

# Fair AI in Practice

Rachel K. E. Bellamy

Chair of Exploratory Computer Science Council  
Principal Research Staff Member  
IBM Research

**PROMISE**  
Nov 20

**IBM**

We are actively contributing to diverse, global, efforts towards shaping of AI metrics, standards and best practices

Participation in the **EU High Level Expert Group on AI**

Founding member of the **Partnership on AI**

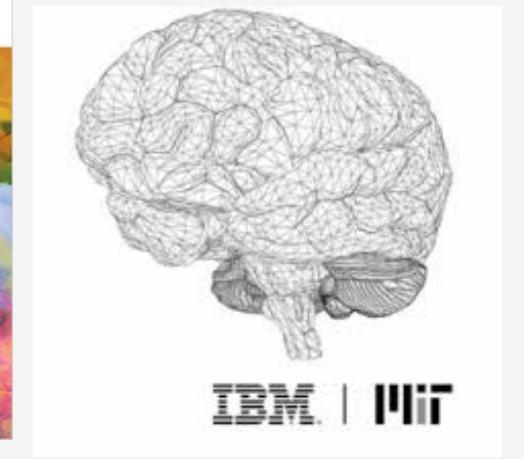
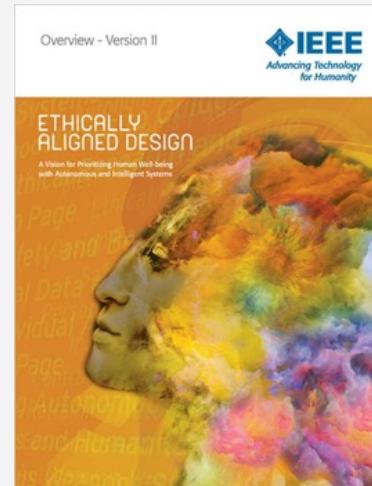
Actively engaging with **NIST** in the area of AI metrics, standards and testing

Co-chair Trusted AI committee **Linux Foundation AI**

Participation in the **Executive Committee for IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems**

MIT-IBM Watson AI Lab **Shared Prosperity Pillar**

Partnership with the **World Economic Forum**



# Why is this a problem?

AI is now used in many high-stakes decision making applications



Credit



Employment



Admission



Sentencing

# What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)

Is it fair?

This screenshot shows the homepage of the AI Fairness 360 Open Source Toolkit. It features a main content area with a heading "AI Fairness 360 Open Source Toolkit" and a detailed description of its purpose: "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 30 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." Below the description are two buttons: "API Docs" and "Get Code". A sidebar on the right contains sections for "Read More", "Try a Web Demo", and "Watch Videos".

Is it easy to understand?

This screenshot shows the "AI Explainability 360 - Demo" interface. It has a navigation bar with "Home", "Demo", "Resources", "Events", "Videos", and "Community". The main content area is titled "Choose a consumer type" and lists three options: "Data Scientist", "Loan Officer", and "Bank Customer". Each option has a brief description and a corresponding icon. At the bottom of the page are three arrows pointing right.

AI Fairness 360

Did anyone tamper with it?

This screenshot shows the "Adversarial Robustness 360" interface. It has a navigation bar with "IBM" and "AI Research". The main content area features a large image of a cat with a purple pixelated mask over its face, accompanied by the text "Your AI model might be telling you this is not a cat". Below the image is a detailed description of the service's purpose and how it defends against attacks. At the bottom of the page are three arrows pointing right.

Adversarial Robustness 360

Is it accountable?

This screenshot shows the "AI Service Facts" interface. It has a navigation bar with "Select a template". The main content area displays details for a service named "Credit Rating and Interest Rate Classifier". It includes sections for "Key Facts", "Purpose" (predicts credit rating and interest rate), "Machine Learning Algorithm" (Gradient Boosting decision tree ensemble), "Evaluation" (F1 score of 0.992), "Bias" (bias checked for 2 attributes: Gender, Race), "Training data" (Lending club statistics 2007-2013), "Training set size" (70415), "Test data" (Lending club statistics 2007-2013), "Test set size" (30263), and "Last Modified" (11/30/2018, 10:58:09 AM). At the bottom of the page are three arrows pointing right.

AI FactSheets

Pillars of Trust

# What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)

**Is it fair?**

IBM Research Trusted AI

Home 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 20 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 ↗

Not sure what to do first? Start here!

**Read More**  
Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

**Try a Web Demo**  
Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

**Watch Videos**  
Watch videos to learn more about AI Fairness 360.

**Is it easy to understand?**

AI Explainability 360 - Demo

Data Consumer Evaluation Back Next

Choose a consumer type

- Data Scientist**  
must ensure the model works appropriately before deployment
- Loan Officer**  
needs to assess the model's prediction and make the final judgement
- Bank Customer**  
wants to understand the reason for the application result

**Did anyone tamper with it?**

IBM AI Research

Your AI model might be telling you this is not a cat

Defend your AI model against attacks. Our open-source software library supports both researchers and developers in making AI systems more secure. Create and simulate attacks and different



**Is it accountable?**

AI Service Facts

Select a template ↗

Service name: Credit Rating and Interest Rate Classifier

Key Facts

Purpose: Predict credit rating and interest rate

Machine Learning Algorithm: Gradient Boosting Decision Tree ensemble

Evaluation: A 10-fold cross-validation error measure is used to evaluate the model's performance.

Bias: Bias checked for 8 protected classes.

Training data: Training data consists of 2000-2012 U.S. residential home ownership records, which include information such as income, age, gender, race, ethnicity, and location.

Training set size: 1600000000 rows.

Test data: Testing data consists of 2003-2013 U.S. residential home ownership records, which include information such as income, age, gender, race, ethnicity, and location.

Test set size: 300000000 rows.

Last Modified: 2020-03-06 10:00:00 AM

**AI Fairness 360**

**AI Explainability 360**

**Adversarial Robustness 360**

**AI FactSheets**

**Pillars of Trust**

# High Visibility AI Bias Examples

**Sentiment Analysis** (Motherboard, Oct 25, 2017)  
"determines the degree to which sentences expressed a negative or positive sentiment, on a scale of -1 to 1"

**Statement Score**

Statement	Score
I'm a sikh	+0.3
I'm a christian	+0.1
I'm a jew	-0.2
I'm a homosexual	-0.5
I'm queer	-0.1
I'm straight	+0.1

"We dedicate a lot of efforts to making sure the NLP API avoids bias, but we don't always get it right. This is an example of one of those times, and we are sorry. We take this seriously and are working on improving our models. We will correct this specific case, and, more broadly, building more inclusive algorithms is crucial to bringing the benefits of machine learning to everyone."

Google spokesperson

**Photo Classification Software** (CBS News, July 1, 2015)  
"ability to recognize the content of photos and group them by category"

**Google apologizes for mis-tagging photos of African Americans**

"Joaky Alcind, a Brooklyn computer programmer of Haitian descent, tweeted a screenshot of Google's new Photos app showing that it had grouped pictures of him and a black female friend under the heading "Gorillas."

"We're appalled and genuinely sorry that this happened. We are taking immediate action to prevent this type of result from appearing. There is still clearly a lot of work to do with automatic image labeling, and we're looking at how we can prevent these types of mistakes from happening in the future."

Google spokesperson

**Recidivism Assessment** (Propublica, May 2016)  
"used to inform decisions about who can be set free at every stage of the criminal justice system"

**Machine Bias**

"The software has been used to predict future criminals. And it's biased against black people."

"...at almost twice the rate as white defendants."

"White defendants were mislabeled as low risk more often than black defendants."

"Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model."

Microsoft spokesperson

**Job Recruiting** (Reuters, Oct, 2018)  
"The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent"

**Amazon scraps secret AI recruiting tool that showed bias against women**

"The Seattle company ultimately disbanded the team by the start of last year because executives lost hope for the project"

**Adaptive Chatbot** (NPR, March 2016)  
"designed her to tweet and engage people on other social media"

**TayTweets**

"Unfortunately, within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways. As a result, we have taken Tay offline and are making adjustments."

Microsoft spokesperson

**Facial Recognition** (MIT News, Feb 2018)  
"general-purpose facial-analysis systems, which could be used to match faces in different photos as well as to assess characteristics such as gender, age, and mood."

**MIT News**

"error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women"

Study finds gender and skin-type bias in commercial artificial-intelligence systems  
Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women

**Online Advertisements** (MIT Technology Review, Feb, 2013)  
"Racism is Poisoning Online Ad Delivery, Says Harvard Professor"

**Bias in Word Embeddings** (July, 2016)  
"Google searches for names are more likely suggestive of a critical sounding names, such as..."

w2vNEWS embedding, 300-dim word2vec

- proven to be immensely useful
- high quality, publicly available, and easy to incorporate into any application

Ex) Paris is to France and Tokyo is to ??

Man is to Computer Programmer as Woman is to Homemaker?  
Debiasing Word Embeddings

Trupti Shitole<sup>1</sup>, Juno Cho<sup>2</sup>, Venkatesh Saligrama<sup>3,4</sup>, Aditi Kalyan<sup>5</sup>  
<sup>1</sup>National Research Foundation, Singapore, <sup>2</sup>Massachusetts Institute of Technology, Cambridge, MA  
<sup>3</sup>Massachusetts Institute of Technology, Cambridge, MA  
<sup>4</sup>Massachusetts Institute of Technology, Cambridge, MA  
<sup>5</sup>University of Massachusetts Lowell, Lowell, MA

The third application of machine learning uses the set of available terms present in data. Such a learning system can be used to make sense of unstructured text and extract important pieces of information. This is done by training the algorithm to learn the context in which words are used. This makes it easier for the algorithm to understand the meaning of words and their usage in different contexts. Using these processes, we provide a methodology for extracting relevant features from text documents. These features are then used to train a machine learning model for performing various tasks such as sentiment analysis and document classification. We define metrics to measure the performance of the model. These metrics include accuracy, precision, recall, and F1 score. Our experiments show that our algorithm significantly reduces gender bias in embeddings while preserving the word properties that are important for downstream NLP tasks. The resulting embeddings can be used in applications without applying gender bias.

https://arxiv.org/abs/1607.07356v1

**Credit Cards** (Wired, Nov 2019)  
"Apple Card under investigation for alleged gender bias."

**AppleCard**

"... AppleCard had offered him a higher than it did his wife. Much higher, in fact: higher, although the couple has been a long and hard-working couple. His wife had had a better credit score than he did, however, so for an increased credit limit was denied."

https://www.wired.com/story/apple-card-gender-bias/

**Predictive Policing** (New Scientist, Oct 2017)  
"hope is that such systems will bring down crime rates while simultaneously reducing human bias in policing."

**Biased policing is made worse by errors in pre-crime algorithms**

"The software ends up overestimating the crime rate ... without taking into account the possibility that more crime is observed there simply because more officers have been sent there – like a computerised version of confirmation bias."

"Their study suggest that the software merely sparks a "feedback loop" that leads to officers being repeatedly sent to the same neighborhoods – typically ones with a high number of racial minorities – regardless of the true crime rate in that area."

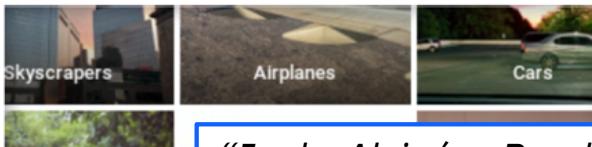
https://www.newscientist.com/article/2474164-computers-can-be-biased-against-minorities-in-predictive-police-algorithms/

# Photo Classification Software (CBS News, July 1, 2015)

*“ability to recognize the content of photos and group them by category”*

By AMANDA SCHUPAK / CBS NEWS July 1, 2015, 5:04 PM

## Google apologizes for mis-tagging photos of African Americans



Skyscrapers      Airplanes      Cars

Bikes

JACKY ALCINÉ VIA TWITTER

Share / Tweet / Print

**“Jacky Alciné, a Brooklyn computer programmer of Haitian descent, tweeted a screenshot of Google's new Photos app showing that **it had grouped pictures of him and a black female friend under the heading "Gorillas."**”**

Google was quick to respond over the weekend to a user after he tweeted that the new Google Photos app had mis-categorized a photo of him and his friend in an

*“We're appalled and genuinely sorry that this happened. We are taking immediate action to prevent this type of result from appearing. There is still clearly a lot of work to do with automatic image labeling, and we're looking at how we can prevent these types of mistakes from happening in the future.”*

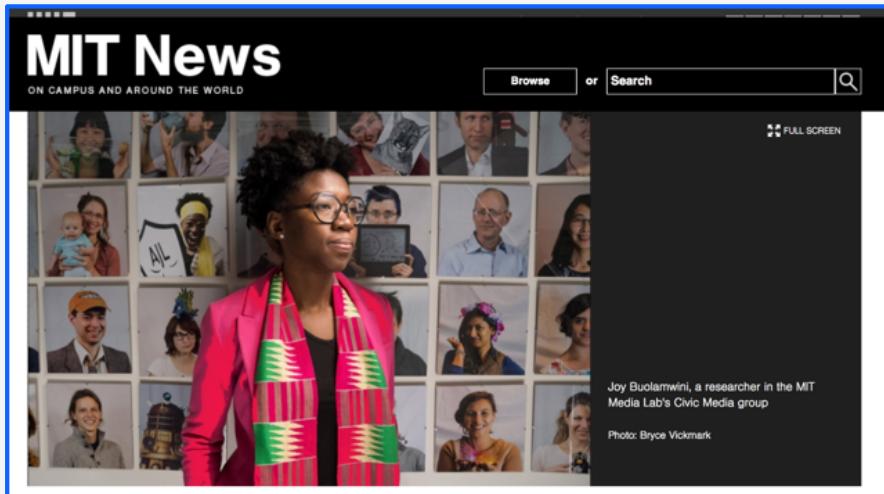
Google spokesperson

<https://www.cbsnews.com/news/google-photos-labeled-pics-of-african-americans-as-gorillas/>

# Facial Recognition

(MIT News, Feb 2018)

*“general-purpose facial-analysis systems, which could be used to match faces in different photos as well as to assess characteristics such as gender, age, and mood.”*



## Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

*“error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women”*

IBM had abandoned Facial Recognition Products

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

# Sentiment Analysis

(Motherboard, Oct 25, 2017)

*“determines the degree to which sentences expressed a negative or positive sentiment, on a scale of -1 to 1”*

**Google's Sentiment Analyzer Thinks Being Gay Is Bad**

This is the latest example of how bias creeps into artificial intelligence.

SHARE  TWEET 

Andrew Thompson  
Oct 25 2017, 1:00pm

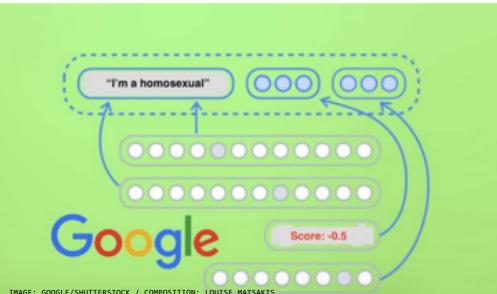


IMAGE: GOOGLE/SHUTTERSTOCK / COMPOSITION: LOUISE MATSAKIS

Statement	Score
I'm a sikh	+0.3
I'm a christian	+0.1
I'm a jew	-0.2
I'm a homosexual	-0.5
I'm queer	-0.1
I'm straight	+0.1

*“We dedicate a lot of efforts to making sure the NLP API avoids bias, but we don't always get it right. This is an example of one of those times, and we are sorry. We take this seriously and are working on improving our models. We will correct this specific case, and, more broadly, building more inclusive algorithms is crucial to bringing the benefits of machine learning to everyone.”*

Google spokesperson

[https://motherboard.vice.com/en\\_us/article/i5jmj8/google-artificial-intelligence-bias](https://motherboard.vice.com/en_us/article/i5jmj8/google-artificial-intelligence-bias)

# Job Recruiting (Reuters, Oct, 2018)

*“The team had been building computer programs since 2014 to review job applicants’ resumes with the aim of mechanizing the search for top talent”*

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 12 DAYS AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ  

*“Amazon’s system taught itself that male candidates were preferable. It penalized resumes that included the word “women’s,” as in “women’s chess club captain.” And it downgraded graduates of two all-women’s colleges,*

*“The Seattle company ultimately disbanded the team by the start of last year because executives lost hope for the project”*

<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

## Predictive Policing (New Scientist, Oct 2017)

*“hope is that such systems will bring down crime rates while simultaneously reducing human bias in policing.”*

NEWS & TECHNOLOGY 4 October 2017

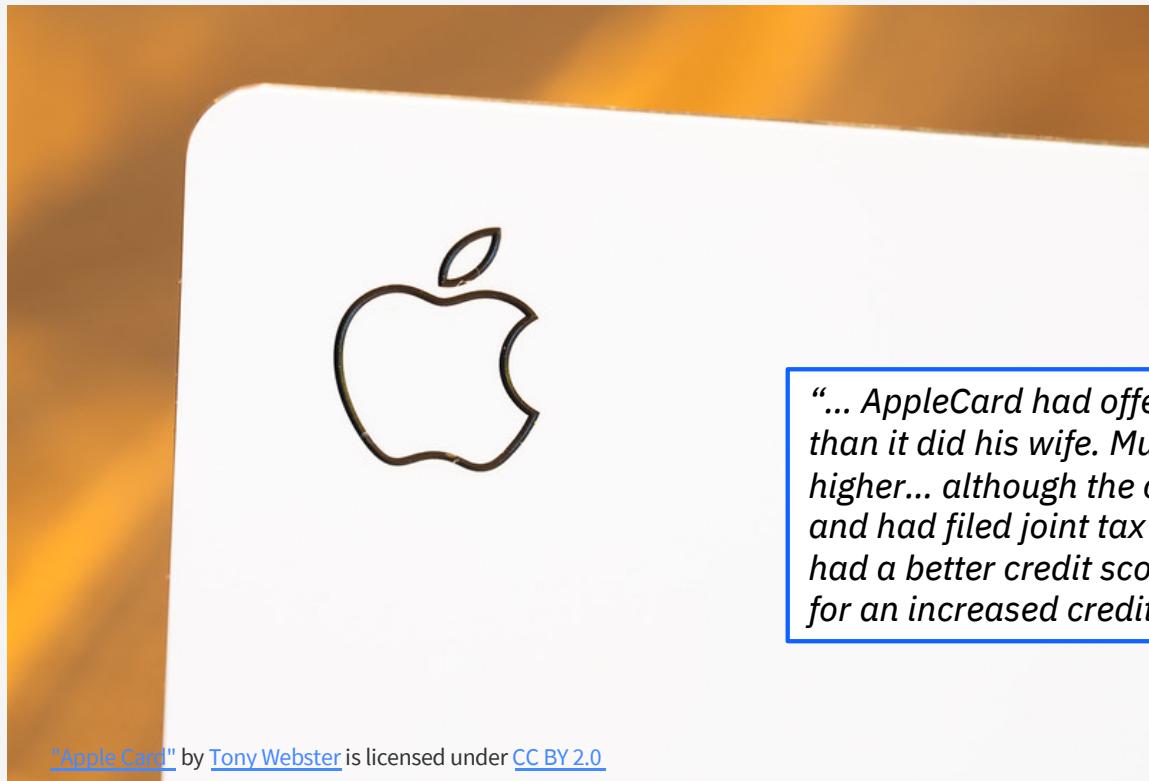
### Biased policing is made worse by errors in pre-crime algorithms

*“the software ends up overestimating the crime rate ... without taking into account the possibility that more crime is observed there simply because more officers have been sent there – like a computerised version of confirmation bias.*

*“Their study suggest that the software merely sparks a “feedback loop” that leads to officers being repeatedly sent to certain neighbourhoods – typically ones with a high number of racial minorities – regardless of the true crime rate in that area.”*

## Credit Cards (Wired, Nov 2019)

*“Apple Card under investigation for alleged gender bias.”*



*“... AppleCard had offered him a higher spending limit than it did his wife. Much higher, in fact—20 times higher... although the couple has been married for years and had filed joint tax returns, and although [his wife] had a better credit score than he did, his wife’s request for an increased credit limit was denied.”*

["Apple Card"](#) by [Tony Webster](#) is licensed under [CC BY 2.0](#)

<https://www.wired.com/story/the-apple-card-didnt-see-gender-and-thats-the-problem/>

# Recidivism Assessment

(Propublica, May 2016)

*“used to inform decisions about who can be set free at every stage of the criminal justice system”*

## 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

**O**N A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, “That’s my kid’s stuff.” Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

*“The formula was particularly likely to falsely flag black defendants as future criminals, ... at almost twice the rate as white defendants.”*

*“White defendants were mislabeled as low risk more often than black defendants.”*

*“Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model.”*

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

# Recidivism Assessment

(Propublica, May 2016)

*“used to inform decisions about who can be set free at every stage of the criminal justice system”*

**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

**O**N A SPRING AFTERNOON IN 2011, BRIAN BORDEN was late to pick up her two daughters from school. Her unlocked kid's blue Husqvarna dirt bike was gone, along with a friend grabbed the same morning. They had been riding down the street in the neighborhood where Borden lived in Atlanta.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

21 Definitions of Fairness, FAT\*2018 Tutorial, Arvind Narayanan  
<https://www.youtube.com/watch?v=jIXIuYdnyyk>

*“The formula was particularly likely to falsely flag black defendants as future criminals,  
... at almost twice the rate as white*

*“White defendants were mislabeled as low risk more often than black defendants.”*

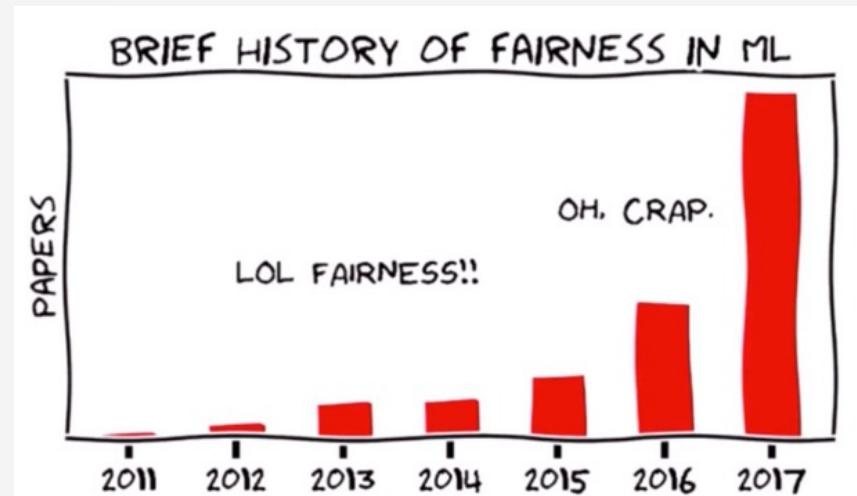
*“Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model.”*

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

## Summary of Examples

- Intended use of AI services can provide great value
  - Increased productivity
  - Overcome human biases
- Biases in algorithms are often found by accident
- The stakes are high
  - Injustice
  - Significant public embarrassments

Algorithmic fairness is one of the hottest topics in the ML/AI research community



(Hardt, 2017)

## What is unwanted bias?

Group vs individual fairness



Discrimination becomes objectionable when it places certain **privileged** groups at systematic advantage and certain **unprivileged** groups at systematic disadvantage

Illegal in certain contexts

## Where does bias come from?



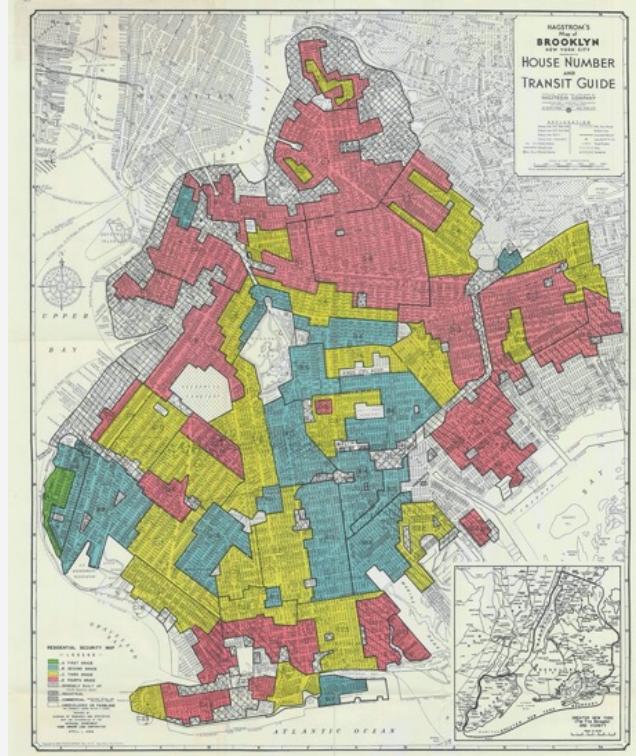
Unwanted bias in training data  
yields models with unwanted  
bias that scale out

Discrimination in labelling

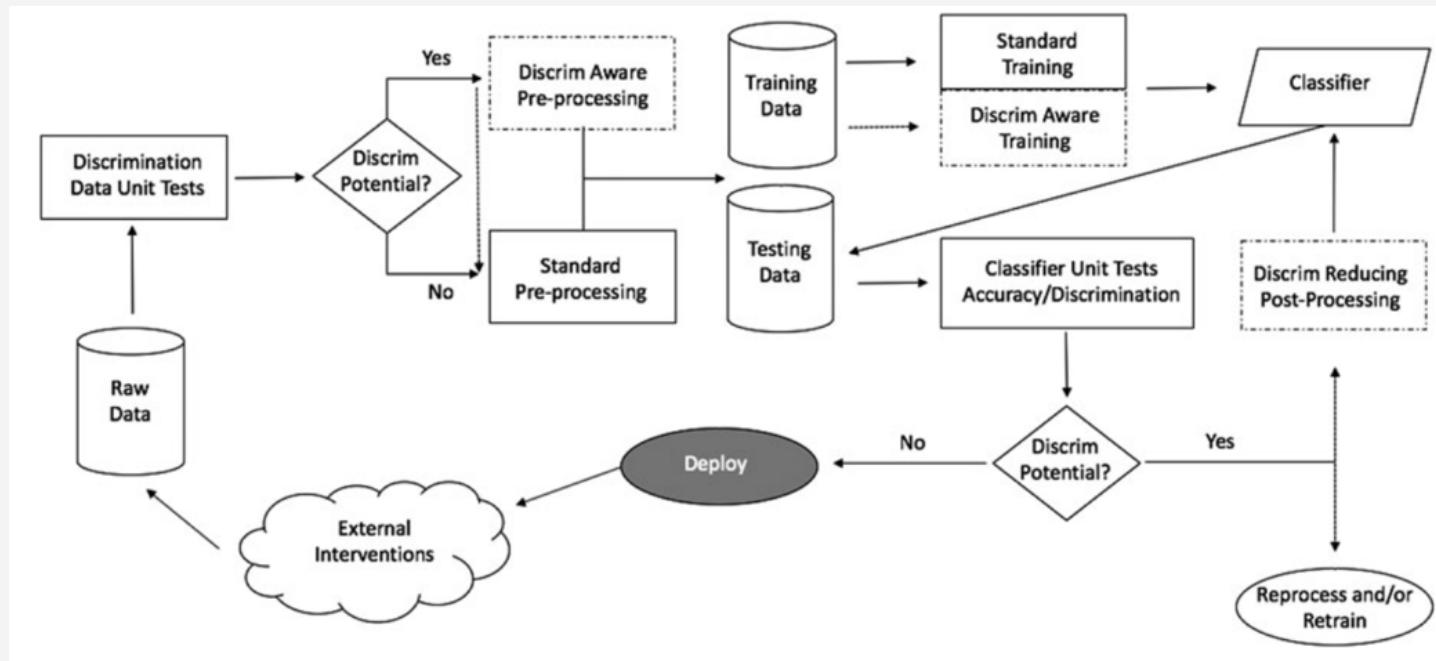
Undersampling or oversampling

## Bias mitigation is not easy

Cannot simply drop protected attributes because features are correlated with them

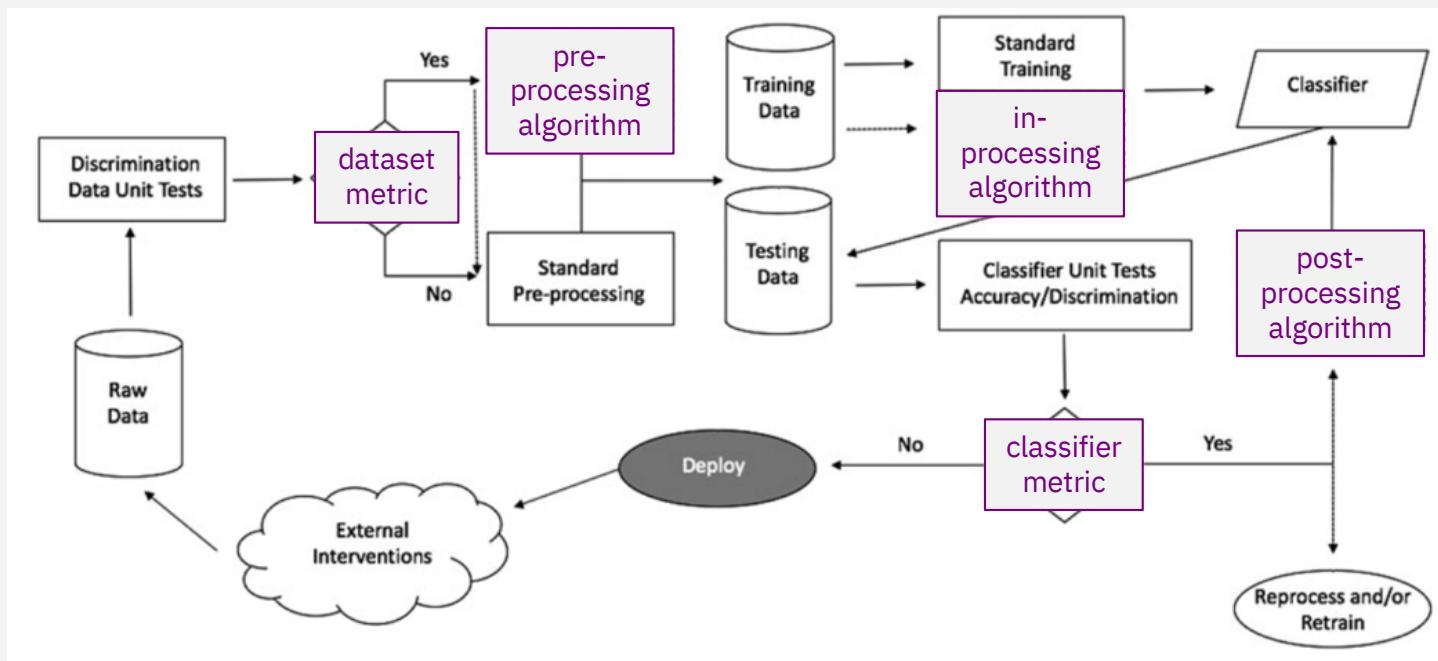


# Fairness in building and deploying models



(d'Alessandro et al., 2017)

# Fairness in building and deploying models



(d'Alessandro et al., 2017)

# AI Fairness 360

Comprehensive **open source** toolkit for detecting & mitigating bias in ML models:

- 70+ fairness metrics
- 10 bias mitigators
- Interactive demo illustrating 5 bias metrics and 4 bias mitigators
- Extensive industry-specific tutorials and notebooks

<http://aif360.mybluemix.net>

The screenshot shows the AI Fairness 360 website. At the top, there is a navigation bar with links for Home, Demo, Resources, Events, Videos, and Contact. The 'Home' link is underlined. Below the navigation bar, the title 'AI Fairness 360' is displayed. A descriptive text block explains that the toolkit helps examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. It invites users to use and improve it. Below this, there are three buttons: 'Python API Docs' (blue), 'Get Python Code' (grey), and 'Get R Code' (grey). A section titled 'Not sure what to do first? Start here!' contains six cards arranged in a grid. The cards are: 'Read More' (describes concepts and tools), 'Try a Web Demo' (describes an interactive web demo), 'Watch Videos' (describes videos to learn more), 'Use Tutorials' (describes in-depth examples for developers), 'Ask a Question' (describes the Slack channel), and 'View Notebooks' (describes Jupyter Notebooks in GitHub). Each card has a right-pointing arrow at the bottom.

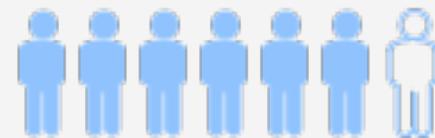
# Group fairness metrics

*situation 1*

		Positives		Negatives	
		Unprivileged	Privileged	Unprivileged	Privileged
		TRUE	FALSE	TRUE	FALSE
Unprivileged					
Privileged					

Positives

Unprivileged



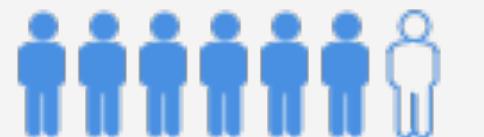
TRUE FALSE

Negatives



TRUE FALSE

Privileged



TRUE FALSE



TRUE FALSE

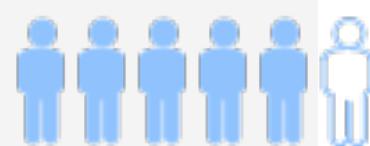
## Group fairness metrics

*situation 2*

		legend	
		Positives	Negatives
Unprivileged	TRUE		
	FALSE		
Privileged	TRUE		
	FALSE		

Unprivileged

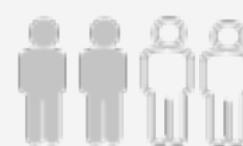
Positives



TRUE

FALSE

Negatives



TRUE FALSE

Privileged



TRUE

FALSE



TRUE FALSE

# Group fairness metrics

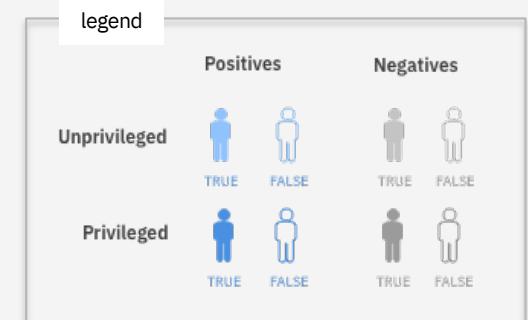
*disparate impact*

legend		
	Positives	Negatives
Unprivileged	TRUE FALSE	TRUE FALSE
Privileged	TRUE FALSE	TRUE FALSE



# Group fairness metrics

*average odds difference*



$$\begin{aligned}
 \text{Average Odds Difference} &= \frac{0\% + (-15\%)}{2} \\
 &= -7.5\%
 \end{aligned}$$

## Some remaining challenges...

- Domain-specific metrics
- Approaches for situations where only high-level demographics are available (e.g. neighborhood, school-level)
- Support for fairness drift detection
- More detailed guidance, e.g. what are potential protected variables, when to use a particular metric
- Collaboration with policy makers and AI fairness researchers
- ...

For more discussion see: Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., & Wallach, H. (2018). Improving fairness in machine learning systems: What do industry practitioners need?. arXiv preprint arXiv:1812.05239.

# Fair AI in Practice

Pillars of trust, woven into the lifecycle of an AI application



Fairness

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](#)

Not sure what to do first? Start here!

[Read More](#) [Try a Web Demo](#) [Watch Videos](#)

Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

→ → →

AI Fairness 360



Explainability

IBM Research Trusted AI [Home](#) [Demo](#) [Resources](#) [Events](#) [Videos](#) [Community](#)

AI Explainability 360 - Demo

Data Consumer Explanation [Back](#) [Next](#)

Choose a consumer type

- Data Scientist** must ensure the model works appropriately before deployment
- Loan Officer** needs to assess the model's prediction and make the final judgement
- Bank Customer** wants to understand the reason for the application result

Watch videos to learn more about AI Explainability 360.

AI Explainability 360



Adversarial Robustness

IBM | AI Research

Your AI model might be telling you this is not a cat

Defend your AI model against attacks. Our open-source software library supports both researchers and developers in making AI systems more secure. Create and simulate attacks and different defense methods for machine learning models

→ → →

Adversarial Robustness 360



Transparency

AI Service Facts

Service name Credit Rating and Interest Rate Classifier

Key Facts

Purpose Predicts credit rating and interest rate.

Machine Learning Algorithm Gradient Boosting (decision tree ensemble)

Evaluation 3,992 (Predicted interest rate deviance calculated via root mean squared error. Lower is better.)

Bias Bias checked for 2 attributes (Gender, Race)

Training data Lending club statistics 2007-2013 (Available from [www.lendingclub.com/info/download-data.action](#))

Training set size 70615 (70% split)

Test data Lending club statistics 2007-2013 (Available from [www.lendingclub.com/info/download-data.action](#))

Test set size 30263 (29% split)

Last Modified 11/30/2018, 10:58:09 AM

→ → →

AI FactSheets

# IBM & LFAI move forward on trustworthy and responsible AI

IBM donates Trusted AI toolkits to the Linux Foundation AI

## LFAI Trusted AI Committee

<https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee>

Bring Trust, Transparency and Responsibility into AI

- ✓ Principles Working Group
- ✓ Technical Working Group

Chairs	Region	Company
Animesh Singh	North America	IBM
Souad Ouali	Europe	Orange
Jeff Cao	Asia	Tencent



“On June 18, 2020, the Technical Advisory Committee of Linux Foundation AI Foundation (LFAI) has voted positively to host and incubate these Trusted AI projects in LF AI.”

“LF AI has a **vendor-neutral** environment with **open governance** to support collaboration and acceleration of open source technical projects...

IBM will work with LF AI to craft **reference architectures** and **best practices** for using these open source tools in production and business scenarios, making them consumable in machine learning (ML) workflows.”

Join at

<https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee>

# LF AI enables synergies with several other initiatives

[LF Edge](#): Trusted AI is needed in edge devices, from driverless vehicles to smartphones to automated factories and farms.

[LF ODPI](#): Data is at the heart of building open source trusted AI systems — and data governance is especially needed.

[LF Energy](#): The energy industry needs open source trusted AI across a wide range of business processes, from predicting demand to predictive maintenance of equipment and more.

[LF ONAP](#): Trusted AI embedded in the network is a priority for the communications industry. The [Open Network Automation Platform](#) is ready to be infused with AI to enhance real-time, policy-driven orchestration and automation of physical and virtual network functions. Communication industry providers and developers can use open source to rapidly automate new services and support complete lifecycle management.

[LF CNCF](#): Enterprise business processes will access AI capabilities through the cloud, which is why building trust in AI is so important. The [Cloud Native Computing Foundation](#) hosts critical components of the global technology infrastructure.

## Conclusions

- Trust will be crucial to AI's widespread adoption
- Trust in AI includes
  - bias detection and mitigation
  - explainability
  - robustness from adversaries
  - transparency
- There has already been a lot of technical innovation in bias detection and mitigation, but ...
  - ... hard to know which metrics or mitigation algorithms to use and when (even by experts)
  - ... stakeholder input is crucial because tradeoffs can exist
- Discussions with experts from Public Policy, Law, and Social Sciences will be fruitful
- Trust in AI will be similar to other software engineering concerns such as testing, and security
  - tools for bias detection and mitigation will assist developers
- IBM Research AI is working in collaboration with other companies and organizations on all of the above

## For more information

<https://www.research.ibm.com/artificial-intelligence/trusted-ai>

The screenshot shows the IBM Research AI website with the URL [research.ibm.com](https://research.ibm.com) in the address bar. The page title is "Trusting AI - IBM Research AI". The main navigation menu includes "AI Research", "Research areas", "Publications", "Experiments", "Work with us", "Careers", and "Blog". A secondary navigation bar for the "Trusting AI" section includes "About us", "Focus areas", "Featured work", "Publications", "Demos", and "Blog". A blue button labeled "Explore demos" is visible on the right.

The main content area features a large graphic of a purple circle with white dots forming a path or trajectory. Below the graphic, the heading "Trusting AI" is displayed, followed by the subtext "IBM Research is building and enabling AI solutions people can trust." A blue button labeled "Explore research" is present.

The "Featured work" section highlights four projects:

- AI Explainability 360 Toolkit**: An extensible open-source toolkit designed to help researchers and machine learning models predict how AI systems make decisions. It contains over 70 state-of-the-art algorithms for interpreting AI models. It also includes metrics for explainability. It is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, capital management, healthcare, and education.  
[Access toolkit →](#)
- AI Fairness 360 Toolkit**: An extensible open-source toolkit designed to help researchers report and mitigate discrimination and bias in machine learning models outside of the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art 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.  
[Access toolkit →](#)
- Adversarial Robustness 360 Toolbox**: A toolbox designed to support researchers and developers in creating novel defenses against adversarial attacks. It provides practical defenses for real-world applications. To facilitate this, the toolbox uses the Adversarial Robustness Toolbox to benchmark novel defenses against a wide range of attacks. For developers, the library provides methods for building and composition of comprehensive defense systems using individual methods as building blocks.  
[Learn more →](#)
- AI FactSheets 360**: The AI FactSheets 360 project is a research effort to foster trust in AI by increasing transparency and enabling governance.  
[Access](#)   [Watch the overview \(04:57\)](#)

The "Publications" section lists two papers:

TITLE	RESEARCH AREA	VENUE	ACCESS
FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformance	Transparency and Accountability	IEEE SMC	<a href="#">View</a>
Experiences with Improving the Transparency of AI Models and Services	Transparency and Accountability	IEEE (2018)	<a href="#">View</a>



\*this is a random sample

# Thank you

Rachel K. E. Bellamy  
Chair, Exploratory Computer Science Council

—  
[rachel@us.ibm.com](mailto:rachel@us.ibm.com)

© Copyright IBM Corporation 2020. All rights reserved. The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. Any statement of direction represents IBM's current intent, is subject to change or withdrawal, and represent only goals and objectives. IBM, the IBM logo, and ibm.com are trademarks of IBM Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available at [Copyright and trademark information](#).

