

Fair AI in Practice

Rachel K. E. Bellamy

Chair of Exploratory Computer Science Council

Principal Research Staff Member

IBM Research

PROMISE
Nov 20



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**

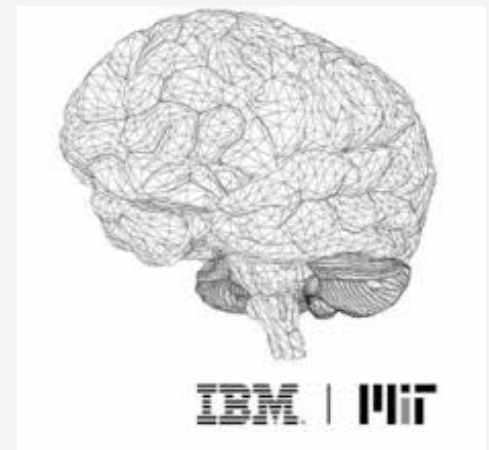
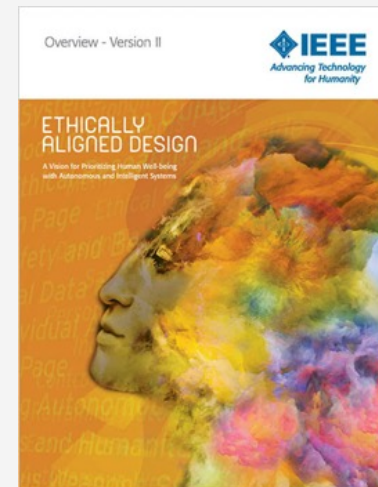
Actively engaging with **NIST** in the area of AI metrics, standards and testing

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

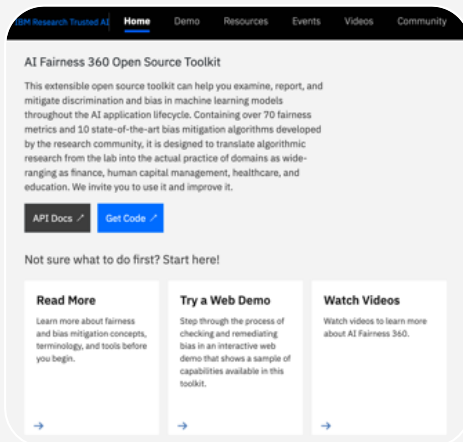


Sentencing

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

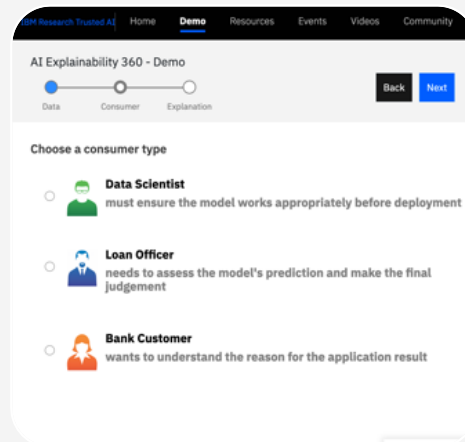
(Other than that it is 99% accurate)

Is it fair?



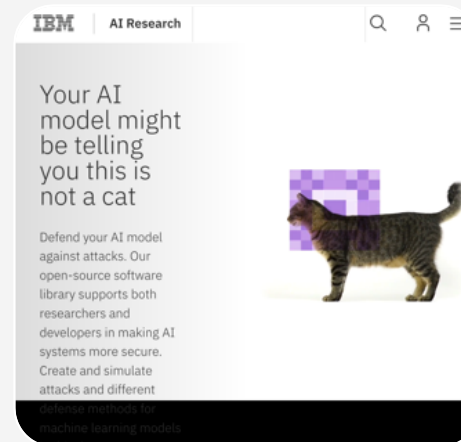
AI Fairness 360

Is it easy to understand?



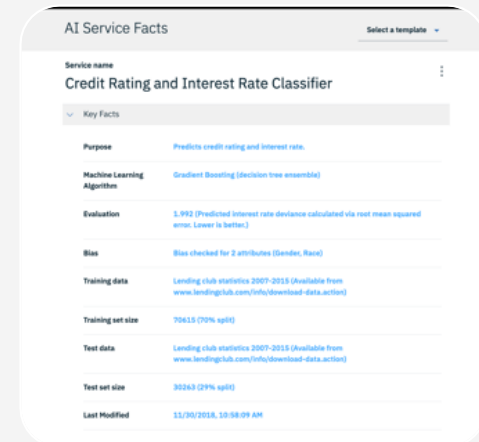
AI Explainability 360

Did anyone tamper with it?



Adversarial Robustness 360

Is it accountable?



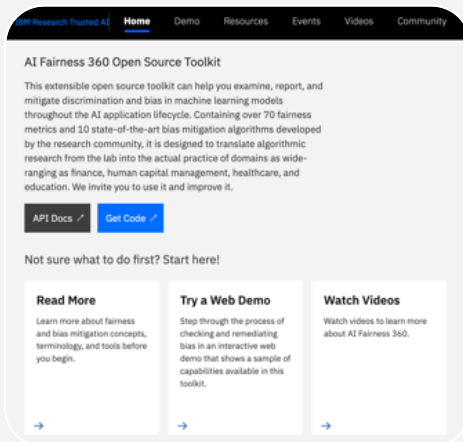
AI FactSheets

Pillars of Trust

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

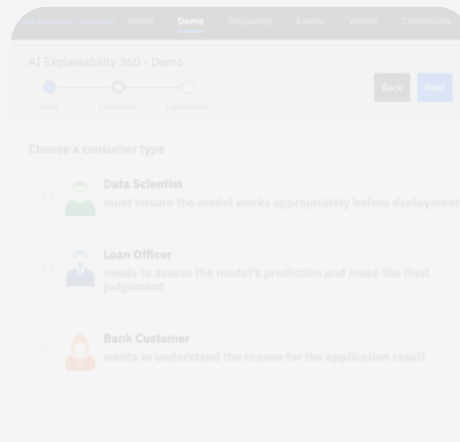
(Other than that it is 99% accurate)

Is it fair?



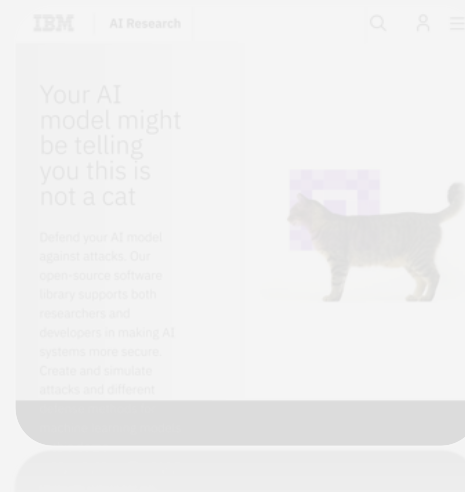
AI Fairness 360

Is it easy to understand?



AI Explainability 360

Did anyone tamper with it?



Adversarial Robustness 360

Is it accountable?

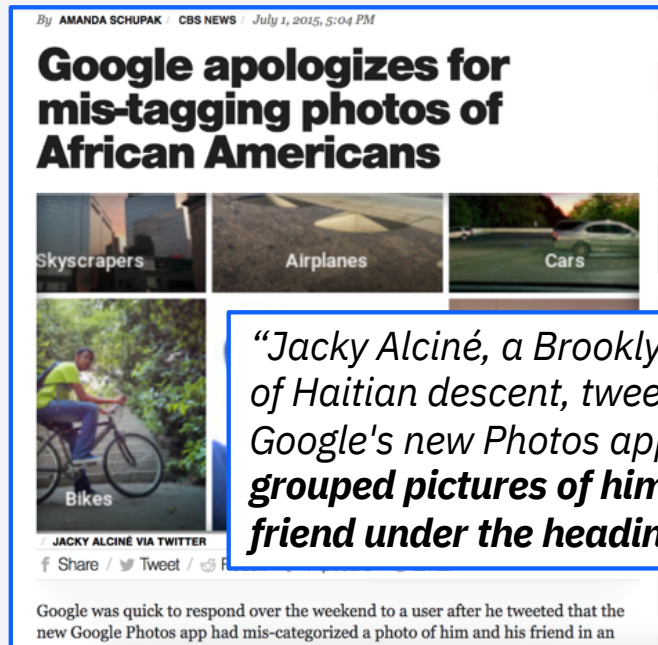


AI FactSheets

Pillars of Trust

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

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



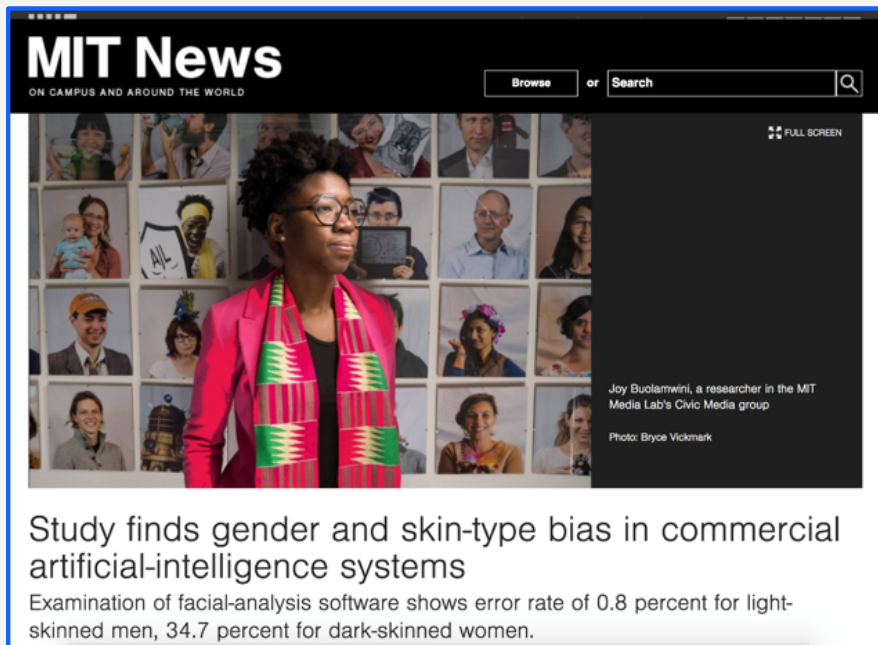
"We're appalled and genuinely sorry that this happened. We are taking immediate action to prevent this type of result from appearing. There is still clearly a lot of work to do with automatic image labeling, and we're looking at how we can prevent these types of mistakes from happening in the future."

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.”



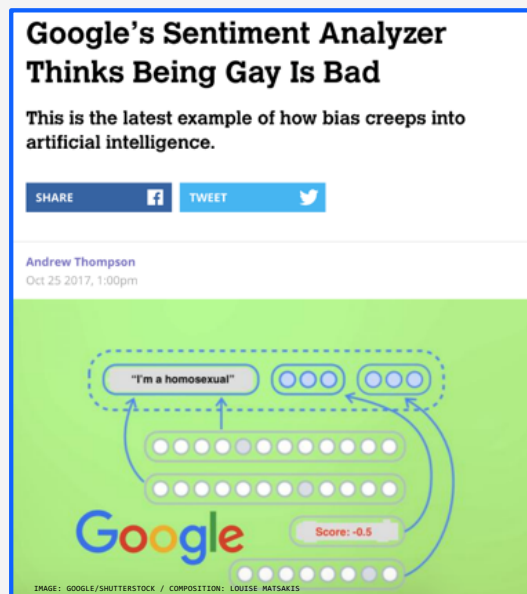
“error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women”

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”



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/i5jmi8/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

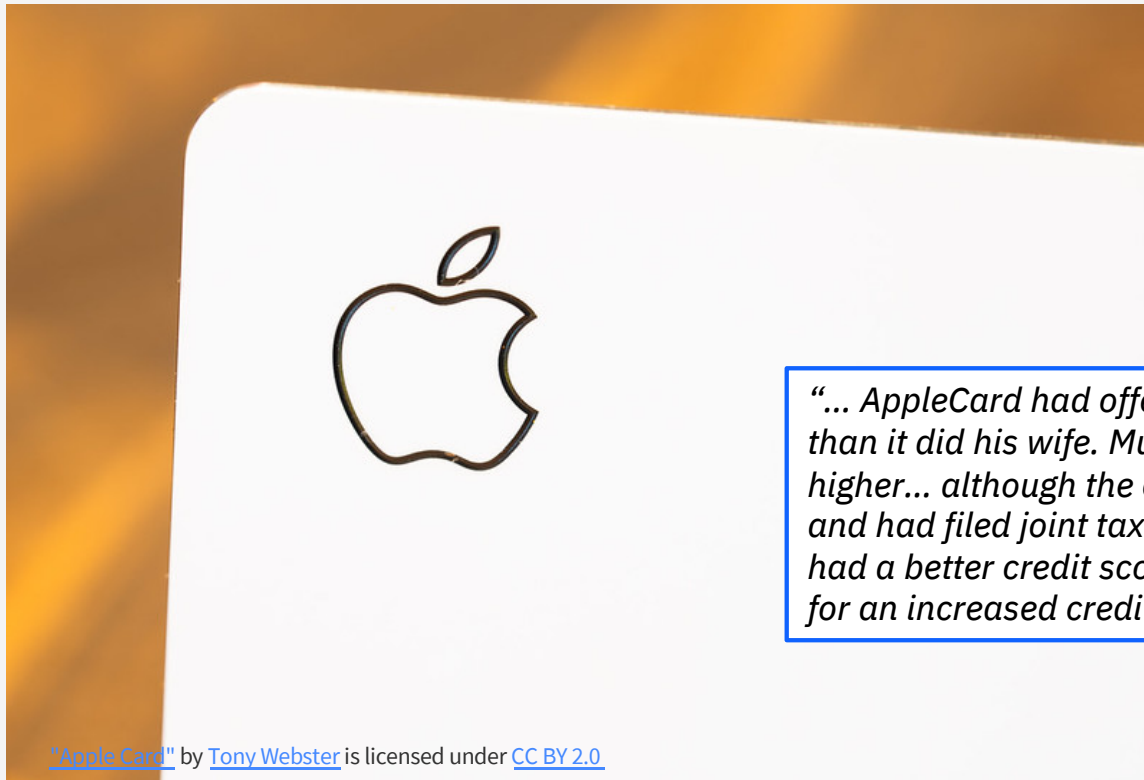
“the software ends up overestimating the crime rate ... without taking into account the possibility that more crime is observed there simply because more officers have been sent there – like a computerised version of confirmation bias.

“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.”

<https://www.newscientist.com/article/mg23631464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/>

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-genderand-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

ON 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

ON A SPRING AFTERNOON IN 2012, Brianna Borden was running late to pick up her 6-year-old brother. She unlocked his blue Honda and a friend grabbed the bike and scooter and rode down the street in the

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

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

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

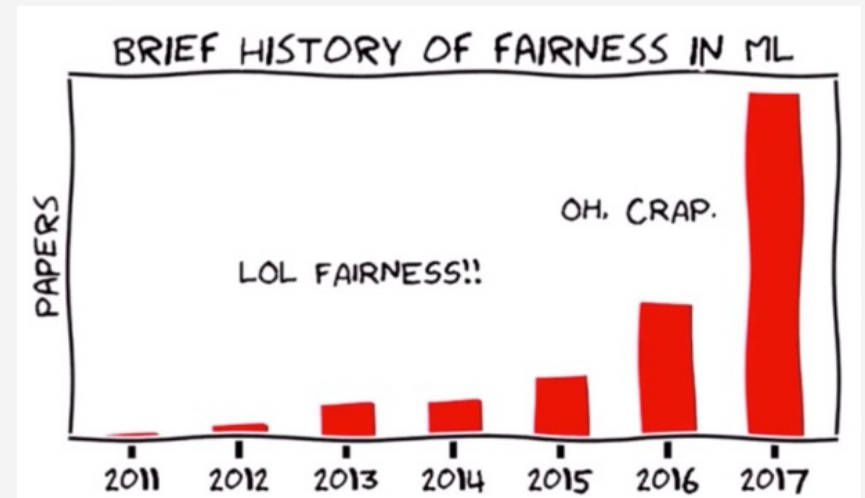
“Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model.”

<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?



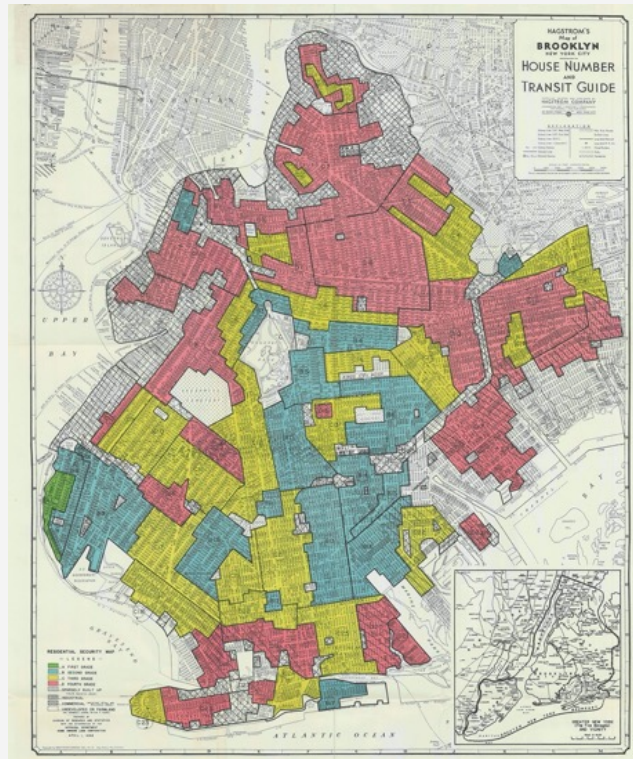
Unwanted bias in training data
yields models with unwanted
bias that scale out

Discrimination in labelling

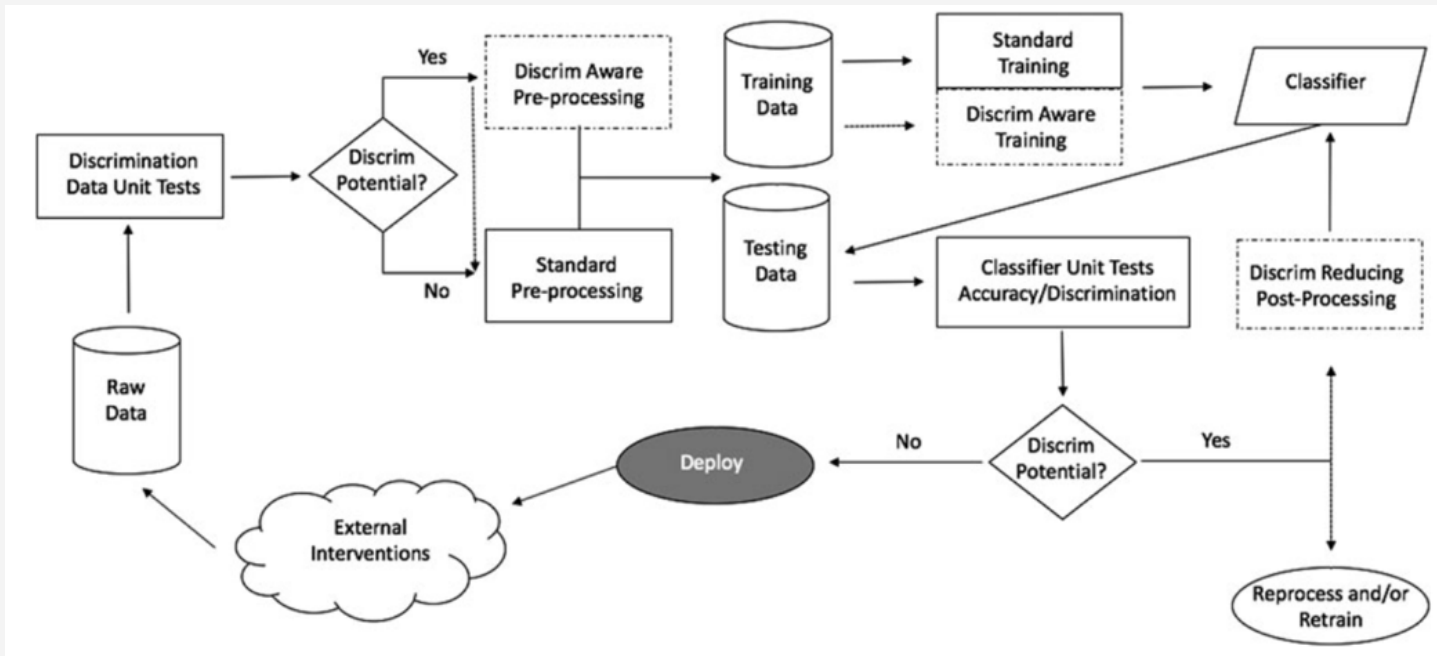
Undersampling or oversampling

Bias mitigation is not easy

Cannot simply drop protected attributes because features are correlated with them

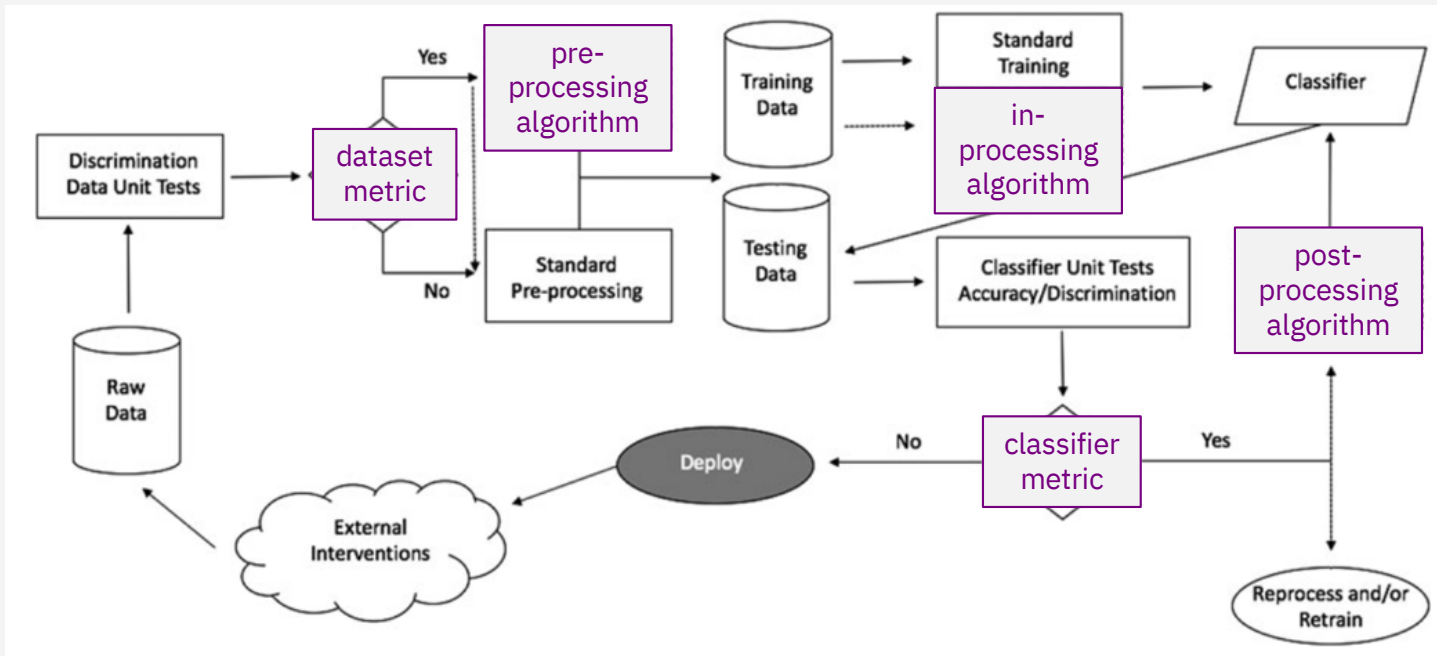


Fairness in building and deploying models



(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>

IBM Research Trusted AI

[Home](#)[Demo](#)[Resources](#)[Events](#)[Videos](#)

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

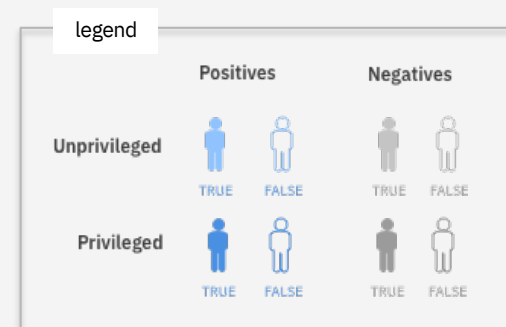
[Python API Docs ↗](#)[Get Python Code ↗](#)[Get R Code ↗](#)

Not sure what to do first? Start here!

<h3>Read More</h3> <p>Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.</p> →	<h3>Try a Web Demo</h3> <p>Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.</p> →	<h3>Watch Videos</h3> <p>Watch videos to learn more about AI Fairness 360.</p> →	<h3>Read the Docs</h3> <p>Read the documentation for AI Fairness 360.</p> →
<h3>Use Tutorials</h3> <p>Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.</p>	<h3>Ask a Question</h3> <p>Join our AIF360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.</p>	<h3>View Notebooks</h3> <p>Open a directory of Jupyter Notebooks in GitHub that provide working examples of bias detection and mitigation in sample datasets. Then share your own notebooks!</p>	<h3>Contribute</h3> <p>You can contribute to the project by submitting pull requests, issues, or documentation updates.</p>

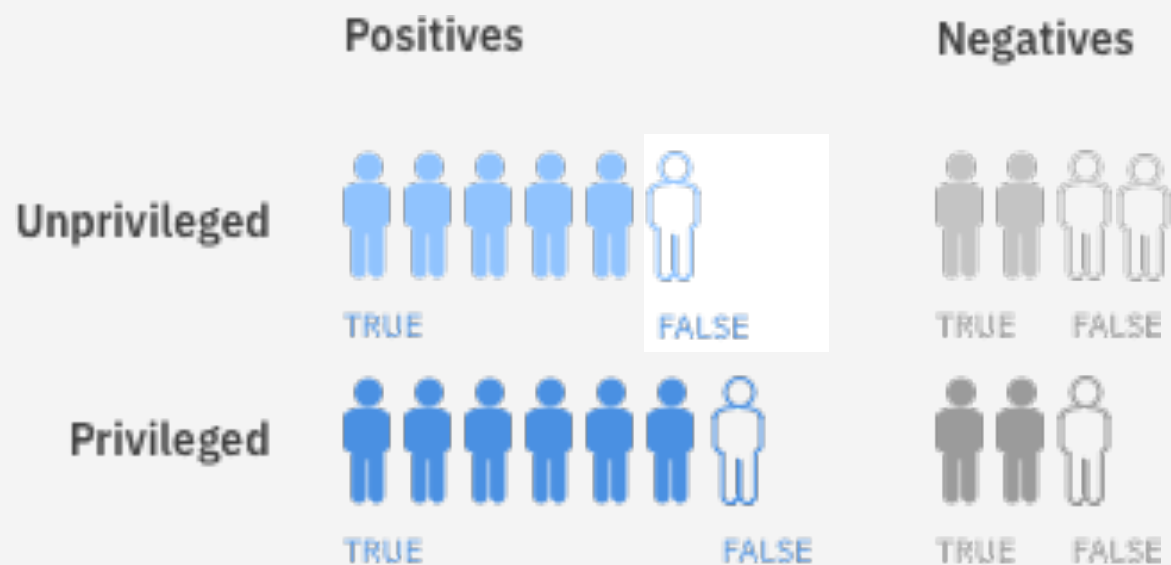
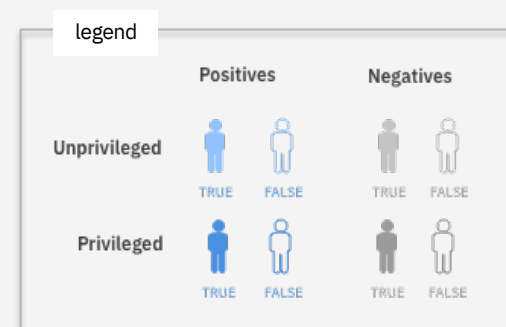
Group fairness metrics

situation 1



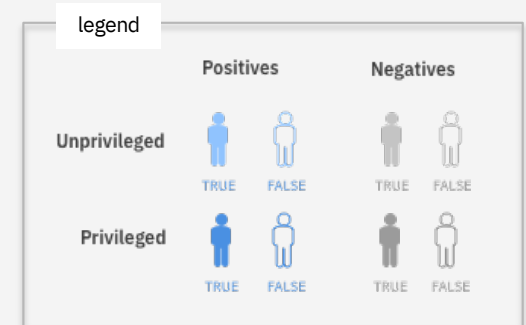
Group fairness metrics

situation 2



Group fairness metrics

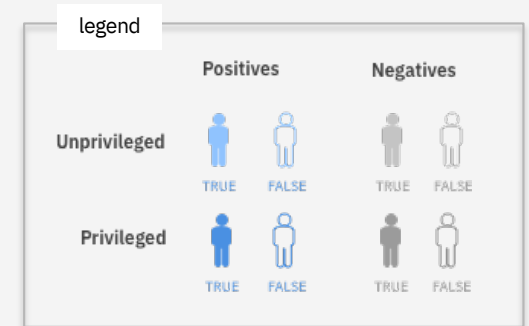
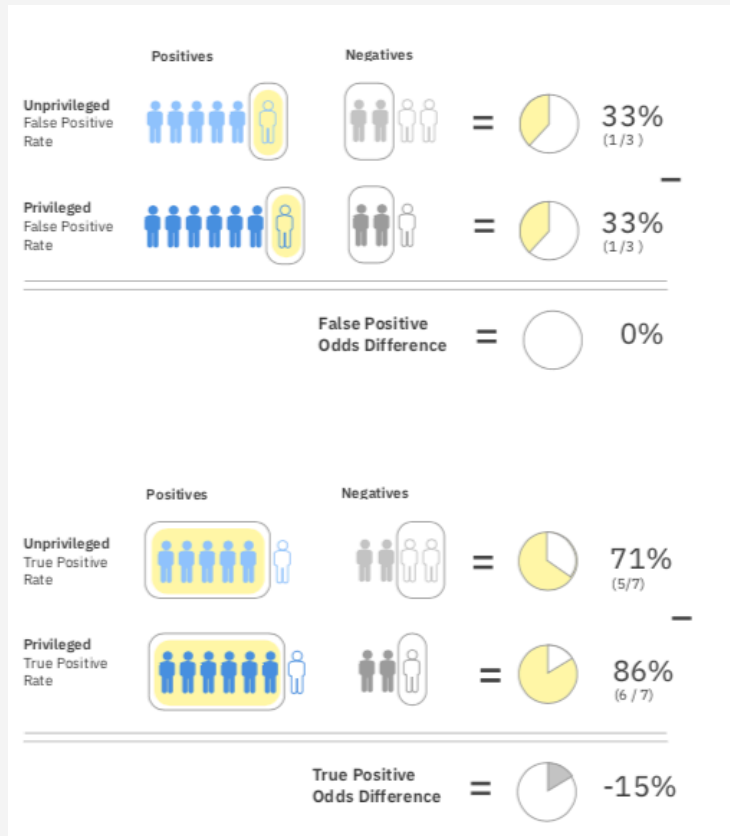
disparate impact



$$\text{Disparate Impact} = 0.86$$

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

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.

[→](#)

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

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 [Select a template](#)

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	1.912 (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 2007-2015 (Available from www.lendingclub.com/info/download-data.action)
Training set size	75618 (70% split)
Test data	Lending club statistics 2007-2015 (Available from www.lendingclub.com/info/download-data.action)
Test set size	30243 (20% split)
Last Modified	11/04/2018, 10:08:09 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>

research.ibm.com

Trusting AI - IBM Research AI

AI Research Research areas Publications Experiments Work with us Careers Blog

Trusting AI About us Focus areas Featured work Publications Demos Blog Explore demos

Trusting AI

IBM Research is building and enabling AI solutions people can trust.

Explore research

Featured work

AI Explainability 360 Toolkit

This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. Covering eight state-of-the-art algorithms for interpretable machine learning as well as metrics for explainability, 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 →

AI Fairness 360 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 50 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.

Access toolkit →

Adversarial Robustness 360 Toolbox

The Adversarial Robustness Toolbox is designed to support researchers and developers in creating novel defense techniques, as well as in deploying practical defenses of real-world AI systems. Researchers can use the Adversarial Robustness Toolbox to benchmark novel defenses against the state-of-the-art. For developers, the library provides interfaces which support the composition of comprehensive defense systems using individual methods as building blocks.

Learn more →

AI FactSheets

IBM scientists suggest that AI services be accompanied with a factsheet outlining the details about how it operates, how it was trained and tested, its performance metrics, fairness and robustness checks, intended uses, maintenance, and other critical details.

Access →

Publications

IBM Research AI is developing diverse approaches for how to achieve fairness, robustness, explainability, accountability, value alignment, and how to integrate them throughout the entire lifecycle of an AI application.

Please explore all of our trusting AI related research papers.

TITLE	RESEARCH AREA	VENUE	ACCESS
FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity	Transparency and Accountability	IEEE 2019	Access
Experiences with Improving the Transparency of AI Models and Services	Transparency and Accountability	IEEE 2019	Access

Watch the overview (04:57)



*this is a random sample

Thank you

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

—
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](#).

