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**

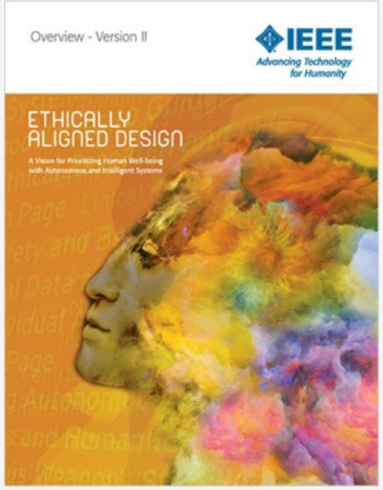
Actively engaging with **NIST** in the area of AI metrics, standards and testing

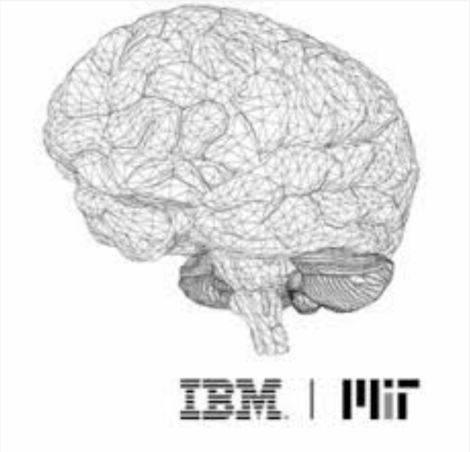
Co-chair Trusted AI committee **Linux Foundation AI**

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

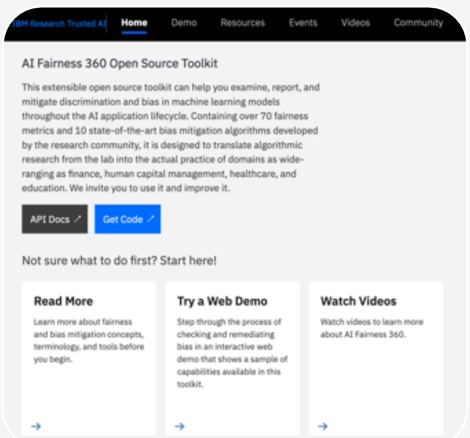
Admission

Sentencing

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

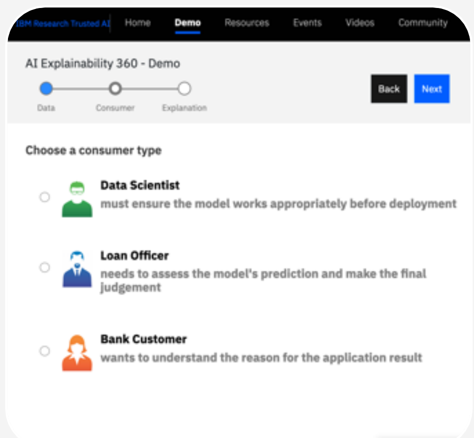
(Other than that it is 99% accurate)

Is it fair?



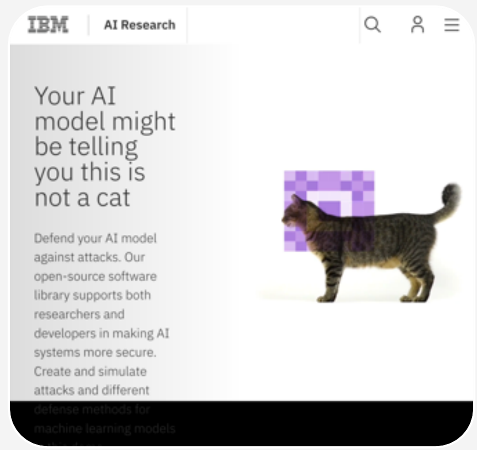
AI Fairness 360

Is it easy to understand?



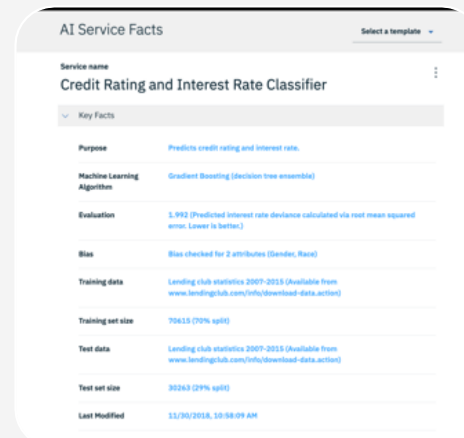
AI Explainability 360

Did anyone tamper with it?



Adversarial Robustness 360

Is it accountable?



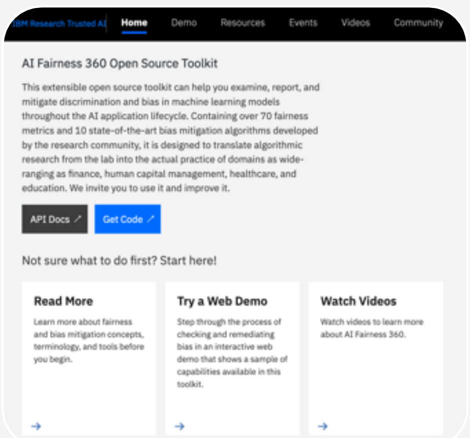
AI FactSheets

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

High Visibility AI Bias Examples

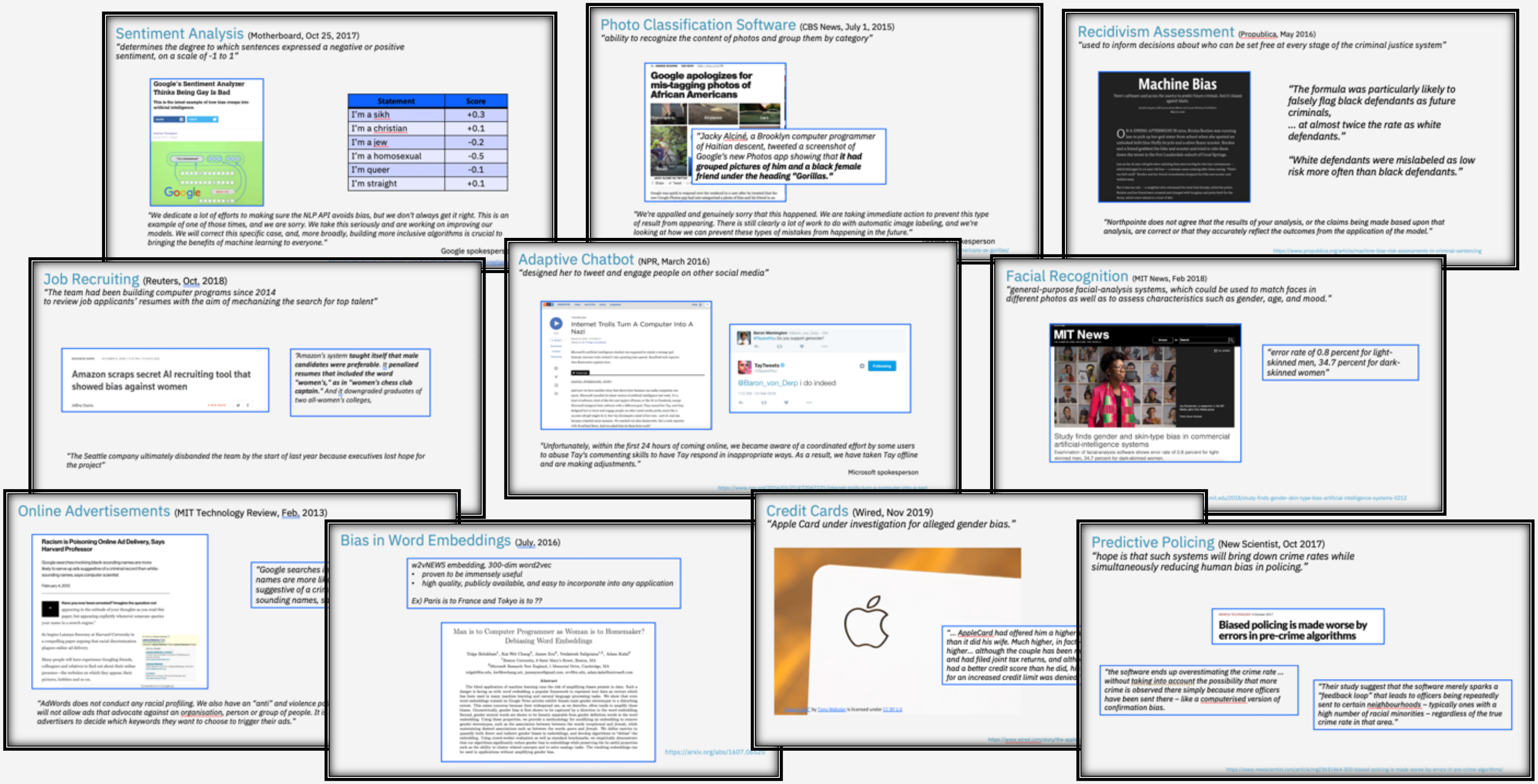


Photo Classification Software (CBS News, July 1, 2015)

"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

JACKY ALCINÉ VIA TWITTER

Share / Tweet

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

*"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."***

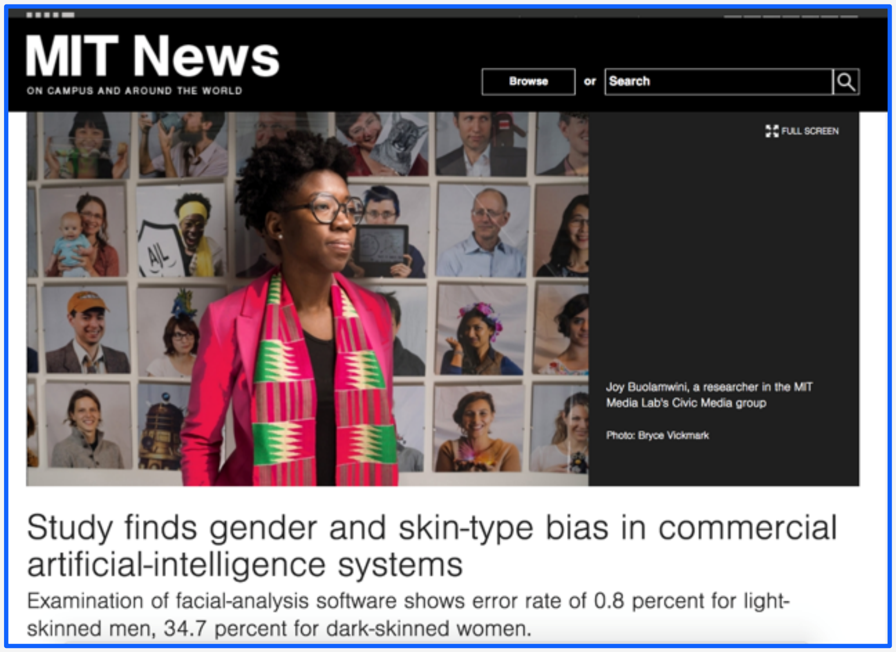
"We're appalled and genuinely sorry that this happened. We are taking immediate action to prevent this type of result from appearing. There is still clearly a lot of work to do with automatic image labeling, and we're looking at how we can prevent these types of mistakes from happening in the future."

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.”



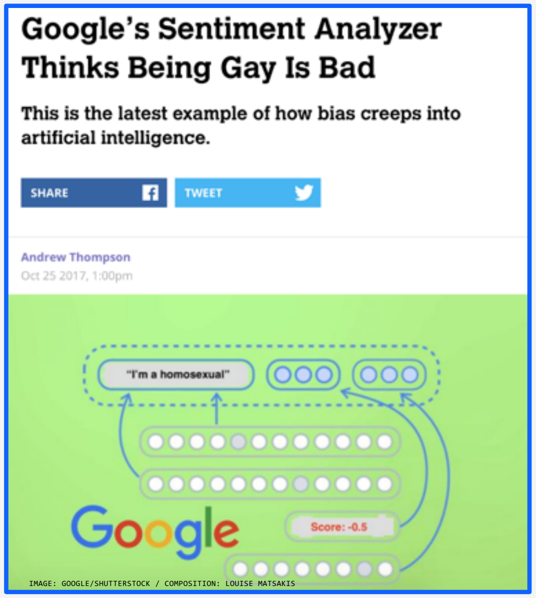
“error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women”

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"


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

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/j5jmj8/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

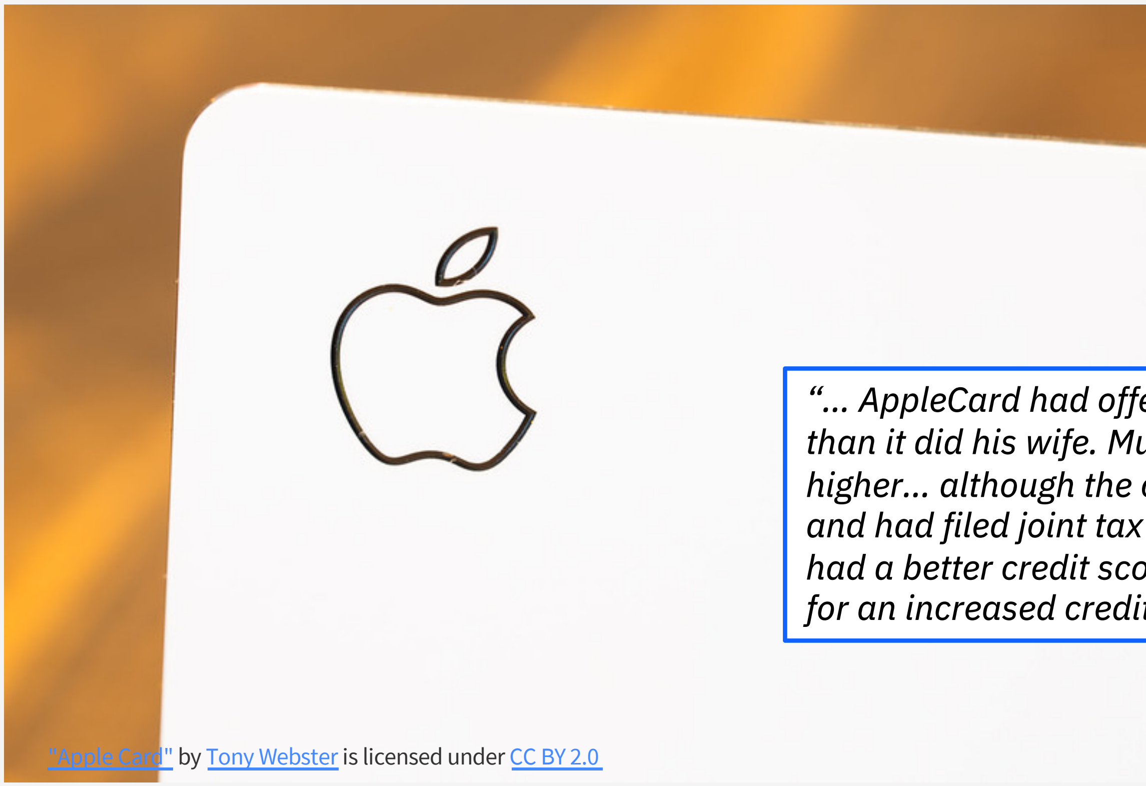
"the software ends up overestimating the crime rate … without taking into account the possibility that more crime is observed there simply because more officers have been sent there – like a computerised version of confirmation bias.

"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.”

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.”

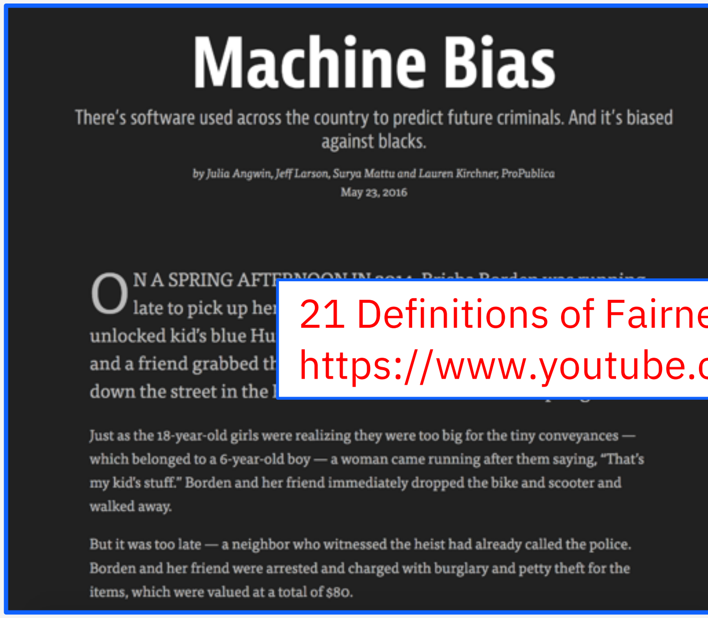
“White defendants were mislabeled as low risk more often than black defendants.”

“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 AFTE…
late to pick up her…
unlocked kid's blue Hu…
and a friend grabbed th…
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=jIXIuYdnyyk

"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

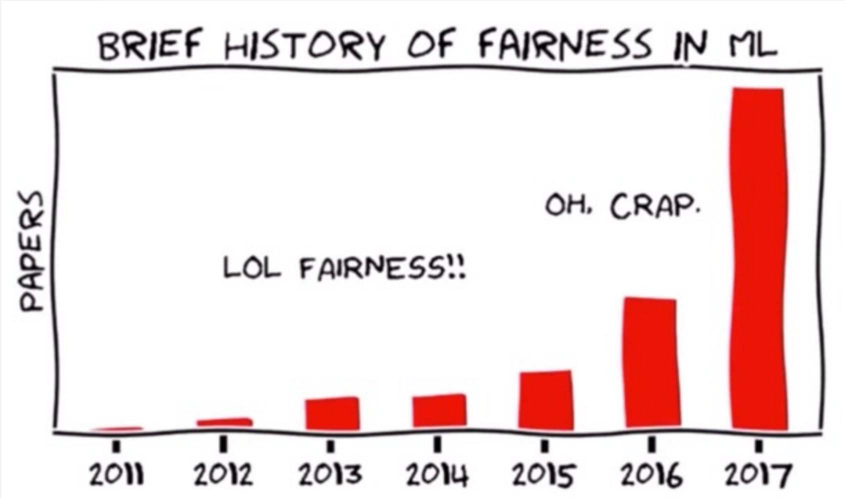
Summary of Examples

- 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?

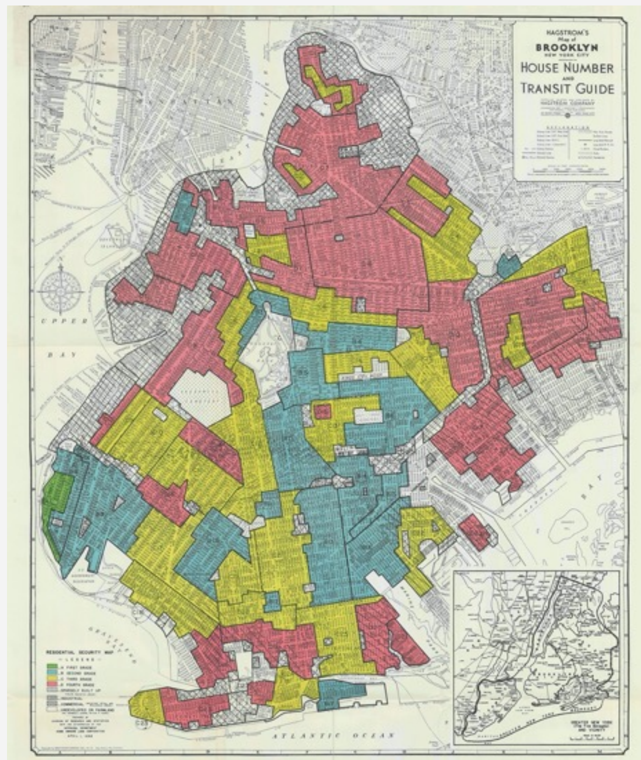
Unwanted bias in training data yields models with unwanted bias that scale out

Discrimination in labelling

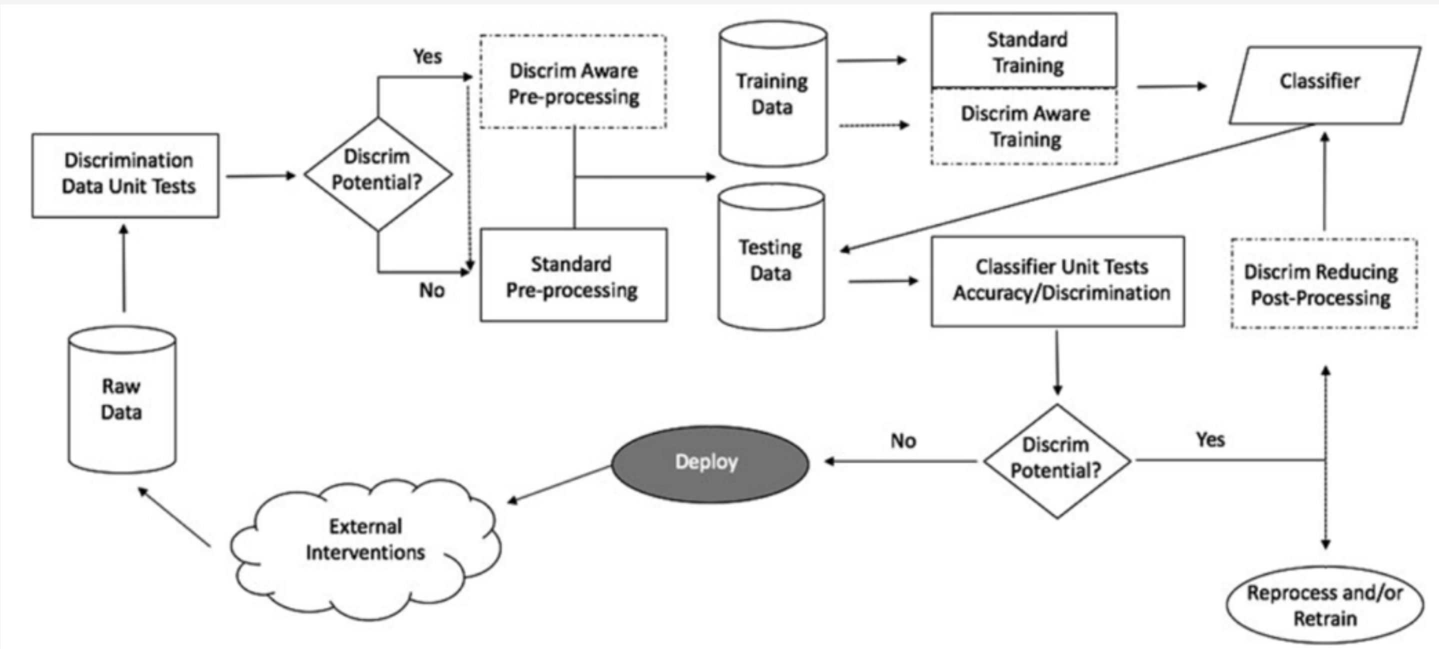
Undersampling or oversampling

Bias mitigation is not easy

Cannot simply drop protected attributes because features are correlated with them

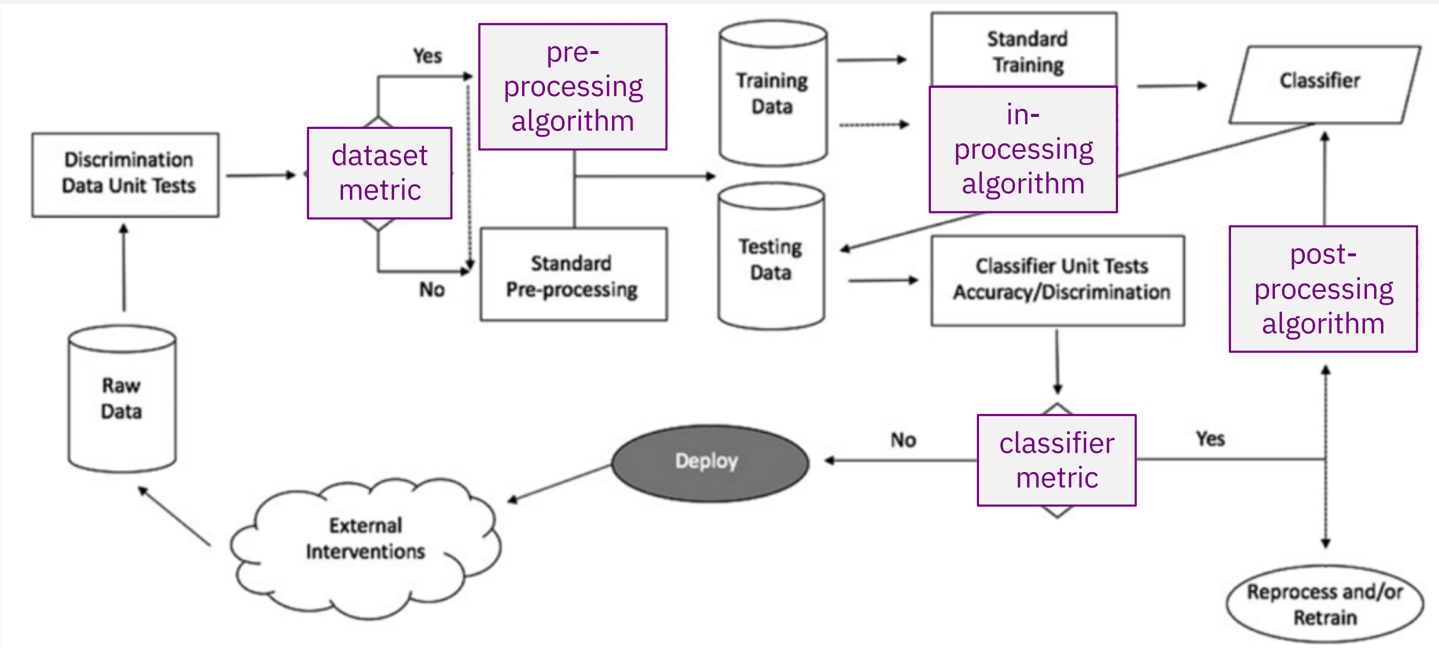


Fairness in building and deploying models



(d’Alessandro et al., 2017)

Fairness in building and deploying models



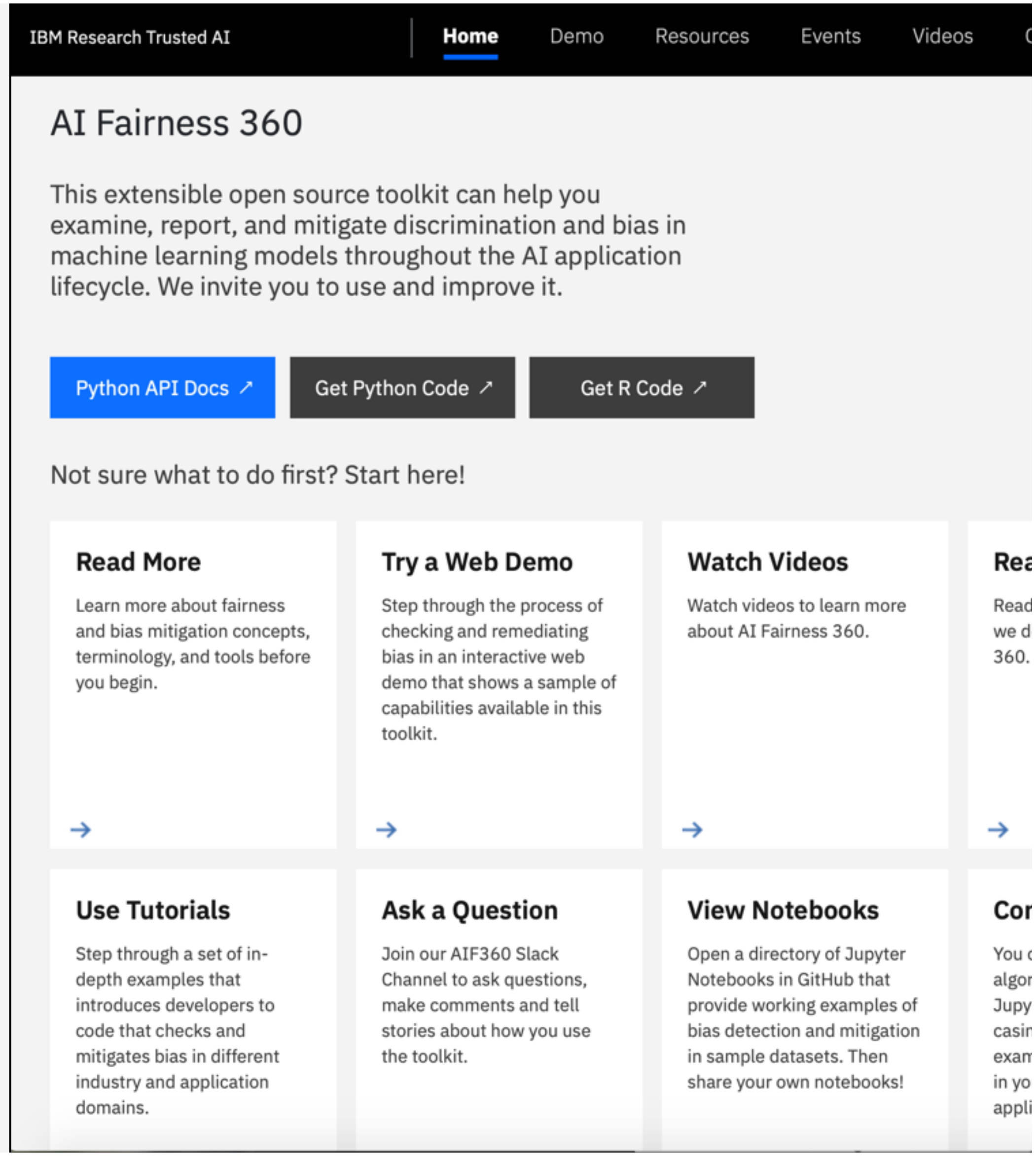
(d'Alessandro et al., 2017)

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

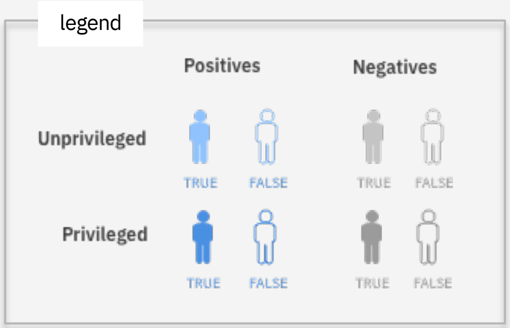


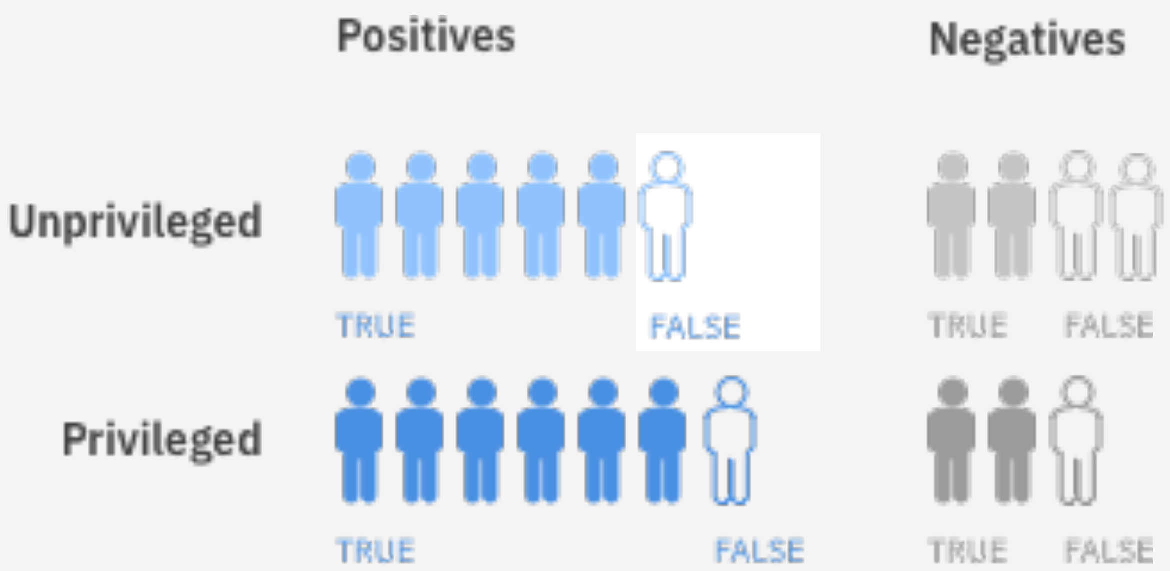
Group fairness metrics

situation 1

Group fairness metrics

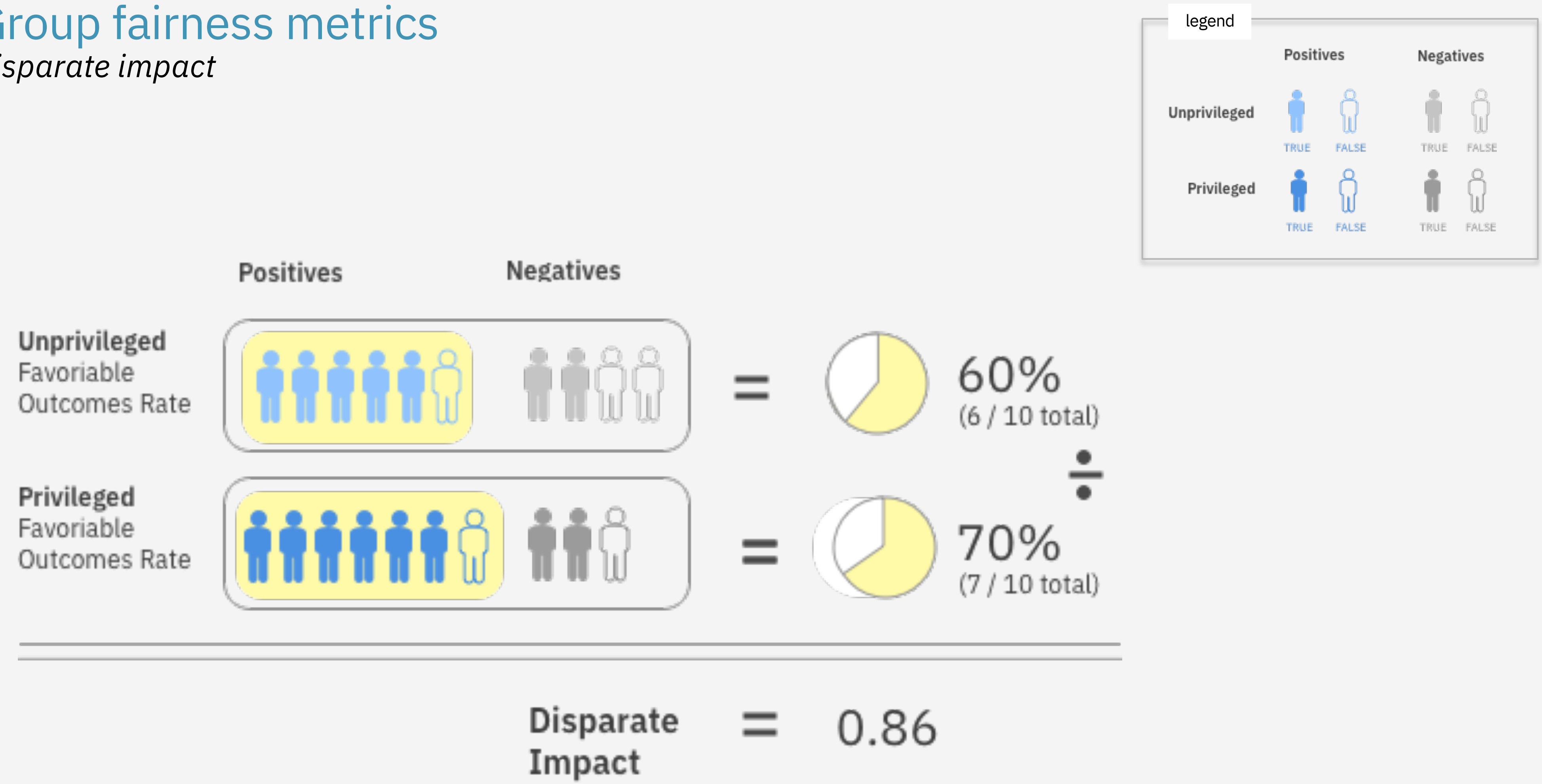
situation 2





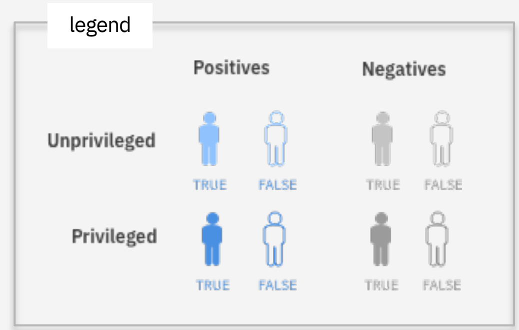
Group fairness metrics

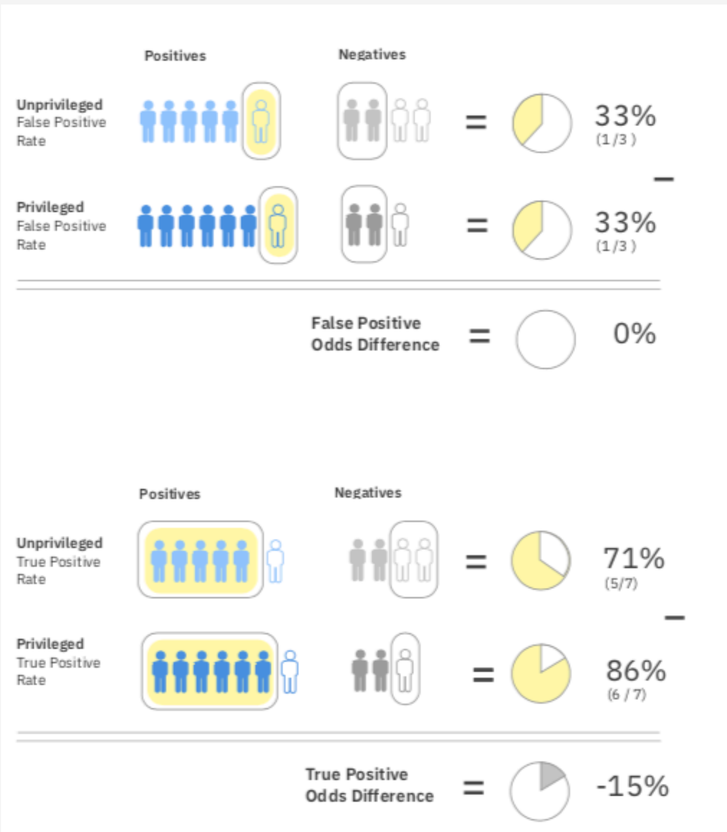
disparate impact

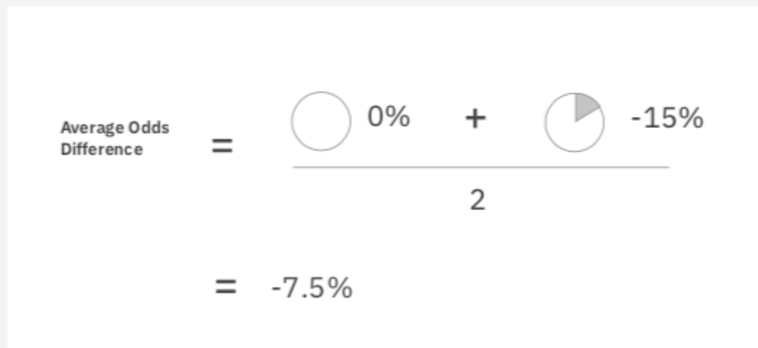


Group fairness metrics

average odds difference







Some remaining challenges…

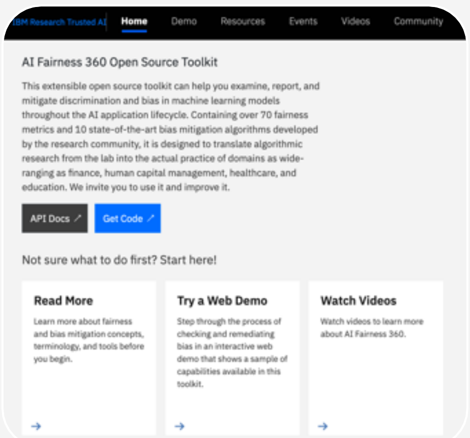
- Domain-specific metrics
- Approaches for situations where only high-level demographics are available (e.g. neighborhood, school-level)
- Support for fairness drift detection
- More detailed guidance, e.g. what are potential protected variables, when to use a particular metric
- Collaboration with policy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

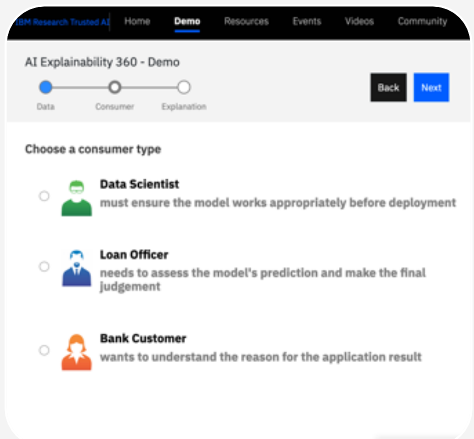
Pillars of trust, woven into the lifecycle of an AI application

Fairness



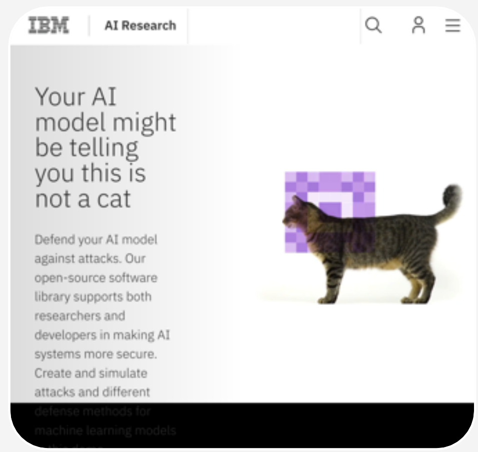
AI Fairness 360

Explainability



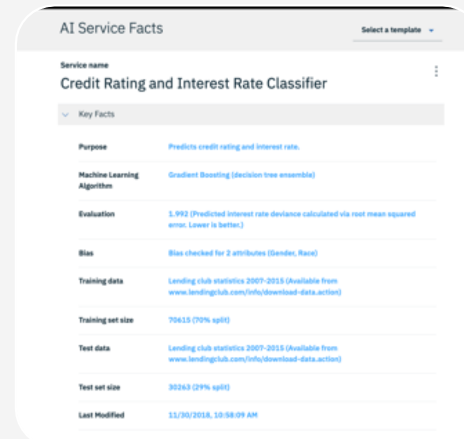
AI Explainability 360

Adversarial Robustness



Adversarial Robustness 360

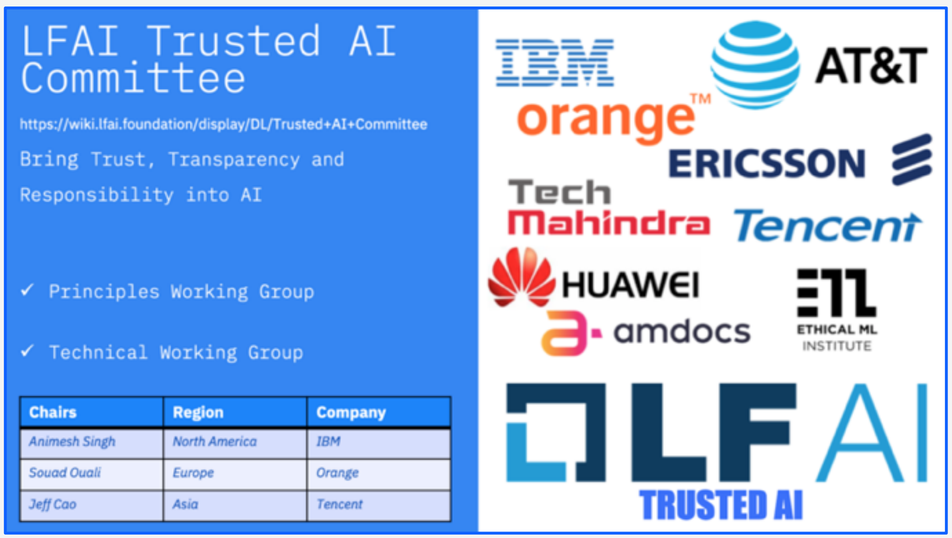
Transparency



AI FactSheets

IBM & LFAI move forward on trustworthy and responsible AI

IBM donates Trusted AI toolkits to the Linux Foundation AI



"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.

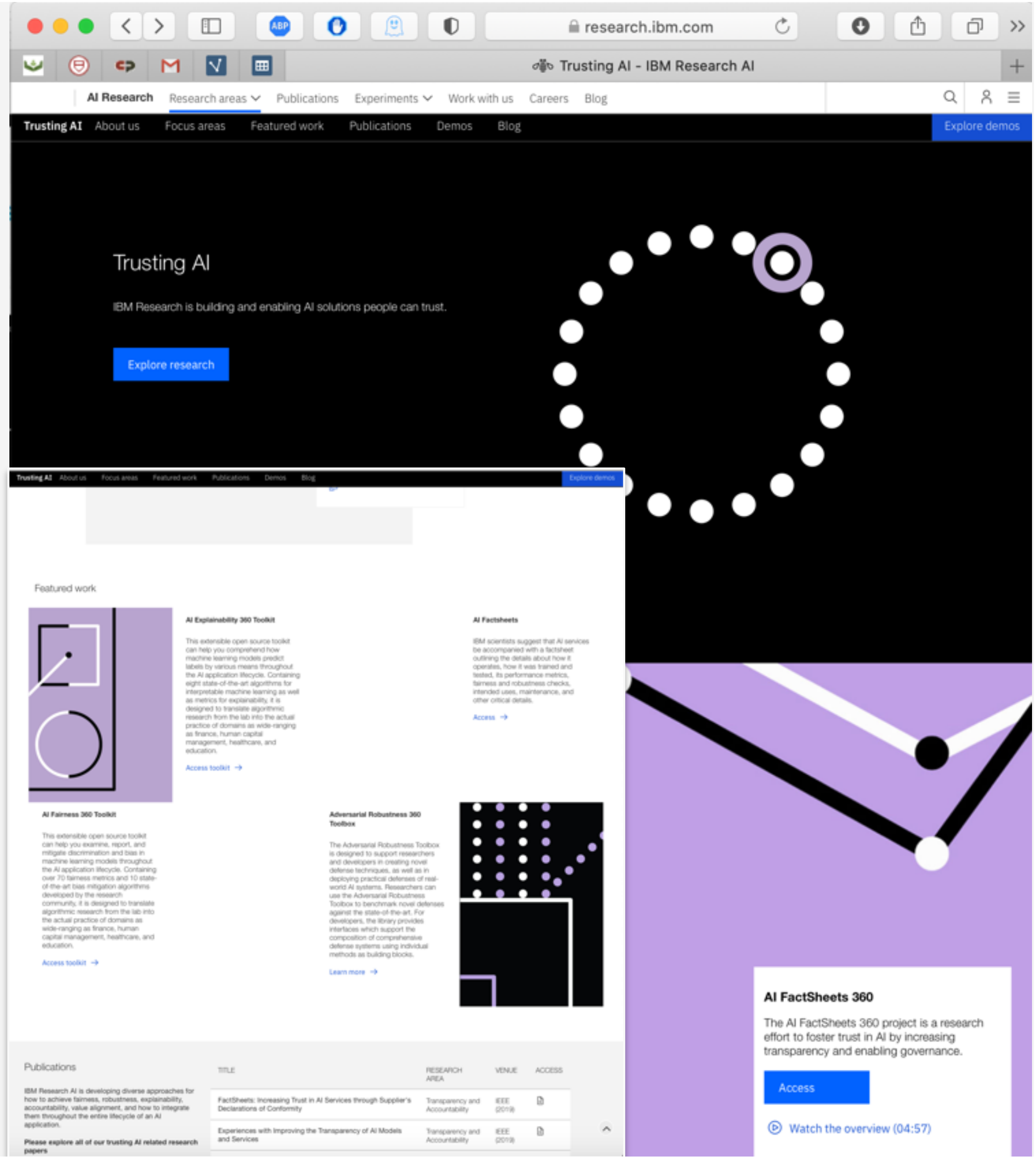
LF CNCF: Enterprise business processes will access AI capabilities through the cloud, which is why building trust in AI is so important. The Cloud Native Computing Foundation hosts critical components of the global technology infrastructure.

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



*this is a random sample

Thank you

Rachel K. E. Bellamy
Chair, Exploratory Computer Science Council
—
rachel@us.ibm.com