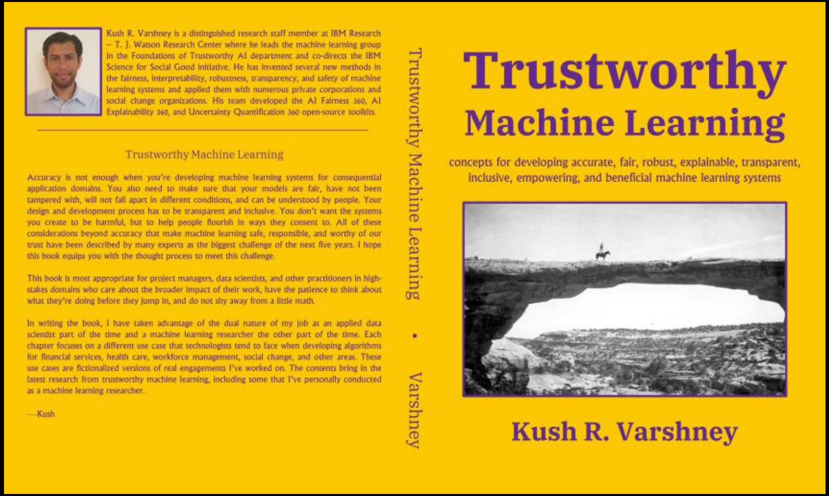
Trustworthy Machine Learning

—

Kush R. Varshney
Distinguished Research Staff Member and Manager

krvarshn@us.ibm.com | @krvarshney

http://www.trustworthymachinelearning.com



Kush R. Varshney is a distinguished research staff member at IBM Research – T. J. Watson Research Center where he leads the machine learning group in the Foundations of Trustworthy AI department and co-directs the IBM Science for Social Good initiative. He has invented several new methods in the fairness, interpretability, robustness, transparency, and safety of machine learning systems and applied them with numerous private corporations and social change organizations. His team developed the AI Fairness 360, AI Explainability 360, and Uncertainty Quantification 360 open-source toolkits.

Trustworthy Machine Learning

Accuracy is not enough when you're developing machine learning systems for consequential application domains. You also need to make sure that your models are fair, have not been tampered with, will not fall apart in different conditions, and can be understood by people. Your design and development process has to be transparent and inclusive. You don't want the systems you create to be harmful, but to help people flourish in ways they consent to. All of these considerations beyond accuracy that make machine learning safe, responsible, and worthy of our trust have been described by many experts as the biggest challenge of the next five years. I hope this book equips you with the thought process to meet this challenge.

This book is most appropriate for project managers, data scientists, and other practitioners in high-stakes domains who care about the broader impact of their work, have the patience to think about what they're doing before they jump in, and do not shy away from a little math.

In writing the book, I have taken advantage of the dual nature of my job as an applied data scientist part of the time and a machine learning researcher the other part of the time. Each chapter focuses on a different use case that technologists tend to face when developing algorithms for financial services, health care, workforce management, social change, and other areas. These use cases are fictionalized versions of real engagements I've worked on. The contents bring in the latest research from trustworthy machine learning, including some that I've personally conducted as a machine learning researcher.

—Kush

Trustworthy Machine Learning • Varshney

Trustworthy
Machine Learning

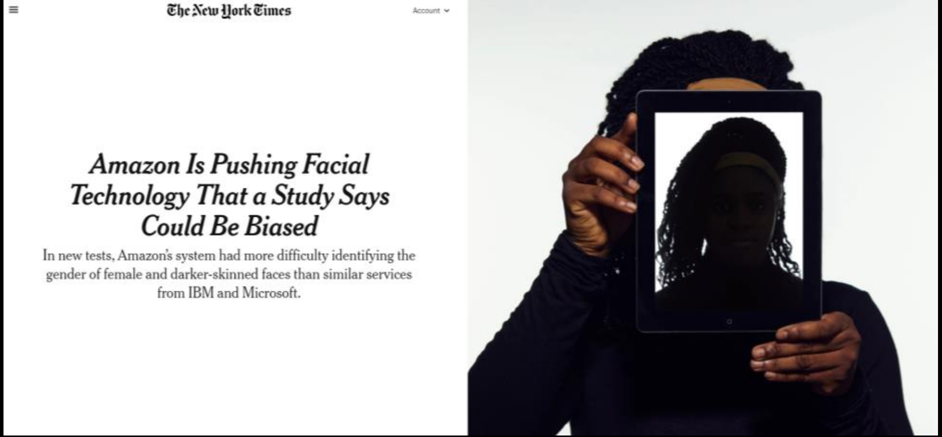
concepts for developing accurate, fair, robust, explainable, transparent, inclusive, empowering, and beneficial machine learning systems

Kush R. Varshney

Research

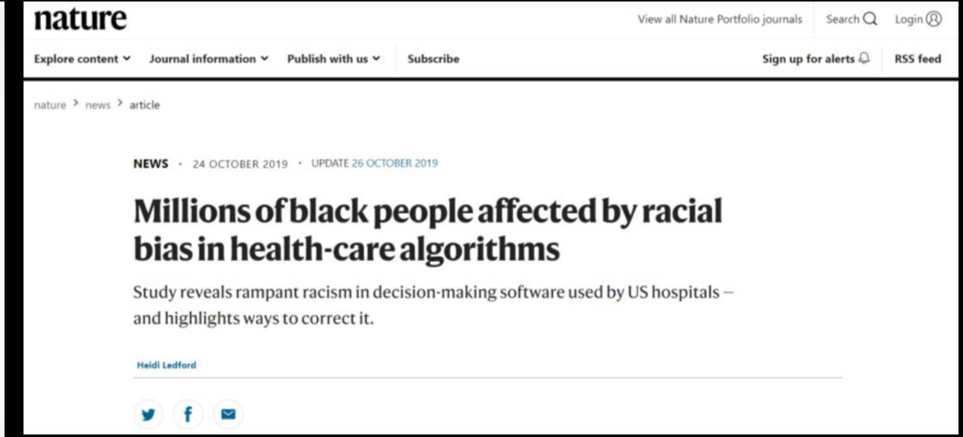
IBM

Decision making supported by machine learning can have unwanted bias



Amazon Is Pushing Facial Technology That a Study Says Could Be Biased

In new tests, Amazon's system had more difficulty identifying the gender of female and darker-skinned faces than similar services from IBM and Microsoft.

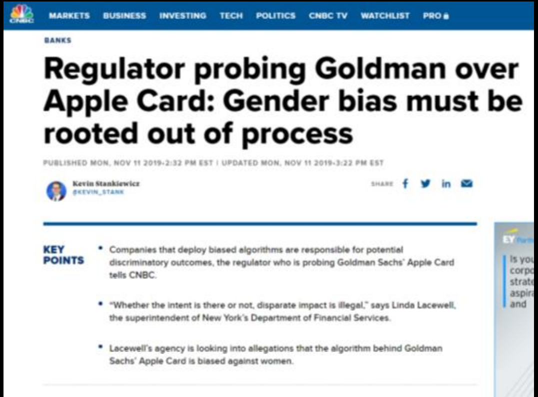


NEWS · 24 OCTOBER 2019 · UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Heidi Ledford



BANKS

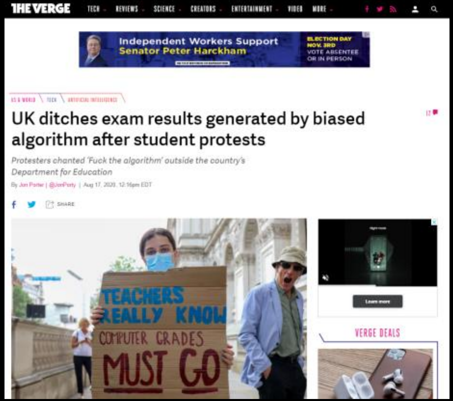
Regulator probing Goldman over Apple Card: Gender bias must be rooted out of process

PUBLISHED MON, NOV 11 2019-2:32 PM EST | UPDATED MON, NOV 11 2019-3:22 PM EST

Kevin Stankiewicz
@KEVIN_STANK

KEY POINTS

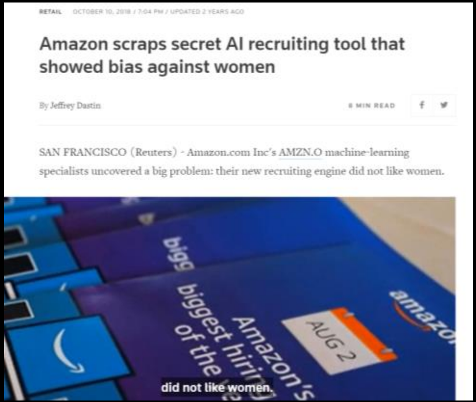
- Companies that deploy biased algorithms are responsible for potential discriminatory outcomes, the regulator who is probing Goldman Sachs' Apple Card tells CNBC.
- "Whether the intent is there or not, disparate impact is illegal," says Linda Lacewell, the superintendent of New York's Department of Financial Services.
- Lacewell's agency is looking into allegations that the algorithm behind Goldman Sachs' Apple Card is biased against women.



UK ditches exam results generated by biased algorithm after student protests

Protesters chanted 'Fuck the algorithm' outside the country's Department for Education

By Jon Porter | @JonPorty | Aug 17, 2020, 12:16pm EDT

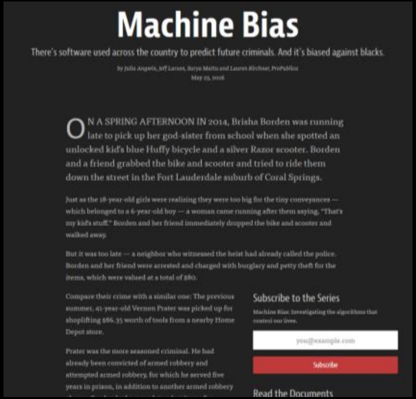


OCTOBER 10, 2018 / 7:04 PM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.



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.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store.

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery

Subscribe to the Series

Machine Bias: Investigating the algorithms that control our lives.

Read the Documents

“Non-traditional” fairness use cases

Infrastructure rollout by telecommunications providers

Selecting people to check at retail self-checkouts

Tree-planting decisions by forest managers

Delinquency collections

Recommendations in fantasy football

Trustworthy AI is not just about bias

EXCLUSIVE AUTONOMOUS VEHICLES UBER/LYFT

Uber Finds Deadly Accident Likely Caused By Software Set to Ignore Objects On Road

By Amir Efrati May 07, 2018 9:48 AM PDT · Comments by Noah David, Michael D. Geer and 4 others

Subscribe now

Uber has determined that the likely cause of a fatal collision involving one of its prototype self-driving cars in Arizona in March was a problem with the software that decides how the car should react to objects it detects, according to two people briefed about the matter.

A shot from an ABC TV station in Tempe, Arizona, after an Uber self-driving car killed a pedestrian. Photo by AP.

The car's sensors detected the pedestrian, who was crossing the street with a bicycle, but Uber's software decided it didn't need to react right away. That's a result of how the software was tuned. Like other autonomous vehicle systems, Uber's software has the ability to ignore "false positives," or objects in its path that wouldn't actually be a problem for the vehicle, such as a plastic bag floating over a road. In this case, Uber executives believe the company's system was tuned so that it reacted less to such objects. But the tuning went too far, and the car didn't react fast enough, one of these people said.

THE TAKEAWAY

- Software in car was set to ignore some objects
- Safety driver took eyes off road at critical moment

PulmonologyAdvisor

ASTHMA COPD INFLUENZA LUNG CANCER OSA PNEU

NEWS CALCULATORS CHARTS CME

December 12, 2017

The Potential Pitfalls of Machine Learning Algorithms in Medicine

Tafari Mbadiwe

Back in the 1990s an intrepid group of researchers out of the University of Pittsburgh set out to write a computer program that could do a better job than doctors of predicting whether serious complications would develop in patients who presented with pneumonia.[1] Success may have been a long shot, but it was definitely a shot worth taking. After all, the researchers figured that if they pulled it off, they could both lower costs *and* improve patient outcomes in one fell swoop. So they built a neural network — basically a computer program that responds dynamically to external inputs — and turned it loose on a database covering three-quarters of a million patients in 78 hospitals across 23 states.

Machine learning programs can process enormous quantities of information and make meaningful and actionable predictions about future behaviors and outcomes.

The results were curious, to say the least. The program seemed to have determined that patients with pneumonia and asthma had *better* outcomes than those who did not have asthma. Asthma, it appeared, was somehow providing some sort of protection.[2] The neural net, which was by many measures

Attributes of trustworthiness

	Source	Attribute 1	Attribute 2	Attribute 3	Attribute 4
trustworthy people	Mishra	competent	reliable	open	concerned
	Maister et al.	credibility	reliability	intimacy	low self-orientation
	Sucher and Gupta	competent	use fair means to achieve its goals	take responsibility for all its impact	motivated to serve others' interests as well as its own
trustworthy AI	Toreini et al.	ability	integrity	predictability	benevolence
	Ashoori and Weisz	technical competence	reliability	understandability	personal attachment

accuracy

distributional robustness; fairness; adversarial robustness

explainability; transparency; uncertainty quantification; value alignment

social good; empowering

What does it take to trust an AI system?

accuracy

fairness

explainability

uncertainty quantification

robustness

privacy

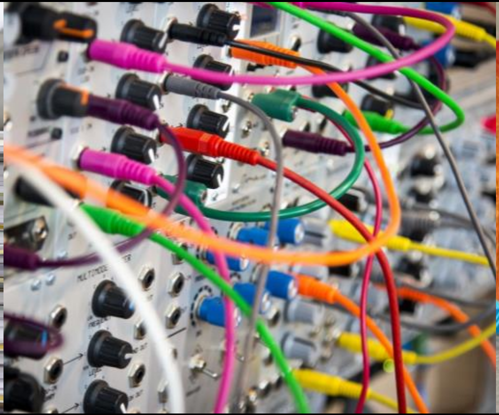
data quality

testing

Multiple factors are placing trust in AI as a top priority

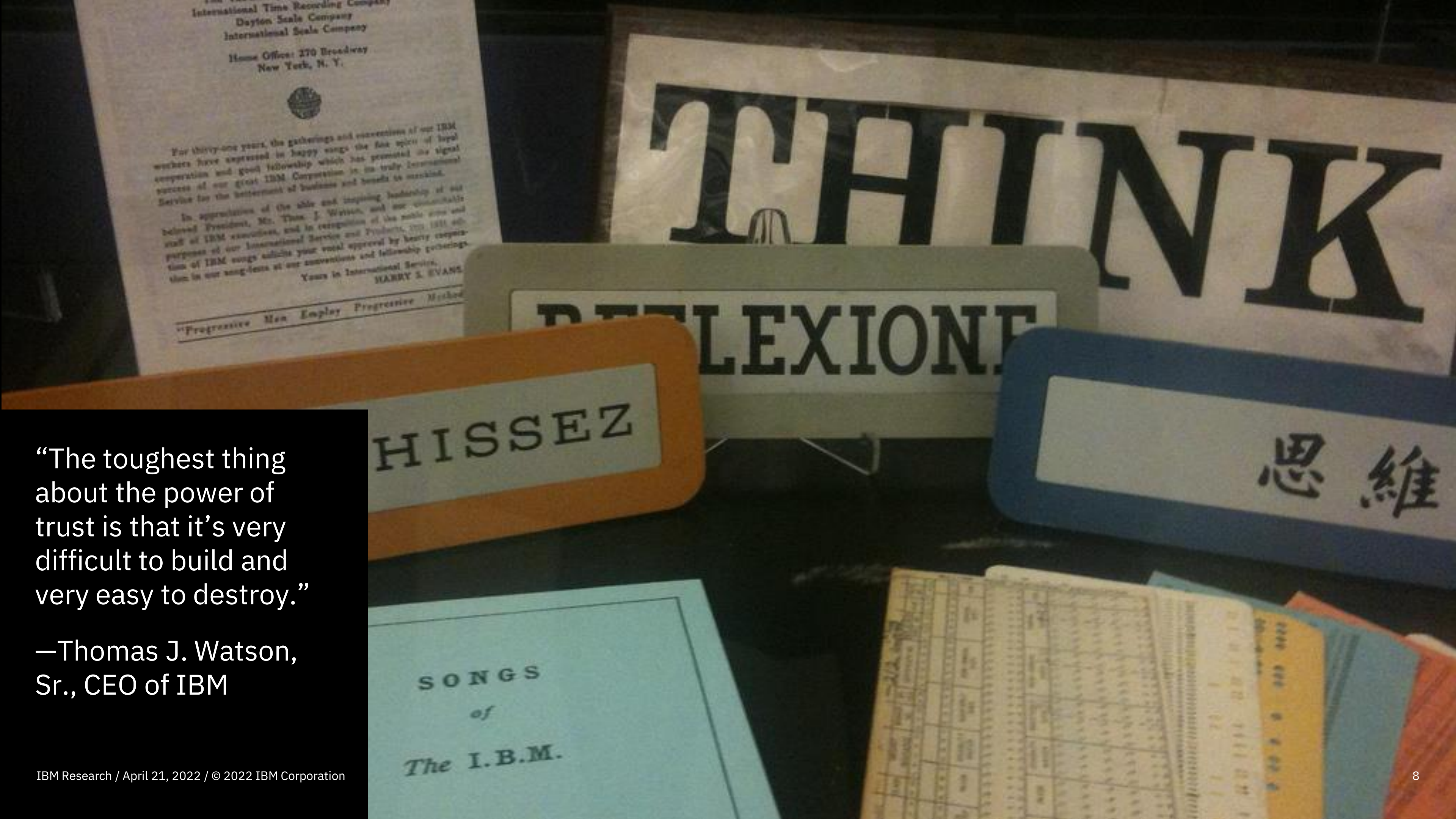
brand reputation

increased regulation



complexity of AI deployments

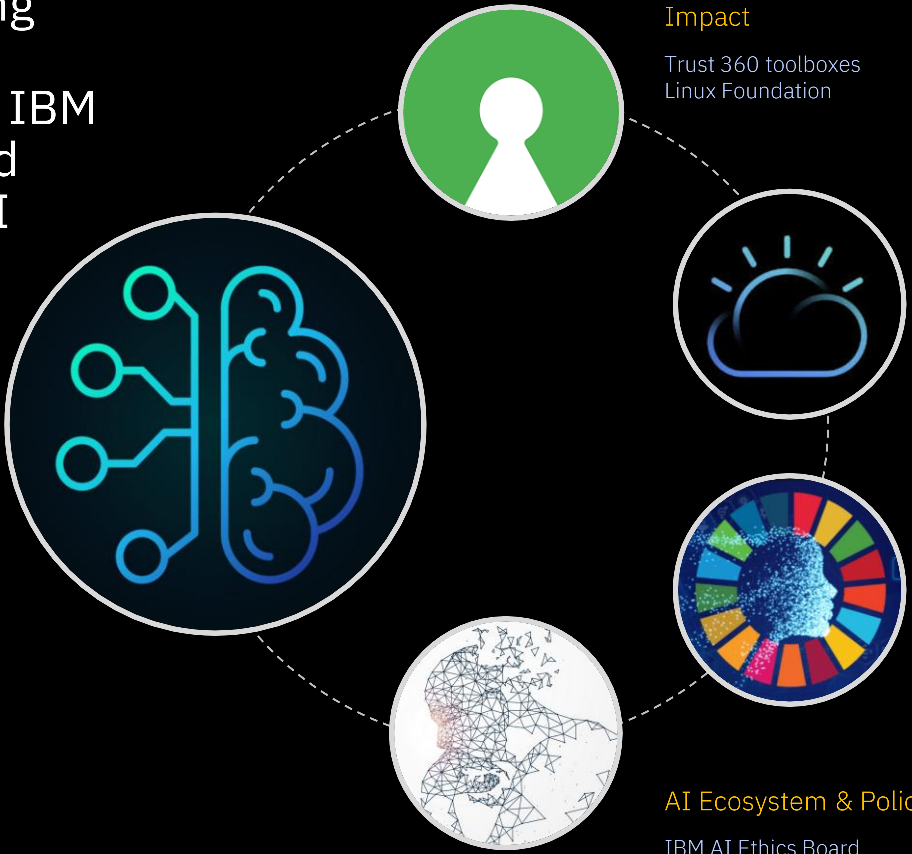
focus on social justice


International Time Recording Company
Dayton Scale Company
International Scale Company
Home Office: 270 Broadway
New York, N. Y.
Yours in International Service,
HARRY S. EVANS
"Progressive Men Employ Progressive Methods"
THINK
REFLEXIONE
HISSEZ
思維
SONGS of The I.B.M.

"The toughest thing about the power of trust is that it's very difficult to build and very easy to destroy."

—Thomas J. Watson, Sr., CEO of IBM

From groundbreaking science, to differentiating assets/technologies, to innovative applications, IBM Research is a recognized leader in Trustworthy AI



Science of Trustworthy AI

Foundational theoretical work in fairness, explainability, robustness, uncertainty quantification, transparency, generative modeling

Open Source & Community Impact

Trust 360 toolboxes
Linux Foundation

Product Contributions

Pipeline of innovations to IBM products

Beneficial AI Deployments

Science for Social Good

AI Ecosystem & Policy

IBM AI Ethics Board
PAI, EU Commission High Level Expert Group on AI, NIST, AI Caucus, National AI Strategy, …

Open-source toolkits

AI Fairness 360 http://aif360.mybluemix.net/

AI Explainability 360 http://aix360.mybluemix.net/

Adversarial Robustness 360 http://art360.mybluemix.net/

Uncertainty Quantification 360 http://uq360.mybluemix.net/

AI Privacy 360 http://aip360.mybluemix.net/

Causal Inference 360 http://ci360.mybluemix.net/

AI FactSheets 360 http://aifs360.mybluemix.net/

Thank you

Kush R. Varshney
Distinguished Research Staff Member and Manager
—
krvarshn@us.ibm.com