



# Fair & Explainable Machine Learning

## Advances in Information Retrieval

---

**Hinda Haned**

October 7th, 2019

[h.haned@uva.nl](mailto:h.haned@uva.nl)

# About me

2015

Lead Data  
Scientist



Ahold  
Delhaize



2018

Chair  
Data Science  
Special appointment

ILPS



Best practices for safe and  
responsible applications of  
machine learning

# Lecture goals


This lecture will cover important aspects of both topics, with some practical examples

- Part 1: explanations
- Part 2: fairness
- Q&A


## Background & Motivation

---


# Ubiquity of ML

 Spotify


Albums for You




Musas (Un Homenaje ...)



Postales

 van Hinda




**BONUS**

AH Aubergine

0.59  
per stuk

+







**25% KORTING**


Coca-Cola Regular


4.08  
3.96  
6 x 0,33 l


+


 Restaurants   


[Back to results](#)


 Spuistraat 294-302, 1012 VX Amsterdam


 9V9Q+P8 Amsterdam


 vijfsvlieghen.nl


 020 530 4060

 **Closing soon:** 6–10pm ▾

 Add a label

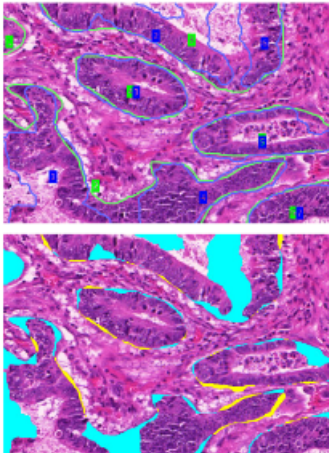
 [Suggest an edit](#)

**Popular times** Mondays ▾ 

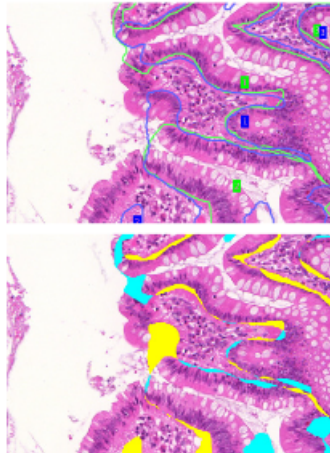


6a 9a 12p 3p 6p 9p

## Classification tumor glands



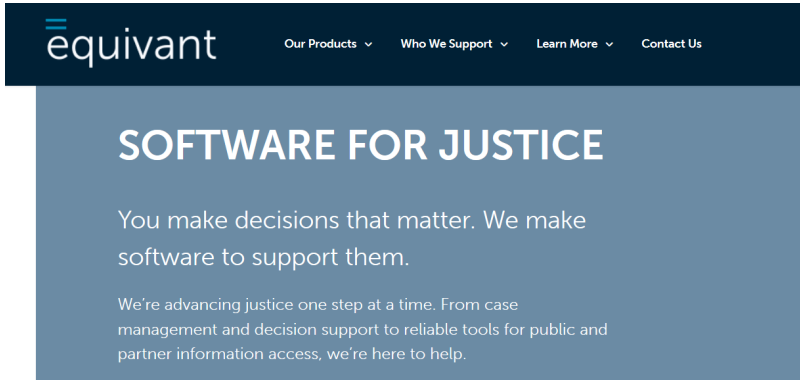
(e) malignant



(f) benign

Kainz et al, 2017

## Sentencing software (USA)



The image is a screenshot of the Equivant website. The top navigation bar is dark blue with the Equivant logo on the left and four menu items: 'Our Products', 'Who We Support', 'Learn More', and 'Contact Us'. The main content area has a light blue background. It features the heading 'SOFTWARE FOR JUSTICE' in large, bold, white capital letters. Below the heading is a paragraph: 'You make decisions that matter. We make software to support them.' At the bottom of the section is another paragraph: 'We're advancing justice one step at a time. From case management and decision support to reliable tools for public and partner information access, we're here to help.'

**equivant** Our Products ▾ Who We Support ▾ Learn More ▾ Contact Us

## SOFTWARE FOR JUSTICE

You make decisions that matter. We make software to support them.

We're advancing justice one step at a time. From case management and decision support to reliable tools for public and partner information access, we're here to help.

# Controversial “AI stories”

2,780 views | Jan 28, 2019, 07:20pm

**Forbes**

## Amazon Refuses To Quit Selling 'Flawed' And 'Racially Biased' Facial Recognition

Zak Doffman Contributor @

## Facebook's ad-serving algorithm discriminates by gender and race

Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.

by Karen Hao

Apr 5, 2019

**BBC**

Sign in

News

Sport

Reel

Worklife

Travel

Future

M

**NEWS**

Home

Video

World

UK

Business

Tech

Science

Stories

Entertainment & Arts

## Amazon scrapped 'sexist AI' tool

© 10 October 2018

Share

- Increased public awareness
- Risk for business and society



Compliance with General Data Protection Regulation (GDPR – 2018) – articles 13-15

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



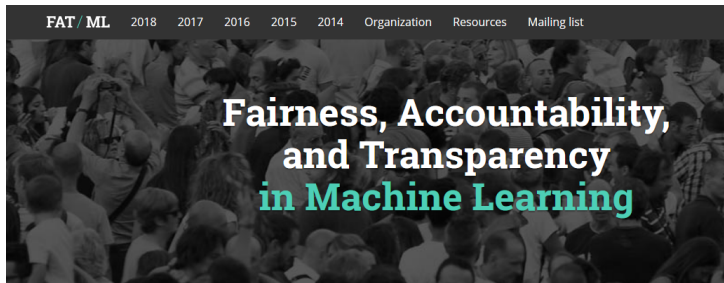
[European Commission](#) > [Strategy](#) > [Digital Single Market](#) > [Reports and studies](#) >

Digital Single Market

REPORT / STUDY | 8 April 2019

## Ethics guidelines for trustworthy AI

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.



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

- <https://www.fatml.org/>
- <https://fatconference.org/>

- **FACT** framework: Fair, Accurate, Confidential and Transparent
- Dutch universities consortium: <https://redasci.org/>

<b>Fair</b>	How to avoid unfair conclusions even if they are true?
<b>Accurate</b>	How to answer questions with a guaranteed level of accuracy?
<b>Confidential</b>	How to answer questions without revealing secrets?
<b>Transparent</b>	How to clarify answers such that they become indisputable?

- **FACT** framework: Fair, Accurate, Confidential and Transparent
- Dutch universities consortium: <https://redasci.org/>

<b>F</b> air	How to avoid unfair conclusions even if they are true?
<b>A</b> ccurate	How to answer questions with a guaranteed level of accuracy?
<b>C</b> onfidential	How to answer questions without revealing secrets?
<b>T</b> ransparent	How to clarify answers such that they become indisputable?

- **Explainability**: algorithm outputs must be understandable and transparent to the decision makers and the subjects impacted by them,
- **Fairness**: algorithms must be designed to avoid inaccurate and unfair outcomes against a given group of individuals.

**Fair**

How to avoid unfair conclusions even if they are true?

**Transparent**

How to clarify answers such that they become indisputable?

# Part 1: Explainability

---

Why should we care about  
Explainability?



- Have some idea of what the system does
  - How do the personalized recommendations work?
- Able to challenge the system: actionable recourse
  - What can I do to change my application's outcome?
- Able to understand when the system makes errors
  - How do I trust the self-driving car?

Model verification

Compliance

User trust

Actionability

## **Approaches to explainability**

---

# Explainable vs. Interpretable

- Interpretability and Explainability often used interchangeably
- Interpretability is a goal
- Explainability **is one way** of achieving Interpretability

# Approaches to explainability

- Build (simple) interpretable models
  - e.g. linear models, simple decision-trees
- Build post-hoc explainer on top of a black-box model
  - Surrogate models for local explanations: feature attribution methods
  - Counterfactual explanations

Extensive survey: [Guidotti et al \(2018\)](#)

# Types of explanations

## Global explanations

Explain a models' decision making process in general.

Typically: feature importance

*Treeinterpreter, PDP, feature importance*

## Local explanations

Explain a single prediction. Since it remains challenging to establish fidelity to black box models in globally interpretable approximations, much attention is put on local explanations.

*LIME, SHAP, Skater*

# LIME: Locally Interpretable Model-Agnostic Explanations

- Feature attribution method: attribute model output to feature input
- Designed for classification tasks, but also works for regression
- Intuition: explains behavior of the black-box model by perturbing the input and learning how the output changes

Ribeiro et al, SIGKDD 2016

Explanation model for instance  $x$ : minimise loss AND model complexity

$$\xi(x) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

- $x$ : the instance being explained
- $f$  : target model which we want to interpret
- $g \in G$  : an interpretable model such as linear models, decision trees
- $\pi_x(z)$ : proximity measure between an instance  $z$  to  $x$
- $\Omega(g)$ : a measure of complexity of the explanation  $g$
- $\mathcal{L}$ : locality loss

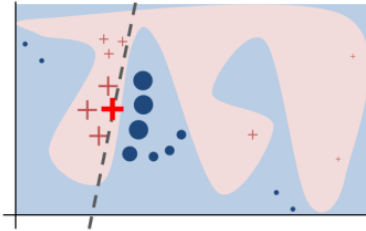


Minimise loss AND model complexity

$$\xi(x) = \arg \min_{\underbrace{g \in G}_{\text{family}}} \underbrace{\mathcal{L}(f, g, \pi_x)}_{\text{faithfulness}} + \underbrace{\Omega(g)}_{\text{complexity}}$$

- faithfulness: how close the explanation is to the prediction of the original model  $f$
- explanation family:  $G$  is the family of possible explanations, for example all possible linear regression models
- complexity: LIME only optimises loss part, the user has to determine the complexity

# Local explanations



Decision function  $f$  (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using  $f$ , and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

- Create a new dataset by perturbing instances around  $x$
- Weigh samples according to distance to  $x$  (neighborhoods)
- Get the prediction's the new perturbed data set
- Train a weighted, interpretable model on the perturbed dataset
- Explain the prediction by interpreting the local model

# Lime on image data

Superpixels explanations for the top 3 predicted classes.

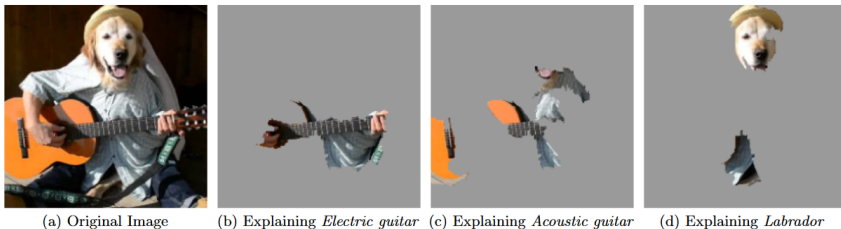


Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ( $p = 0.32$ ), "Acoustic guitar" ( $p = 0.24$ ) and "Labrador" ( $p = 0.21$ )

What the neural network picks up on for each of the classes is quite natural to humans. Figure 4b in particular provides insight as to why acoustic guitar was predicted to be electric: due to the fretboard.

Ribeiro et al, SIGKDD 2016

### Advantages

- human-friendly explanations
- works for tabular data
- viable python package easy to use

### Disadvantages

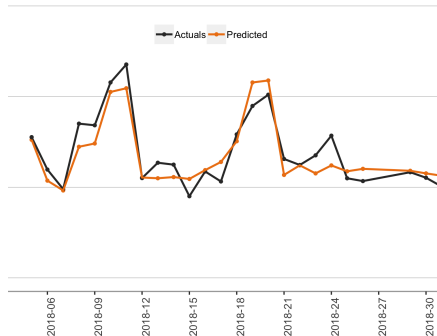
- local model built on large neighborhoods
- feature correlation ignored
- instability of explanations

# Illustration

---

- Historical transactions
- Special events
- Promotions,
- Weather
- In total: 94 features

## LSTM model



Example: predicting next week's sales

- Current model

- transaction history
- auto-regressors

- New model

- LSTM
- 90+ features

Example: predicting next week's sales

- Current model

- transaction history
- auto-regressors

- New model

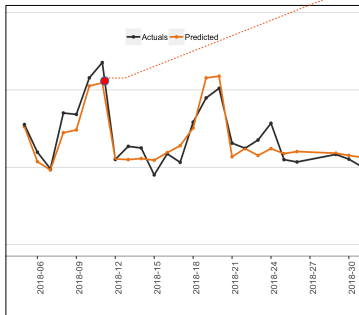
- LSTM
- 90+ features

## User feedback

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



# Explaining forecasts



Give me an explanation!

Data Scientist



How can I improve the forecast?

Stakeholder



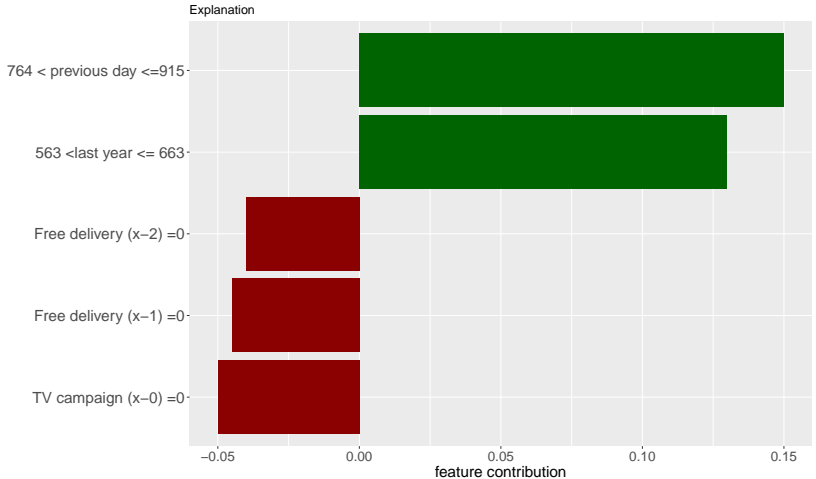
Can I trust this forecast?

End-user



When does the forecast fail?

# Explaining forecasts

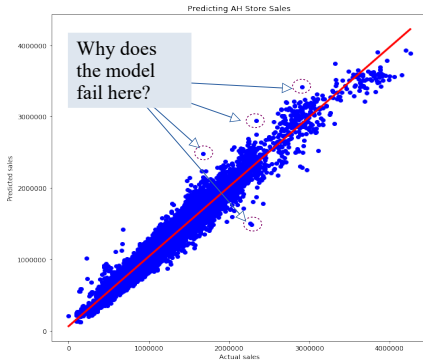


- Feature attribution helps understand forecasts
- Not clear if it increases trust (see what follows!)
- LIME was modified to account for times series data
- Work in progress!

## Contextualised explanations

---

# Explain large forecasting errors



Give me an explanation!

Data Scientist



How can I improve the forecast?

Stakeholder



Can I trust this forecast?

End-user



When does the forecast fail?

- **MC-BRP:** Monte Carlo Bounds for Reasonable Predictions.
- Intuitively, MC-BRP is based on 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 XAI workshop 2019.

# Contrastive explanations for large forecasting errors

**RQ** How does providing contrastive explanations for large errors impact users' perception of the model?

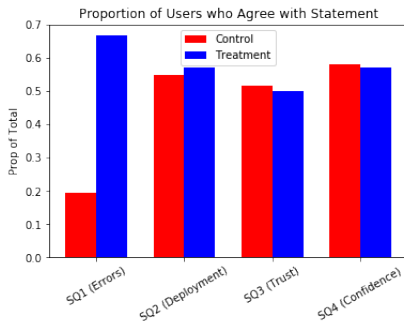
Input	Trend	Value	Reasonable range
A	As input increases, sales increase	9628.00	[4140,6565]
B	As input increases, sales increase	18160.67	[8290,15322]
C	As input increases, sales increase	97332.00	[51219,75600]
D	As input increases, sales increase	226.00	[95,153]
E	As input increases, sales increase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

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

# Contrastive explanations for large errors

We ask our users the following subjective questions:

- **SQ1:** I understand why the model makes large errors in predictions.
- **SQ2:** I would support using this model as a forecasting tool.
- **SQ3:** I trust this model.
- **SQ4:** In my opinion this model produces mostly reasonable outputs.



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



We find that explanations generated by our method help users understand why models make large errors in predictions, but do not have a significant impact on support in deploying the model, trust in the model, or perceptions of the model's performance.

“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. 2015

## More on available XAI methods

<https://christophm.github.io/interpretable-ml-book>

## Part 2: Fairness

---

- Fairness is linked to discrimination
- Discrimination: treating someone differently because of their group membership, not individual merit or other characteristics
- Group membership: gender, age, ethnicity, location, topical interests, social-economic status

Why should we care about Fairness in machine learning?

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

# Types of harm

- **Harm of allocation** when a system allocates or withholds certain groups, an opportunity or resource. Economically oriented view (e.g. who gets a discount, who gets hired)
- **Harms of representation** systems reinforce the subordination of certain groups along the lines of identity like race, class, gender etc.

Kate Crawford's NIPS 2017 Keynote presentation: Trouble with Bias



# Harm of allocation

## THE WALL STREET JOURNAL.

Subscribe Now | Sign In  
**SPECIAL OFFER: JOIN NOW**

Home World U.S. Politics Economy Business **Tech** Markets Opinion Arts Life Real Estate

Search

**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

**Holt Ritz Carlton** Gets Hold Up to 30 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




Lower prices offered to buyers who live in more affluent neighborhoods

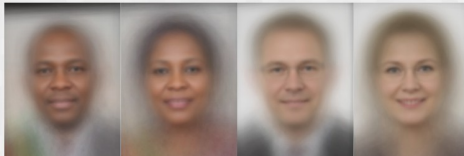
# Harm of representation

<http://gendershades.org/overview.html>

When we analyze the results by intersectional subgroups - darker males, darker females, lighter males, lighter females - we see that all companies perform worst on darker females.

IBM and Microsoft perform best on lighter males. Face++ performs best on darker males.

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



$$f(\text{trash can}) = \text{trash can}$$

# Sources of unfairness



- Data collection: skewed samples, unreliable over certain groups
  - Diversity of sources
  - Protected attributes
- redundantly encoded in observables (e.g. zipcode)
  - Algorithms trained on biased data yield biased results that may build up over time (Selbst 2016)

- Task binary classification
- Outcomes binary (hire/don't hire, loan/no loan, discount/no discount)
- Sensitive attribute known and protected
- Fairness goals
  - Group fairness: subjects in protected and unprotected groups have equal probability of being assigned to to the positive prediction class
  - Individual fairness: similar people experience similar outcomes

# Testing fairness

More than 20 definitions for fairness testing

- Defined based on predicted outcomes (e.g. credit score)
- Defined on predicted and actual outcomes (e.g. sentencing)

- Group fairness or statistical parity
- Conditional statistical parity
- Predictive parity
- False positive error rate balance
- False negative error rate balance
- Equalised odds
- Conditional use accuracy equality
- Overall accuracy equality
- Treatment equality

Verma & Rubin, FairWare, 2018

## German credit score data: fairness testing

- Features: credit history, gender, zipcode, employment history, nationality,...
- Classifier: logistic regression
- Predicted category  $d$ : credit score (good/bad)
- Protected group  $G$ : subjects younger than 25 years old
- Group fairness: individuals in both protected and unprotected groups have equal probability of being assigned to the positive predicted class

$Pr(d = \text{positive} | G = \text{age} \geq 25) = 0.72$  and  $Pr(d = \text{positive} | G = \text{age} < 25) = 0.58$

Privileged group is getting 14% more positive outcomes in the training dataset

## Mitigation methods

---



## Methods for fair classification

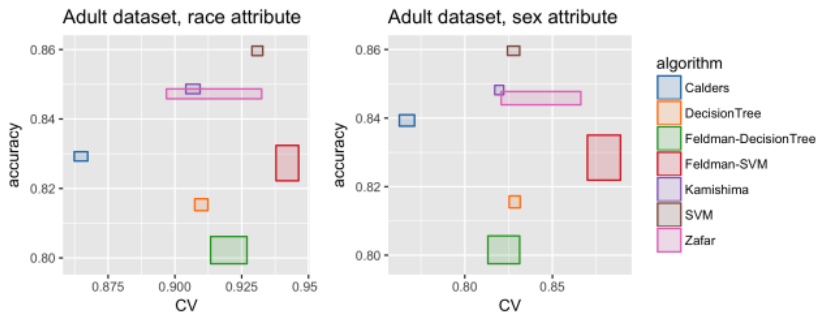
- Pre-processing: modify the train data
- In-processing: modify the algorithm's objective function to incorporate fairness constraints/penalty
- Post-processing: modifies the predictions produced by the algorithm

Pre-processing	Re-weighting (Kamiran & Calders, 2012)
	Optimized pre-processing (Calmon et al., 2017)
	Learning fair representations (Zemel et al., 2013)
	Disparate impact remover (Feldman et al., 2015)
In-processing	Adversarial debiasing (Zhang et al., 2018)
	Prejudice remover (Kamishima et al., 2012)
Post-processing	Equalized odds post-processing (Hardt et al., 2016)
	Calibrated eq. odds postprocessing (Pleiss et al., 2017)
	Reject option classification (Kamiran et al., 2012)

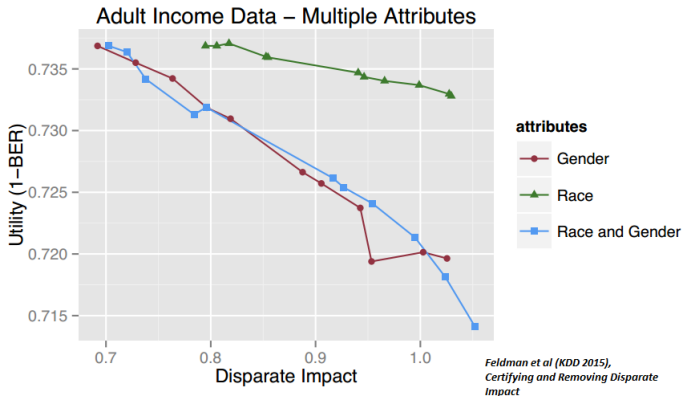
- For all methods, there is a trade-off to be found between utility and a desired measure of fairness
- Sensitive attributes are known
- Ground truth or observable outcomes are available
- Fairness-preserving algorithms tend to be sensitive to fluctuations in dataset composition, and to different forms of pre-processing (Friedler et al, 2019)

# Fairness algorithms: no easy fix

Fairness-preserving algorithms tend to be sensitive to fluctuations in dataset composition, and to different forms of pre-processing (Friedler et al, 2019)

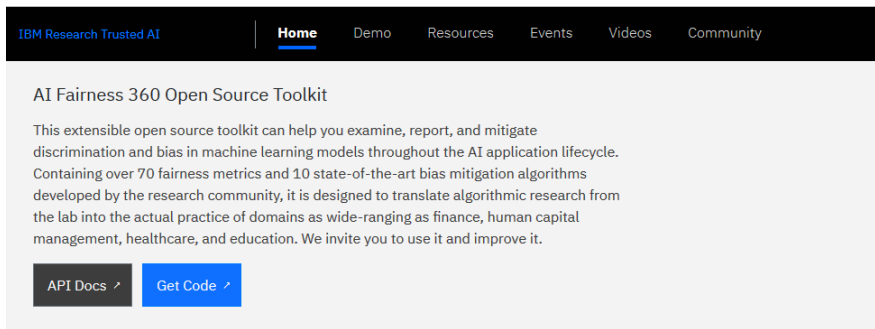


# Cost of fairness intervention example



Disparate impact is measured here by  $\frac{Pr(d = positive | G = protected)}{Pr(d = positive | G = unprotected)}$

# SOTA: experiment!



The screenshot shows the top navigation bar of the IBM Research Trusted AI website. The navigation links are: IBM Research Trusted AI, Home (underlined), Demo, Resources, Events, Videos, and Community. Below the navigation bar, the main heading is "AI Fairness 360 Open Source Toolkit". The text describes the toolkit as an extensible open source tool for examining, reporting, and mitigating discrimination and bias in machine learning models. It mentions over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms. At the bottom of the section, there are two buttons: "API Docs" and "Get Code".

IBM Research Trusted AI | Home Demo Resources Events Videos Community

## AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs ↗ Get Code ↗

<https://aif360.mybluemix.net/>

- Explainable ML: context- and user-dependent
- Fair ML: there is no easy algorithmic fix, problem to be scoped as a multidimensional problem
- Many open research questions
- Issues relevant for both industry and academic worlds!

[h.haned@uva.nl](mailto:h.haned@uva.nl)

## References

---



## Relevant references

- Dietvorst et al, 2015. [Algorithmic aversion](#)
- Miller 2018. [Insights from the social sciences](#)
- Ribeiro et al 2016. [LIME](#)
- Friedler et al 2019. [Fairness mitigation](#)
- Lucic et al, 2019. [Contextualised explanations](#)