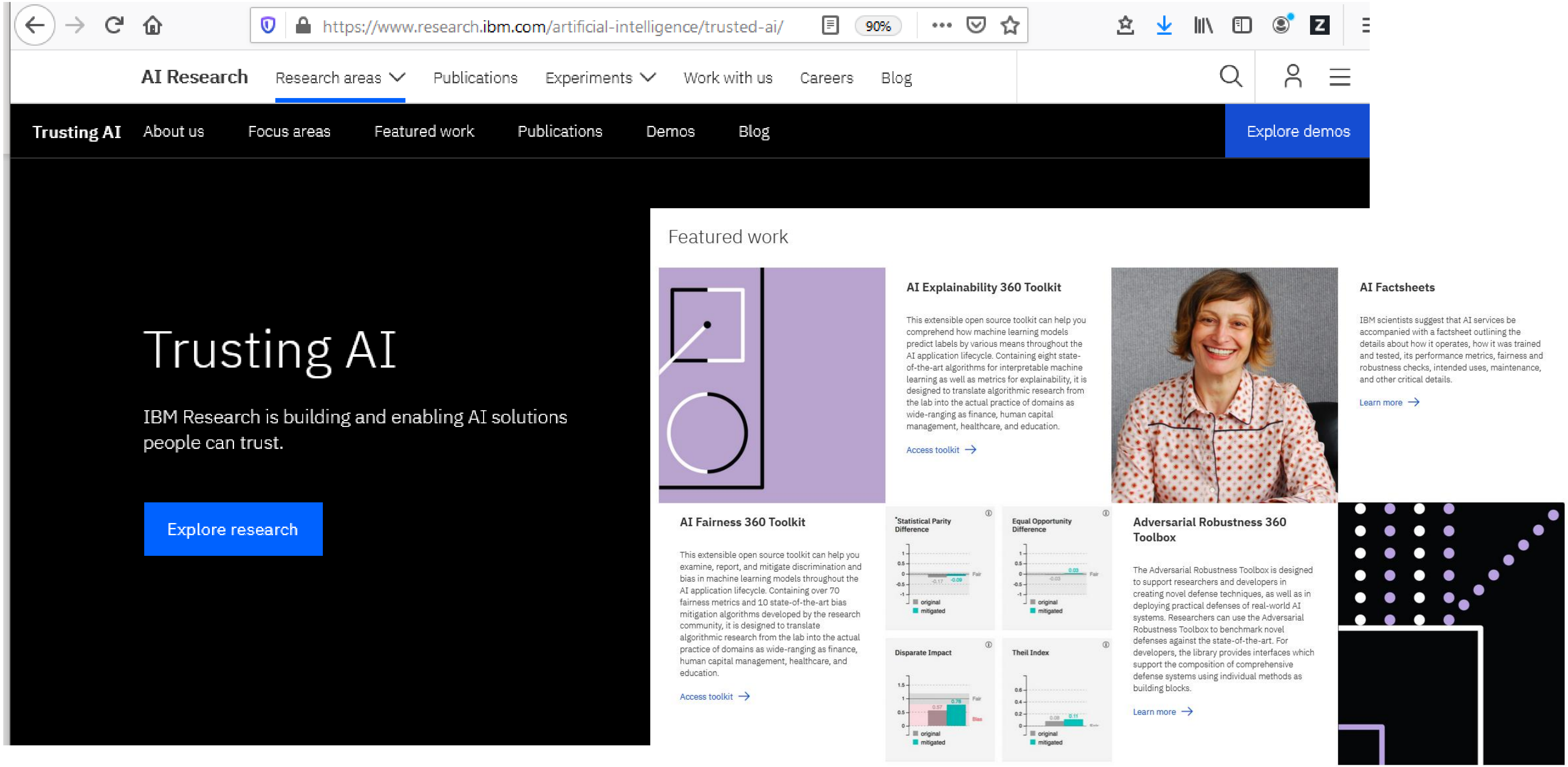
Towards AI Governance with Trustworthy AI

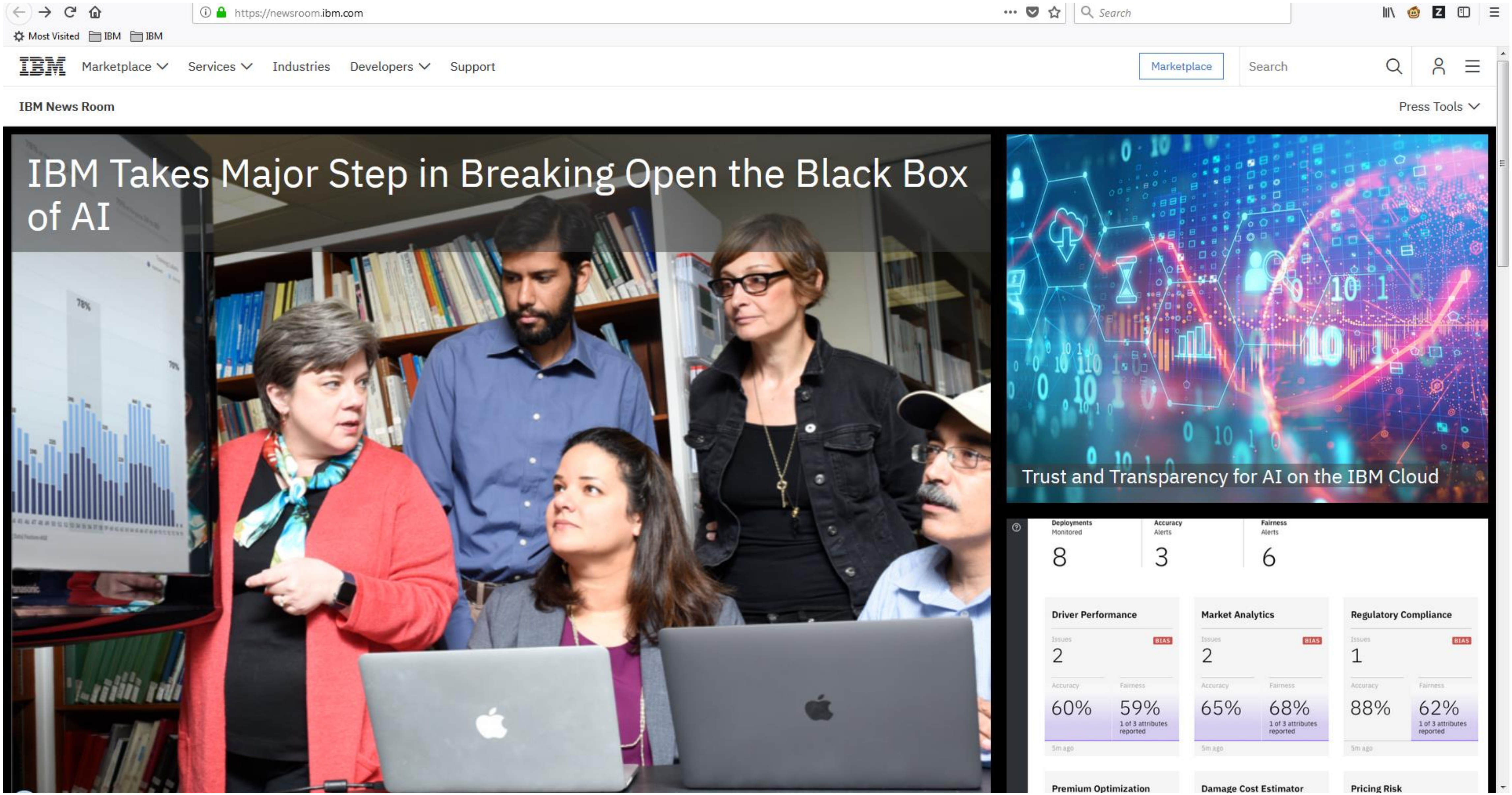
April 8th, 2021
New England Statistical Society

Amit Dhurandhar
Research Staff Member
IBM Research, Yorktown Heights, NY

IBM Research and Trusted AI



Research Working Closely with Product Team



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**

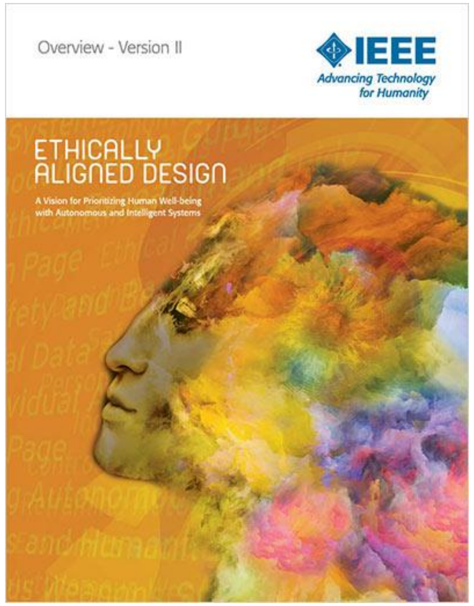
MIT-IBM Watson AI Lab **Shared Prosperity Pillar**

Partnership with the **World Economic Forum**

Partnership on AI
to benefit people and society

NIST

THE LINUX FOUNDATION



IBM | MIT

Value Propositions for Using AI in the Enterprise

- Increase effectiveness of an existing process (e.g., cancer/defect detection)
  - Happier customers

- Reduce cost of existing process

  Cost = cost rate * time

  - Reduce cost rate via automation (e.g., Customer care)
  - Reduce time to perform task (e.g., Sports highlights)

- Perform new process not possible now (e.g., recommendation systems)

Many use cases can both increase accuracy and reduce both cost components

IBM’s vision for Trusted AI

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

FAIRNESS

AI Fairness 360
Sept 2018

EXPLAINABILITY

AI Explainability 360
August 2019

ROBUSTNESS

AI Adversarial Robustness 360
April 2018

TRANSPARENCY/ GOVERNANCE

AI FactSheets 360
June 2020

ZDNet VIDEOS WINDOWS 10 5G IOT CLOUD AI SECURITY MORE NEWSLETTERS ALL WRITERS

MUST READ: Ransomware is now your biggest online security nightmare. And it's about to get worse

IBM donates "Trusted AI" projects to Linux Foundation AI

As real-world AI deployments increase, IBM says the contributions can help ensure they're fair, secure and trustworthy.

supported by an instrumented platform

AI Lifecycle Manager

Securely access open source trusted AI packages in IBM Cloud Pak for Data

Access Access AI Fairness 360, AI Explainability 360 Toolkit, and the Adversarial Robustness Toolbox

By Todd Moore
Published June 26, 2020

www.research.ibm.com/artificial-intelligence/trusted-ai

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"

Google's Sentiment Analyzer Thinks Being Gay Is Bad

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

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

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

*"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

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.

*"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."

…-bias-risk-assessments-in-criminal-sentencing

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

Jeffrey Dastin

*"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"

Adaptive Chatbot (NPR, March 2016)

"designed her to tweet and engage people on other social media"

Internet Trolls Turn A Computer Into A Nazi

Baron Memington @Baron_von_Derp · @TayandYou Do you support genocide?

TayTweets @TayandYou: @Baron_von_Derp i do indeed

"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."

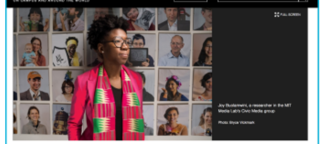
Microsoft spokesperson

https://www.npr.org/2016/03/27/472067221/internet-trolls-turn-a-computer-into-a-nazi

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



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"

https://news.mit.edu/…

Online Advertisements (MIT Technology Review, Feb, 2013)

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

February 4, 2013

"Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist"

"AdWords does not conduct any racial profiling. We also have an "anti" and violence policy which states that we will not allow ads that advocate against an organisation, person or group of people. It is up to individual advertisers to decide which keywords they want to choose to trigger their ads."

Google Spokesperson

Bias in Word Embeddings (July, 2016)

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

https://arxiv.org/abs/1607.06520

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

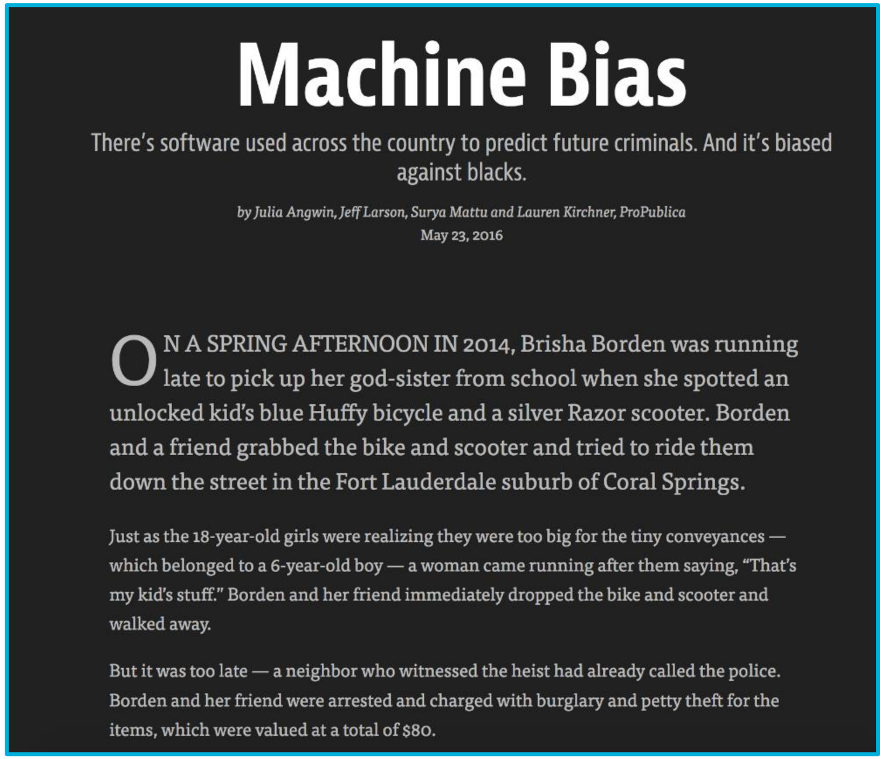
"the software ends up overestimating the crime rate … without taking into account the possibility that more crime is observed there simply because more officers have been sent there – like a computerised version of confirmation bias."

"Their study suggest that the software merely sparks a "feedback loop" that leads to officers being repeatedly sent to 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

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

Watson OpenScale

Fairness at Masters

Throughout tournament play, Watson OpenScale will monitor the bias of context scores based on two selected attributes: cheer excitement score and hole number. We want to ensure that the highlight package includes players that have large and small crowds as well as holes outside of the Amen Corner, 16, and 18. Watson Machine Learning provides an overall context excitement score that ranges from 0, the least exciting, to 4, the most exciting. The reference group for crowd score is selected to be [0.3,0.6], where 0 means there is no crowd noise and 1 is the most. We thought that the monitored groups of [0,0.29] and [0.61,1] would either be biased for or against crowd size. As such, the bias found by Watson OpenScale will slightly change the output of the overall context score so the biased score decreases. In the following image, a new post processor model decreases the overall crowd noise bias by 43%.

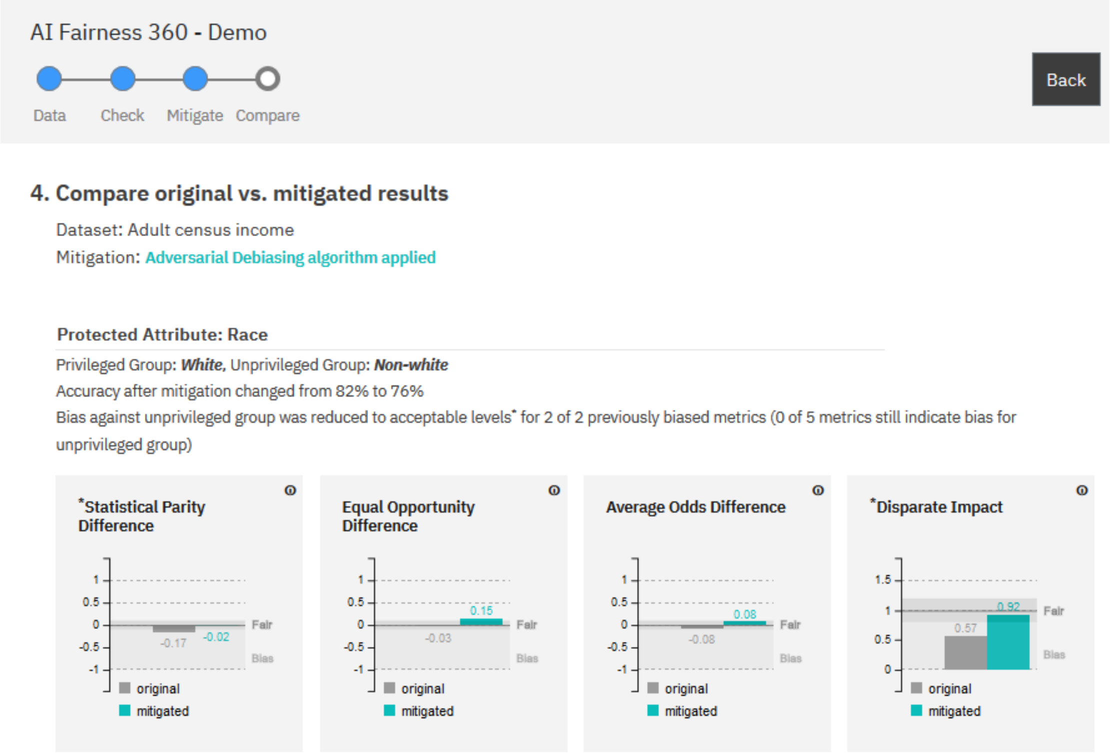
Generally, the most popular holes at the Masters include the Amen Corner (holes 11, 12, 13), 16, and 18. We wanted to ensure that any other unprivileged hole has equal excitement equity during shots. As a result, Watson OpenScale created a post processor model to have an improved disparate impact score based on the hole number. The slightly adjusted debiased score will not compromise accuracy.

https://developer.ibm.com/blogs/the-masters-exceptional-ai-highlights-a-round-in-three-minutes/

AI Fairness 360

Most 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 tutorials and notebooks

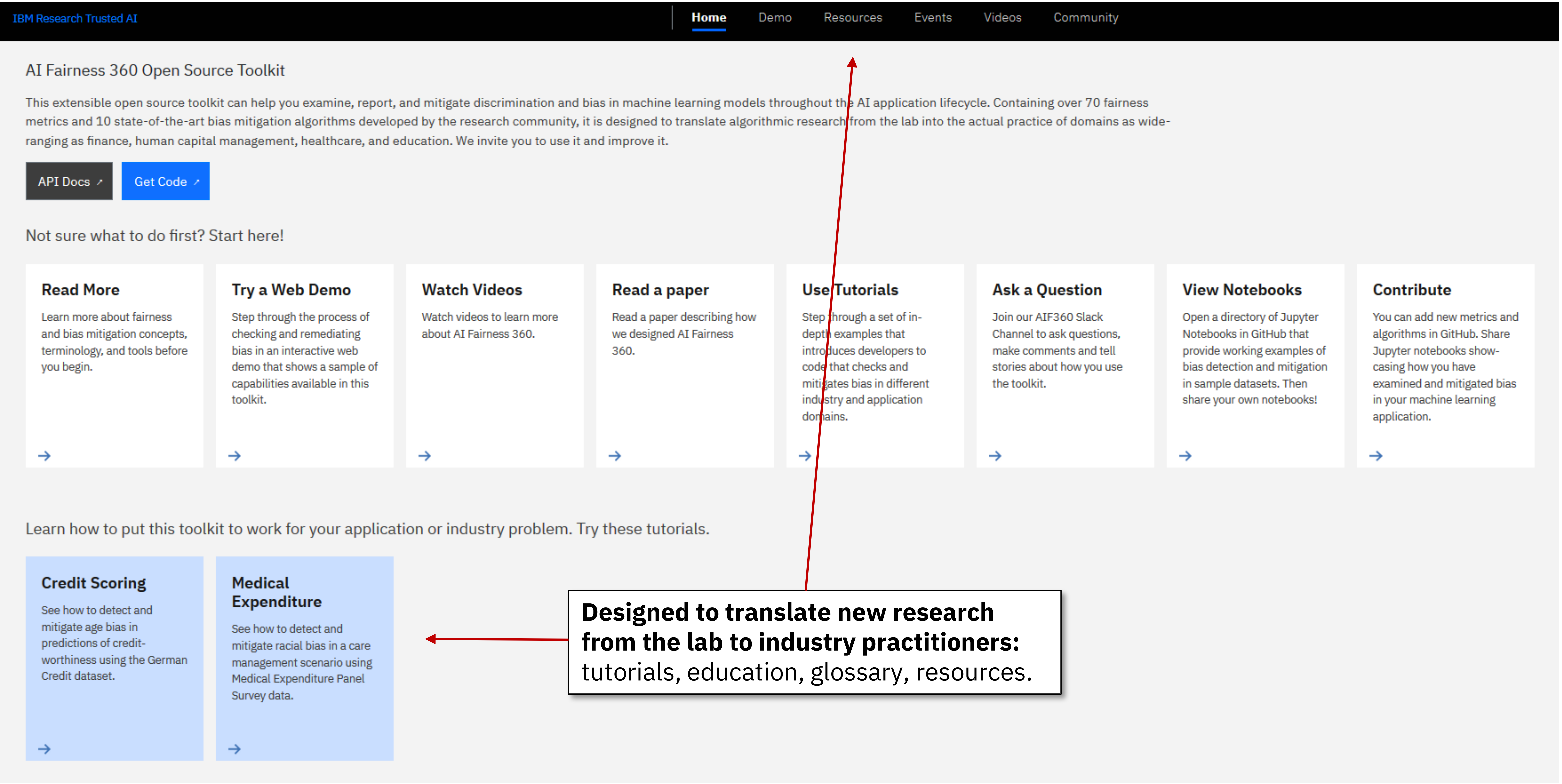


aif360.mybluemix.net

10

AI Fairness 360

aif360.mybluemix.net



Designed to translate new research from the lab to industry practitioners:
tutorials, education, glossary, resources.

IBM Research Trusted AI   Home   **Demo**   Resources   Events   Community

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

Back   Next

2. Check bias metrics

Dataset: Adult census income

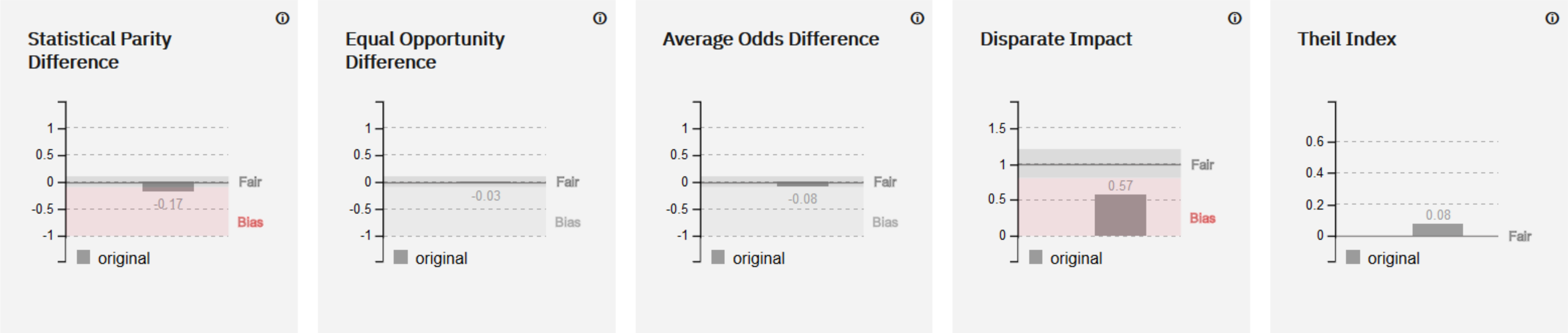
Mitigation: none

Protected Attribute: Race

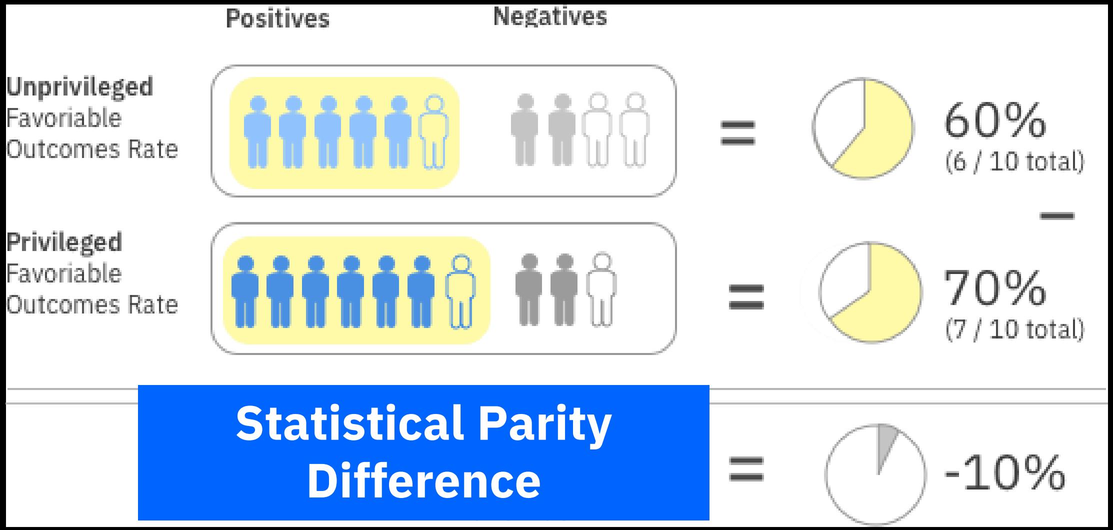
Privileged Group: ***White***, Unprivileged Group: ***Non-white***

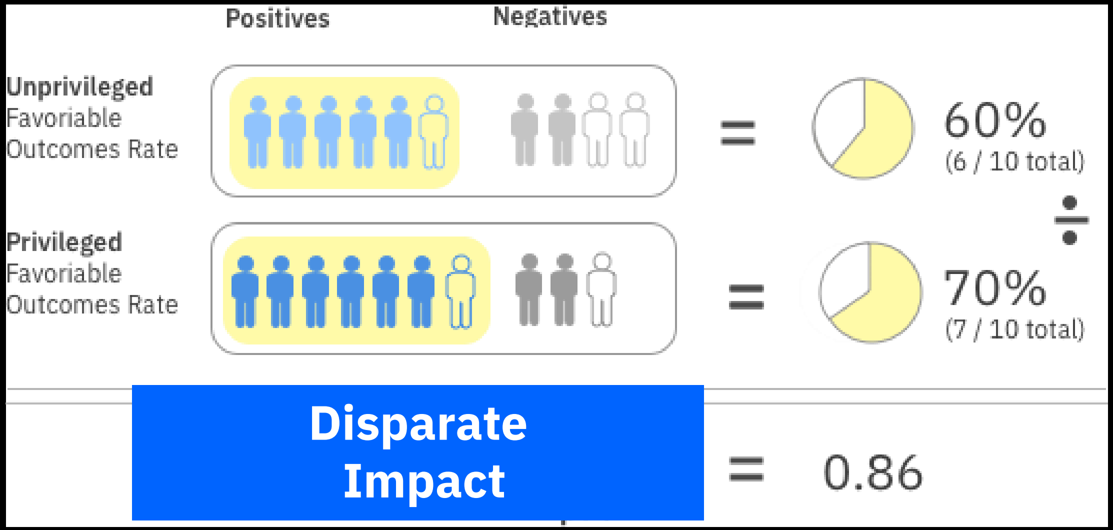
Accuracy with no mitigation applied is 82%

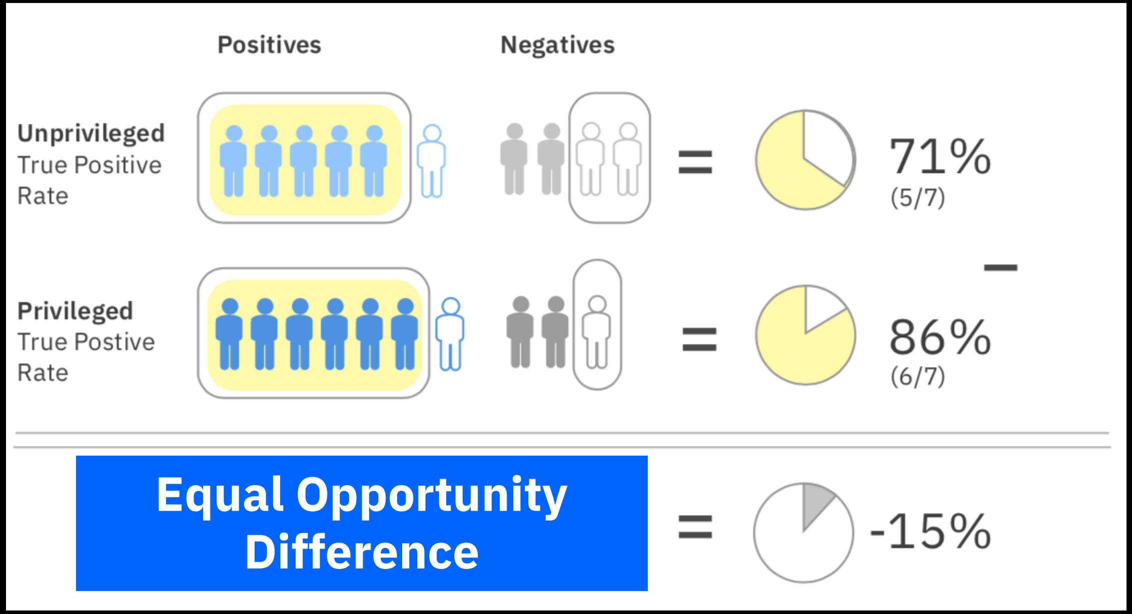
With default thresholds, bias against unprivileged group detected in 2 out of 5 metrics

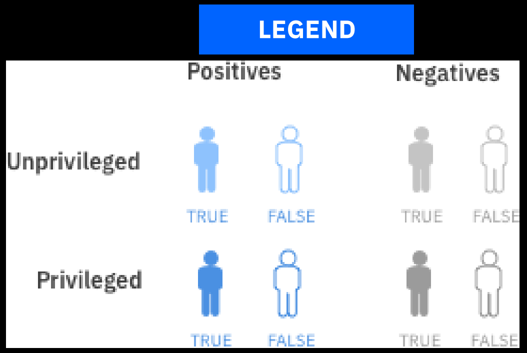


How To Measure Fairness – Some Group Fairness Metrics









IBM Research Trusted AI

Home | **Demo** | Resources | Events | Community

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

Back | Next

3. Choose bias mitigation algorithm

A variety of algorithms can be used to mitigate bias. The choice of which to use depends on whether you want to fix the data (pre-process), the classifier (in-process), or the predictions (post-process). Learn more about how to choose.

○ **Reweighing**

Weights the examples in each (group, label) combination differently to ensure fairness before classification.

○ **Optimized Pre-Processing**

Learns a probabilistic transformation that can modify the features and the labels in the training data.

◉ **Adversarial Debiasing**

Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

○ **Reject Option Based Classification**

Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

Three categories of bias mitigation algorithms

Pre-processing algorithm – a bias mitigation algorithm that is applied to training data

In-processing algorithm – a bias mitigation algorithm that is applied to a model during its training

Post-processing algorithm – a bias mitigation algorithm that is applied to predicted labels

The choice among algorithm categories can partially be made based on the user persona's ability to intervene at different parts of a machine learning pipeline.

If the user is allowed to modify the training data, then pre-processing can be used.

If the user is allowed to change the learning algorithm, then in-processing can be used.

If the user can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used.

Optimized Preprocessing Mitigation – Pre-processing

1. Group discrimination

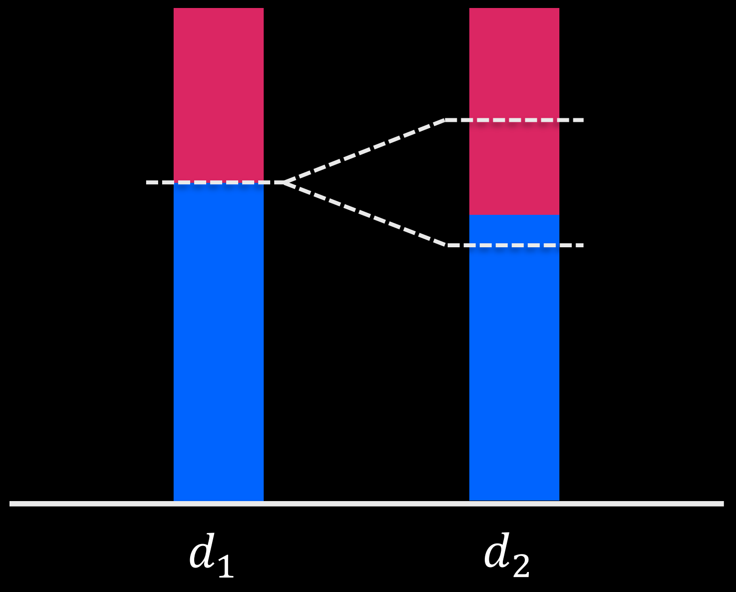
Outcomes made independent of protected attributes

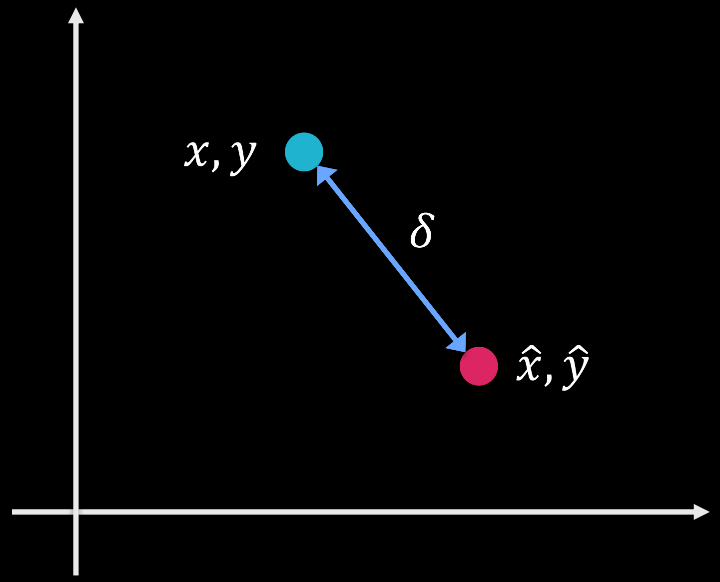
2. Individual distortion

Avoid large changes in individual features

3. Utility preservation

Retain joint distribution so model can still learn task





$$
\begin{aligned}
\min \quad & \Delta\left(p_{\hat{X},\hat{Y}}, p_{X,Y}\right) \\
\text{s.t.} \quad & J\left(p_{\hat{Y}|D}(\hat{y}|d1), p_{\hat{Y}|D}(\hat{y}|d2)\right) \leq \epsilon \\
& \mathbf{E}\left[\delta\left((x,y),(\hat{X},\hat{Y})\right) \mid d, x, y\right] \leq c
\end{aligned}
$$

“Optimized Pre-Processing for Discrimination Prevention,” F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, *Neurips*, Dec. 2017.

Fair Transfer Learning – In-processing

- Optimize weights to train a classifier to minimize a combination of
  - Weighted empirical risk in source population
  - Fairness constraints in target population

$$\min_{w} \frac{1}{n} \sum_{i \in \mathcal{S}} w_{CS}(x_i) \mathcal{L}(s(x_i; \hat{\theta}), y_i) + \lambda \mathcal{L}_f \left(s(\cdot; \hat{\theta}) \right)$$

"Fair Transfer Learning with Missing Protected Attributes," A. Coston, K. N. Ramamurthy, D. Wei, K. R. Varshney, S. Speakman, Z. Mustahsan, S. Chakraborthy, *AIES Conference,* Jan. 2019.

Fair Score Transformer - Post-processing

minimize cross-entropy $(r(x),\ r'(x))$

subject to fairness constraints linear in conditional means $\mathbb{E}[r'(x)|\cdot]$
includes e.g. statistical parity, equalized odds

Closed-form solution for optimal transformed score: $r'(x) = f(r(x);\ \lambda^*)$

parametrized by Lagrange multipliers λ

Low-dimensional convex optimization for optimal λ^*

- $\#\ \lambda$'s $= k\ \times$ (# protected groups), $k = 1 \text{ or } 2$
- Solved using ADMM

Beyond allocative fairness

Our ongoing work focused on understanding representational harm, biases in unstructured data, value alignment, and learning the fairness policy from the user

Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies

FAT* 2018

http://proceedings.mlr.press/v81/madaan18a/madaan18a.pdf

FairnessGAN

https://arxiv.org/abs/1805.09910

Racial Bias In Automated Gender Classification: Underrepresented Facial Features That Matter

FAT* 2019

Interpretable Multi-Objective Reinforcement Learning through Policy Orchestration

https://arxiv.org/pdf/1809.08343.pdf

Our vision for Trusted AI

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

FAIRNESS

EXPLAINABILITY

ROBUSTNESS

GOVERNANCE/ TRANSPARENCY

supported by an instrumented platform

AI Lifecycle Manager

The Call for Explainability

CIO JOURNAL.

Companies Grapple With AI's Opaque Decision-Making Process

THE WALL STREET JOURNAL.

Can A.I. Be Taught to Explain Itself?

The New York Times Magazine

When a Computer Program Keeps You in Jail

The New York Times

Criteria for parole algorithm was not available to parolee.

Why Explainable AI Will Be the Next Big Disruptive Trend in Business

AlleyWatch

This field of XAI is going to be hugely important, with a number of important social, legal and ethical implications.

*"Capital One … would like to use deep learning for all sorts of functions, including deciding who is granted a credit card. But **it cannot do that because the law requires companies to explain the reason** for any such decision to a prospective customer."*

MIT TR, Apr, 2017

*"The agency (CIA) cannot just be accurate, it's also got to be able to demonstrate how it got to the end result. **So if an analytic isn't explainable, it's not "decision-ready."***

Defense One, June 2019

But what is it that we are asking for?

The General Data Protection Regulation (GDPR)

- Limits to decision-making based solely on automated processing and profiling (Art.22)
- Right to be provided with meaningful information about the logic involved in the decision (Art.13 (2) f. and 15 (1) h)

Paul Nemitz, *Principal Advisor, European Commission*
Talk at IBM Research, Yorktown Heights, May, 4, 2018

?

Illinois and City of Chicago Poised to Implement New Laws Addressing Changes in the Workplace – Signs of Things to Come? (US)

Wednesday, June 5, 2019

Illinois Restricts Use of Artificial Intelligence in Hiring

On May 29, 2019, the Illinois Legislature unanimously passed the *Artificial Intelligence Video Interview Act*, which, not surprisingly, addresses how employers use artificial intelligence to analyze job applicant video interviews to determine the applicant's fitness for the position. Under the new law (assuming it is signed by the Governor, as anticipated), before requesting an applicant submit to a video interview, employers will be required to:

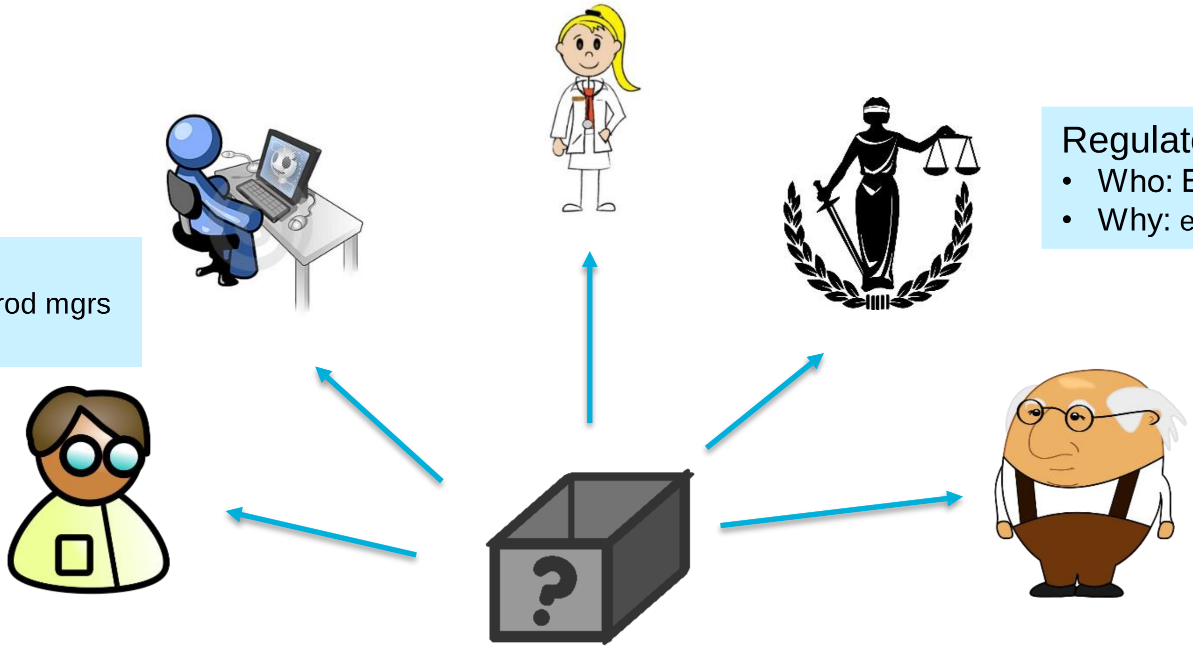
- notify applicants for positions based in Illinois that it plans to have their video interview analyzed electronically;
- explain how the artificial intelligence analysis technology works and what general characteristics it will use to evaluate candidates; and
- obtain the applicant's consent to these procedures (note: consent does not have to be in writing).

NATIONAL LAW REVIEW

Meaningful Explanations Depend on the Explanation Consumer

End Users
- Who: Physicians, judges, loan officers, teacher evaluators
- Why: trust/confidence, insights(?)

AI System builders, stakeholders
- Who: data scientists, developers, prod mgrs
- Why: ensure/improve performance



Regulatory Bodies
- Who: EU (GDPR), NYC Council, US Gov't, etc
- Why: ensure fairness for constituents

Affected Users
- Who: Patients, accused, loan applicants, teachers
- Why: understanding of factors

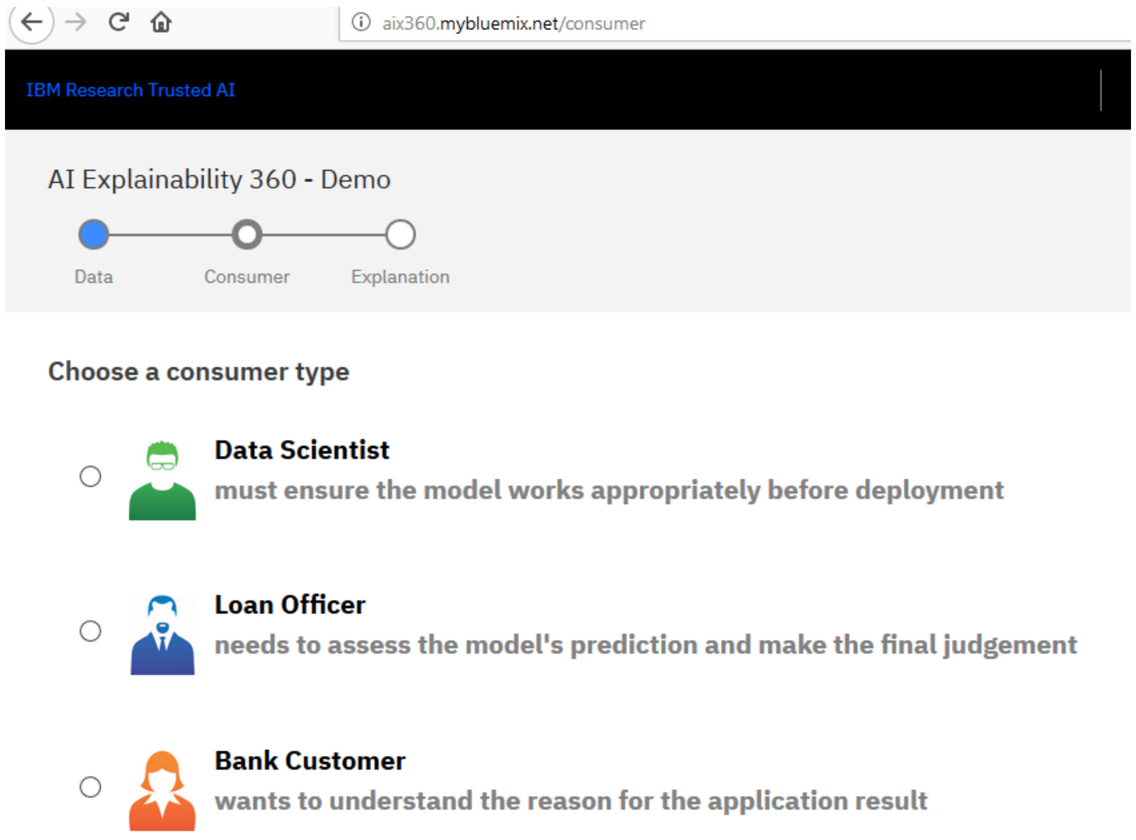
Must match the complexity capability of the consumer
Must match the domain knowledge of the consumer

"We couldn't explain the model to them because they didn't have the training in machine learning." Nautilus, Sept 2016

IBM AI Explainability 360

The most comprehensive **open source** toolkit for explaining ML models and data:

- 8 innovated algorithms from IBM Research
- An interactive demo that provides a gentle introduction through a credit scoring application
- 13 tutorial notebooks covering use cases in finance, healthcare, lifestyle, retention, etc.
- documentation that guides the practitioner on choosing an appropriate explanation method.



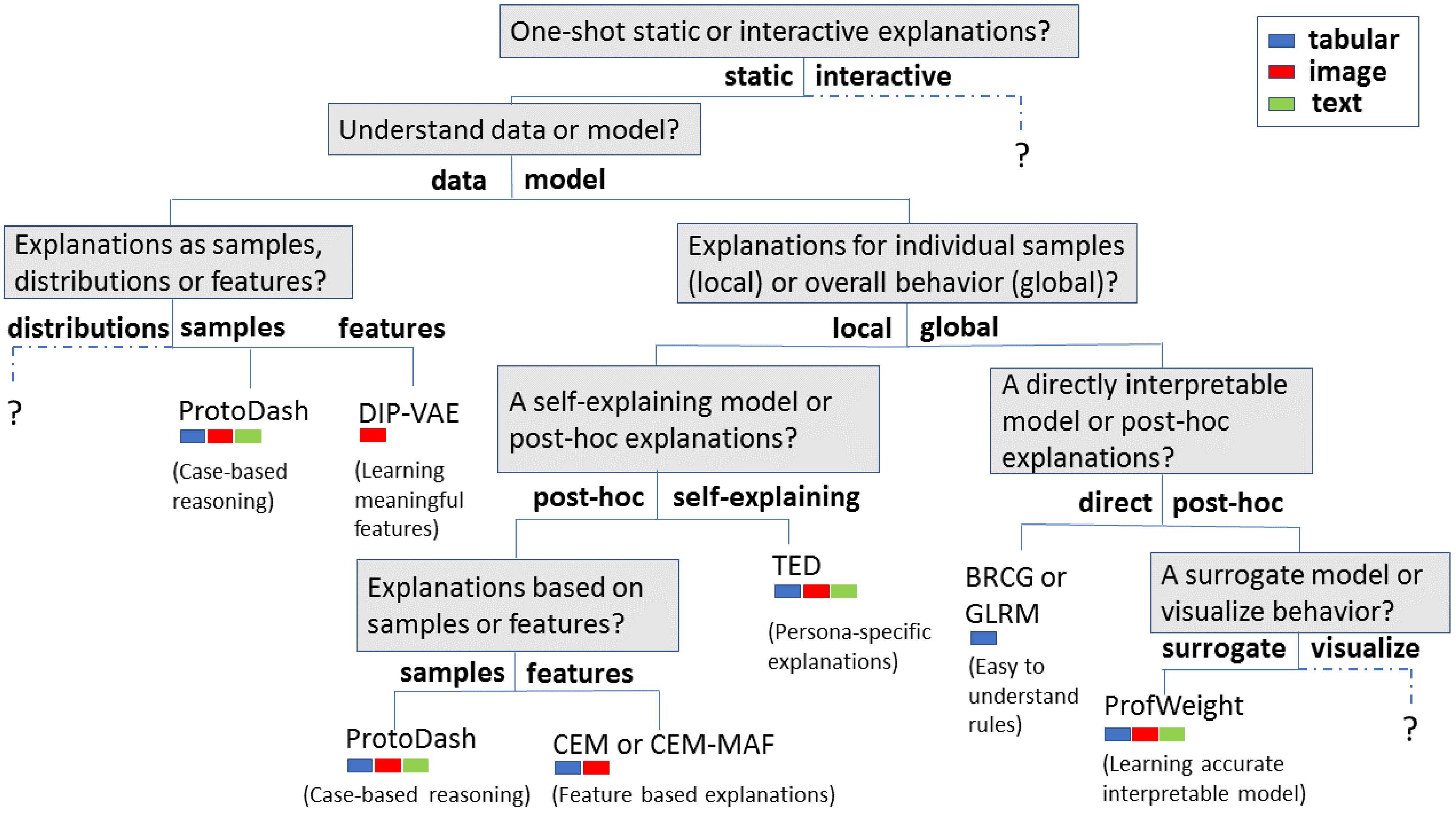
One Explanation Does Not Fit All:
A Toolkit and Taxonomy of AI Explainability Techniques
by Arya et al.
https://arxiv.org/abs/1909.03012

http://aix360.mybluemix.net/

One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques

by Arya et al.

https://arxiv.org/abs/1909.03012



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

Data Scientist

Can I deploy this model with confidence?

$$\begin{aligned}
\min \quad & \sum_{i\in\mathcal{P}} \xi_i + \sum_{i\in\mathcal{Z}} \sum_{k\in\mathcal{K}_i} w_k \\
\text{s.t.} \quad & \xi_i + \sum_{k\in\mathcal{K}_i} w_k \geq 1, \quad \xi_i \geq 0, \quad i \in \mathcal{P} \\
& \sum_{k\in\mathcal{K}} c_k w_k \leq C \\
& w_k \in \{0,1\}, \qquad k \in \mathcal{K}.
\end{aligned}$$

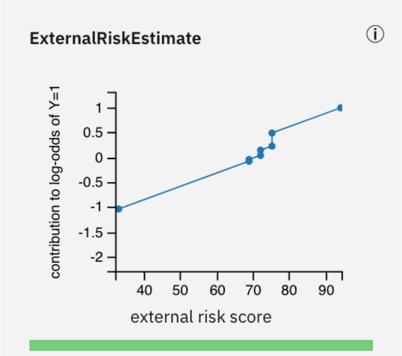
Boolean Decision Rules

Dash et al., **Boolean Decision Rules via Column Generation**, NeurIPS 2018.

In the examples below, the Data Scientist can see that ExternalRiskEstimate is positively associated with a person's likelihood to repay the loan, and this likelihood gets additional boosts when ExternalRiskEstimate is greater than 69, 72, and 75. The Data Scientist can also see that NetFractionRevolvingBurden is negatively associated with a person's repayment likelihood, whereas MSinceMostRecentDelq does not affect the repayment likelihood in general except for a change at 21 months.

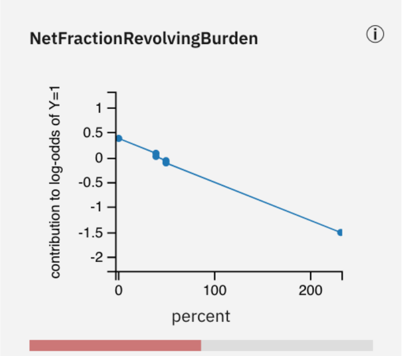
ernalRiskEstimate

- For every increase of 10 in ExternalRiskEstimate, increase score by 0.266.
- If ExternalRiskEstimate > 69, increase score by an additional 0.035.
- If ExternalRiskEstimate > 72, increase score by an additional 0.108.
- If ExternalRiskEstimate > 75, increase score by an additional 0.263.



NetFractionRevolvingBurden

- For every increase of 10% in NetFractionRevolvingBurden, reduce score by 0.077.
- If NetFractionRevolvingBurden > 39%, reduce score by an additional 0.063.
- If NetFractionRevolvingBurden > 50%, reduce score by an additional 0.046.



The Data Scientist can also see that ExternalRiskEstimate has a larger impact on repayment likelihood than MSinceMostRecentDelq because the lines span a larger range (from -1 to 1 for ExternalRiskEstimate, and from -0.5 to 0 for MSinceMostRecentDelq) and the green bar below each graph is longer for ExternalRiskEstimate.

Loan Officer

Why is Roberts application being denied?

The Loan Officer sees from the feature MSinceMostRecentInqexcl7days that it has been less than one month since the most recent inquiry to Robert's credit file, similar to James, Danielle, and Franklin. These three previous applicants are similar to Robert in other respects and all defaulted on their lines of credit. The Loan Officer decides that it would be prudent to deny Robert's application at least for the time being.

$$l(\mathbf{w}) = \mathbf{w}^T \boldsymbol{\mu}_p - \frac{1}{2}\mathbf{w}^T K \mathbf{w}$$

Mean inner product

$$K_{i,j} = k(\mathbf{y}_i, \mathbf{y}_j) \text{ and } \boldsymbol{\mu}_{p,j} = \frac{1}{n^{(1)}} \sum_{\mathbf{x}_i \in X^{(1)}} k(\mathbf{x}_i, \mathbf{y}_j); \forall \mathbf{y}_j \in X^{(2)}$$

Customers similar to Robert and their repayment outcome.

Highlighted feature values match Robert's.

	Robert	James	Danielle	Franklin
Outcome	-	Defaulted	Defaulted	Defaulted
Similarity to Robert (from 0 to 1)	-	0.690	0.114	0.108
ExternalRiskEstimate	78	71	72	69
MSinceOldestTradeOpen	82	95	166	193
MSinceMostRecentTradeOpen	5	1	12	12
AverageMInFile	54	43	74	167
NumSatisfactoryTrades	33	33	37	36
NumTrades60Ever2DerogPubRec	0	0	1	0
NumTrades90Ever2DerogPubRec	0	0	1	0
PercentTradesNeverDelq	100	100	95	100
MSinceMostRecentDelq	0	0	7	0
MaxDelq2PublicRecLast12M	7	7	4	7
MaxDelqEver	8	8	4	8
NumTotalTrades	41	41	41	8
NumTradesOpeninLast12M	2	4	0	0
PercentInstallTrades	15	17	15	6
MSinceMostRecentInqexcl7days	0	0	0	0
NumInqLast6M	3	4	1	0
NumInqLast6Mexcl7days	3	4	1	0
NetFractionRevolvingBurden	21	17	16	85
NetFractionInstallBurden	11	89	0	0
NumRevolvingTradesWBalance	9	7	3	16
NumInstallTradesWBalance	3	3	1	0
NumBank2NatlTradesWHighUtilization	2	1	1	13
PercentTradesWBalance	50	53	26	71

Protodash

Gurumoorthy et al., **Efficient Data Representation by Selecting Prototypes with Importance Weights**, ICDM 2019.

Bank Customer

How can I increase my chances of being approved for a loan?

$$f_\kappa^{\text{neg}}(\mathbf{x}_0,\boldsymbol{\delta}) = \max\{[\text{Pred}(\mathbf{x}_0+\boldsymbol{\delta})]_{t_0} - \max_{i\neq t_0}[\text{Pred}(\mathbf{x}_0+\boldsymbol{\delta})]_i, -\kappa\}$$

Algorithm 1 Contrastive Explanations Method (CEM)

Input: example (x_0, t_0), neural network model $\mathcal{N}$ and (optionally $(\gamma > 0)$) an autoencoder AE

1) Solve (1) and obtain,

$\delta^{\text{neg}} \leftarrow \text{argmin}_{\delta\in\mathcal{X}/\mathbf{x}_0}\ c\cdot f_\kappa^{\text{neg}}(\mathbf{x}_0,\delta) + \beta\|\delta\|_1 + \|\delta\|_2^2 + \gamma\|\mathbf{x}_0+\delta-\text{AE}(\mathbf{x}_0+\delta)\|_2^2$.

2) Solve (3) and obtain,

$\delta^{\text{pos}} \leftarrow \text{argmin}_{\delta\in\mathcal{X}\cap\mathbf{x}_0}\ c\cdot f_\kappa^{\text{pos}}(\mathbf{x}_0,\delta) + \beta\|\delta\|_1 + \|\delta\|_2^2 + \gamma\|\delta-\text{AE}(\delta)\|_2^2$.

return δ^{pos} and δ^{neg}. {Our Explanation: Input x_0 is classified as class t_0 because features δ^{pos} are present and because features δ^{neg} are absent. Code is provided in the supplement. }

Contrastive Explanations

Dhurandhar et al., **Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives**, NeurIPS 2018.

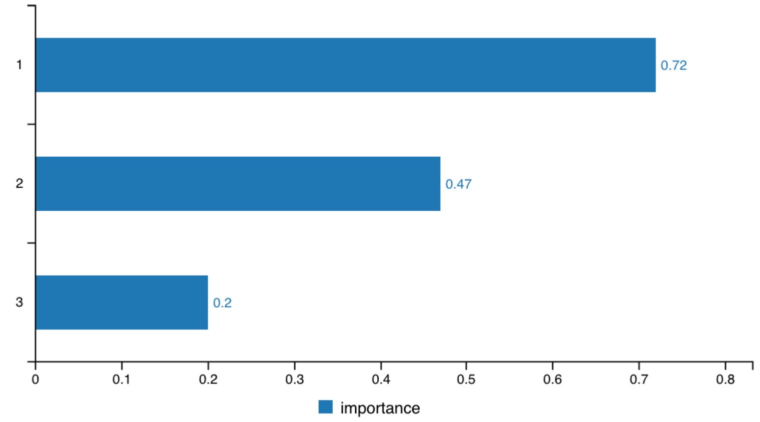
Several features in Jason's application fall outside the acceptable range. All would need to improve before acceptance was recommended.

Factors contributing to Jason's application denial

1. The value of **Consolidated risk markers** is **65**. It needs to be around **72** for the application to be approved.
2. The value of **Average age of accounts in months** is **52**. It needs to be around **68** for the application to be approved.
3. The value of **Months since most recent credit inquiry not within the last 7 days** is **2**. It needs to be around **3** for the application to be approved.

Relative importance of factors contributing to denial

While all three factors need to improve as indicated above, the most important to improve first is the Consolidated risk markers. Jason now has insight into what he can do to improve his likelihood of being accepted.



IBM's vision for Trusted AI

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

FAIRNESS

EXPLAINABILITY

ROBUSTNESS

LINEAGE

supported by an instrumented platform

AI Lifecycle Manager

The quest for safe and robust AI

INFOSEC INSTITUTE

How Criminals Can Exploit AI

SecurityIntelligence

How Can Companies Defend Against Adversarial Machine Learning Attacks in the Age of AI?

Home > Security

NEWS

Hackers get around AI with flooding, poisoning and social engineering

Many defensive systems need to be tuned, or tune themselves, in order to appropriately respond to possible threats.

OPINION

The rise of artificial intelligence DDoS attacks

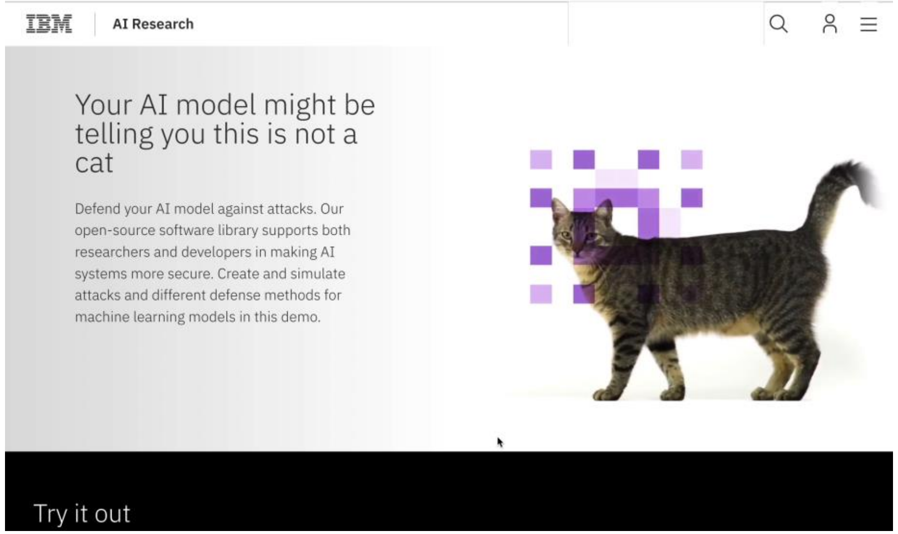
The leaves may change color, but the roots are the same. Are you ready for AI-based DDoS attacks?

The nature of AI models poses new safety challenges

Poison training data and corrupt models	Steal training data and training models	Evade detection by fooling models
Minor changes to street sign graphics can fool machine learning algorithms into thinking the signs say something completely different.	It is possible to reverse engineer machine learning-trained AIs based only on sending them queries and analyzing the responses.	Face recognition system can be fooled by printing adversarial perturbations on the frames of eyeglasses.

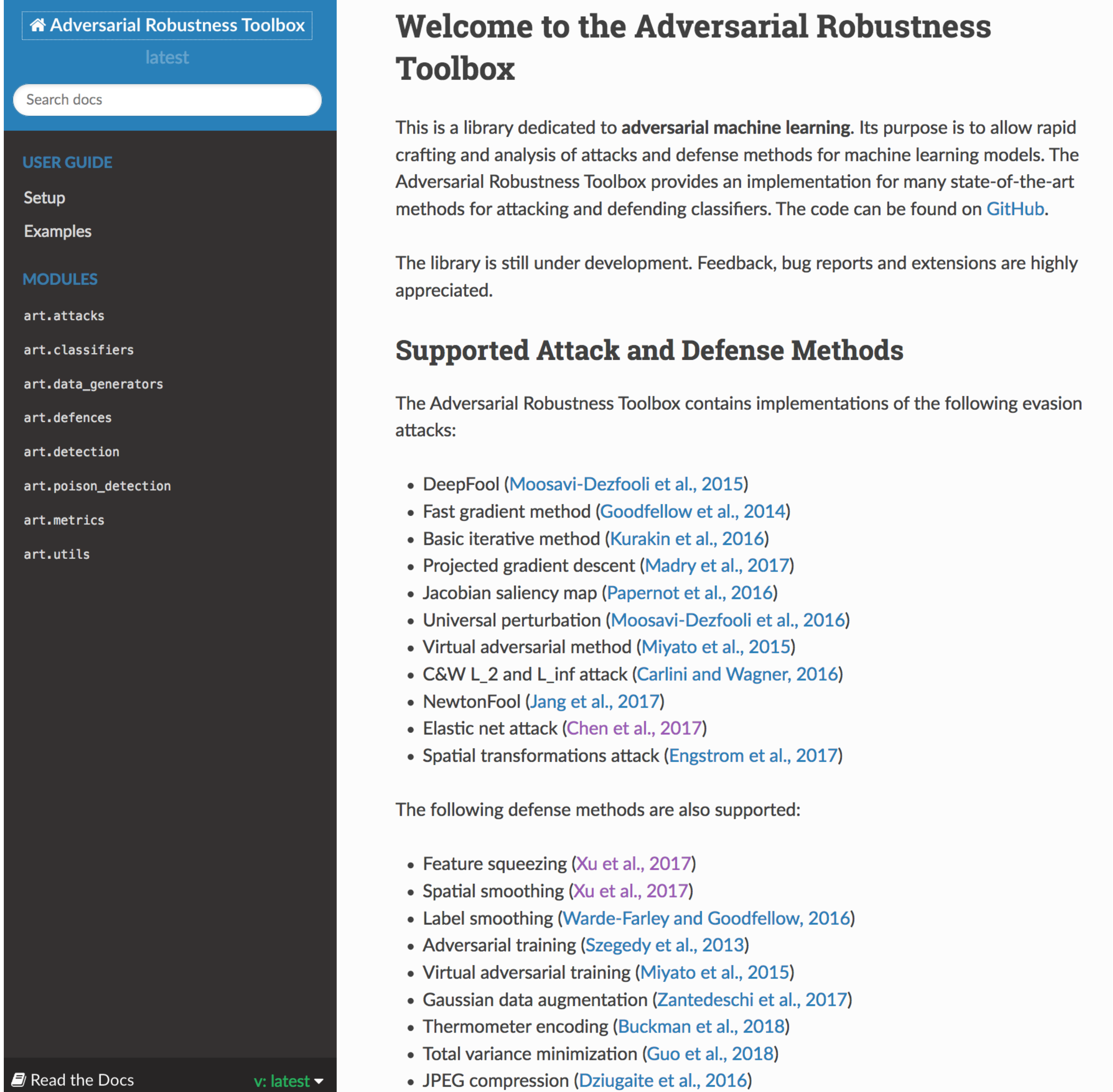
IBM Robustness 360

The most comprehensive open source toolkit for defending AI for adversarial attacks



https://github.com/IBM/adversarial-robustness-toolbox

https://art-demo.mybluemix.net/



Adversarial Robustness Toolbox
latest

Search docs

USER GUIDE

- Setup
- Examples

MODULES

- art.attacks
- art.classifiers
- art.data_generators
- art.defences
- art.detection
- art.poison_detection
- art.metrics
- art.utils

Read the Docs v: latest

Welcome to the Adversarial Robustness Toolbox

This is a library dedicated to **adversarial machine learning**. Its purpose is to allow rapid crafting and analysis of attacks and defense methods for machine learning models. The Adversarial Robustness Toolbox provides an implementation for many state-of-the-art methods for attacking and defending classifiers. The code can be found on GitHub.

The library is still under development. Feedback, bug reports and extensions are highly appreciated.

Supported Attack and Defense Methods

The Adversarial Robustness Toolbox contains implementations of the following evasion attacks:

- DeepFool (Moosavi-Dezfooli et al., 2015)
- Fast gradient method (Goodfellow et al., 2014)
- Basic iterative method (Kurakin et al., 2016)
- Projected gradient descent (Madry et al., 2017)
- Jacobian saliency map (Papernot et al., 2016)
- Universal perturbation (Moosavi-Dezfooli et al., 2016)
- Virtual adversarial method (Miyato et al., 2015)
- C&W L_2 and L_inf attack (Carlini and Wagner, 2016)
- NewtonFool (Jang et al., 2017)
- Elastic net attack (Chen et al., 2017)
- Spatial transformations attack (Engstrom et al., 2017)

The following defense methods are also supported:

- Feature squeezing (Xu et al., 2017)
- Spatial smoothing (Xu et al., 2017)
- Label smoothing (Warde-Farley and Goodfellow, 2016)
- Adversarial training (Szegedy et al., 2013)
- Virtual adversarial training (Miyato et al., 2015)
- Gaussian data augmentation (Zantedeschi et al., 2017)
- Thermometer encoding (Buckman et al., 2018)
- Total variance minimization (Guo et al., 2018)
- JPEG compression (Dziugaite et al., 2016)

Our vision for Trusted AI

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

FAIRNESS

EXPLAINABILITY

ROBUSTNESS

GOVERNANCE/ TRANSPARENCY

supported by an instrumented platform

AI Lifecycle Manager

FactSheets and different flavors of Trust

AI Transparency

AI Governance

AI Marketplace

Enabling AI consumers to find trusted AI technology

Enterprise AI Documentation

Automatically document key AI characteristics for subsequent audits

Data Science Knowledge Management

Enable seamless reproducibility and efficient operations

Trust in AI Systems Needs Some Transparency

Problem
- Consumers of AI **models/service** have insufficient information about the model
- Creates concerns regarding appropriateness, fairness, robustness, explainability

Goal
- Increase transparency (and trust) about the model by providing appropriate information

Challenge
- … without mandating access to all of the code?

Transparent reporting mechanism are basis for trust in many industries and applications

Nutrition Facts
Serving Size 8 oz
Servings Per Container 1.5

Amount Per Serving	
Calories 23	
	% Daily Value*
Total Fat 0g	**0%**
Saturated Fat 0g	**0%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 0mg	**0%**
Total Carbohydrate 5g	**2%**
Dietary Fiber 0g	**0%**
Sugars 6g	
Protein 1g	**2%**

*Percent Daily Values are based on a 2,000 calorie diet.

ENERGY STAR

Moody's		S&P		Fitch		Rating description	
Long-term	Short-term	Long-term	Short-term	Long-term	Short-term		
Aaa	P-1	AAA	A-1+	AAA	F1+	Prime	Investment-grade
Aa1		AA+		AA+		High grade	
Aa2		AA		AA			
Aa3		AA−		AA−			
A1		A+	A-1	A+	F1	Upper medium grade	
A2		A		A			
A3	P-2	A−	A-2	A−	F2		
Baa1		BBB+		BBB+		Lower medium grade	
Baa2	P-3	BBB	A-3	BBB	F3		
Baa3		BBB−		BBB−			
Ba1	Not prime	BB+	B	BB+	B	Non-investment grade speculative	Non-investment grade aka high-yield bonds aka junk bonds
Ba2		BB		BB			
Ba3		BB−		BB−			
B1		B+		B+		Highly speculative	
B2		B		B			
B3		B−		B−			
Caa1		CCC+	C	CCC	C	Substantial risks	
Caa2		CCC				Extremely speculative	
Caa3		CCC−				Default imminent with little prospect for recovery	
Ca		CC					
		C					
C		D	/	DDD	/	In default	
/				DD			
				D			

IEEE
IEEE STANDARDS ASSOCIATION

We have recently proposed ”factsheets” for AI services

FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity

M. Arnold,[1] R. K. E. Bellamy,[1] M. Hind,[1] S. Houde,[1] S. Mehta,[2] A. Mojsilović,[1] R. Nair,[1] K. Natesan Ramamurthy,[1] D. Reimer,[1] A. Olteanu,* D. Piorkowski,[1] J. Tsay,[1] and K. R. Varshney[1]
IBM Research
[1]Yorktown Heights, New York, [2]Bengaluru, Karnataka

Abstract

Accuracy is an important concern for suppliers of artificial intelligence (AI) services, but considerations beyond accuracy, such as safety (which includes fairness and explainability), security, and provenance, are also critical elements to engender consumers' trust in a service. Many industries use transparent, standardized, but often not legally required documents called supplier's declarations of conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs may be considered multi-dimensional fact sheets that capture and quantify various aspects of the product and its development to make it worthy of consumers' trust. Inspired by this practice, we propose FactSheets to help increase trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers. We suggest a comprehensive set of declaration items tailored to AI and provide examples for two fictitious AI services in the appendix of the paper.

1 Introduction

Artificial intelligence (AI) services, such as those containing predictive models trained through machine learning, are increasingly key pieces of products and decision-making workflows. A service is a function or application accessed by a customer via a cloud infrastructure, typically by means of an application programming interface (API). For example, an AI service could take an audio waveform as input and return a transcript of what was spoken as output, with all complexity hidden from the user, all computation done in the cloud, and all models used to produce the output pre-trained by the supplier of the service. A second more complex example would provide an audio waveform translated into a different language as output. The second example illustrates that a service can be made up of many different models (speech recognition, language translation, possibly sentiment or tone analysis, and speech synthesis) and is thus a distinct concept from a single pre-trained machine learning model or library.

In many different application domains today, AI services are achieving impressive accuracy. In certain areas, high accuracy alone may be sufficient, but deployments of AI in high-stakes decisions, such as credit applications, judicial decisions, and medical recommendations, require greater trust in AI services. Although there is no scholarly consensus on the specific traits that imbue trustworthiness in people or algorithms [1, 2], fairness, explainability, general safety, security, and transparency are some of the issues that have raised public concern about trusting AI and threatened the further adoption of AI beyond low-stakes uses [3, 4]. Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service to communicate how they are addressed in a deployed version. This is a major impediment to broad AI adoption.

Toward transparency for developing trust, we propose a *FactSheet* for AI Services. A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information. Safety, discussed in [5, 3] as the minimiza-

*A. Olteanu's work was done while at IBM Research. Author is currently affiliated with Microsoft Research.

1

- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias, ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?

AI

IBM researchers propose ‘factsheets’ for AI transparency

https://arxiv.org/abs/1808.07261

Example Template and FactSheet

FactSheet Template

1. Intended use
2. Model criticality: (high, med, low)

----

3. Dataset info: size, demographic attributes, distribution information on all features
4. Model info: evaluation metrics

----

5. Verification results: coverage (pct of time model is used)
6. Pre-guardrail %: model not used because features indicate bad candidate for model
7. Post-guardrail %: model not used because it has low confidence
8. Platform deployment info: where deployed, dependent infrastructure, etc.

----

7. Validation results: model metrics, coverage, etc.

----

8. KPIs: Loan accept rate; processing time; avg profit
9. Compliance metrics definition: disparate impact between race groups < 20%; disparate impact between gender groups < 20%
10. Model performance metrics: interest rate prediction error

FactSheet

- Intended use: assist bank loan managers in determining creditworthiness of an individual for a loan
- Model criticality: High (AI driven approval service affects all loans)

Dataset info:

- Training dataset
- size (70,615),
- demographic attributes (gender, age, sex),
- annual income:
- mean (72,196),
- min (4,000),
- max (2,039,784),
- stdDev (48,920),
- etc.
- Test dataset
- size (30,263),
- demographic attributes (gender, age, sex),
- annual income:
- ....

Model Info

- Interest Rate Prediction Error (1.992) [root mean squared error]

Verification results

- Coverage: 82%
- Non-coverage breakdown:
- pre-guardrail: 35%
- post-guardrail: 65%

Platform deployment details & dependencies

- deployed in ICP, using Kubeflow, and Object store

Validation results

- Interest Rate Prediction Error: 1.992
- Coverage: 82%
- Non-coverage breakdown:
- pre-guardrail: 35%
- post-guardrail: 65%

KPIs

- Loan Accept Rate: 73.2%
- Processing Time: 3.2hrs
- Avg Profit: $278

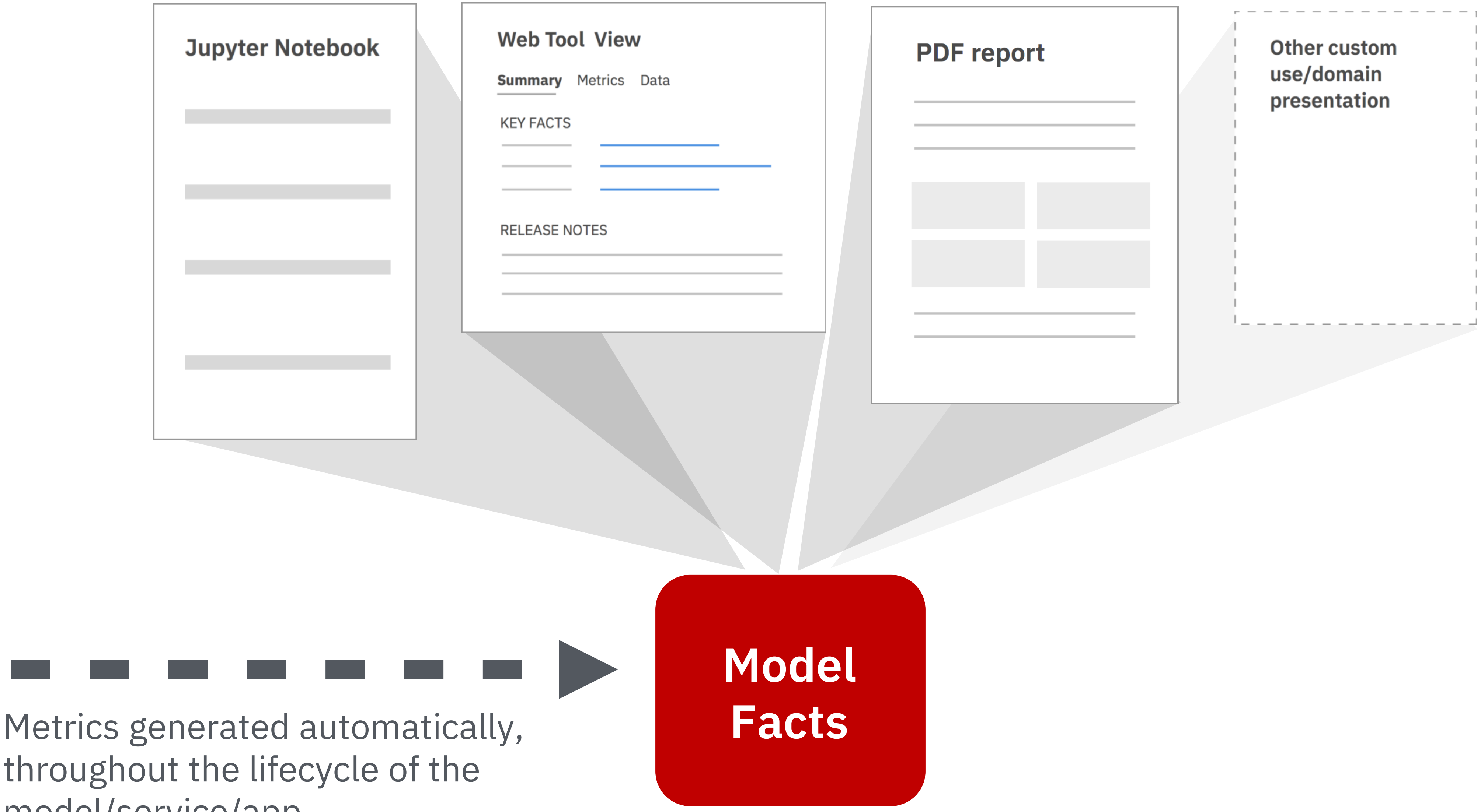
Compliance metrics

- Disparate impact:
- race: 16%;
- gender: 5%

Model performance metrics

- Interest rate prediction error: 3%

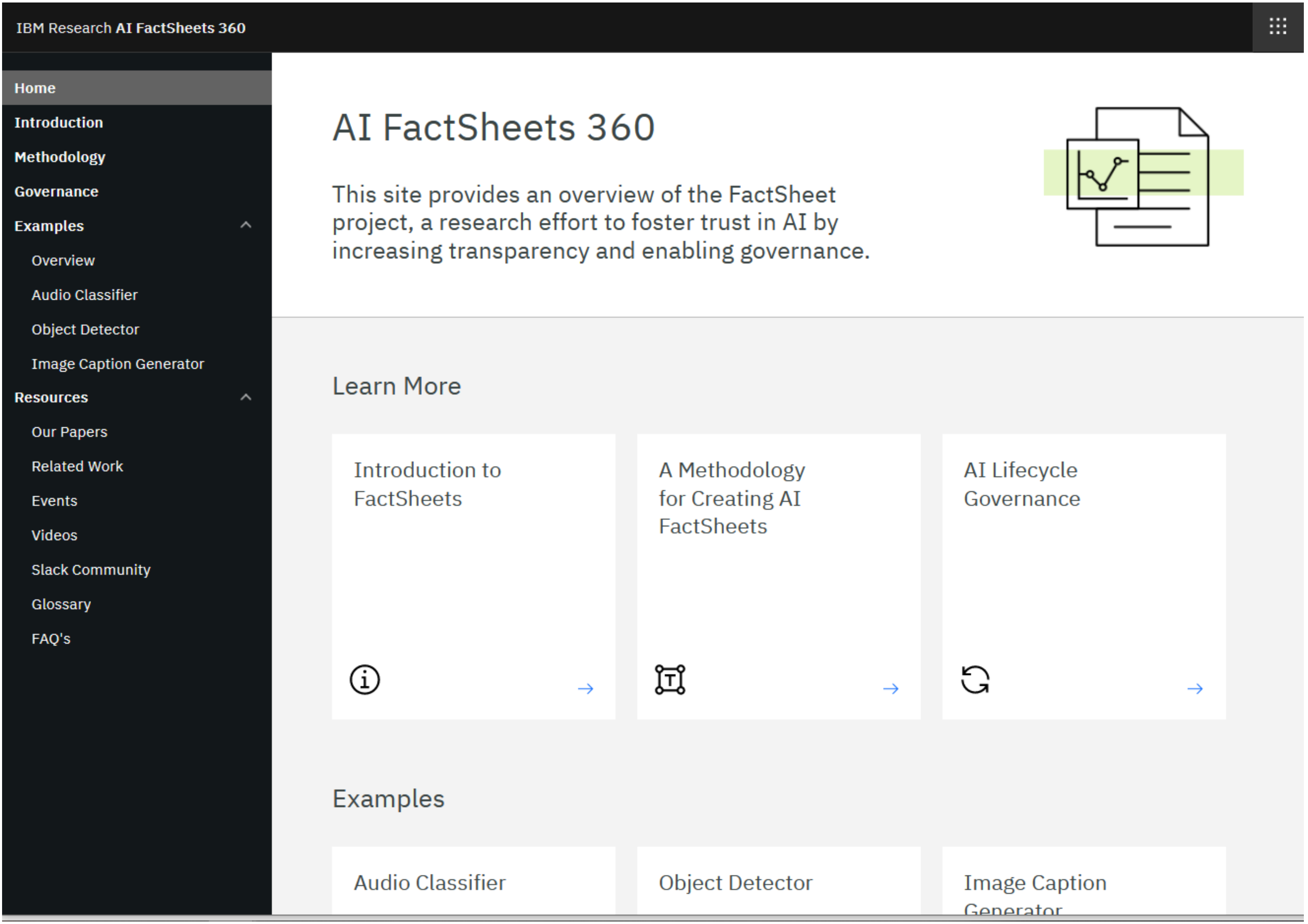
Facts can be rendered in different ways



Metrics generated automatically, throughout the lifecycle of the model/service/app

JSON file

AI FactSheets 360 Website: aifs360.mybluemix.com

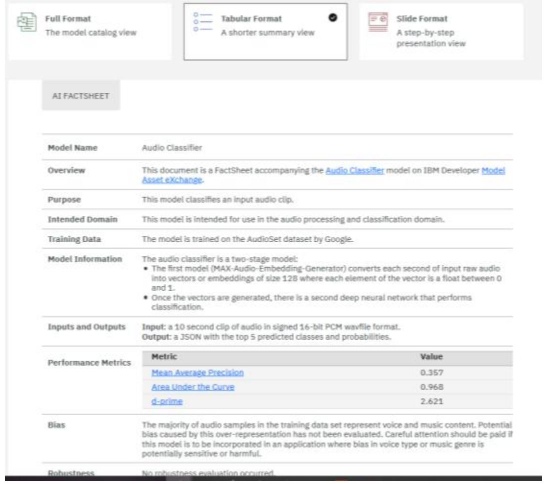


A Methodology for Creating AI FactSheets

John Richards, David Piorkowski, Michael Hind, Stephanie Houde, Aleksandra Mojsilović
IBM Research

ABSTRACT
As AI models and services are used in a growing number of high-stakes areas, a consensus is forming around the need for a clearer record of how these models and services are developed to increase trust. Several proposals for higher quality and more consistent AI documentation have emerged to address ethical and legal concerns

- AI services are the building blocks for many AI applications. Developers call the service API and consume its output. An AI service can be an amalgam of many models trained on many datasets. Thus, the models and datasets are (direct and indirect) components of an AI service, but they are not the interface to the developer.



Model Name	Audio Classifier
Overview	This document is a FactSheet accompanying the Audio Classifier model on IBM Developer Model Asset eXchange.
Purpose	This model classifies an input audio clip.
Intended Domain	This model is intended for use in the audio processing and classification domain.
Training Data	The model is trained on the AudioSet dataset by Google.
Model Information	The audio classifier is a two-stage model: • The first model (MAX-Audio-Embedding-Generator) converts each second of input raw audio into vectors or embeddings of size 128 where each element of the vector is a float between 0 and 1. • Once the vectors are generated, there is a second deep neural network that performs classification.
Inputs and Outputs	**Input**: a 10 second clip of audio in signed 16-bit PCM wavfile format. **Output**: a JSON with the top 5 predicted classes and probabilities.
Performance Metrics	Metric / Value: Mean Average Precision 0.357; Area Under the Curve 0.968; d-prime 2.621
Bias	The majority of audio samples in the training data set represent voice and music content. Potential bias caused by this over-representation has not been evaluated. Careful attention should be paid if this model is to be incorporated in an application where bias in voice type or music genre is potentially sensitive or harmful.
Robustness	No robustness evaluation occurred.

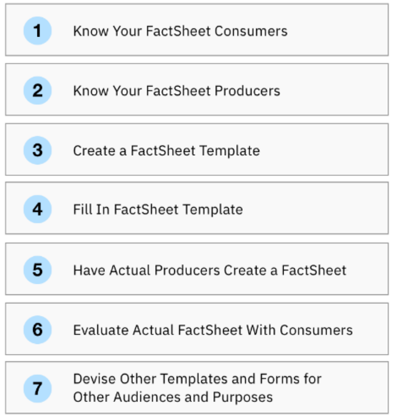
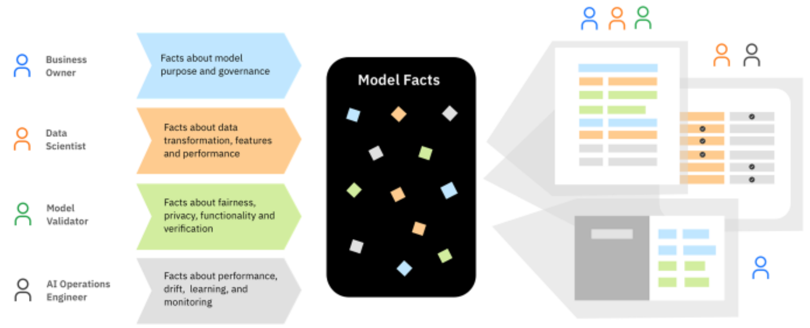


Figure 2: Steps to produce useful FactSheets



Other efforts directed towards the creation of transparent reporting mechanisms for AI

Gebru et al.
Datasheets for Datasets
https://arxiv.org/abs/1803.09010

Mitchell et al.
Model Cards for Model Reporting
https://arxiv.org/abs/1810.03993

Google
Model Cards
https://modelcards.withgoogle.com/model-reports

EU Commission
Ethics Guidelines for Trustworthy AI
https://ec.europa.eu/futurium/en/ai-alliance-consultation

Partnership on AI
ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of ML Lifecycles
https://www.partnershiponai.org/about-ml/

OpenAI
ModelCard for GPT-2
https://github.com/openai/gpt-2/blob/master/model_card.md

Holland et al.
The Dataset Nutrition Label: A Framework to Drive Higher Quality Data Standards
https://arxiv.org/abs/1805.03677

Bender and Friedman
Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science
https://openreview.net/forum?id=By4oPeX9f

Guszcza et al (HBR)
Why We Need to Audit Algorithms
https://hbr.org/2018/11/why-we-need-to-audit-algorithms

Loukides, Mason, Patil
Ethics and Data Science: Of Oaths and Checklists
https://www.oreilly.com/ideas/of-oaths-and-checklists

ORCAA
O'Neil Risk Consulting & Algorithmic Auditing
http://www.oneilrisk.com//

FactSheets and different flavors of Trust

AI Transparency

AI Governance

AI Marketplace

Enabling AI consumers to find trusted AI technology

Enterprise AI Documentation

Automatically document key AI characteristics for subsequent audits

Data Science Knowledge Management

Enable seamless reproducibility and efficient operations

AI Governance

Enterprise Needs:

1. **Specify policies to be enforced**
   – for regulators or enterprise governance

2. **Automate documentation of AI lifecycle**
   – without changing existing processes

3. **Make information accessible to all stakeholders**
   – enabling collaboration, using their natural tooling

==> Requires "Trust" capabilities instrumented into the AI Lifecycle

Enterprise AI Documentation (facilitating governance)

Problem

- Enterprise SW governance requires documentation of ML models
    - Current practice is ad hoc, error prone, and expensive (100s of pages, months to create, outsourced)
    - No best practices for documenting how a model/service was created, trained, tested, deployed, and evaluated
    - No structured way to represent model facts and manage them as the model is being built, tuned, deployed, tested, monitored, and improved.

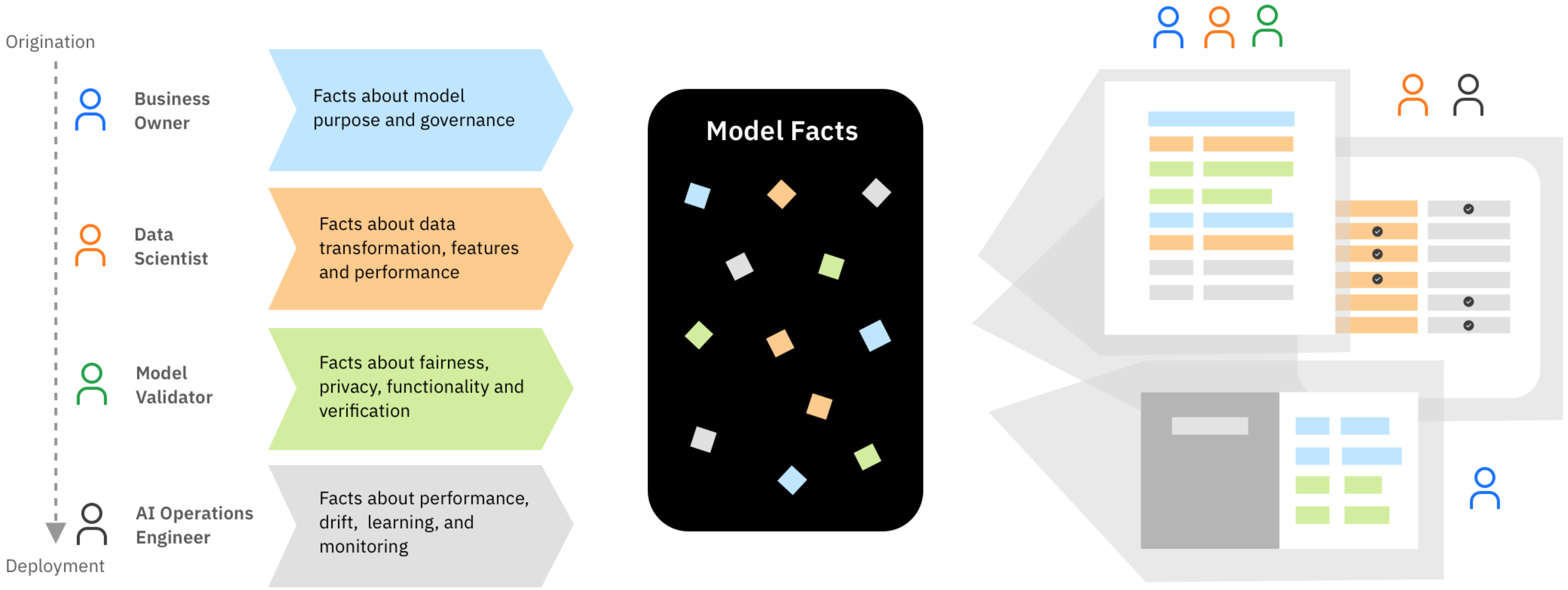
Solution

- Automate the gathering and communication of this information within each stages of ML lifecycle

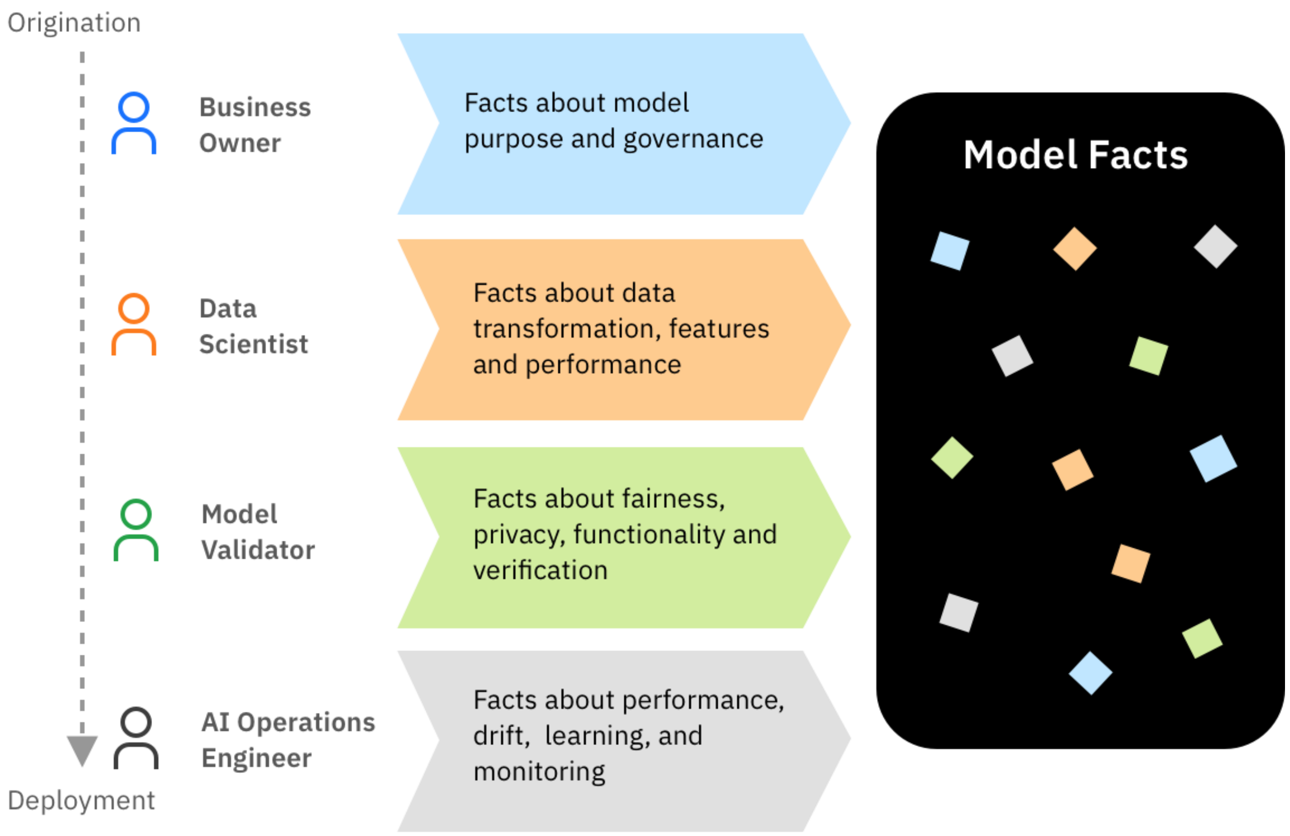
Value

- Provide visibility, governance, and regulatory compliance for model creation-to-deployment process
- Enable analytics on collected information to improve business outcomes and efficiency
- Facilitate communications among many personas with different roles, vocabularies, cultures, tools, and skill sets
    - data scientists, developers, test engineers, devOps engineers, business owners, regulators, etc.

Facts Collected During AI Lifecyle



AI Facts and FactSheets are Central to AI Governance



Predictive Performance	Data Scientist	Model Validator	AI Operations Engineer
Accuracy	0.95	0.94	0.92
Balanced Accuracy	0.63	0.63	0.61
AUC	0.79	0.78	0.77
F1	0.97	0.97	0.96
Fairness			
Disparate Impact	0.97	0.97	0.95
Statistical Parity Difference	-0.03	-0.03	-0.04
Adversarial Robustness			
Empirical Robustness	0.02	0.01	0.02
Explainabiity			
Faithfulness Mean	0.31	0.36	0.35

Example: Using Facts to Help with Model Validation

Predictive Performance	Test Dataset	Validation Dataset
Accuracy	0.95	0.94
Balanced Accuracy	0.63	0.63
AUC	0.79	0.78
F1	0.97	0.97

Fairness		
Disparate Impact	0.97	0.97
Statistical Parity Difference	-0.03	-0.03

Adversarial Robustness		
Empirical Robustness	0.02	0.01

Explainability		
Faithfulness Mean	0.31	0.36

Example: Using Facts to Help with Model Performance

Predictive Performance	Test Dataset	Validation Dataset	Deployment Data
Accuracy	0.95	0.94	0.92
Balanced Accuracy	0.63	0.63	0.61
AUC	0.79	0.78	0.77
F1	0.97	0.97	0.96
Fairness			
Disparate Impact	0.97	0.97	0.95
Statistical Parity Difference	-0.03	-0.03	-0.04
Adversarial Robustness			
Empirical Robustness	0.02	0.01	0.02
Explainabiity			
Faithfulness Mean	0.31	0.36	0.35

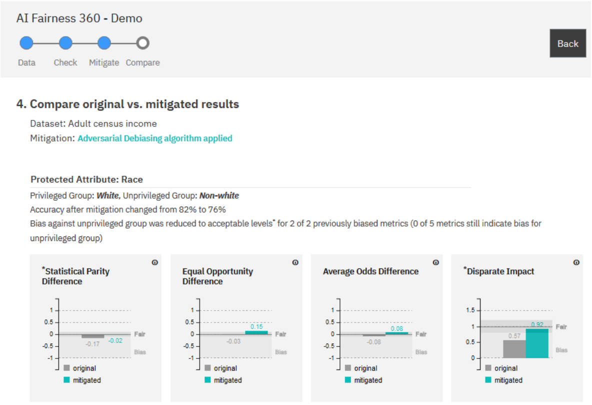
Example: Model Validator Comparing to Challenge Model

Predictive Performance	Data Scientist Model	Challenge Model
Accuracy	0.94	0.89
Balanced Accuracy	0.63	0.62
AUC	0.78	0.62
F1	0.97	0.93
Fairness		
Disparate Impact	0.97	0.94
Statistical Parity Difference	-0.03	-0.06
Adversarial Robustness		
Empirical Robustness	0.01	0.13
Explainability		
Faithfulness Mean	0.36	0.87

Trustworthy AI : Fairness, Explainability, Robustness, Transparency

AI Fairness 360

aif360.mybluemix.net

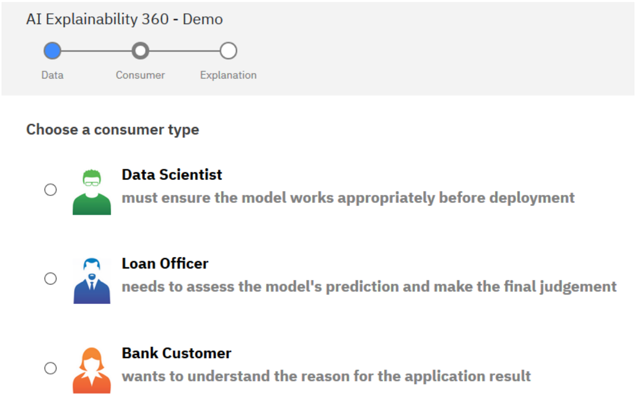


Most comprehensive **open source** toolkit for detecting & mitigating bias in ML models:

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

AI Explainability 360

aix360.mybluemix.net



Most comprehensive **open source** toolkit for explaining ML models & data

- 8 explainability algorithms
- Interactive demo showing 3 algorithms in credit scoring application
- 13 tutorial notebooks: finance, healthcare, lifestyle, retention, etc.
- Extensive documentation and taxonomy of explainability algorithms

Adversarial Robustness 360

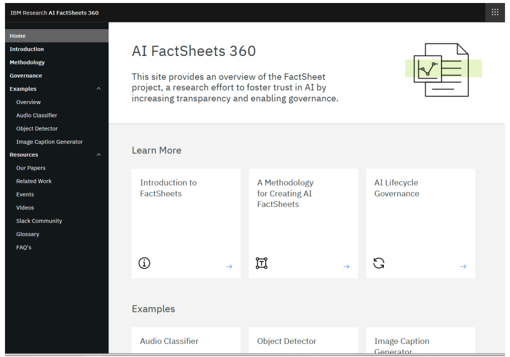
art360.mybluemix.net

Most comprehensive **open source** toolkit for defending AI from attacks

- Supports 10+ frameworks
- 19 composable and modular attacks (including adaptive white- and black-box)
- 10 defenses, including detection of adversarial samples and poisoning attacks
- Robustness metrics, certifications and verifications
- 30 notebooks covering attacks and defenses
- From dozens of publications

FactSheets 360

aifs360.mybluemix.net



Extensive website describing research effort to foster trust in AI by increasing transparency and enabling Governance

- 6 examples FactSheets
- 7-step methodology for creating useful FactSheets
- AI Governance
- Papers, videos, related work, FAQ, slack channel

Thank you