Fairness in automated decision systems: motivations and preliminary researches

Antonio Vetrò

Nexa Center for Internet & Society, DAUIN
Future *Urban Legacy* Lab

Department of Control and Computer Engineering
Politecnico di Torino

antonio.vetro@polito.it

Outline

1. Motivations
2. Fairness
3. Conclusions

We know already: software is everywhere

2011: "Software is eating the world!" [1]

[1] Marc Andreessen, https://on.wsj.com/2lDLhKk

We know already: software is everywhere

2011: "Software is eating the world!" [1]

2016: "Software is programming the world!" [2]

[1] Marc Andreessen, https://on.wsj.com/2lDLhKk
[2] Marc Andreessen, https://bit.ly/2TZW4jJ

Automatic decisions/recommendations

In our societies, institutions and organizations delegate a relevant share of decisions to software, harnessing the large availability of historical data. A few examples:

- Works shift allocators
- Job recruiting
- Assisted driving
- Financial trading
- Justice suport systems
- Patrolling recommenders
- ...

Data and automatic decisions/recommendations

Thanks to the large availability of data and techniques to crunch them, software systems "learn" from historical series (inductive process). To "learn" correctly, though, machines need:

Data and automatic decisions/recommendations

Thanks to the large availability of data and techniques to crunch them, software systems "learn" from historical series (inductive process). To "learn" correctly, though, machines need:

- a sufficiently high number of examples
- a sufficiently diverse set of examples
- examples are annotated with "right answers"

Some well known problems of historical series

When modelling human characteristics and behaviour starting from historical data, there are some well-known potential problems.
For example:

Some well known problems of historical series

When modelling human characteristics and behaviour starting from historical data, there are some well-known potential problems.
For example:

- Reality is always a superset of what that is measurable.

Some well known problems of historical series

When modelling human characteristics and behaviour starting from historical data, there are some well-known potential problems.
For example:

- Reality is always a superset of what that is measurable.
- Some aspects of our society (and of our life) are by no means directly measurable.

Some well known problems of historical series

When modelling human characteristics and behaviour starting from historical data, there are some well-known potential problems. For example:

- Reality is always a superset of what that is measurable.
- Some aspects of our society (and of our life) are by no means directly measurable.
- Assuming that we have all the data of a given social context, then the historical series will also reflect its imperfections, historical prejudices, cultural stereotypes, demographic inequalities. Extracting patterns from these data often implies replicating the same dynamics (or even exacerbating them).

When the software decides for us

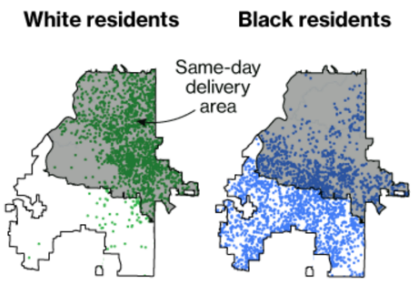
Many software systems often suggest us the "best" decision
(e.g., a navigator recommends the shortest way to go from A to B)

When the software decides for us

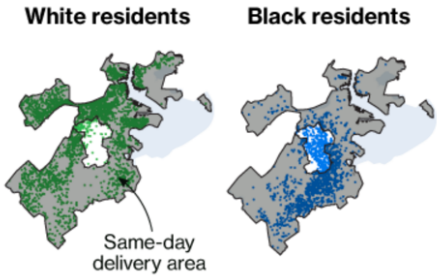
Many software systems often suggest us the "best" decision (e.g., a navigator recommends the shortest way to go from A to B)

- What does *best* mean ?
- On the basis of which criteria ?
- *Best* from which point of view ?

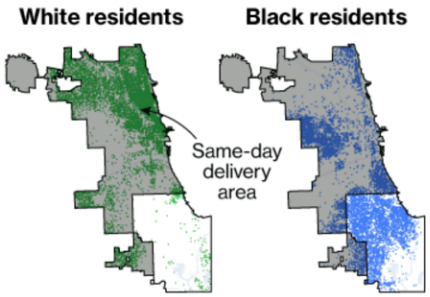
An example: disparities in deciding where to offer same-day delivery service (https://bloom.bg/2Nhgcvb)



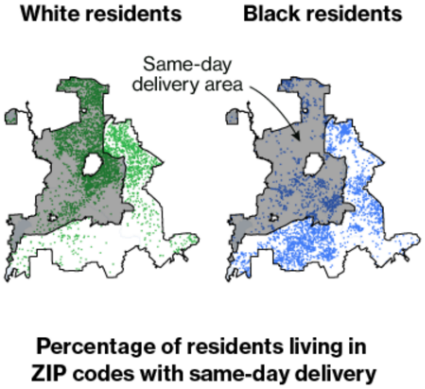
Atlanta



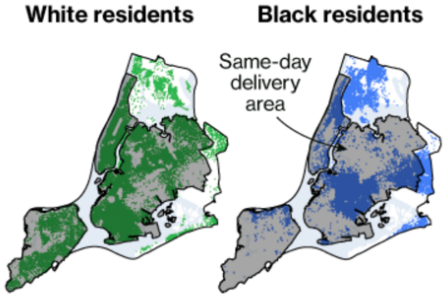
Boston



Chicago



Dallas



New York City

Washington DC

(After the publication of the study, the company extended the service to many of the excluded neighborhoods)

A few considerations

- The source code of the software is not known, and the criteria used by the algorithm are not known
- The data used by the software are also unknown
- The effect revealed by the study was that the service, although most likely unintentionally, discriminated on the basis of the color of the skin
- The practical consequence is that the system incorporated a prejudice, i.e. a **bias**, and behaved **unfairly**.

A few considerations

- The source code of the software is not known, and the criteria used by the algorithm are not known Should they?
- The data used by the software are also unknown
- The effect revealed by the study was that the service, although most likely unintentionally, discriminated on the basis of the color of the skin
- The practical consequence is that the system incorporated a prejudice, i.e. a **bias**, and behaved **unfairly**.

A few considerations

- The source code of the software is not known, and the criteria used by the algorithm are not known Should they?
- The data used by the software are also unknown Should they?
- The effect revealed by the study was that the service, although most likely unintentionally, discriminated on the basis of the color of the skin
- The practical consequence is that the system incorporated a prejudice, i.e. a **bias**, and behaved **unfairly**.

A second example: predictive policing

Predictive policing [3]

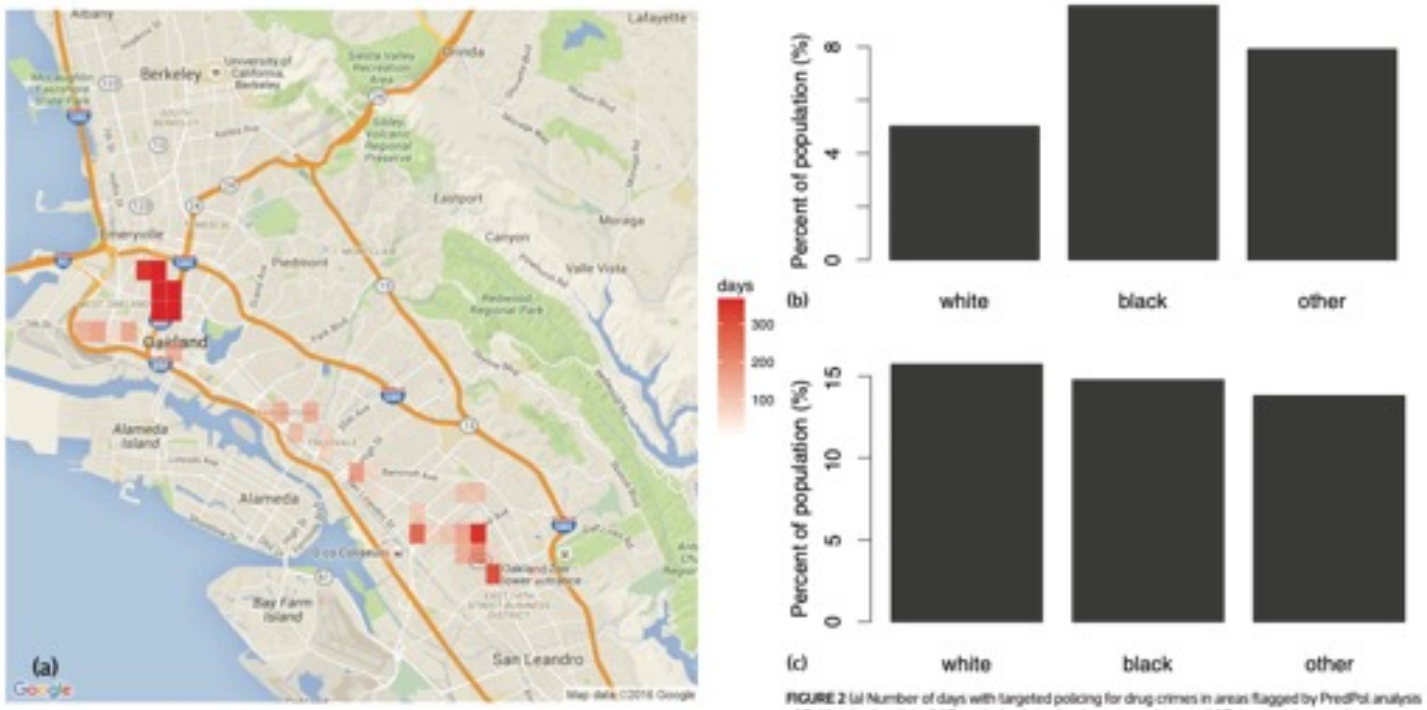


FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race

[3] https://doi.org/10.1111/j.1740-9713.2016.00960.x

A second example: predictive policing

Predictive policing [3]

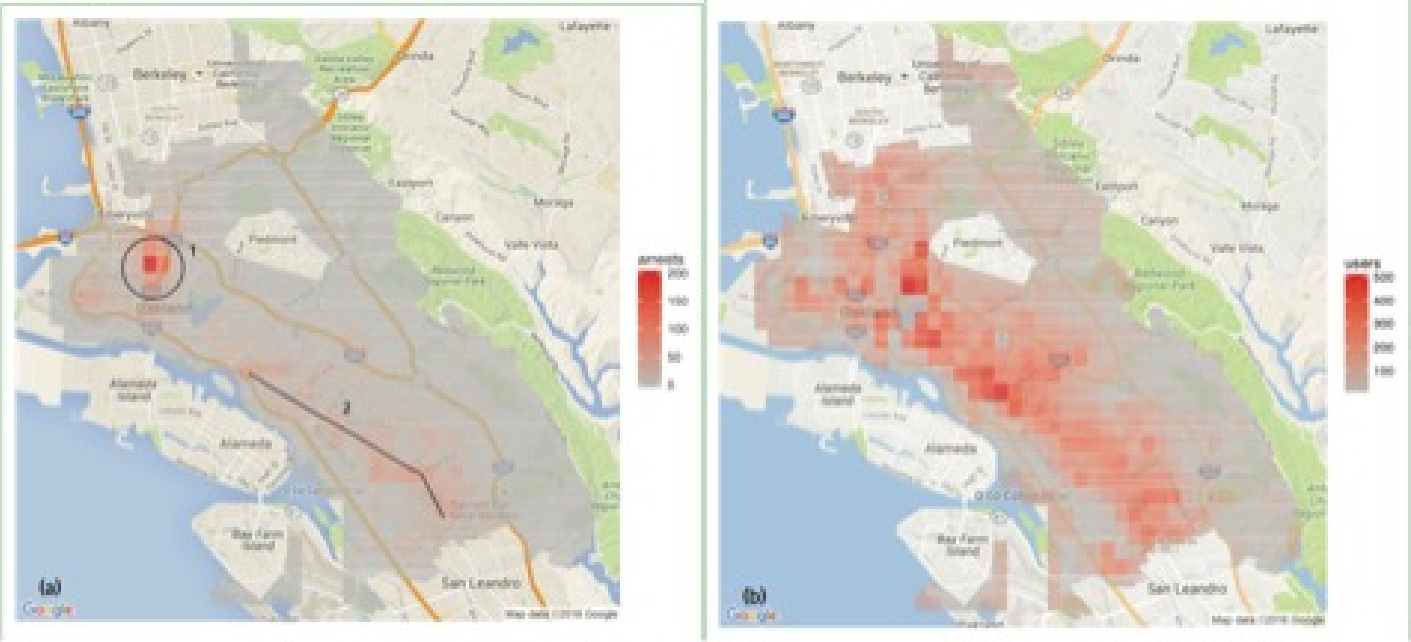


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

[3] https://doi.org/10.1111/j.1740-9713.2016.00960.x

A few considerations

- The source code of the software is not known, and the criteria used by the algorithm are not known Should they?
- The data used by the software are also unknown (except the one shared for the study) Should they?
- The effect revealed by the study was that the service, although most likely unintentionally, discriminated on the basis of the color of the skin
- The practical consequence is that the system incorporated a prejudice, i.e. a **bias**, and behaved **unfairly**.

Third example: the Pro Publica investigation 1/2 (https://bit.ly/1XMKh5R)

Two Petty Theft Arrests

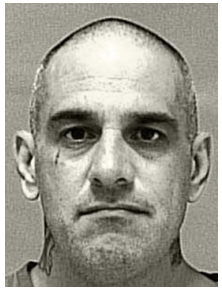
Name	Vernon Prater	Brisha Borden
Priors offenses	2 armed robberies, + 1 attempted	4 juvenile misdemeanors
Recidivism risk	3 (Low-risk)	8 (High-risk)
Subsequent offenses	1 grand theft	None

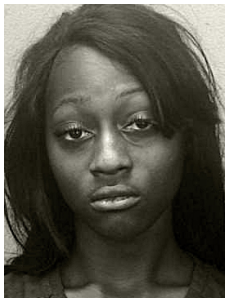
Third example: the Pro Publica investigation 1/2 (https://bit.ly/1XMKh5R)

Two Petty Theft Arrests

Name	Vernon Prater	Brisha Borden
Priors offenses	2 armed robberies, + 1 attempted	4 juvenile misdemeanors
Recidivism risk	3 (Low-risk)	8 (High-risk)
Subsequent offenses	1 grand theft	None

Picture





Third example: the Pro Publica investigation 2/2 (https://bit.ly/1XMKh5R)

Two Drug Arrests

Name	Dylan Fugett	Bernard Parker
Prior offenses	1 attempted burglary	1 resisting arrest without violence
Recidivism risk	3 (Low-risk)	10 (High-risk)
Subsequent offenses	3 drug possessions	None

Third example: the Pro Publica investigation 2/2 (https://bit.ly/1XMKh5R)

Two Drug Arrests

Name	Dylan Fugett	Bernard Parker
Prior offenses	1 attempted burglary	1 resisting arrest without violence
Recidivism risk	3 (Low-risk)	10 (High-risk)
Subsequent offenses	3 drug possessions	None

Picture

Third example: the Pro Publica case

- The formula used by COMPAS was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendant

Third example: the Pro Publica case

- The formula used by COMPAS was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendant
- White defendants were mislabeled as low risk more often than black defendants.

Same considerations as before

- The source code of the software is not known, and the criteria used by the algorithm are not known Should they?
- The data used by the software are also unknown (except the data found by Pro Publica) Should they?
- The effect revealed by the study was that the service, although most likely unintentionally, discriminated on the basis of the color of the skin
- The practical consequence is that the system incorporated a prejudice, i.e. a **bias**, and behaved **unfairly**.

FOCUS: The allegations of the U.S. House and Urban Department

UNITED STATES OF AMERICA
DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT
OFFICE OF ADMINISTRATIVE LAW JUDGES

The Secretary, United States Department of Housing and Urban Development, on behalf of Complainant Assistant Secretary for Fair Housing and Equal Opportunity, Charging Party, v. Facebook, Inc., Respondent	HUD ALJ No. FHEO No. 01-18-0323-8

CHARGE OF DISCRIMINATION

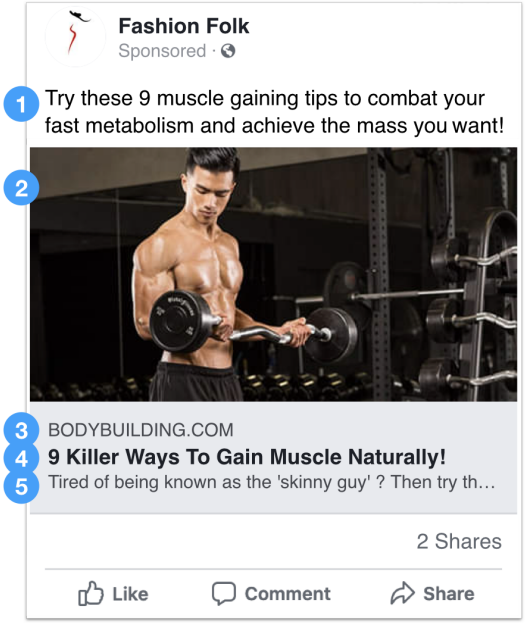
The Facebook advertising platform

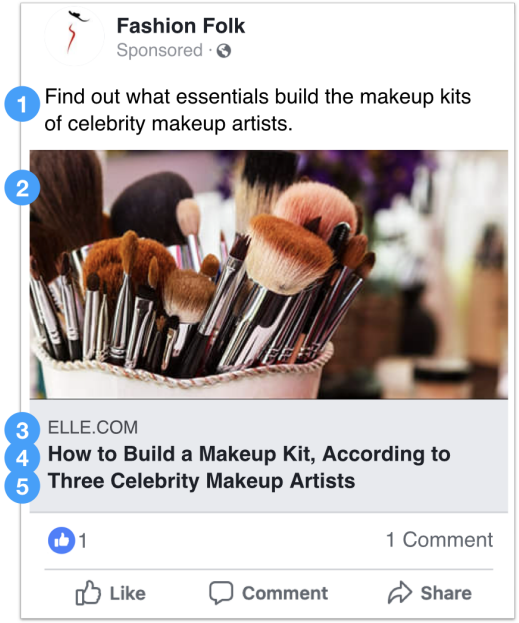
Ali, M. et al. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. arXiv, 1904.02095. Retrieved from https://arxiv.org/abs/1904.02095

https://www.washingtonpost.com/business/2019/03/28/hud-charges-facebook-with-housing-discrimination/

https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html

The Facebook advertising platform





1. Headline and text
2. Images and/or videos
3. Domain (pulled automatically)
4. Title (pulled automatically)
5. Description (pulled automatically)

The Facebook advertising process

1. Ad creation
   a. Ad contents
   b. Audience selection/Targeting
   c. Bidding strategy
2. Ad delivery

Effects evaluated in the study

Effects evaluated in a selected audience:

1. Budget
2. Ad creative
3. Ad image
4. Ad image classification

Effects evaluated in the study

Effects evaluated in a selected audience:

1. Budget
2. Ad creative
3. Ad image
4. Ad image classification

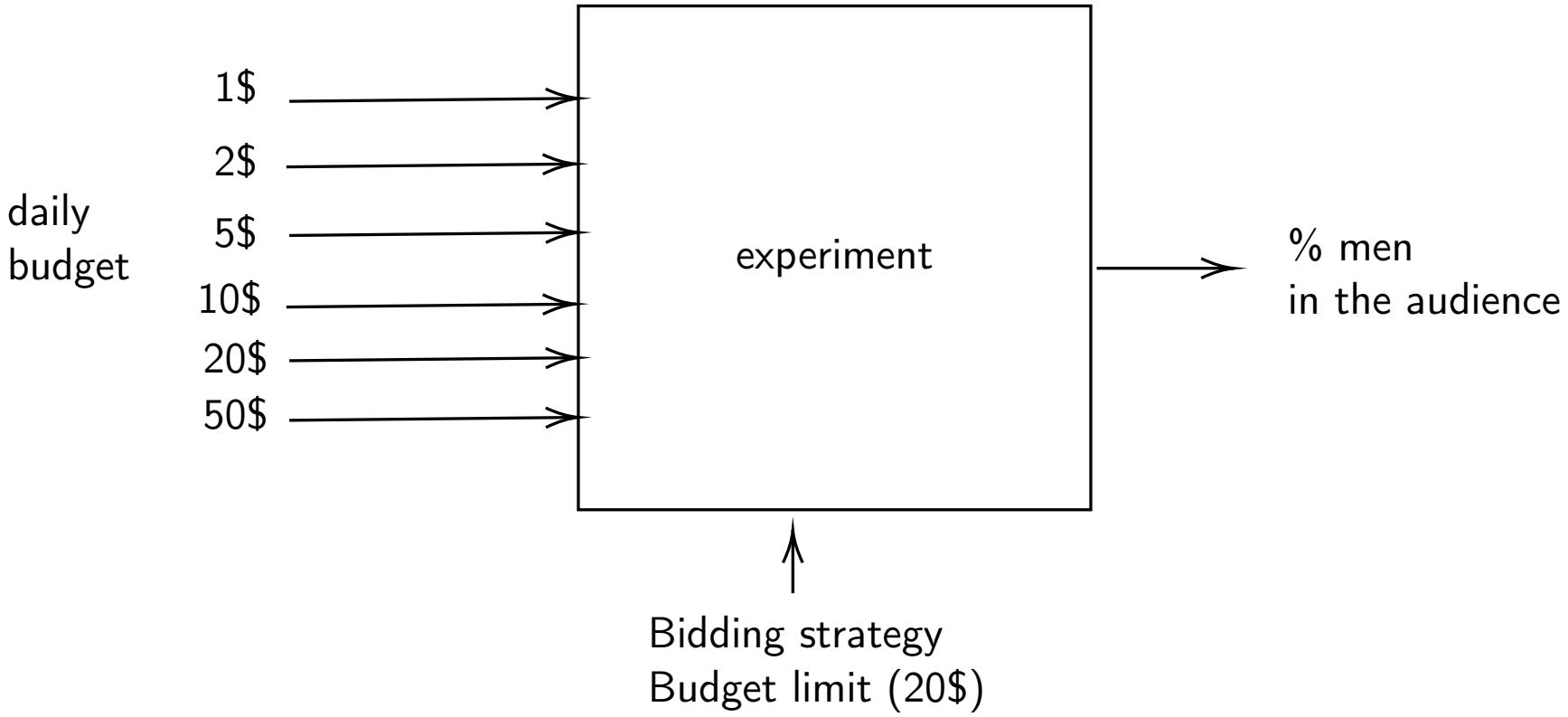
+ 5. Test on real-world ads

Audience selected

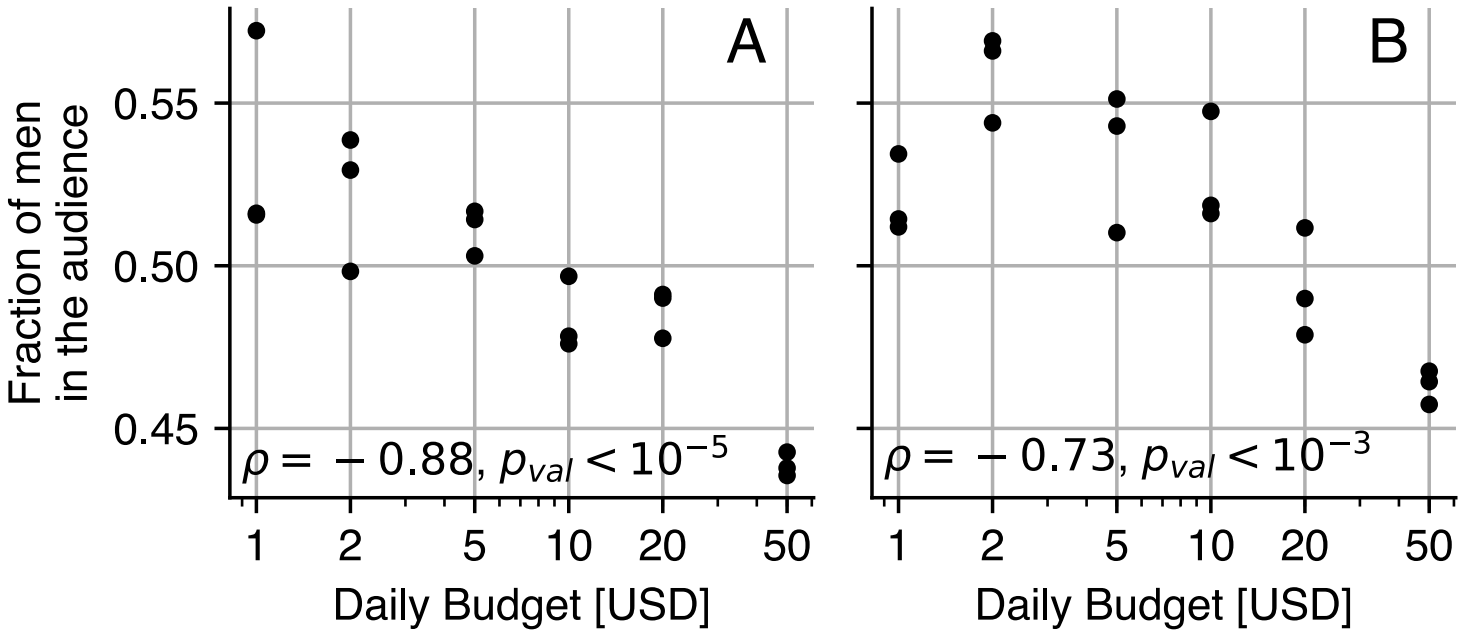
Selected regions in North Carolina (USA)

DMA regions(s)	Aud.	Records
Wilmington	White	450.000
Raleigh-Durham	White	450.002
Greenville-Spartanburg - New Bern	Black	446.047
Charlotte and Greensboro	Black	446.050

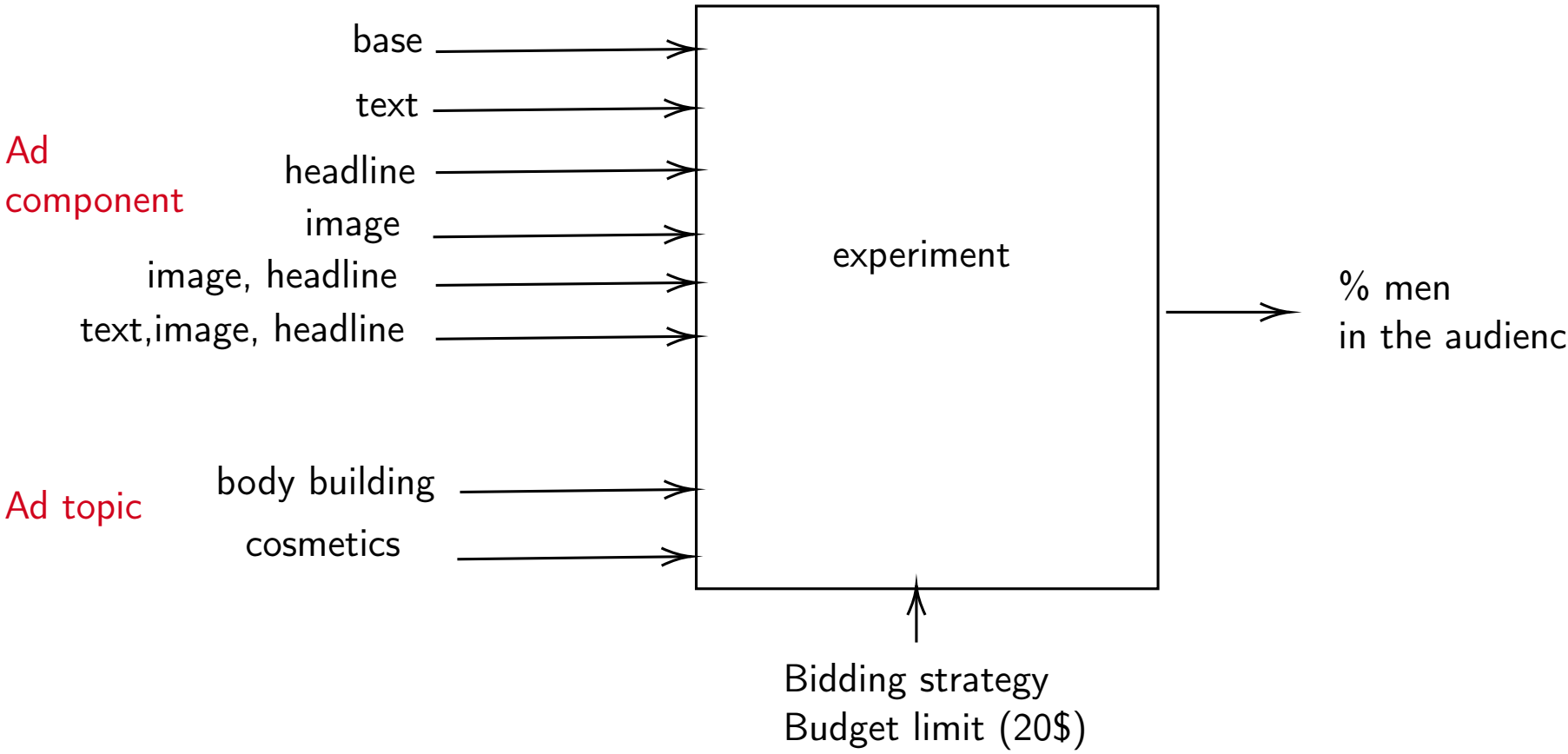
1. Budget effects: design



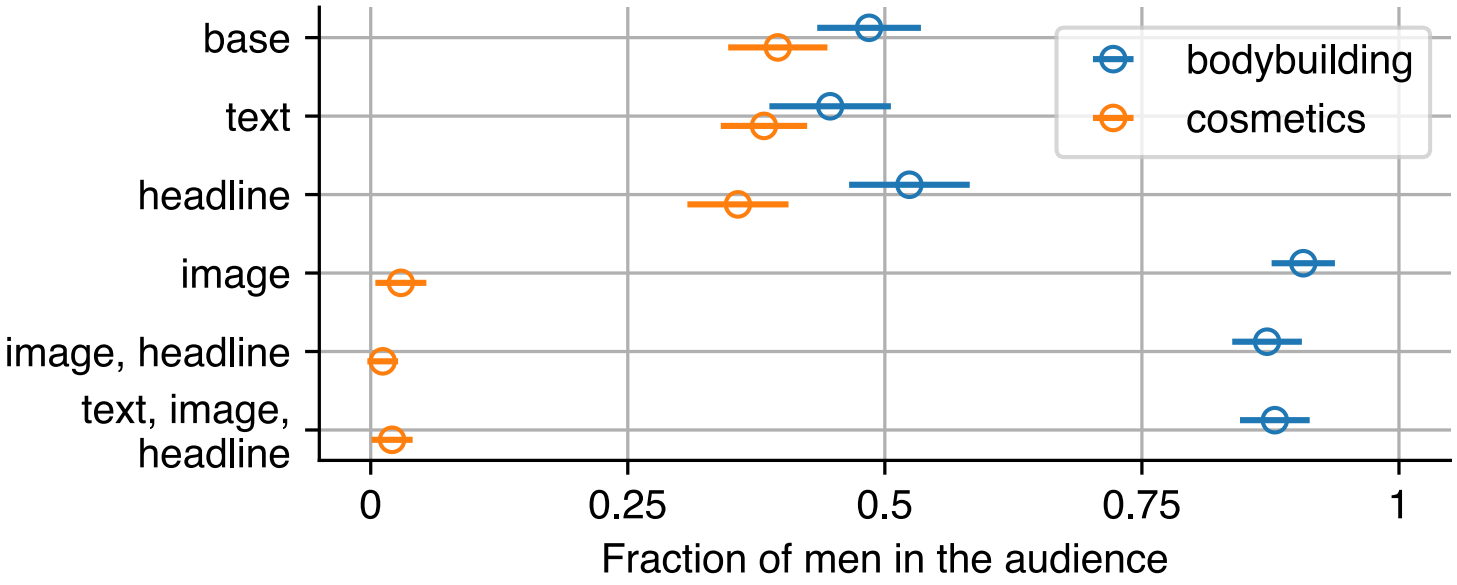
1. Budget effects: results



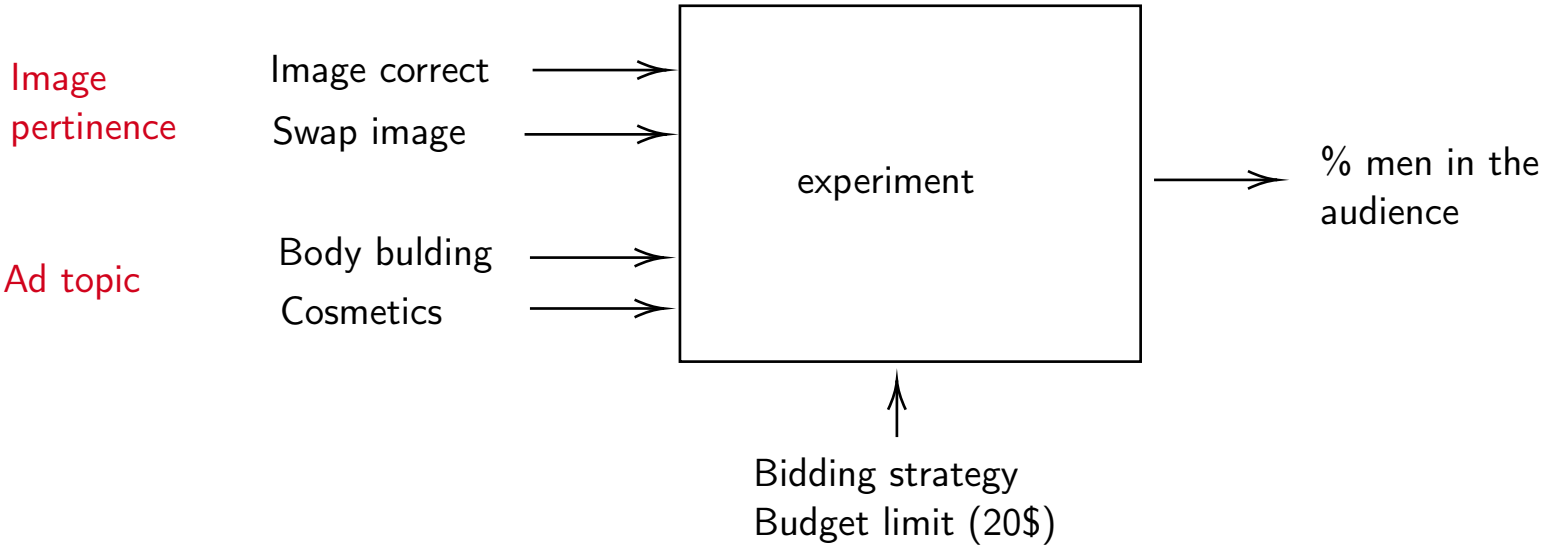
2. Ad creative effect: design



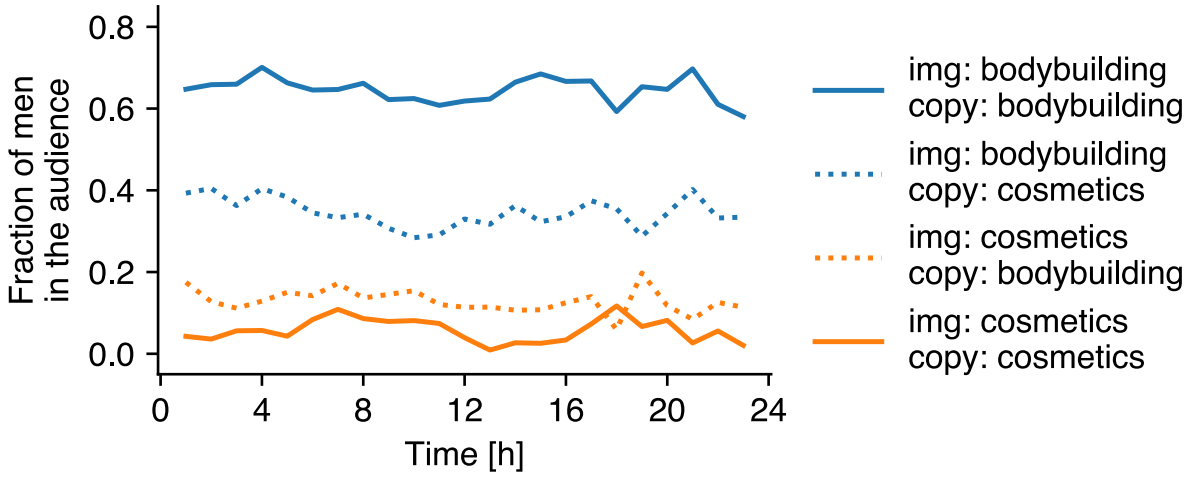
2. Ad creative effect: results



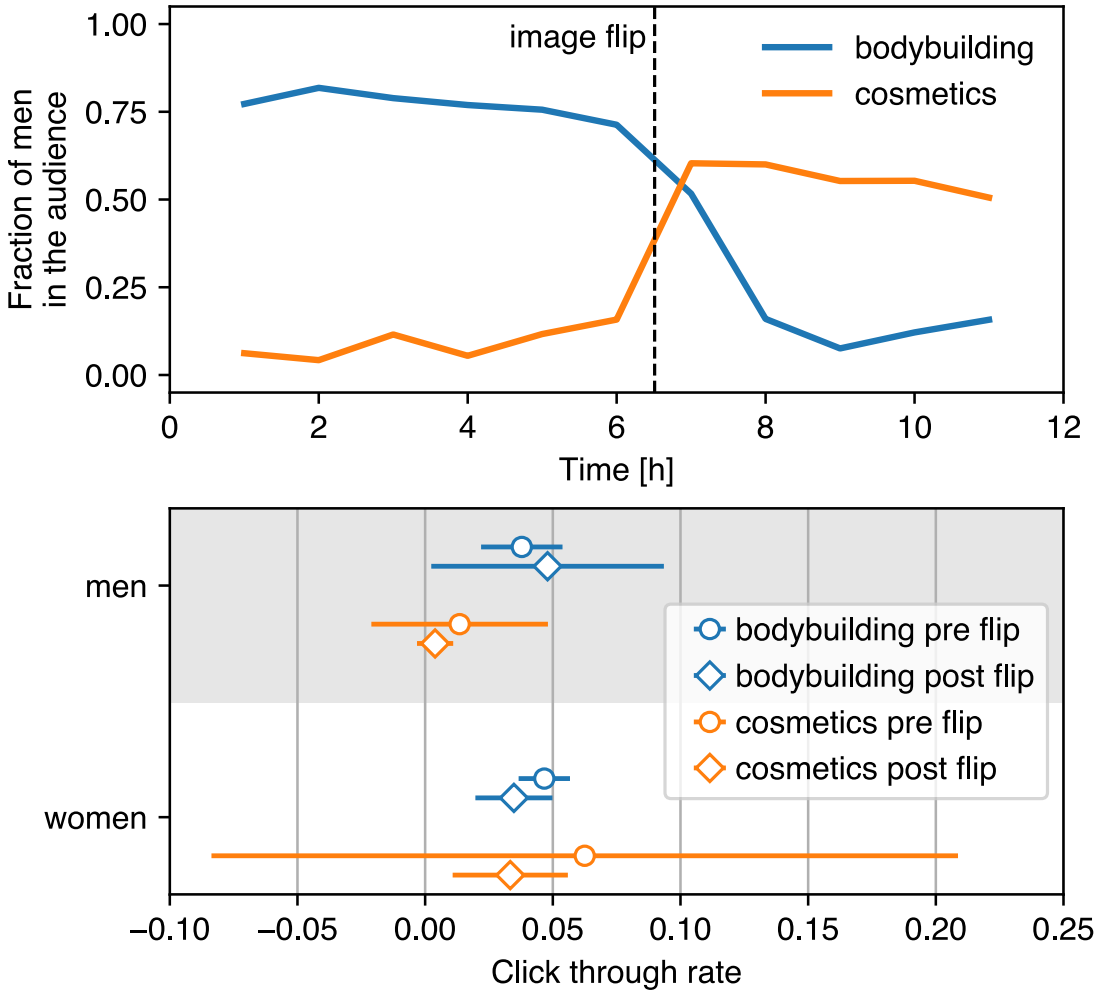
3. Ad image effect: design



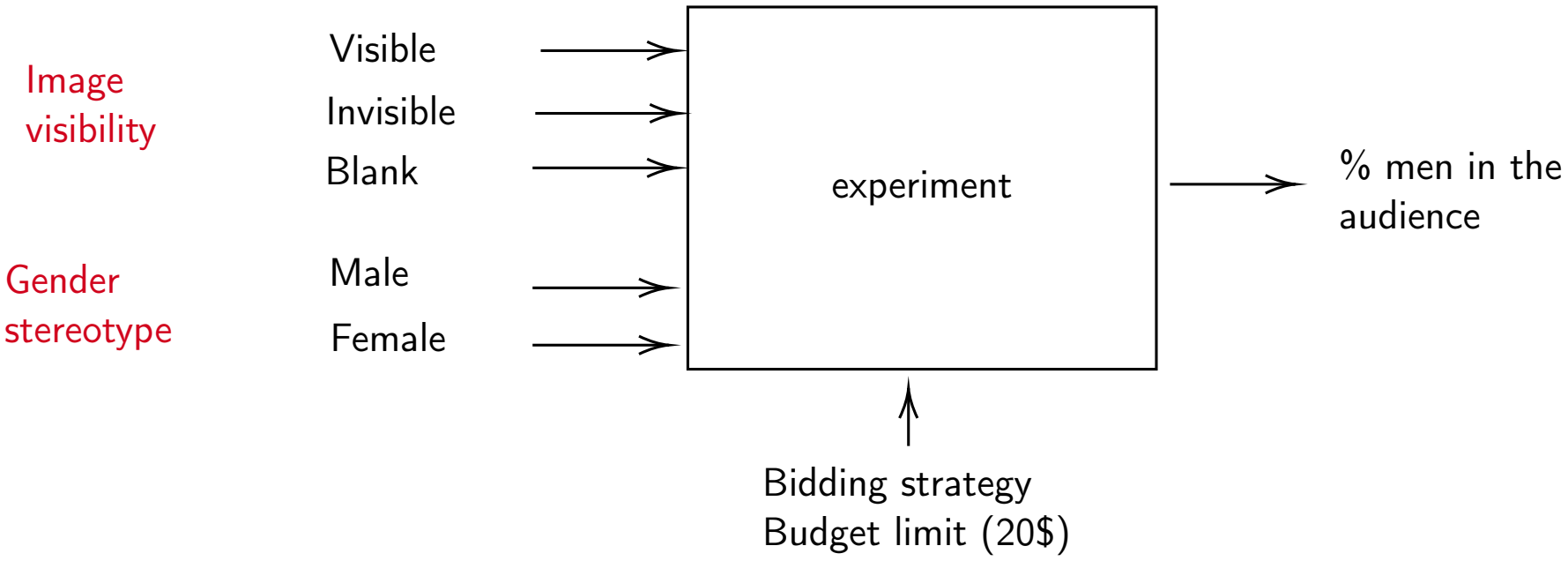
3. Ad image effect: results (1)



3. Ad image effect: results (2)



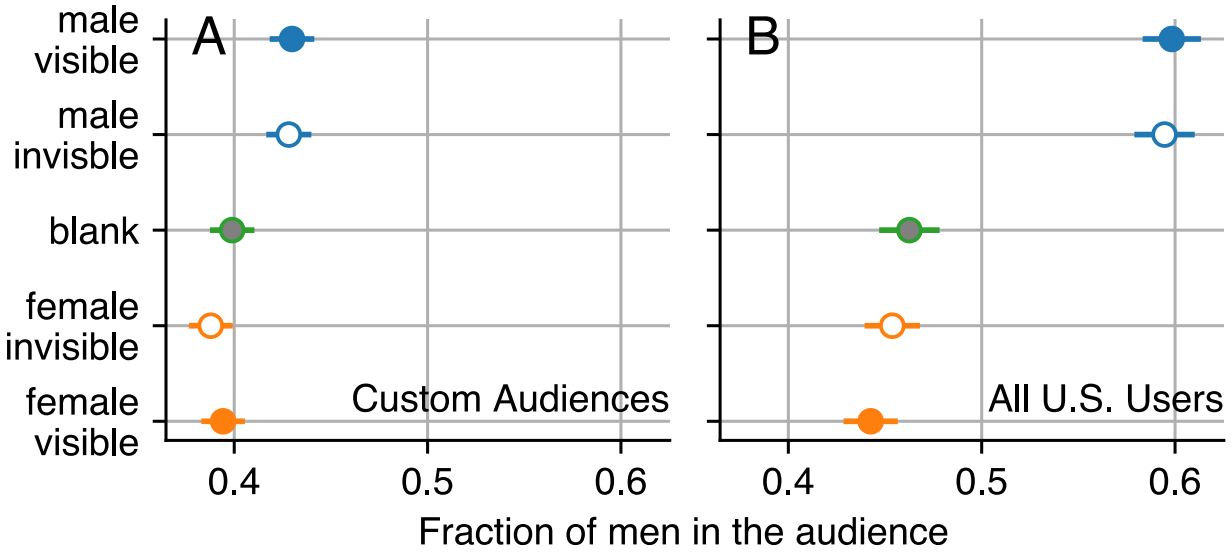
4. Ad image classification: design



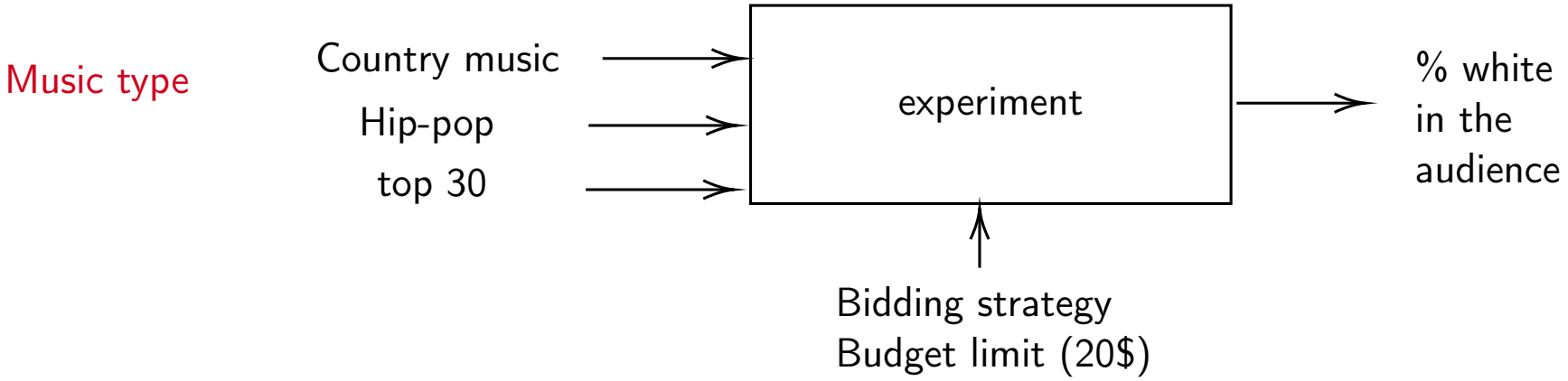
4. Ad image classification: images used

No.	Male		Female	
	Visible	Invisible	Visible	Invisible
1				
2				
3				
4				
5				

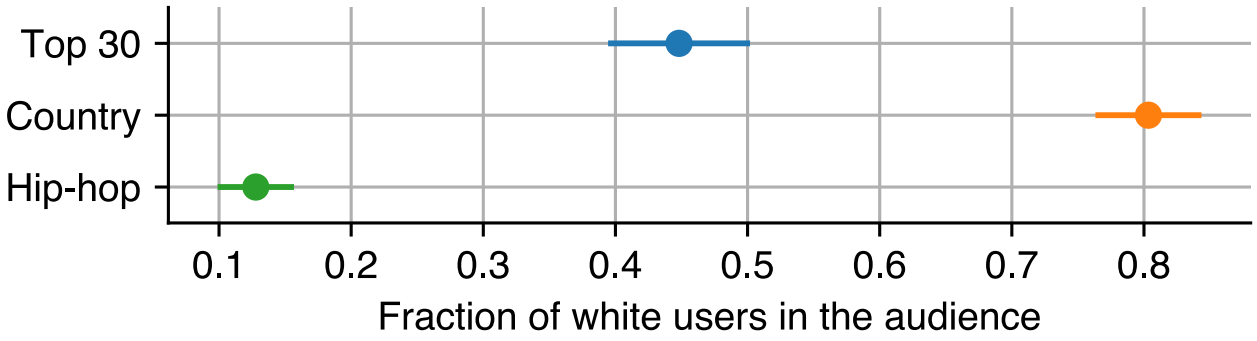
4. Ad image classification: results



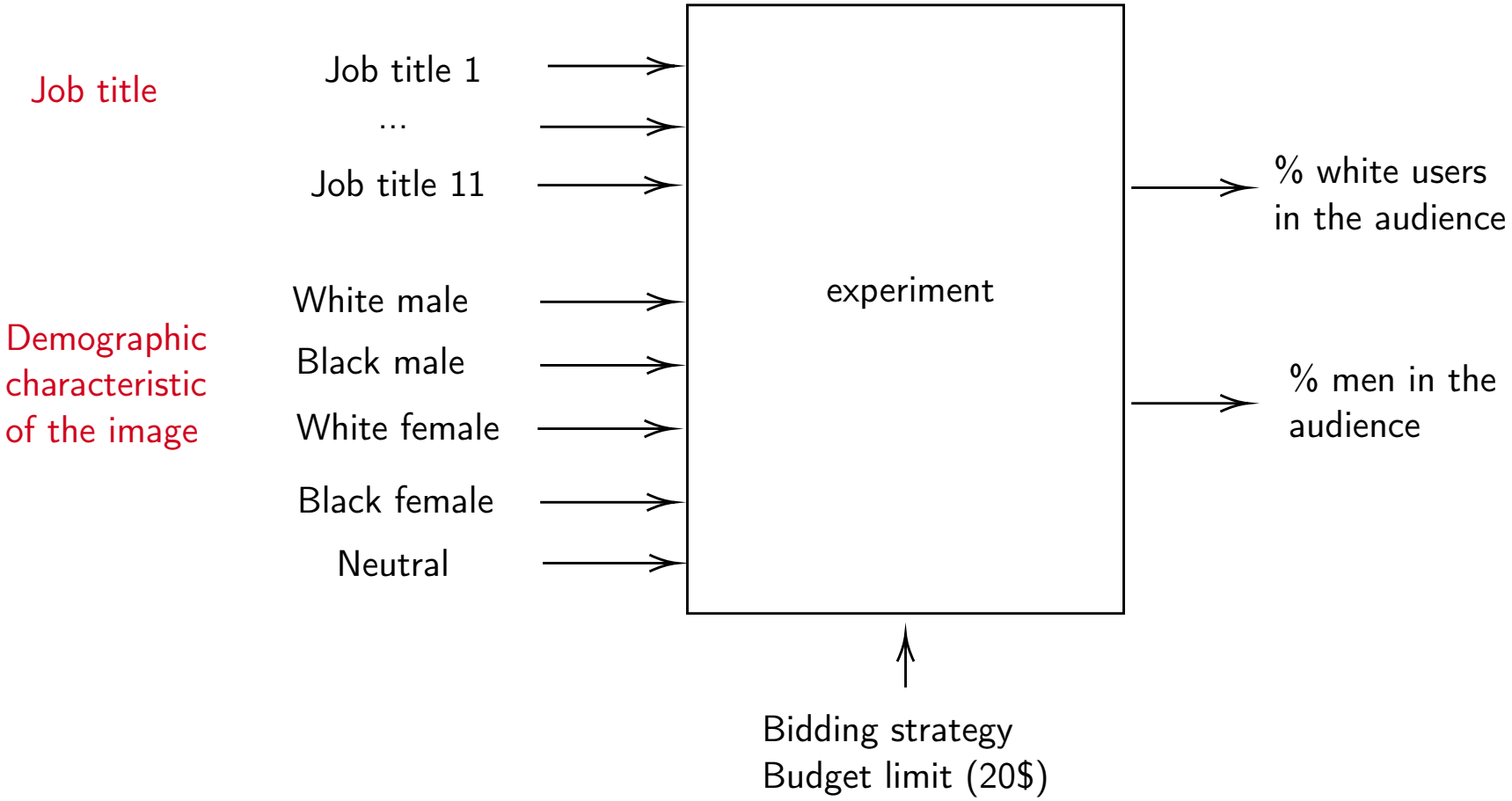
5. Test on real-world ads: entertainment-design



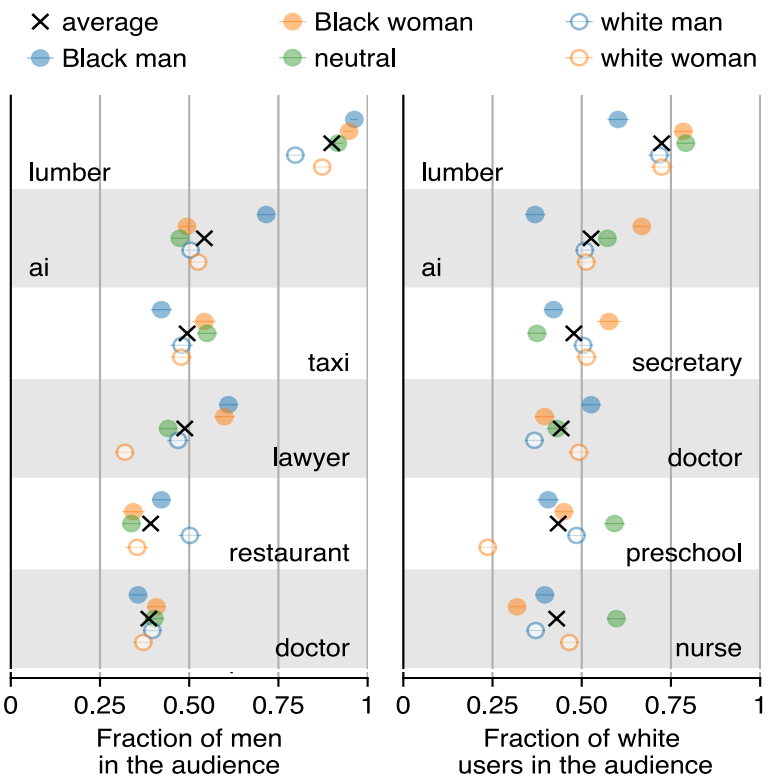
5. Test on real-world ads: entertainment-results

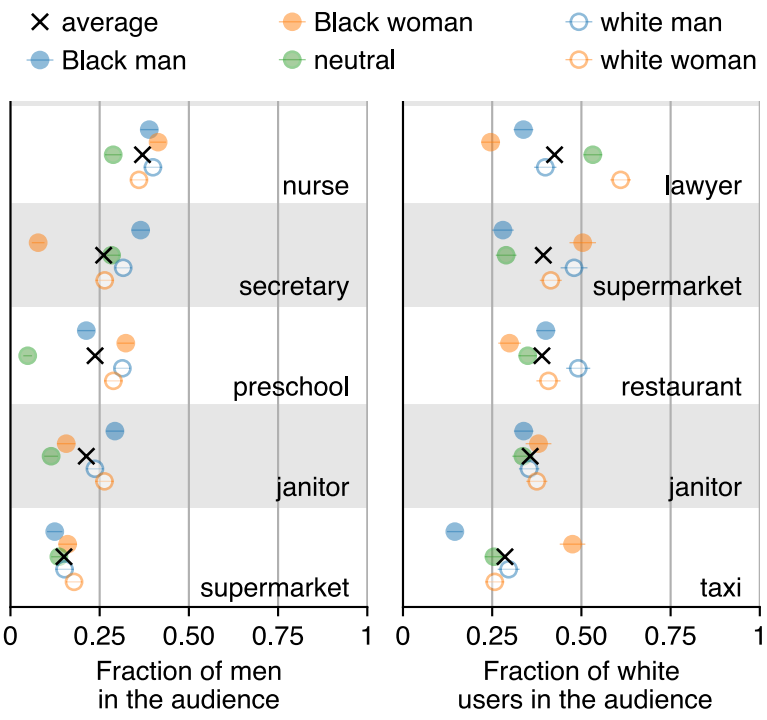


5. Test on real-world ads: employment-design

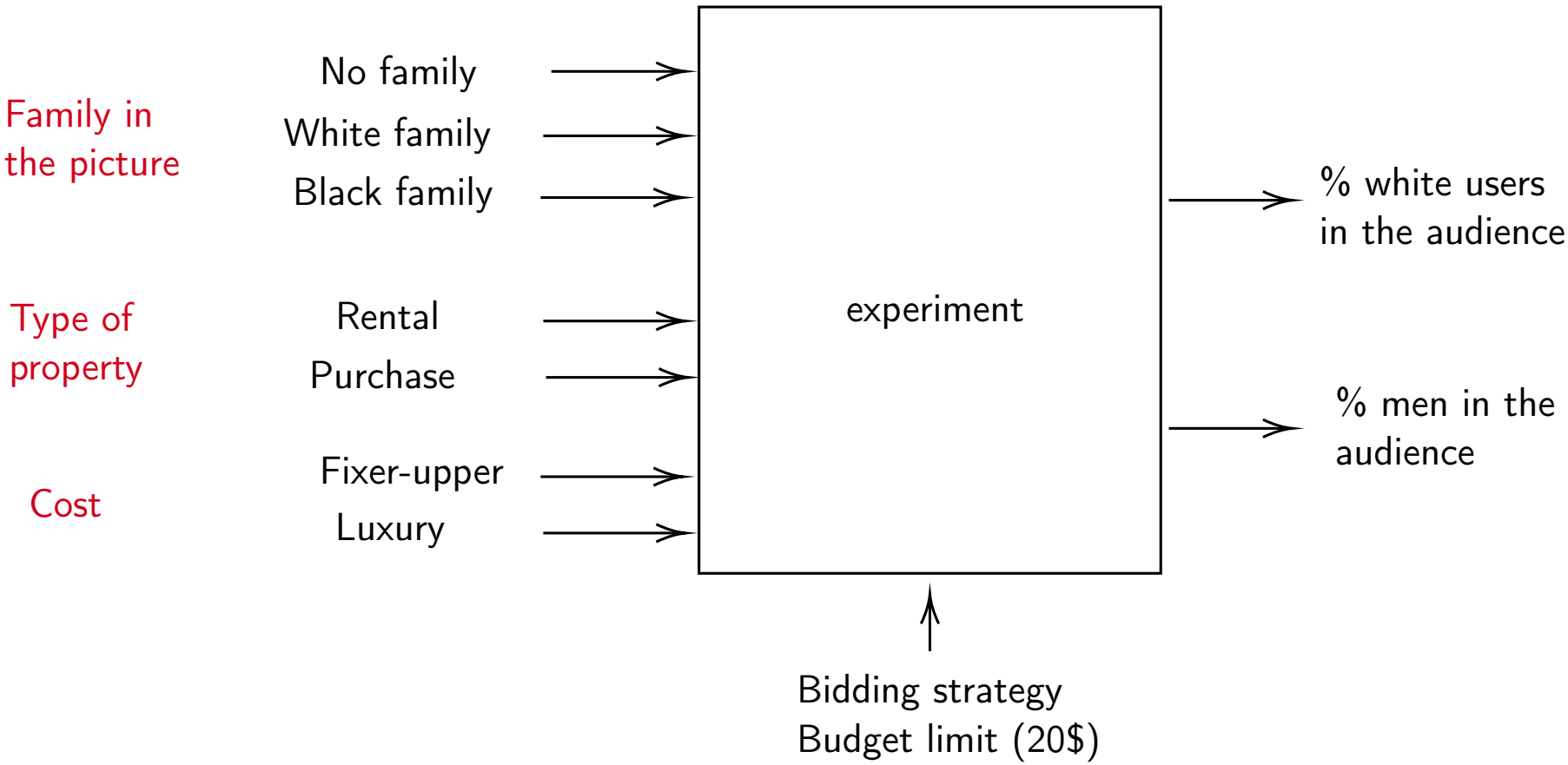


5. Test on real-world ads: employment-results

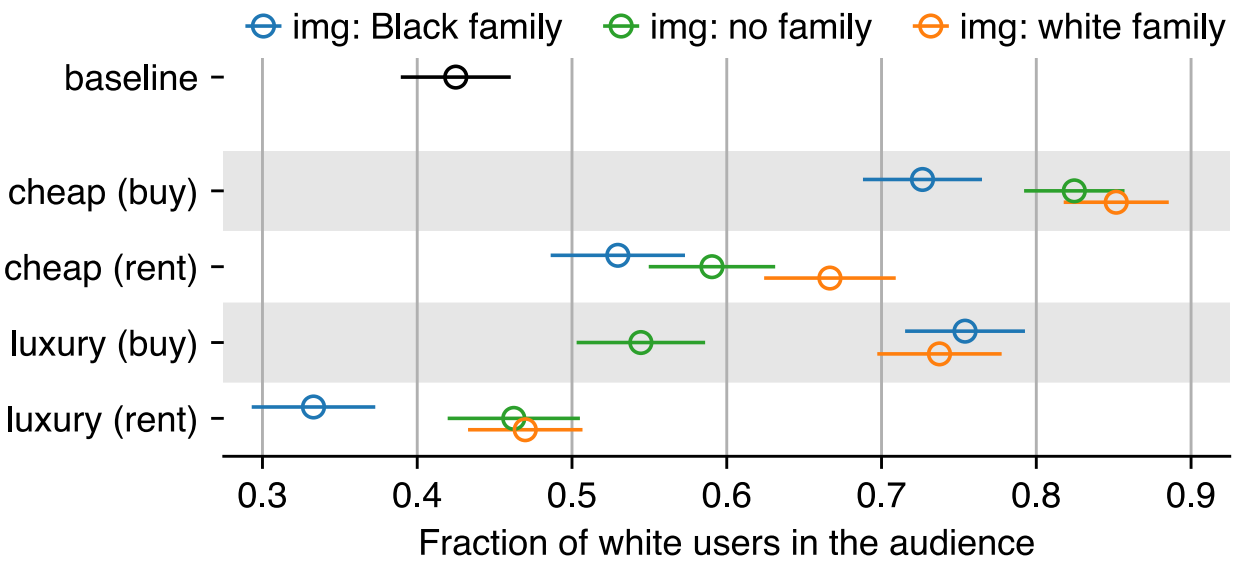




5. Test on real-world ads: housing-design



5. Test on real-world ads: housing-results



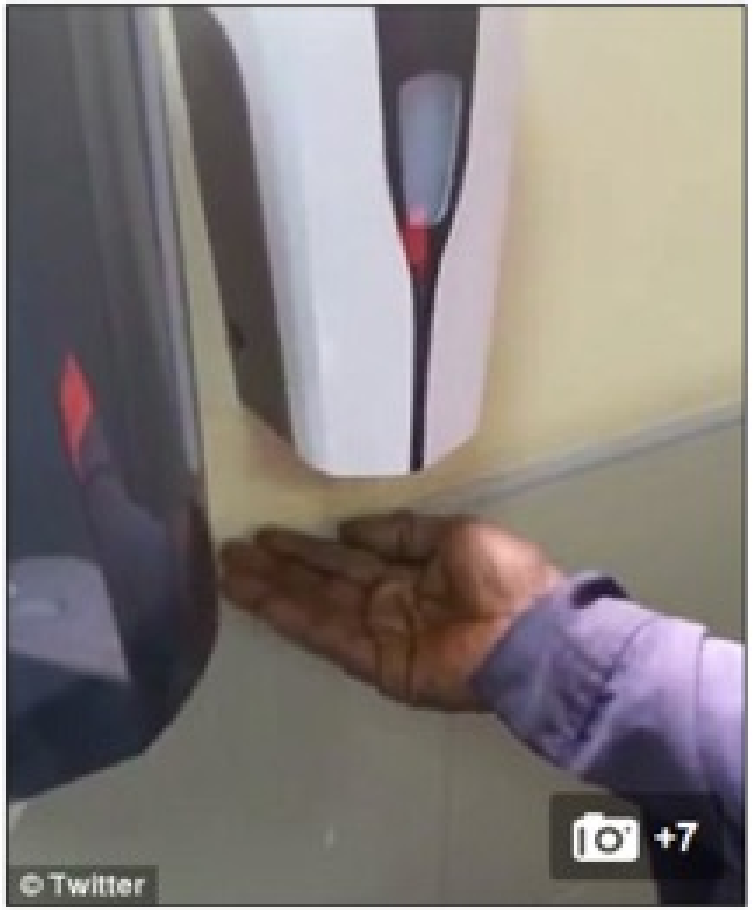
The allegations of the U.S. House and Urban Department

UNITED STATES OF AMERICA
DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT
OFFICE OF ADMINISTRATIVE LAW JUDGES

The Secretary, United States Department of Housing and Urban Development, on behalf of Complainant Assistant Secretary for Fair Housing and Equal Opportunity, Charging Party, v. Facebook, Inc., Respondent	HUD ALJ No. FHEO No. 01-18-0323-8

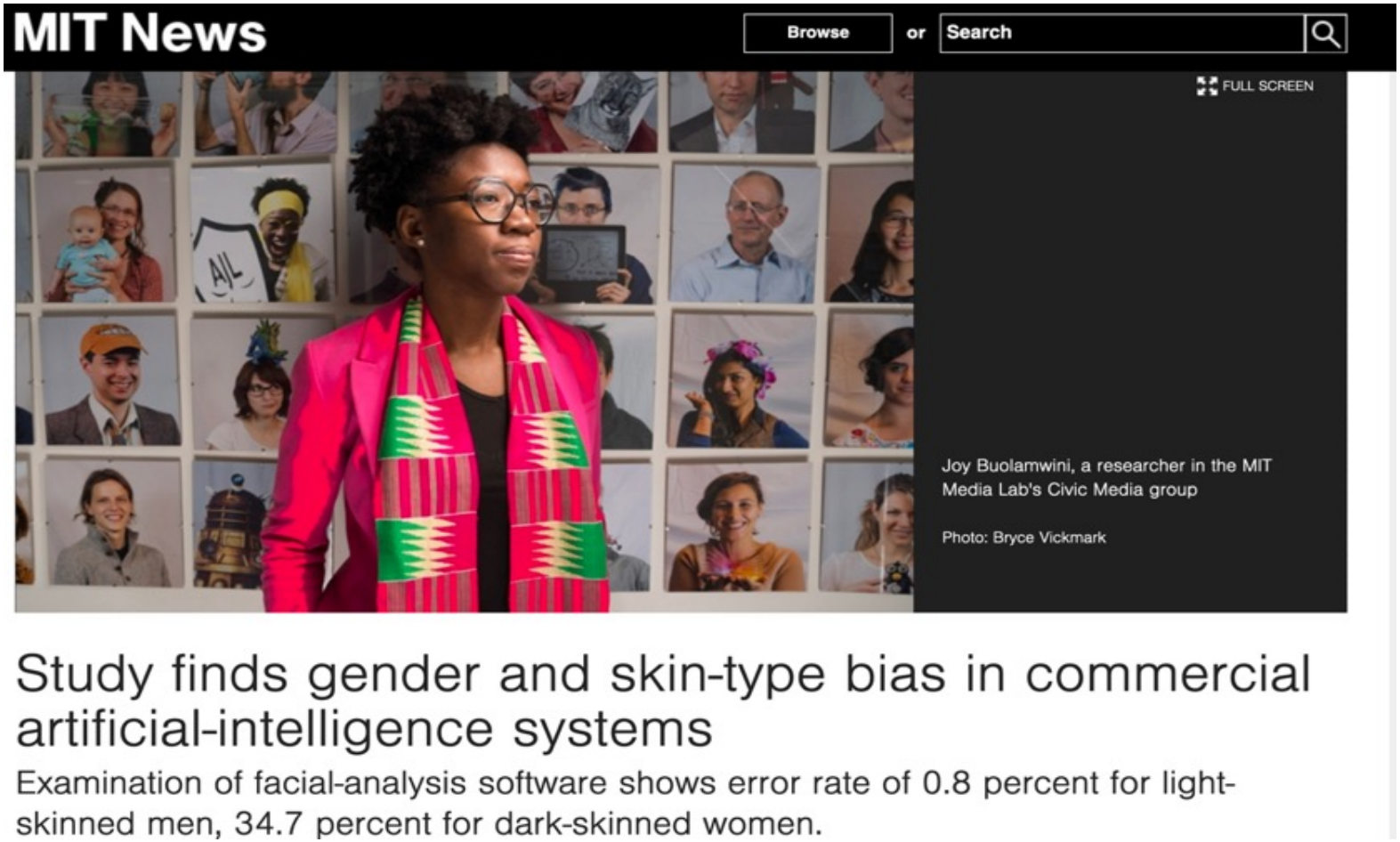
CHARGE OF DISCRIMINATION

Other examples



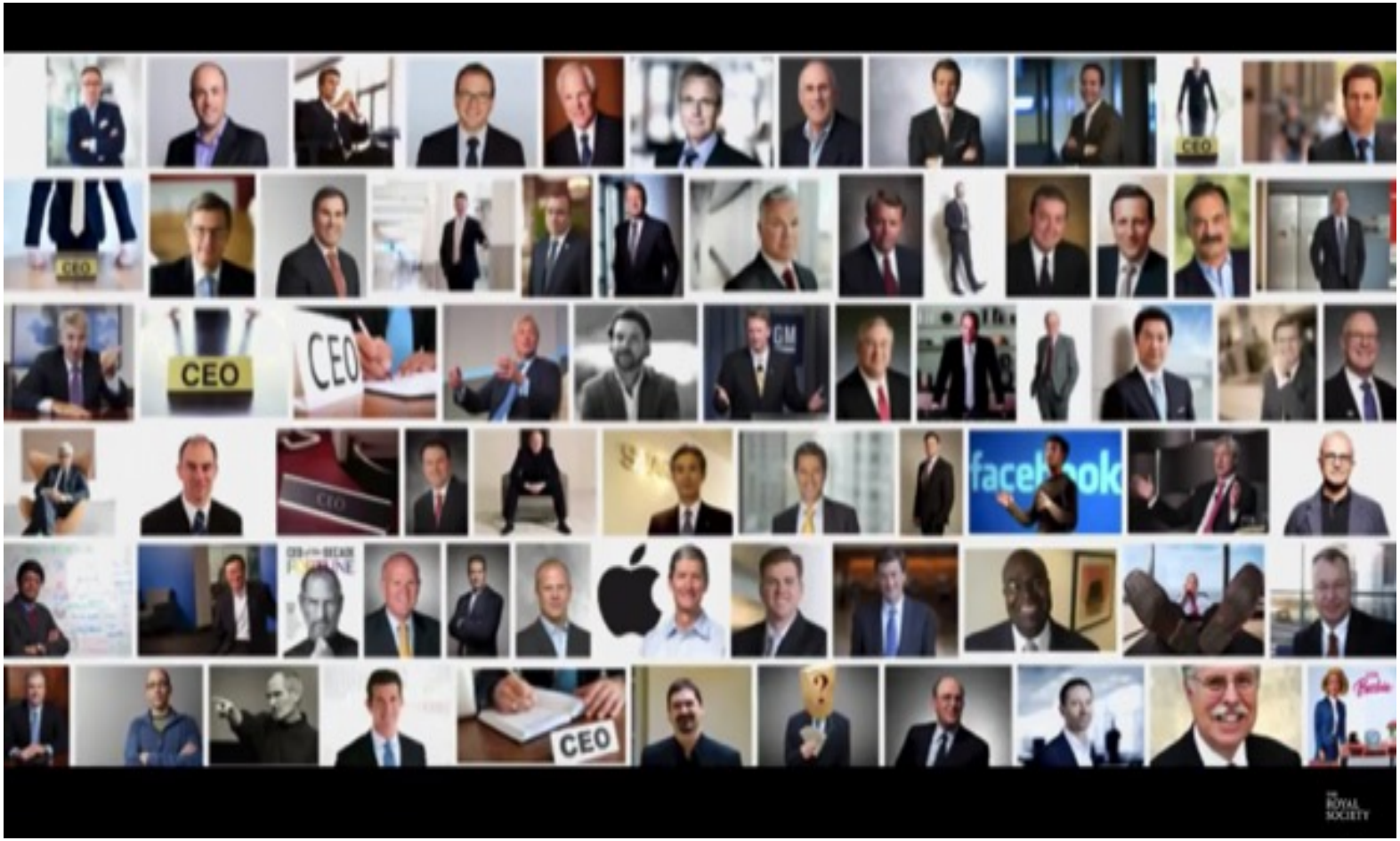
https://twitter.com/twitter/statuses/897756900753891328

Other examples



https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Other examples



A search on Google images with the word "CEO" (Chief Executive Officer)

Other examples

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Reuters

Thu 11 Oct 2018
00.42 BST

337

▲ Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates.
Photograph: Brian Snyder/Reuters

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Other examples

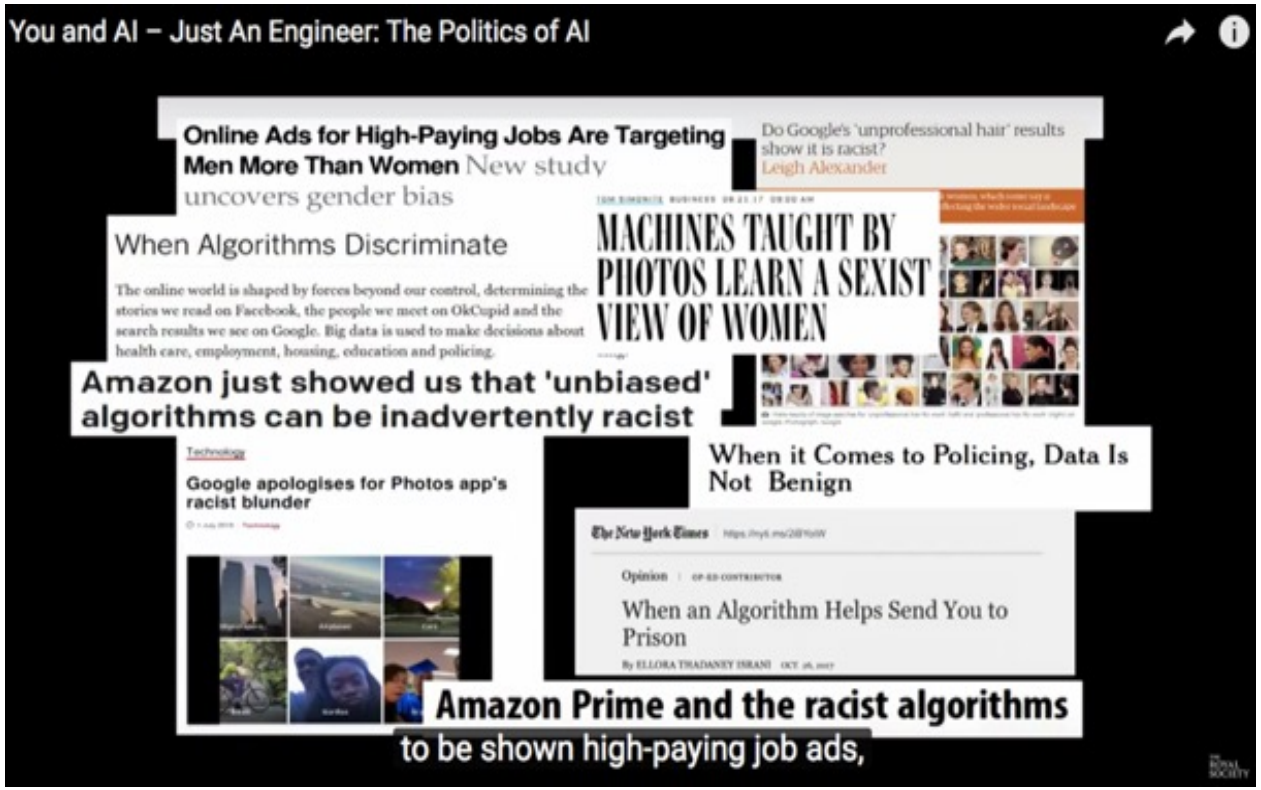
RENEE DIRESTA IDEAS 03.05.19 01:00 PM

HOW AMAZON'S ALGORITHMS CURATED A DYSTOPIAN BOOKSTORE

JOEL SAGET/AFP/GETTY IMAGES

https://www.wired.com/story/amazon-and-the-spread-of-health-misinformation/

Other examples



https://www.youtube.com/watch?v=HPopJb5aDyA

The problem is relevant

116TH CONGRESS
1ST SESSION

S. _____

To direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments.

IN THE SENATE OF THE UNITED STATES

Mr. WYDEN (for himself and Mr. BOOKER) introduced the following bill; which was read twice and referred to the Committee on _____

A BILL

To direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments.

1 *Be it enacted by the Senate and House of Representa-*
2 *tives of the United States of America in Congress assembled,*
3 **SECTION 1. SHORT TITLE.**
4 This Act may be cited as the "Algorithmic Account-
5 ability Act of 2019".

Algorithmic Accountability Act (April 2019, Senate USA)
https://bit.ly/2UCBKZT

Formalizing fairness

Bias types in software systems

Scholars Batya Friedman and Helen Nissenbaum already presented, in 1996, a taxonomy of possible biases in software systems [4]
Computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others. [5]

[4] https://doi.org/10.1145/230538.230561
[5] Friedman and Nissenbaum, https://doi.org/10.1145/230538.230561

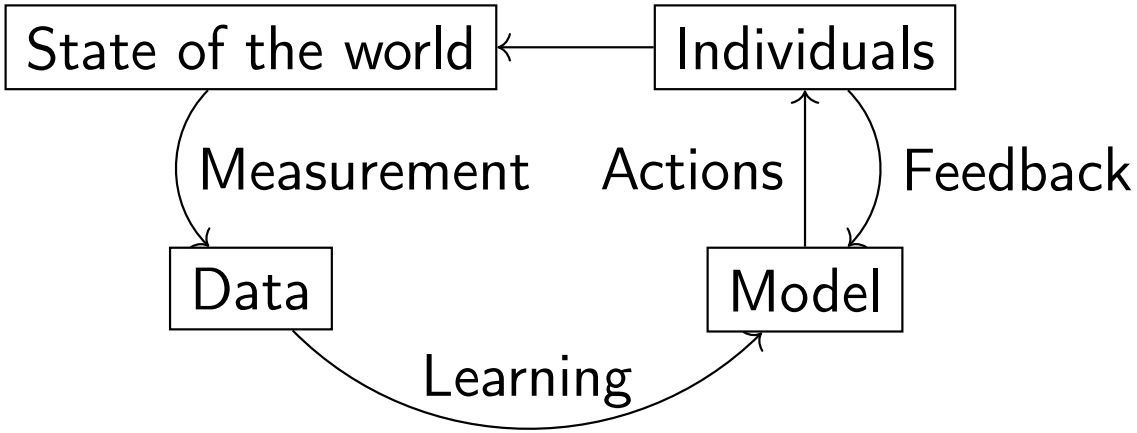
Bias types in software systems

- Pre-existent bias
  - Individual
  - Societal

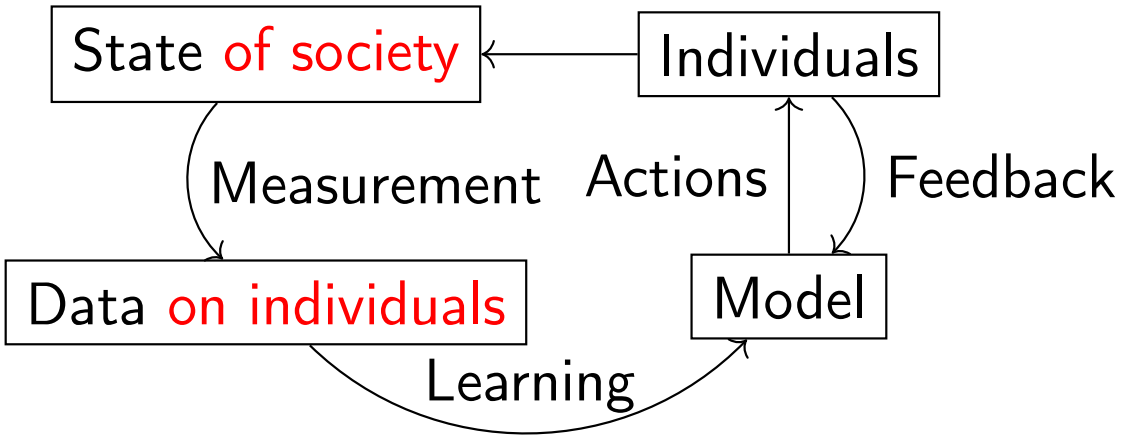
Bias types in software systems

- Pre-existent bias
  - Individual
  - Societal
- Technical Bias
  - Computer tools
  - Decontextualized Algorithms
  - Random Number Generation
  - Formalization of Human Constructs

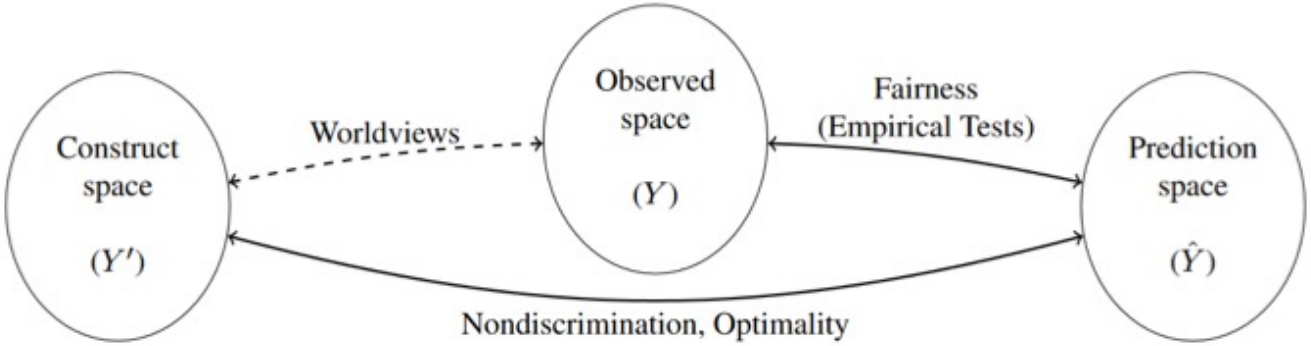
Mapping Bias: the ML loop



Mapping Bias: the ML loop



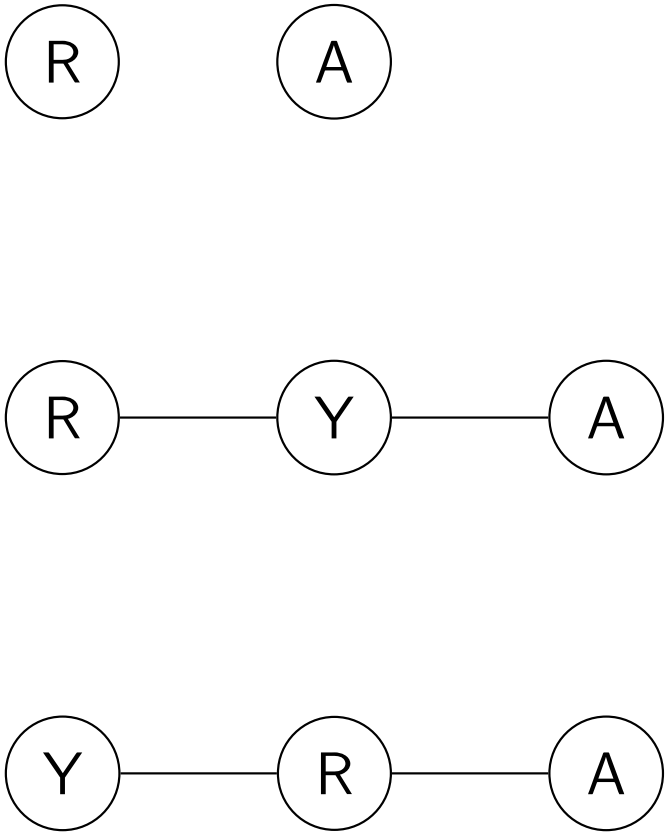
Premise: fairness and the space of observations



Fonte: https://arxiv.org/abs/1808.08619

Formalizing fairness

Independence: $R \perp\!\!\!\perp A$
Separation: $R \perp\!\!\!\perp A \mid Y$
Sufficiency: $Y \perp\!\!\!\perp A \mid R$



Legend

- R classifier
- Y target variable
- A sensitive attribute

$R, Y, A \in [0, 1]$

Independence

Independence: $R \perp\!\!\!\perp A$

$$\mathbb{P}\{R = 1 \mid A = a\} = \mathbb{P}\{R = 1 \mid A = b\}$$

R A

Independence

Independence: $R \perp\!\!\!\perp A$

$$\mathbb{P}\{R = 1 \mid A = a\} = \mathbb{P}\{R = 1 \mid A = b\}$$

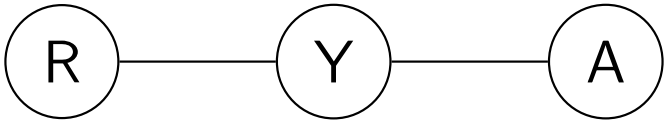
Relaxed constraint:

$$\frac{\mathbb{P}\{R=1|A=a\}}{\mathbb{P}\{R=1|A=b\}} \geq 1 - \epsilon$$

Separation

Separation: $R \perp\!\!\!\perp A \mid Y$

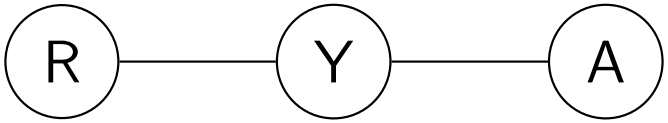
$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}$$



Separation

Separation: $R \perp\!\!\!\perp A \mid Y$

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}$$

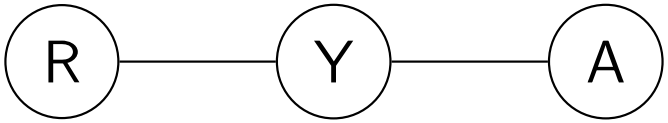


$\mathbb{P}\{R = 1 \mid Y = 1\}$ is true positive ratio
$\mathbb{P}\{R = 1 \mid Y = 0\}$ is false positive ratio

Separation

Separation: $R \perp\!\!\!\perp A \mid Y$

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}$$



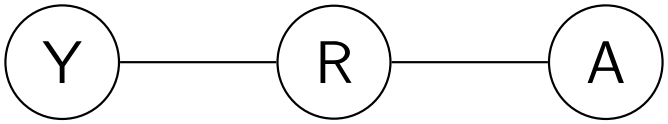
$\mathbb{P}\{R = 1 \mid Y = 1\}$ is true positive ratio
$\mathbb{P}\{R = 1 \mid Y = 0\}$ is false positive ratio

$\Longrightarrow$ Separation requires equity of error rates

Sufficiency

Sufficiency: $Y \perp\!\!\!\perp A \mid R$

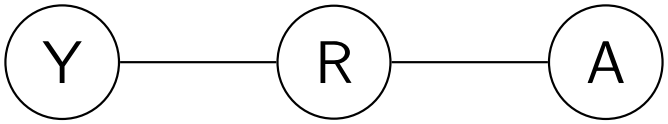
$$\mathbb{P}(Y = 1 \mid R = r, A = a) = \mathbb{P}(Y = 1 \mid R = r, A = b)$$



Sufficiency

Sufficiency: $Y \perp\!\!\!\perp A \mid R$

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = \mathbb{P}(Y = 1 \mid R = r, A = b)$$



Note:

- Interpretation: it is sufficient to use score R to forecast Y, protected attribute A is useless (but R can still be a proxy of A)
- In the binary case, sufficiency implies equity of positive/negative predictive values in all groups

Formalizing fairness: some considerations

Some considerations:

- Many -often mutually exclusive- notions of fairness (70+)

Formalizing fairness: some considerations

Some considerations:

- Many -often mutually exclusive- notions of fairness (70+)
- As a consequence, adjusting data for bias is not enough

Formalizing fairness: some considerations

Some considerations:

- Many -often mutually exclusive- notions of fairness (70+)
- As a consequence, adjusting data for bias is not enough
- Remediations are not a matter of purely technical fixes

Final considerations

1. Problems solved
2. Open challenges now
3. Future horizons

Problems solved

Problem solved:

Problems solved

Problem solved: lack of awareness

Problems solved

Problem solved: ~~lack of awareness~~

- New **scientific communities** gathered around the topic

Examples

- ACM Conference on Fairness, Accountability, and Transparency
- FAT ML (Fairness Accountability and Transparency in Machine Learning) community

Problems solved

Problem solved: ~~lack of awareness~~

- New scientific communities gathered around the topic
- **Institutions** have understood the relevance of the problem

Examples

- High-Level Expert Group on Artificial Intelligence to ethical, legal and societal issues related to AI - European Commission (June 2018; guidelines in April 2019)
- Algorithmic Accountability Bill of 2017 (New York City Council)
- Algorithmic Accountability Act of 2019 (Senate US)
- National Artificial Intelligence strategies

Problems solved

Problem solved: ~~lack of awareness~~

- New scientific communities gathered around the topic
- Institutions have understood the relevance of the problem
- Major **software companies** are developing tools/methodologies to mitigate the problem

Examples

- the IBM AI Fairness 360 Open Source Toolkit
- Microsoft research unit "The Rise of Autonomous Experimentation: Technical, Social, and Ethical Implications of AI"
- The What-IF tool by Google

Problems solved

Problem solved: ~~lack of awareness~~

- New scientific communities gathered around the topic
- Institutions have understood the relevance of the problem
- Major software companies are developing tools/methodologies to mitigate the problem
- New **funding, research labs, classes** have been established on the topic and existing ones have shifted priority towards the issue

Examples

- Ethics and Governance of Artificial Intelligence Fund (Harvard and MIT)
- Responsible Computer Science Challenges
- Diffusion of Tech Ethics courses
- AI NOW Institute (founded in 2017)
- The Alan Turing Institute (founded in 2015 and added AI implications in 2017)

Open challenges now

Open challenges now:

Open challenges now

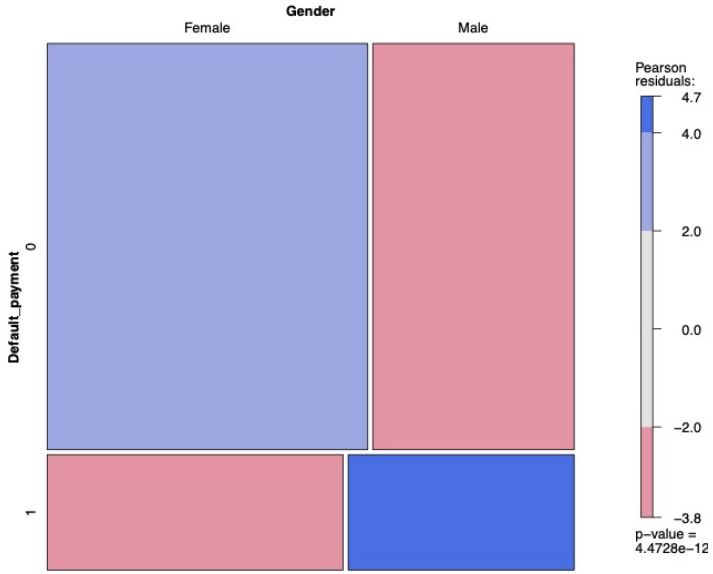
Open challenges now:
Research approaches at Nexa Center and at FULL

- Visualizing bias and data quality
- Mapping inequalities
- The political character of algorithms

Open challenges now

Open challenges now:
Research approaches at Nexa Center and at FULL

- **Visualizing bias and data quality**
- Mapping inequalities
- The political character of algorithms

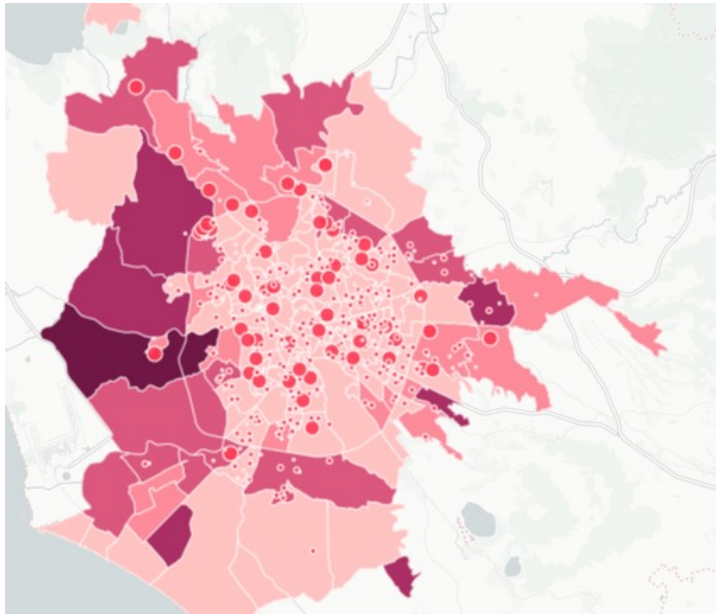


https://doi.org/10.1007/978-3-030-11680-4_30

Open challenges now

Open challenges now:
Research approaches at Nexa Center and at FULL

- Visualizing bias and data quality
- **Mapping inequalities**
- The political character of algorithms



Adapted from Open Polis : https://bit.ly/2XoqYV6

Open challenges now

Open challenges now:
Research approaches at Nexa Center and at FULL

- Visualizing bias and data quality
- Mapping inequalities
- **The political character of algorithms**

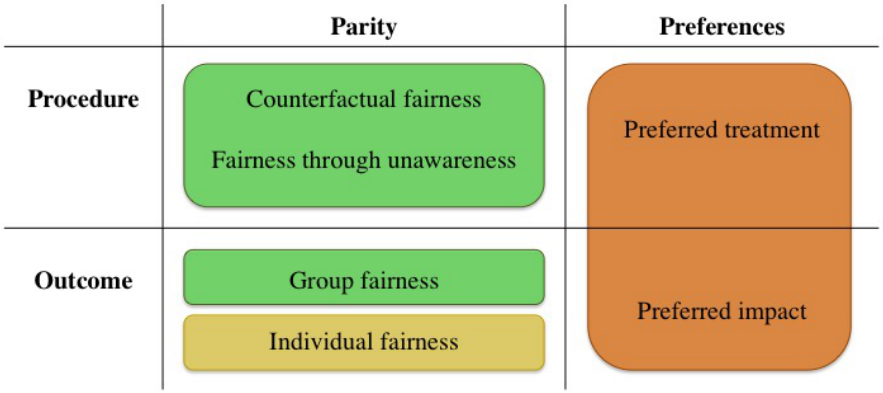


Fig. Trade-off in fairness selection process related to democracy typologies. Legend: Competitive (green), Liberal (yellow), Egalitarian (orange)
In collaboration with Fondazione Bruno Kessler - Work in progress

Future horizons

Future horizons

Future horizons

Future horizons

The **scale** and **impact** of automated decisions in societies pushes us to understand better *which* decisions to delegate to technologies and *to what extent*

Future horizons

Future horizons

The **scale** and **impact** of automated decisions in societies pushes us to understand better *which* decisions to delegate to technologies and *to what extent*

We the engineers need help from **human and social scientists**

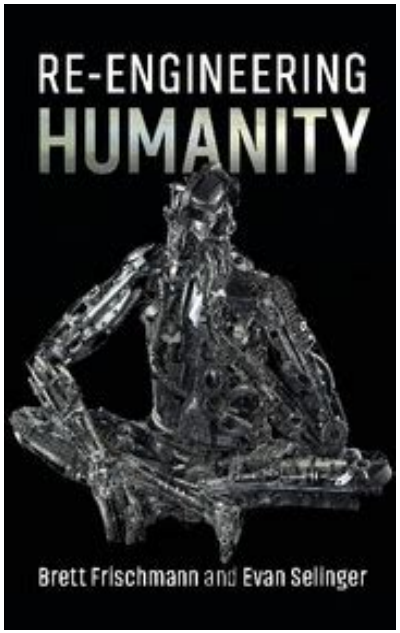
Future horizons

Future horizons

The **scale** and **impact** of automated decisions in societies pushes us to understand better *which* decisions to delegate to technologies and *to what extent*

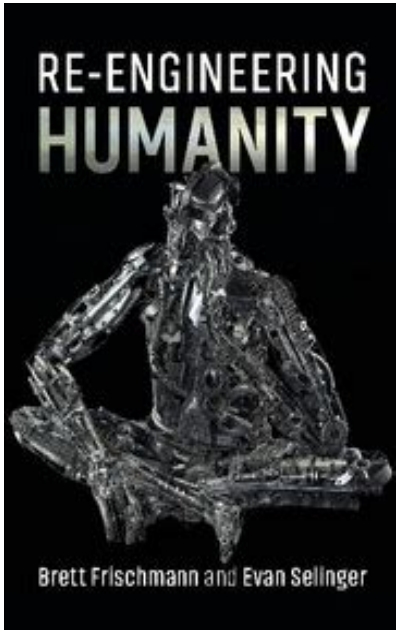
We the engineers need help from **human and social scientists** and viceversa

Final note



Modern techno-social dilemmas like the tragedy of commons, or climate change are different from techno-social engineering: the issue at hand is the fact that in the latter humans are seen by the technocratic forces as programmable objects.
But we too have a role, when we delegate to technology.

Final note



Modern techno-social dilemmas like the tragedy of commons, or climate change are different from techno-social engineering: the issue at hand is the fact that in the latter humans are seen by the technocratic forces as programmable objects.
But we too have a role, when we delegate to technology.

Thank you.

mail: antonio.vetro@polito.it

twitter: @phisaz