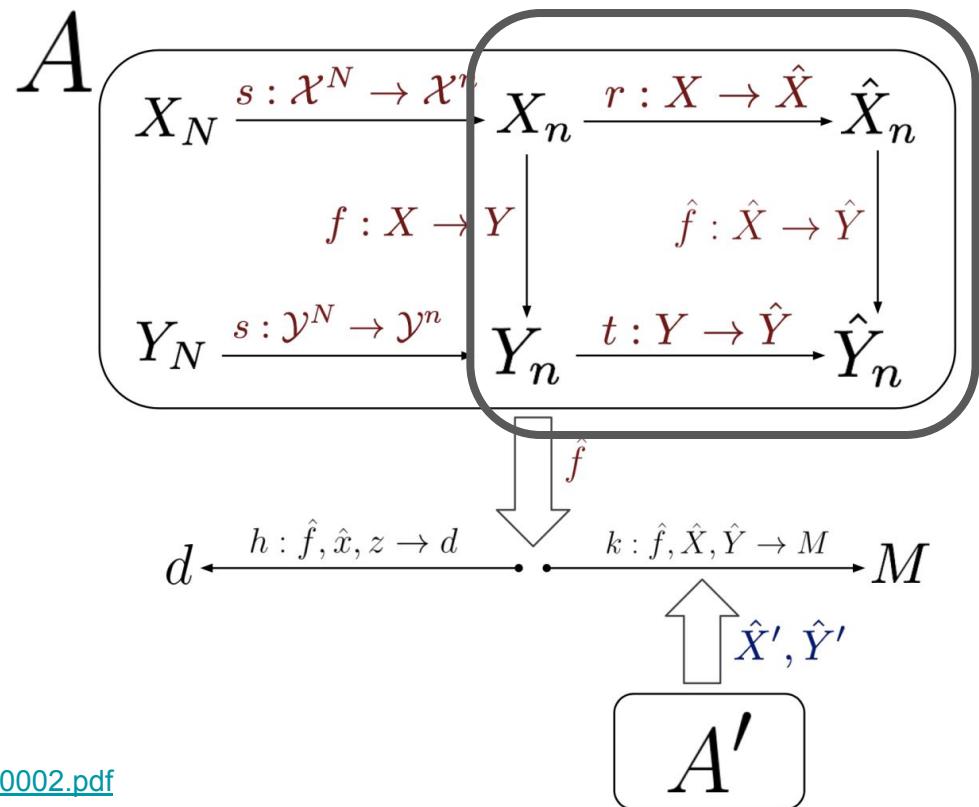
How to reason about AI bias formally?

First, we map from the full population N to a sample n via the function s . This might introduce **representation (sampling) bias**.



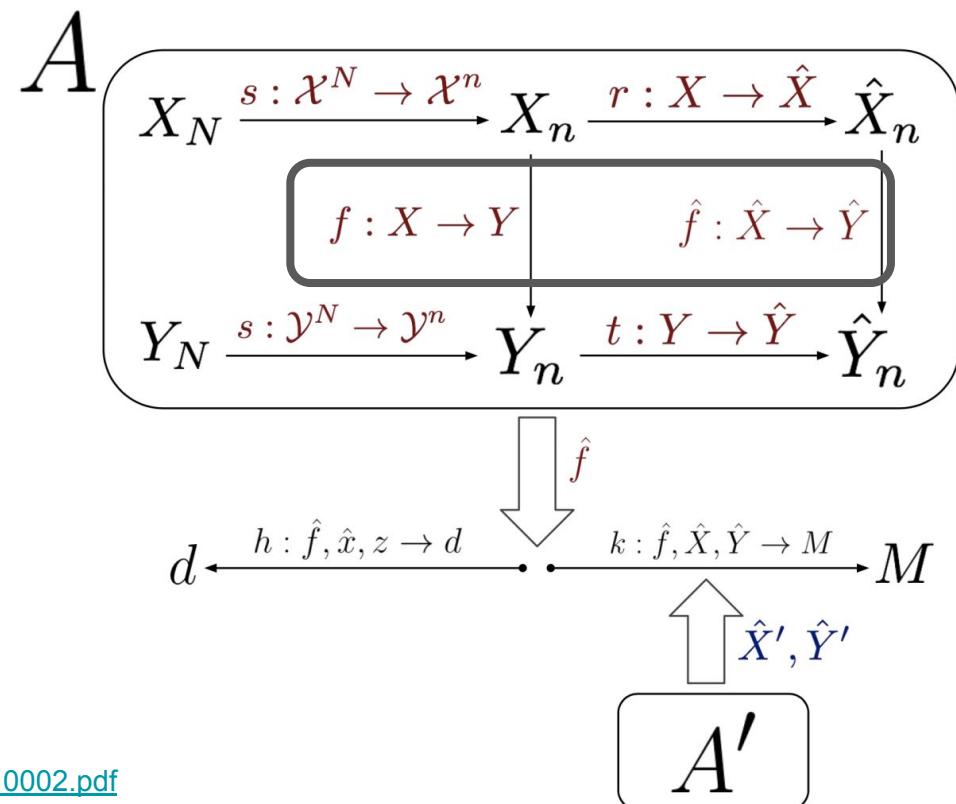
How to reason about AI bias formally?

Then, we measure real-world features (e.g., skill) into machine-readable features (e.g., years of experience), via functions r and t . This might introduce **measurement (and possibly historical) bias**.

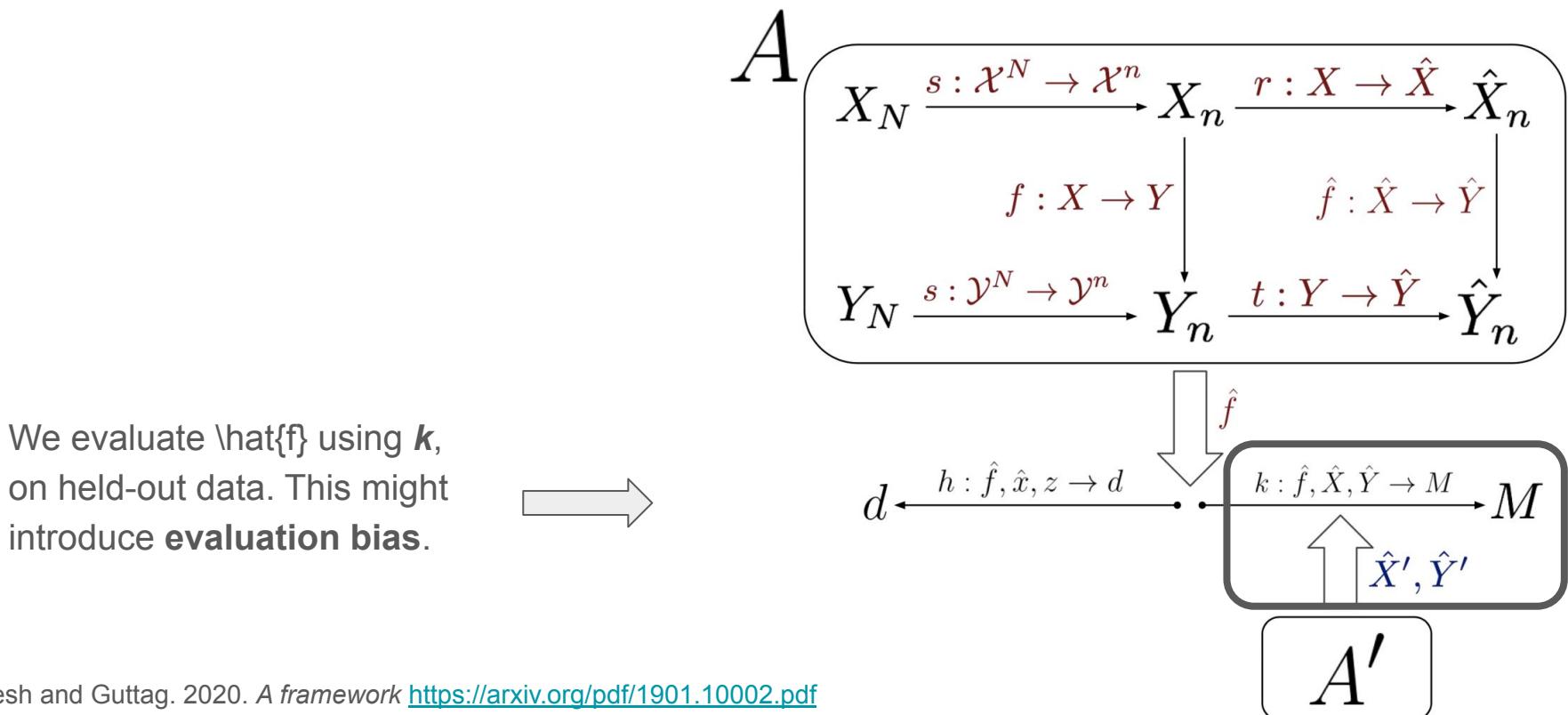


How to reason about AI bias formally?

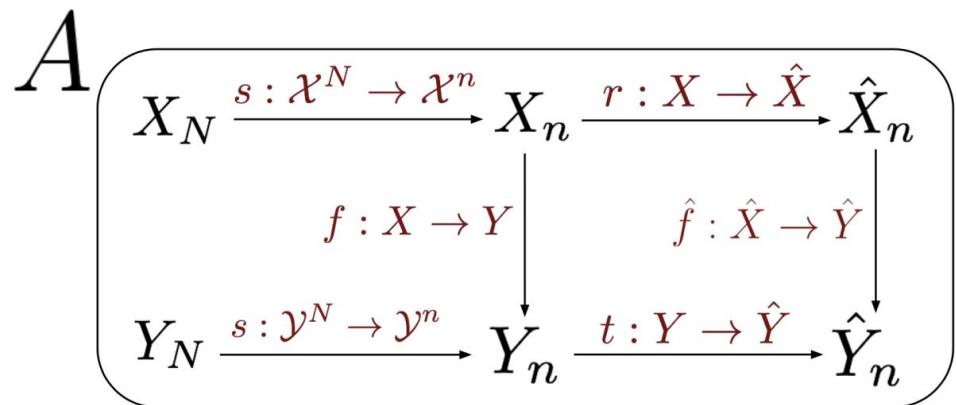
What we want to learn is f , a function from real-world features to outcomes. What we can learn is \hat{f} .



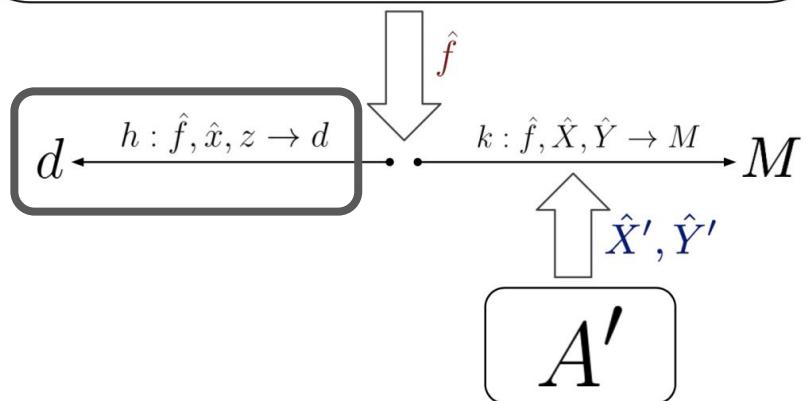
How to reason about AI bias formally?



How to reason about AI bias formally?



Finally, we use \hat{f} to make a decision d on new data points. This might introduce **aggregation and deployment bias**.



How are fairness and bias related?

As we mention at the beginning, fairness could be defined as lack of bias. Which forms of bias to focus on, and thus what form of fairness to implement, is sometimes a technical choice (e.g., statistical bias), sometimes an ethical choice.

Beware! Cognitive bias

What Should We
Remember?

To avoid mistakes,
we aim to preserve autonomy
and group status, and avoid
irreversible decisions

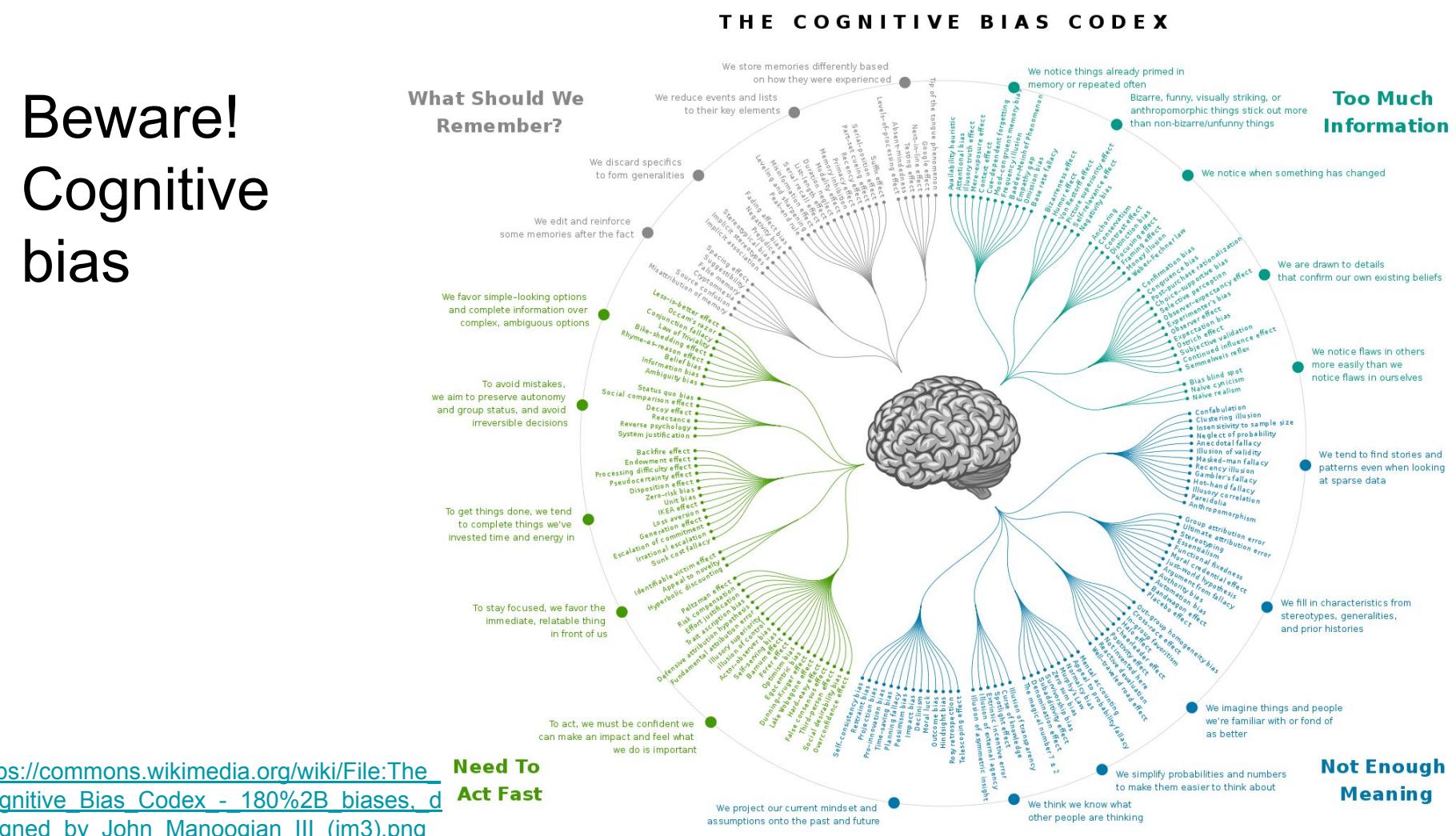
To get things done, we tend
to complete things we've
invested time and energy in

To stay focused, we favor the
immediate, relatable thing
in front of us

To act, we must be confident we
can make an impact and feel what
we do is important

Need To
Act Fast

THE COGNITIVE BIAS CODEX

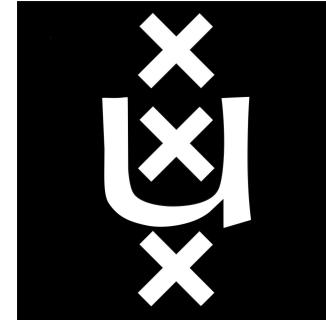


[https://commons.wikimedia.org/wiki/File:The_Cognitive_Bias_Codex_-_180%2B_biases_designed_by_John_Manooian_III_\(jm3\).png](https://commons.wikimedia.org/wiki/File:The_Cognitive_Bias_Codex_-_180%2B_biases_designed_by_John_Manooian_III_(jm3).png)

Q&A

PART 3: Countermeasures

Giovanni Colavizza



How to ‘deal with’ bias?

Start by **defining what you mean by bias**, this should be tightly related with your **problem formulation** (what is it that you are trying to solve or understand?). If your models or insights will become operational (go ‘in production’), also consider which **definition of fairness** you want them to live up to, if any.

How to ‘deal with’ bias?

Start by **defining what you mean by bias**, this should be tightly related with your **problem formulation** (what is it that you are trying to solve or understand?). If your models or insights will become operational (go ‘in production’), also consider which **definition of fairness** you want them to live up to, if any.

If you have **data**, explore/profile your data *before* doing any modelling. Look for skewness, missing values, outliers, unbalance across protected groups, etc. Perform data selection or correction accordingly. If you don’t have data, design a **data acquisition** process which avoids possible bias (as you previously defined).

Data

Collecting and using data is a big deal, and goes a long way in explaining biases.

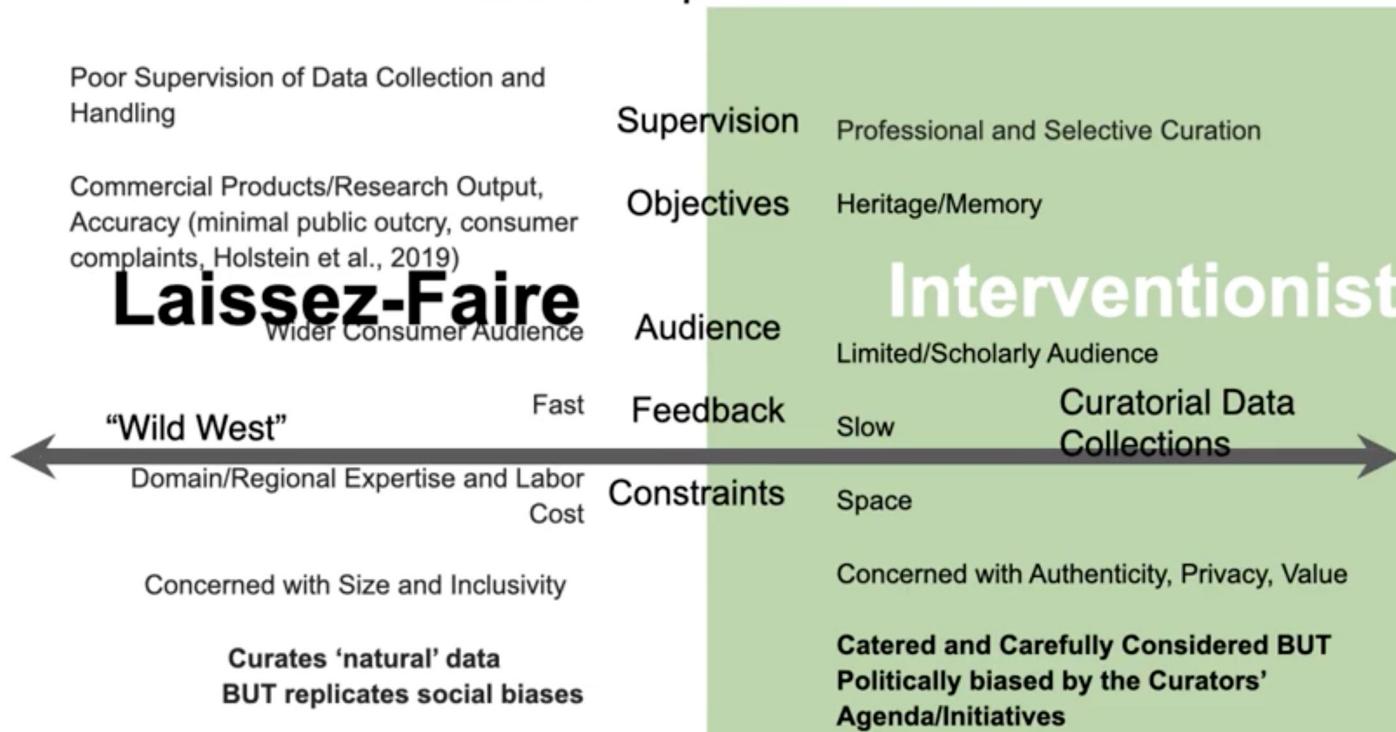
Issues to consider when thinking data include: **consent, power, inclusion, representation (e.g., of protected groups), privacy, transparency, provenance, quality**, and more.

Useful references are:

- Jo and Gebru. 2020. *Lessons from archives* <https://dl.acm.org/doi/abs/10.1145/3351095.3372829>
 - Introduces useful concepts and ideas from archives are applied to AI data. An example is ‘**appraisal**’, or the task of selecting data according to a predefined set of objectives and criteria.
- Gebru et al. 2018. *Datasheets for datasets* <https://arxiv.org/abs/1803.09010>
 - Proposes a simple and effective way to **attach useful documentation to datasets** (e.g., goals and criteria, evaluation, funding, etc.)

Data

Data Collection Spectrum



From https://www.youtube.com/watch?v=v_XBjd1Fxqc (minute 23:42)

Cf. Jo and Gebru. 2020. *Lessons from archives* <https://dl.acm.org/doi/abs/10.1145/3351095.3372829>

How to ‘deal with’ bias?

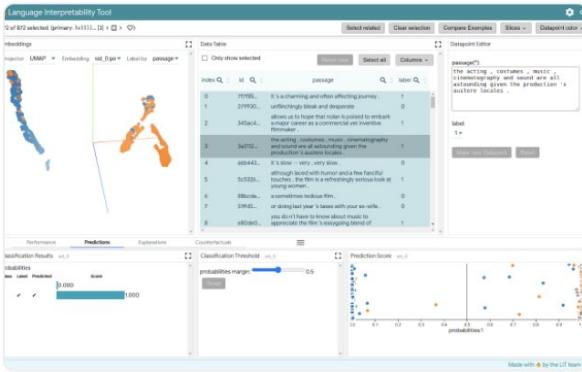
Start by **defining what you mean by bias**, this should be tightly related with your **problem formulation** (what is it that you are trying to solve or understand?). If your models or insights will become operational (go ‘in production’), also consider which **definition of fairness** you want them to live up to, if any.

If you have **data**, explore/profile your data *before* doing any modelling. Look for skewness, missing values, outliers, unbalance across protected groups, etc. Perform data selection or correction accordingly. If you don’t have data, design a **data acquisition** process which avoids possible bias (as you previously defined).

Make sure you put extra care to consider underrepresented and protected groups when doing **model evaluation**, and to consider possibly unintended consequences when **deploying**. Evaluation is never over, don’t sleep on your models.

Above all, keeping an open mind and ask for diverse feedback!

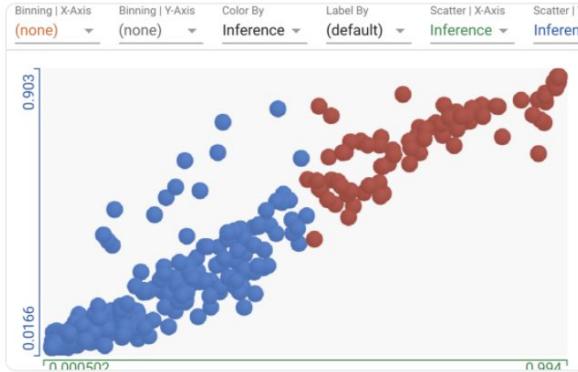
Useful tools: PAIR



LIT

Interactively analyze natural language models in an extensible and framework-agnostic interface.

↗ Explore LIT



What-If Tool

Visually probe the behavior of trained models, with minimal coding.

↗ Explore What-If Tool



Facets

Visualization libraries to explore, understand, and analyze large machine learning datasets.

↗ Explore Facets

A word of caution

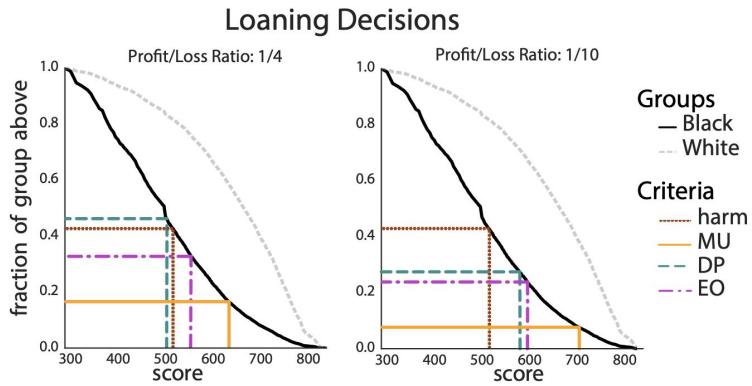


Figure 5: The empirical CDFs of both groups are plotted along with the decision thresholds resulting from **MaxUtil**, **DemParity**, and **EqOpt** for a model with bank utilities set to (a) $\frac{u_-}{u_+} = -4$ and (b) $\frac{u_-}{u_+} = -10$. The threshold for active harm is displayed; in (a) **DemParity** causes active harm while in (b) it does not. **EqOpt** and **MaxUtil** never cause active harm.

A word of caution

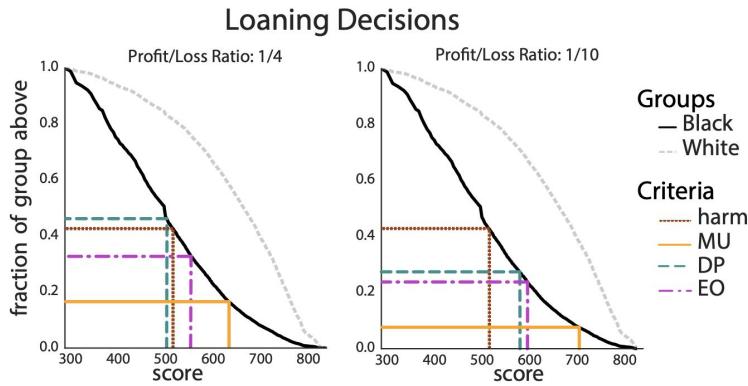


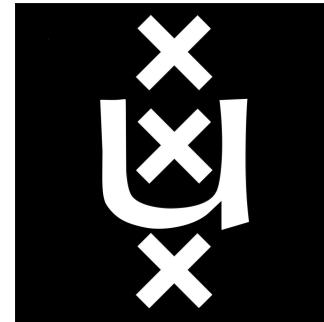
Figure 5: The empirical CDFs of both groups are plotted along with the decision thresholds resulting from `MaxUtil`, `DemParity`, and `EqOpt` for a model with bank utilities set to (a) $\frac{u_-}{u_+} = -4$ and (b) $\frac{u_-}{u_+} = -10$. The threshold for active harm is displayed; in (a) `DemParity` causes active harm while in (b) it does not. `EqOpt` and `MaxUtil` never cause active harm.

Socio-technical issues cannot be solved by technology alone. Sometimes, the urge to ‘do the right thing’ can lead to even worse outcomes.

Q&A

PART 4: Assignment

Giovanni Colavizza



ProPublica's 'Machine Bias'



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ProPublica's 'Machine Bias'

ProPublica (a US newsroom) investigated a commercial system to predict recidivism risk called COMPAS. The **COMPAS system** is a black-box (i.e., its inner workings are proprietary), assigning a score from 1 (low-risk) to 10 (high-risk) to people, which might be used for decision making (e.g., bail or preventive incarceration).

ProPublica performed an analysis (releasing data and code), attempting to replicate COMPAS' predictions using some variables and **claiming that the COMPAS system is biased against African-Americans.**

ProPublica's 'Machine Bias'

ProPublica (a US newsroom) investigated a commercial system to predict recidivism risk called COMPAS. The **COMPAS system** is a black-box (i.e., its inner workings are proprietary), assigning a score from 1 (low-risk) to 10 (high-risk) to people, which might be used for decision making (e.g., bail or preventive incarceration).

ProPublica performed an analysis (releasing data and code), attempting to replicate COMPAS' predictions using some variables and **claiming that the COMPAS system is biased against African-Americans.**

Your task will be, in groups, to a) get to know the dataset ProPublica used; b) replicate their analysis and understand their claims in terms of 'bias' and 'fairness'; c) engage with critical reviews on ProPublica's analysis and convey your own informed opinion on it.

Set-up

Please organise yourselves into groups of 4 students (exceptionally, 3 or 5). You will keep the same group for all assignments. Motivated requests to change group can be made.

Let's check the course repository for more info (assignment 1):

https://github.com/Giovanni1085/UvA_AlforSociety_2021

Note: we will assume you can clone a GitHub repository, set-up a working Python environment (ideally virtual, e.g., via Conda) and work with Jupyter notebooks. If you need some pointers/help, we have included a guide to setting up your working environment in the repo.

Q&A