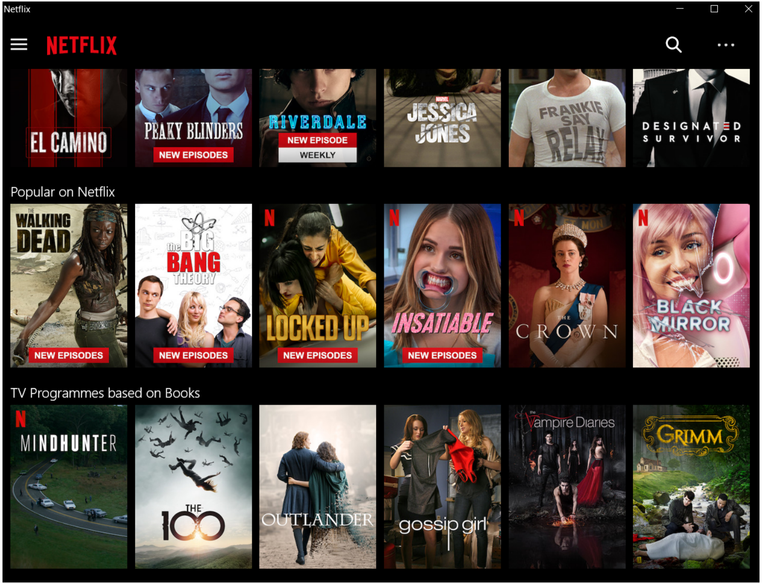
Fair, Transparent and Accountable Data Science

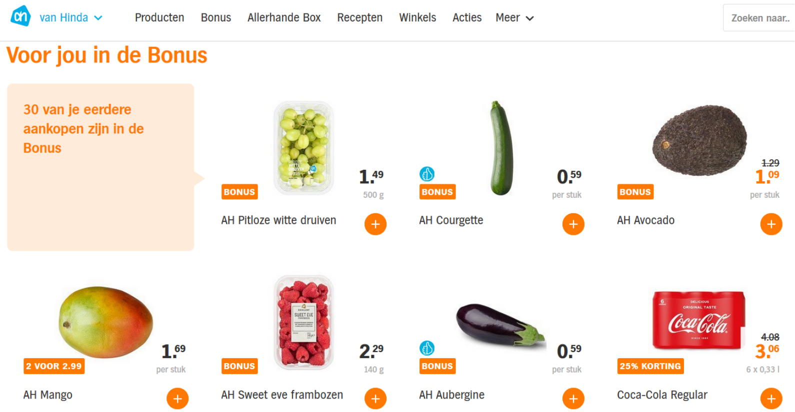
Hinda Haned

October 16th, 2019

h.haned@uva.nl

How does data science help us?


Netflix
NETFLIX
EL CAMINO
PEAKY BLINDERS
NEW EPISODES
RIVERDALE
NEW EPISODE
WEEKLY
JESSICA JONES
FRANKIE SAY
DESIGNATED SURVIVOR
Popular on Netflix
THE WALKING DEAD
NEW EPISODES
the BIG BANG THEORY
NEW EPISODES
LOCKED UP
NEW EPISODES
INSATIABLE
NEW EPISODES
THE CROWN
BLACK MIRROR
TV Programmes based on Books
MINDHUNTER
THE 100
OUTLANDER
gossip girl
the Vampire Diaries
GRIMM


van Hinda
Producten
Bonus
Allerhande Box
Recepten
Winkels
Acties
Meer
Zoeken naar..
Voor jou in de Bonus
30 van je eerdere aankopen zijn in de Bonus
BONUS
1.49
500 g
AH Pitloze witte druiven
BONUS
0.59
per stuk
AH Courgette
BONUS
1.29
1.09
per stuk
AH Avocado
2 VOOR 2.99
1.69
per stuk
AH Mango
BONUS
2.29
140 g
AH Sweet eve frambozen
BONUS
0.59
per stuk
AH Aubergine
25% KORTING
4.08
3.06
6 x 0,33 l
Coca-Cola Regular

The Curious Incident of the Dog in the Night-time

Winnaar van de Costa Book Award 2003

Auteur: Mark Haddon | Taal: Engels | 155 reviews | E-mail deze pagina

Inkijkexemplaar

Kies je bindwijze — Bekijk alle bindwijzen (10)

E-book	Luisterboek	Paperback	Pocket	Losbladig
€ 6,99	€ 11,62	€ 11,99	€ 8,99	€ 7,39

11,99

Op voorraad

Voor 23:59 besteld, morgen in huis

+ **Select** bezorgopties

Verkoop door bol.com

+ In winkelwagen | Op verlanglijstje

Andere verkopers (3)

Bekijk en vergelijk alle verkopers vanaf € 11,28

Vandaag nog in huis? op jouw locatie

Bezorging waar en wanneer het jou uitkomt vanaf € 1,29

Bekijk alle bezorgopties

PREMIUM

Jobs where you're a top applicant

You may have an edge over other candidates

Top 10% of 115 applicants

Data Scientist Oil & Gas (based in Abu Dhabi)
Executive Solutions
Amsterdam Area, Netherlands

3 days ago · Easy Apply

Top 10% of 36 applicants

Senior Data Scientist
Agoda
Amsterdam, NL

1 week ago

Top 10% of 42 applicants

Data Scientist
Robert Walters
Amsterdam, NL

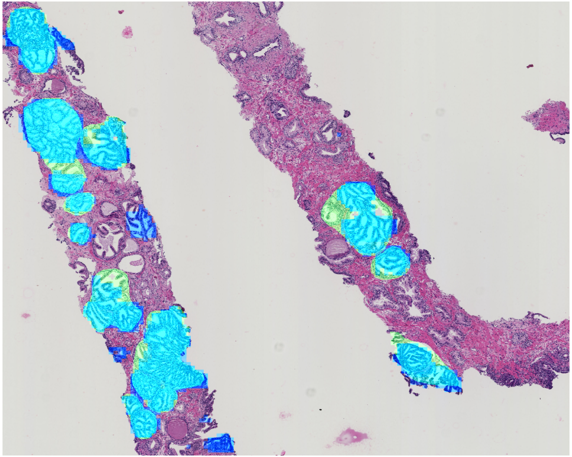
2 alumni

1 week ago

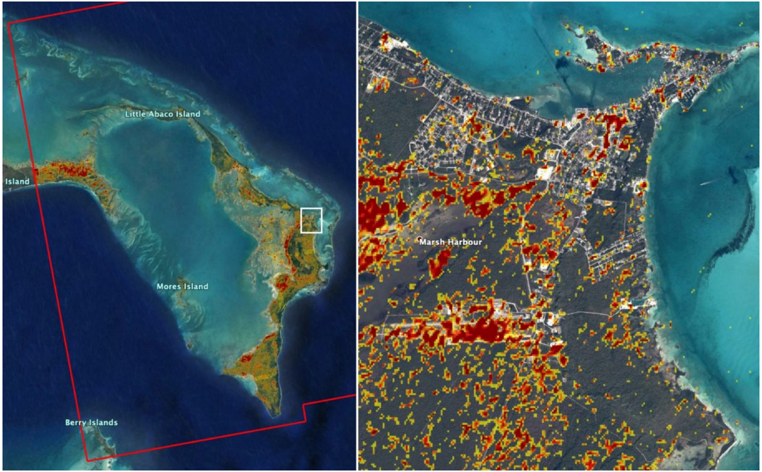
Top 10% of 322 applicants

Data Scientist
Osella Technology
Amsterdam Area, Netherlands

1 day ago · Easy Apply



Tumor growth pattern detection
Source: P. Ambrosini


Little Abaco Island
Island
Mores Island
Berry Islands
Marsh Harbour

Source: NASA

Can data science fail?

Anna Field ~~€ 69,99~~
Cocktailjurk - black € 55,95

Anna Field € 54,99
Galajurk - biking red

Anna Field ~~€ 69,99~~
Cocktailjurk - black € 55,95

MAMALICIOUS € 59,99
MLMIVANA - Cocktailjurk - navy blazer

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

THE WALL STREET JOURNAL.


Samsung Tries to Appease Investors But Delays Big Changes

Big Names Take Hit on Theranos


HP Enterprise Unveils Prototype of Next-Generation Computer

MARKETS
Zenefits Hit With $7 Million Fine by California Insurance Regulator


How Gets Ho Up to Sp Days

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI

December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

Recommended Videos

1. How Jill Stein's Election Recount Efforts Could Play Out

2. Great Barrier Reef Suffers Largest Die-off

The Guardian
@guardian

Digital dystopia: how algorithms punish the poor

Digital dystopia: how algorithms punish the poor

In an exclusive global series, the Guardian lays bare the revolution and the wreckage that is engulfing the welfare state worldwide

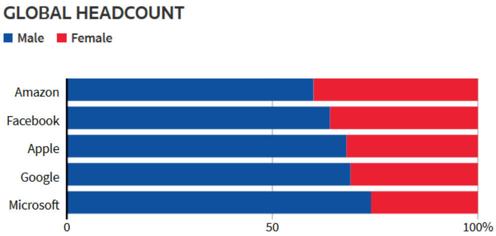
theguardian.com

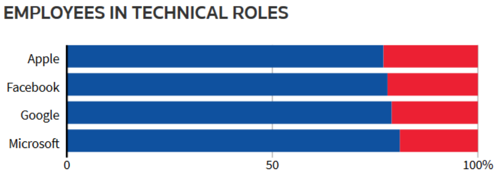
> "Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result algorithms can reinforce human prejudices."

C.C. Miller. When algorithms discriminate, NYT, 2019.

What is bias?

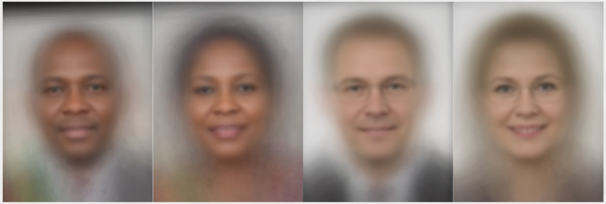
- Systematic errors that create unfair outcomes
- Sources: algorithm design, biased data collection or selection
- Algorithms learn and perpetuate bias





Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

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%



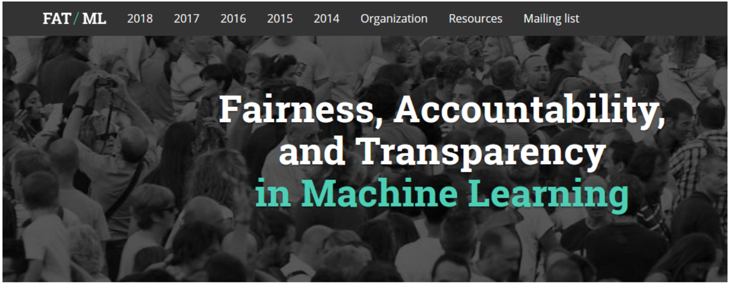
http://gendershades.org/overview.html

Complex systems raise concern

- Why this ad?
- Why this discount?
- Why this recommendation?

- Why was I rejected?
- Can I change the outcome?
- When will the system fail?

FATML/FAT* Field



Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

`https://www.fatml.org/`

Regulation: GDPR

"Data subjects have a right to **meaningful information** about the **logic involved** and to the significance and the **envisaged consequence** of automated decision-making"

Ethics


European Commission
INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION
AI
ETHICS GUIDELINES
FOR TRUSTWORTHY AI

Fair, Transparent and Accountable Data Science

Research questions

Transparency

How can we provide clear and actionable explanations?

Fairness

How do we avoid biased and unfair conclusions?

Accountability

How to evaluate potential harms and enable recourse?

Objective Develop algorithms that are: transparent, fair and actionable, while ensuring utility & performance

Approach Human-centric approach to understanding how users, stakeholders, regulators, data scientists experience a system and how the system impacts them

How can we provide clear and actionable explanations?

How do we avoid biased and unfair conclusions?

How to evaluate potential harms and enable recourse?

How can we provide clear and actionable explanations?

Example: explaining errors for user trust

Transparency through explainability

- Algorithm outputs must be understandable and transparent to the decision makers and the subjects impacted by them
- Explainability: is the extent to which the output can be explained to human subjects to enhance trust and enable feedback

Model verification

Compliance

User trust

Actionability

Example: predicting next week's sales

- Current model
  - auto-regressors
  - transaction history

- New model(s)
  - ensemble learning
  - 40+ features

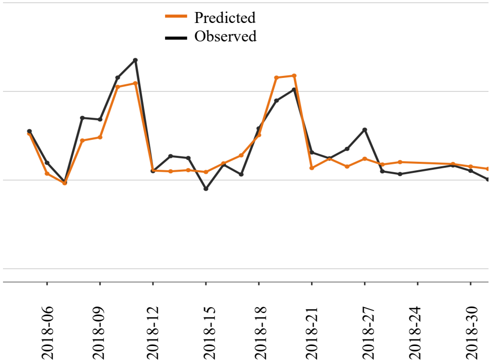
Example: predicting next week's sales

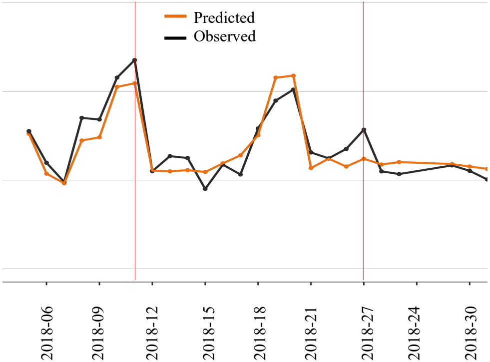
- Current model
  - auto-regressors
  - transaction history

- New model(s)
  - ensemble learning
  - 40+ features

User feedback

- Model perceived as a black-box
- Counter-intuitive results
- Gain in performance vs. loss in interpretability


Predicted
Observed
2018-06
2018-09
2018-12
2018-15
2018-18
2018-21
2018-27
2018-24
2018-30



How can I improve the forecast?

Data Scientist

Can I trust this forecast?

Stakeholder

When does the forecast fail?

End-user

How can we explain the errors of a forecasting model?

A. Lucic, H. Haned, M. de Rijke. Contrastive explanations for large errors in retail. IJCAI, Explainable AI workshop 2019.

What is a good explanation?

"The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case."

"Why A and not B?"

D. J. Hilton. Conversational processes and causal explanation, Psychological Bulletin, 1990.

- **MC-BRP** Monte Carlo Bounds for Reasonable Predictions
- Identifying unusual properties of a particular observation – we assume large errors occur due to unusual features in the test set that are not present in the training set
- Given an erroneous prediction, MC-BRP generates:
  1. Feature values that would result in a reasonable prediction, based on the n most important features
  2. General trends between each feature and the target variable

A. Lucic, H. Haned, M. de Rijke. Contrastive explanations for large errors in retail. IJCAI, Explainable AI workshop 2019.

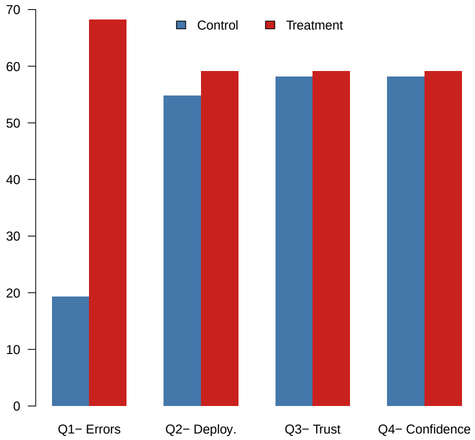
Contrastive explanations for large forecasting errors

Input	Trend	Value	Reasonable range
A	As input increases, sales increase	9628.00	[4140,6565]
B	As input increases, sales increase	18160.67	[8290,15322]
C	As input increases, sales increase	97332.00	[51219,75600]
D	As input increases, sales decrease	226.00	[95,153]
E	As input increases, sales decrease	2013.60	[972,1725]

Contrastive explanations for large errors

We ask our users the following subjective questions:

- **Q1:** I understand why the model makes large errors in predictions
- **Q2:** I would support using this model as a forecasting tool
- **Q3:** I trust this model
- **Q4:** In my opinion this model produces mostly reasonable outputs



Lessons learned

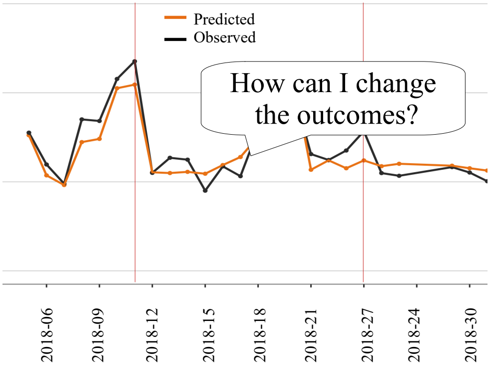
- Explanations generated by our method help users understand why models make large errors
- Explanations do not have a significant impact on support in deploying the model, trust in the model, or perceptions of the model's performance

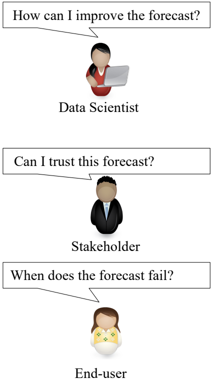
Algorithmic aversion

"We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake"

Dietvorst et al. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology, 2015.

Explanations are not enough





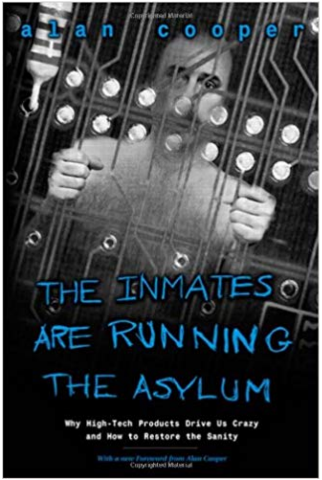
Explanations are not enough

A counterfactual describes the smallest required change to a feature value that changes the prediction to a predefined desired output

- **Model** forecast for next week is 5,000
- **Question** Which feature values must be changed to decrease the forecast to 4,000?
- **Counterfactual** If your delivery on the weekend is no longer free, you will decrease the forecast to below 4,000 transactions

Wachter et al. Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law & Technology, 2018.

"Most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users"



T. Miller et al. Beware of inmates running the Asylum, IJCAI Workshop on explainable AI, 2017.

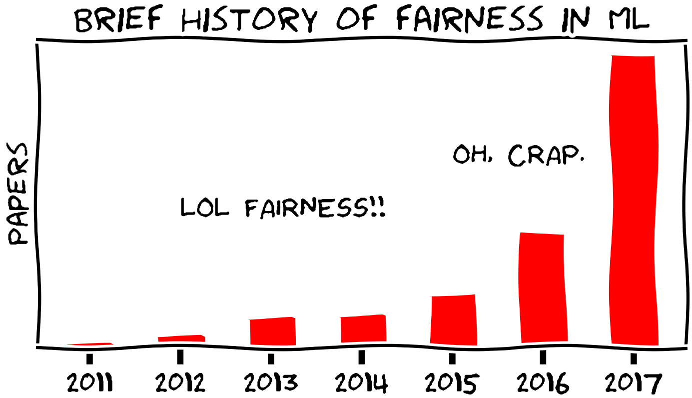
How can we provide clear and actionable explanations?

How do we avoid biased and unfair conclusions?

How to evaluate potential harms and enable recourse?

How do we avoid biased and unfair conclusions?

Example: building fair models



Source: M. Hardt

Fairness

- Fairness is concerned with how outcomes are assigned to particular groups of individuals
- Core principle: avoid bias even if it is supported by data, as to avoid the perpetuation of existing discrimination
- Fairness is a political construct: someone decides

Fairness: avoid harm

- **Harm of allocation** when a system allocates or withholds certain groups, an opportunity or a resource. Economically oriented view: e.g. who gets a discount, who gets hired, who gets assistance
- **Harm of representation** when a system reinforces the subordination of certain groups along the lines of identity like ethnicity, class, gender, etc

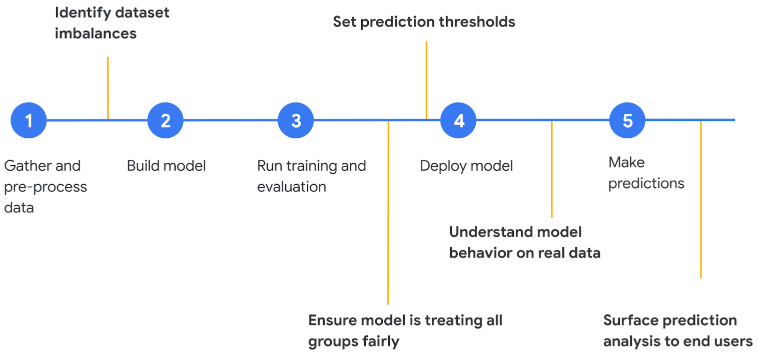
Kate Crawford's NIPS 2017 Keynote presentation: The trouble with Bias.

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

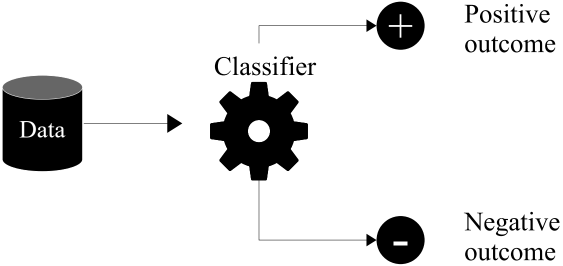


Source: `https://ai.google/`

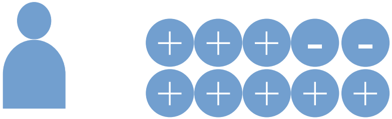
Practical limitations

- Sensitive attributes unknown
- Regulation constraints
- Stakeholders goals vs. fairness goals

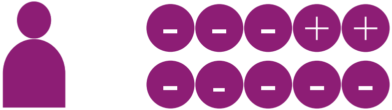
Harm of allocation



Two groups with different outcome distributions

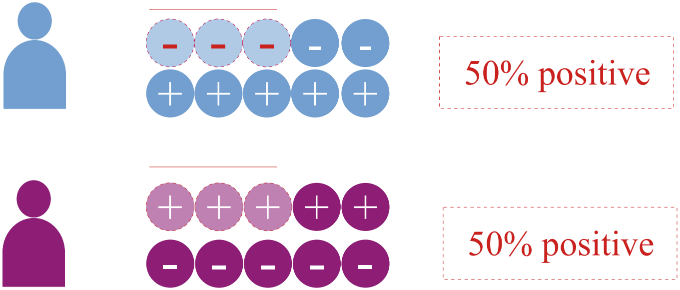


80% positive



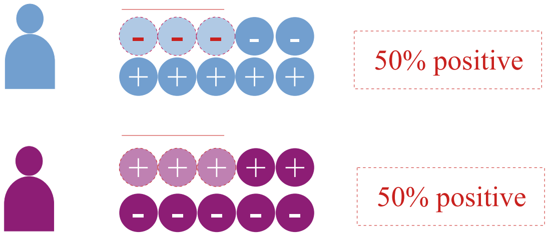
20% positive

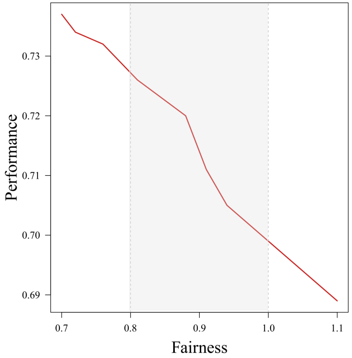
Fairness intervention



Statistical parity: subjects in protected and unprotected groups have equal probability of being assigned to the positive prediction class

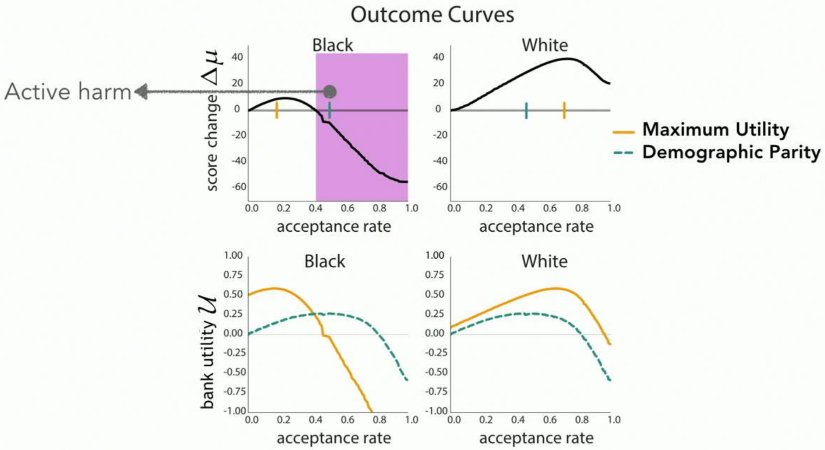
What is the cost of this intervention?





Evaluation is hard

- Sensitive attributes are unknown
- Realised outcomes are unavailable
- Fairness intervention impact is not monitored over time



Liu et al. Delayed impact of fair machine learning, ICML 2017.

Fairness and mitigation toolkits

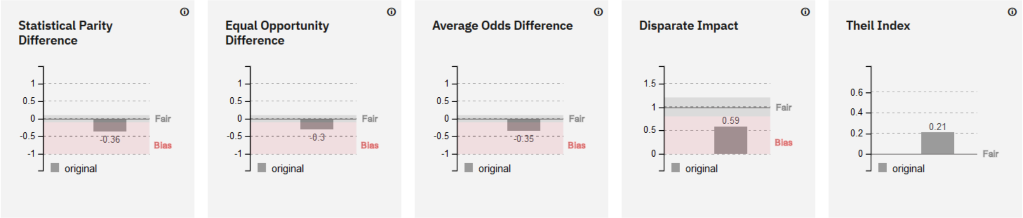
Check bias metrics

Protected Attribute: Sex

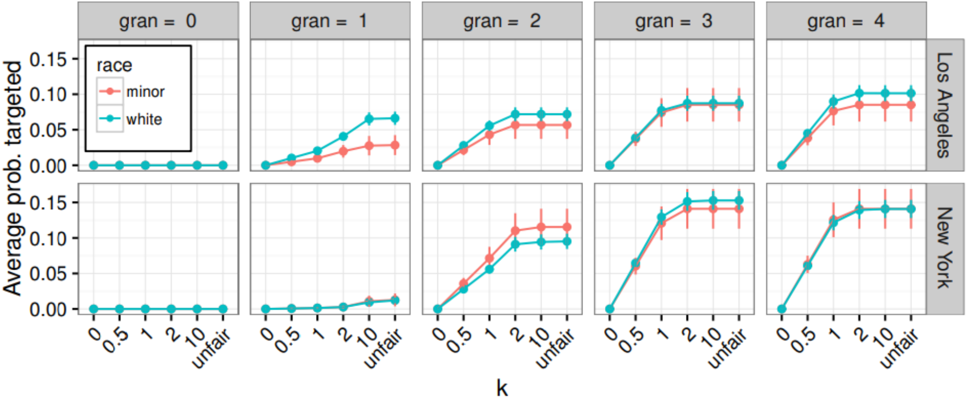
Privileged Group: ***Female***, Unprivileged Group: ***Male***

Accuracy with no mitigation applied is 66%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics



https://aif360.mybluemix.net/



Ridere et al. The price of fairness in location based advertising. FATREC 2017.

> "Any real machine-learning system seeks to make some change in the world. To understand its effects, then, we have to consider it in the context of the larger socio-technical system in which it is embedded."

Barocas et al. Fairness and machine learning, fairmlbook.org, 2019.

Fair machine learning algorithms: what do practitioners (really) need?

Dr. Aysenur Bilgin (CWI)

Dr. Fatih Turkmen (RuG)

How can we provide clear and actionable explanations?

How do we avoid biased and unfair conclusions?

How to evaluate potential harms and enable recourse?

How to evaluate potential harms and enable recourse?

The way forward

Utility
Performance
Feasibility

Transparency
Fairness
Accountability

Diverse
Domains
e.g. aviation, justice

Diverse
Users
e.g. pilots, judges

Diverse
Criteria
e.g. fairness, privacy

Interdisciplinary Research
social science – mathematics – computer science – law – ethics

Aircraft safety

Adopt AI systems while ensuring transparency to stakeholders throughout the algorithmic pipeline.

Prof. Dr. Leon Gommans

Forensic evidence evaluation

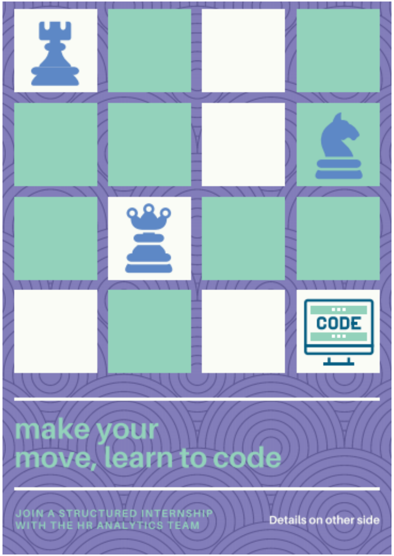
Leveraging more performant models while ensuring transparency through explanations.

Dr. Corina Benschop

Socially aware data science
Empower and connect citizens and communities in a fair and inclusive manner, including those at the margins of society.

Dr. Sennay Ghebreab

Education & Outreach


CODE
make your
move, learn to code
JOIN A STRUCTURED INTERNSHIP
WITH THE HR ANALYTICS TEAM
Details on other side

TRACK #1

A GLIMPSE INTO THE WORLD OF AI

Thank you

Fair, Transparent, and Accountable Data Science

> "Given the limited downside of just one group of people trying to do something different in just one place for a limited time, and the considerable upside if they succeed, my vote is that it is worth the risk."

D.J. Watts. Should social science be more solution-oriented? Nature Human Behaviour, 2017