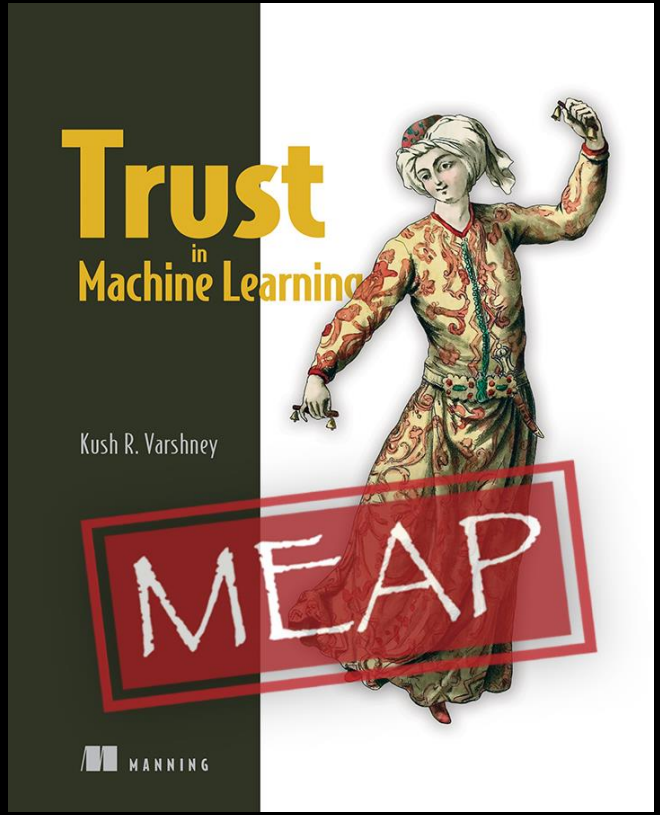
Trustworthy AI

—

Kush R. Varshney
Distinguished Research Staff Member and Manager

krvarshn@us.ibm.com | @krvarshney



https://www.manning.com/books/trust-in-machine-learning

Decision making supported by machine learning can have unwanted bias

The New York Times

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.

nature

View all Nature Portfolio journals | Search | Login

Explore content | Journal information | Publish with us | Subscribe | Sign up for alerts | RSS feed

nature › news › article

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

MARKETS | BUSINESS | INVESTING | TECH | POLITICS | CNBC TV | WATCHLIST | PRO

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

SHARE

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.

THE VERGE

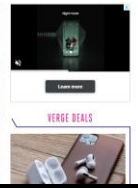
Independent Workers Support Senator Peter Harckham — ELECTION DAY NOV. 3RD VOTE ABSENTEE OR IN PERSON

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

SHARE



VERGE DEALS

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

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin | 6 MIN READ

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.

you@example.com

Subscribe

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

Pulmonology Advisor

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


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

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.

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

Do you trust Eliud Kipchoge to run fast?

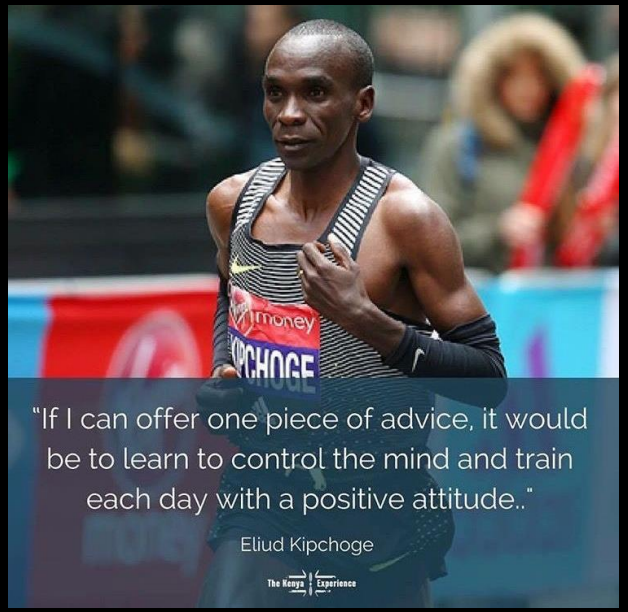
Competent (October 12, 2019)

Reliable

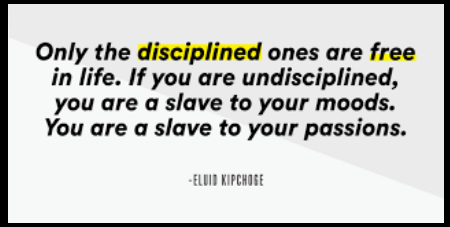
2013 Hamburg Marathon	1st	2:05:30
2013 Berlin Marathon	2nd	2:04:05
2014 Rotterdam Marathon	1st	2:05:00
2014 Chicago Marathon	1st	2:04:11
2015 London Marathon	1st	2:04:42
2015 Berlin Marathon	1st	2:04:00
2016 London Marathon	1st	2:03:05
2016 Summer Olympics	1st	2:08:44
2017 Berlin Marathon	1st	2:03:32
2018 London Marathon	1st	2:04:17
2018 Berlin Marathon	1st	2:01:39
2019 London Marathon	1st	2:02:37
2020 London Marathon	8th	2:06:49
NN Mission Marathon	1st	2:04:30
2020 Summer Olympics	1st	2:08:38

Open


WHEN YOU SMILE AND YOU'RE HAPPY, YOU CAN TRIGGER THE MIND TO NOT FEEL YOUR LEGS.
- ELIUD KIPCHOGE -


"If I can offer one piece of advice, it would be to learn to control the mind and train each day with a positive attitude.."
Eliud Kipchoge
The Kenya Experience

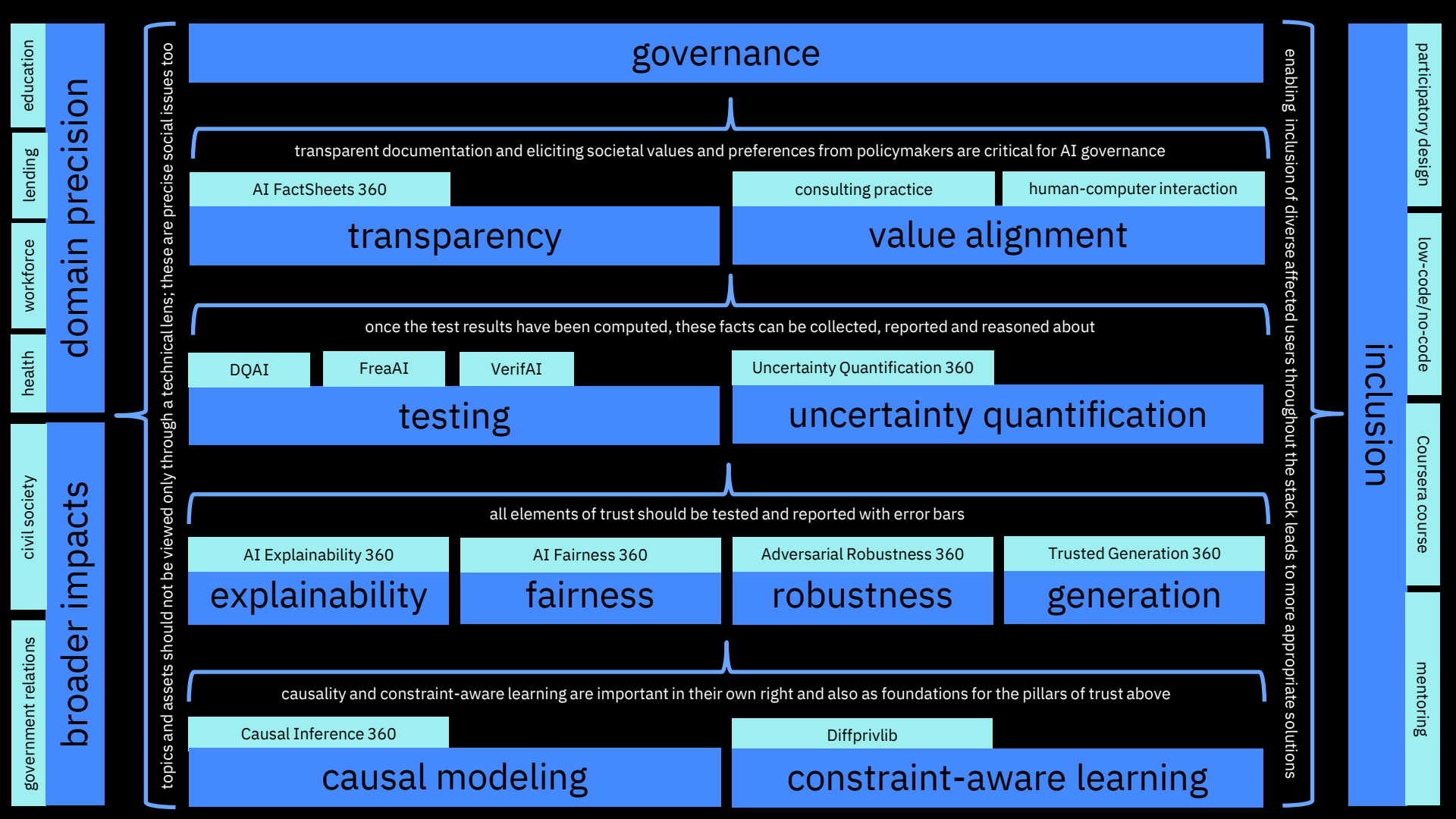

NN running team
"Motivation + Discipline = Consistency"
Eliud Kipchoge

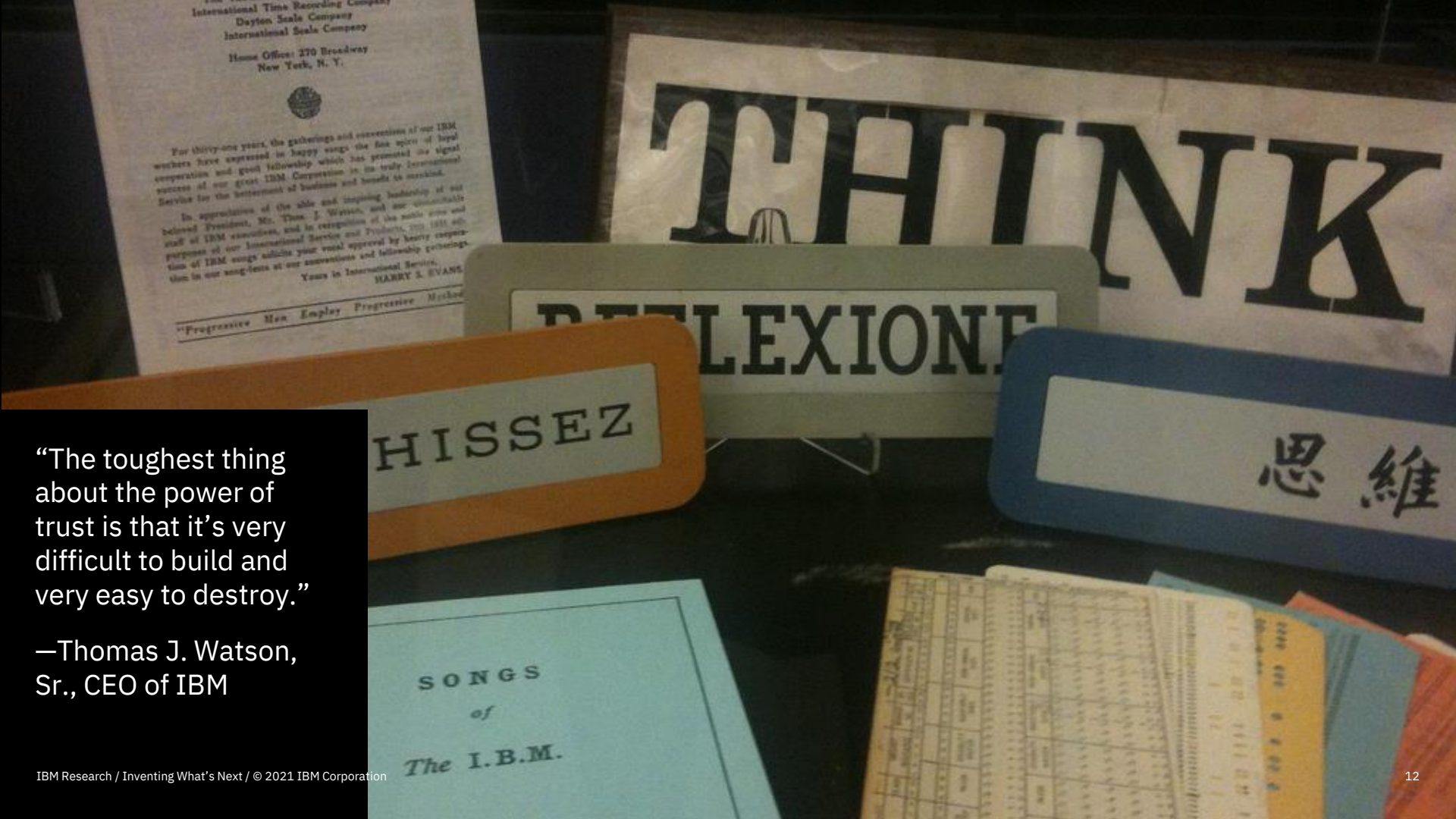

Only the disciplined ones are free in life. If you are undisciplined, you are a slave to your moods. You are a slave to your passions.
-ELIUD KIPCHOGE

Selfless

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; uncertainty quantification transparency; value alignment	social good; empowering


education
lending
workforce
health
domain precision
civil society
government relations
broader impacts
topics and assets should not be viewed only through a technical lens; these are precise social issues too
governance
transparent documentation and eliciting societal values and preferences from policymakers are critical for AI governance
AI FactSheets 360
transparency
consulting practice
human-computer interaction
value alignment
once the test results have been computed, these facts can be collected, reported and reasoned about
DQAI
FreaAI
VerifAI
testing
Uncertainty Quantification 360
uncertainty quantification
all elements of trust should be tested and reported with error bars
AI Explainability 360
explainability
AI Fairness 360
fairness
Adversarial Robustness 360
robustness
Trusted Generation 360
generation
causality and constraint-aware learning are important in their own right and also as foundations for the pillars of trust above
Causal Inference 360
causal modeling
Diffprivlib
constraint-aware learning
enabling inclusion of diverse affected users throughout the stack leads to more appropriate solutions
inclusion
participatory design
low-code/no-code
Coursera course
mentoring



"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

No shortcuts

"I always tell people that this is a really simple deal:

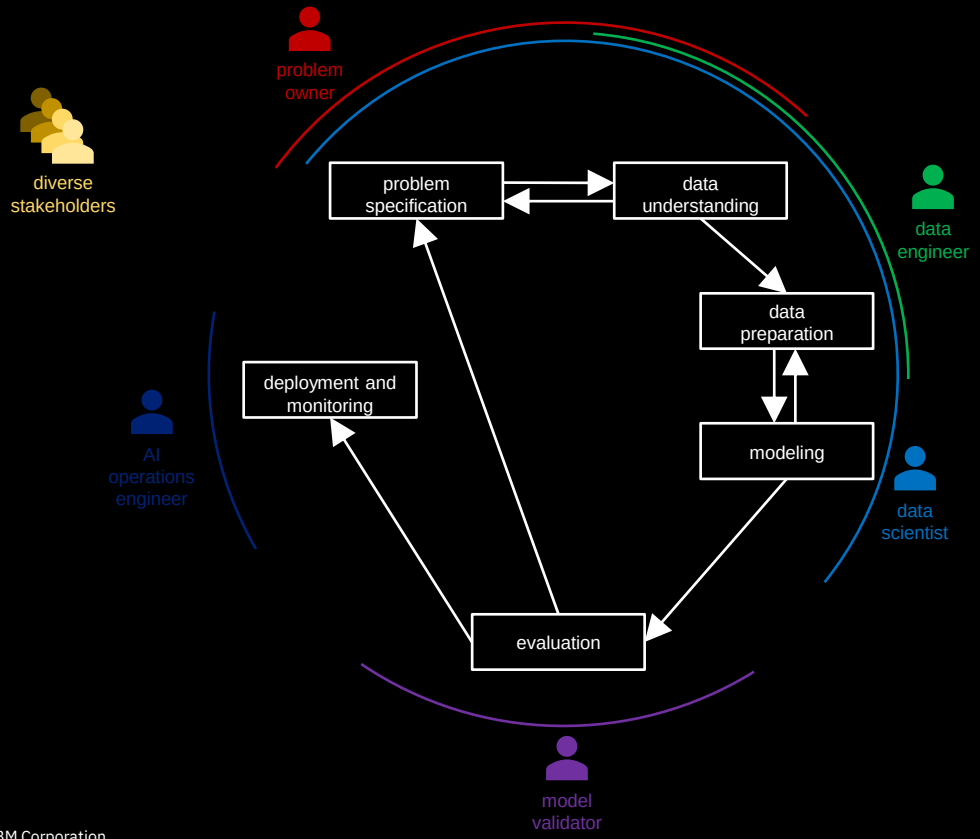
Work hard.

If you work hard, follow what's required and set your priorities right, then you can really perform without taking shortcuts."

Eliud Kipchoge

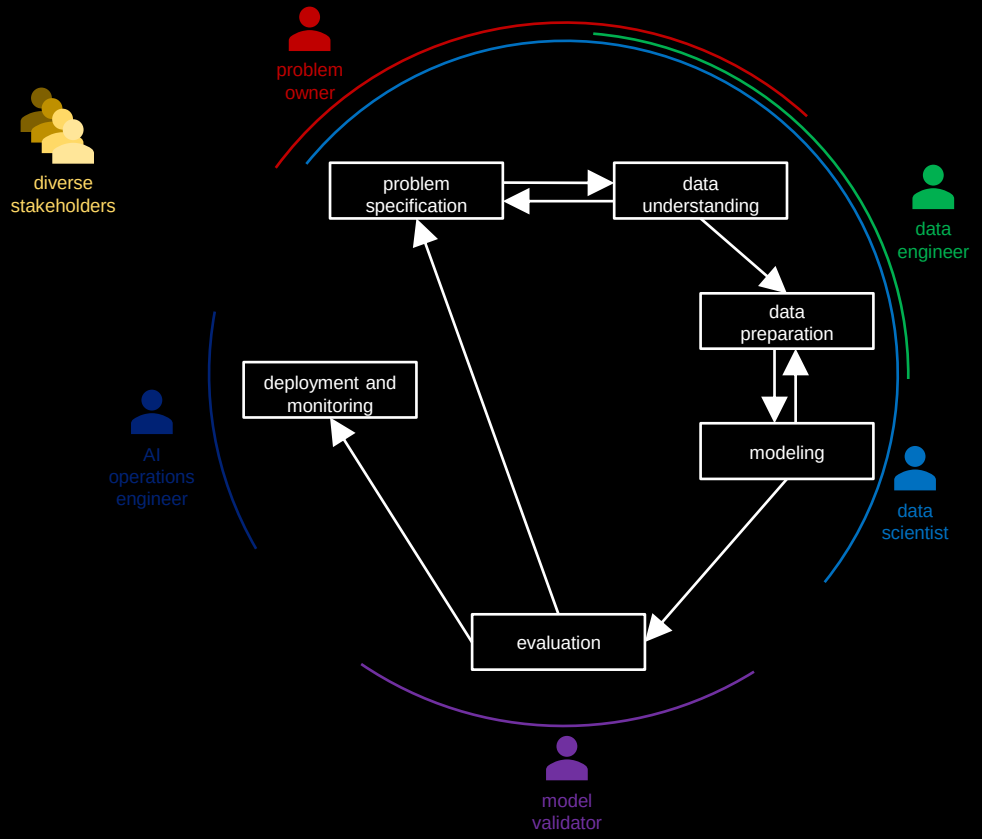
No shortcuts in inclusion

Don’t take shortcuts anywhere in the AI lifecycle

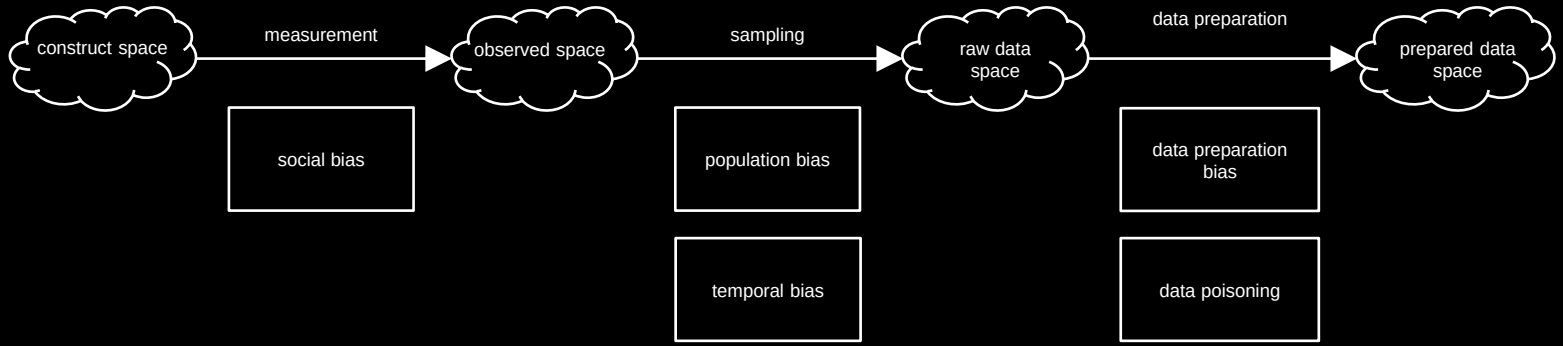


No shortcuts in problem specification

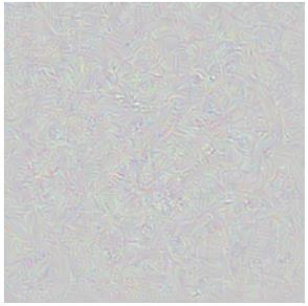
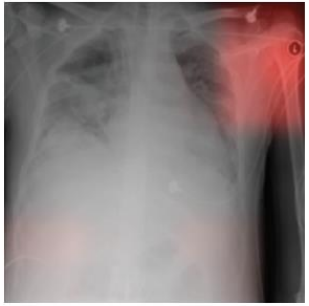
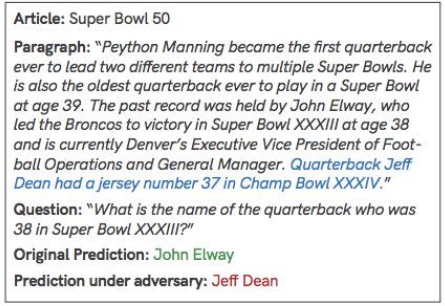
Take advice from a panel of diverse voices



No shortcuts in data understanding and preparation

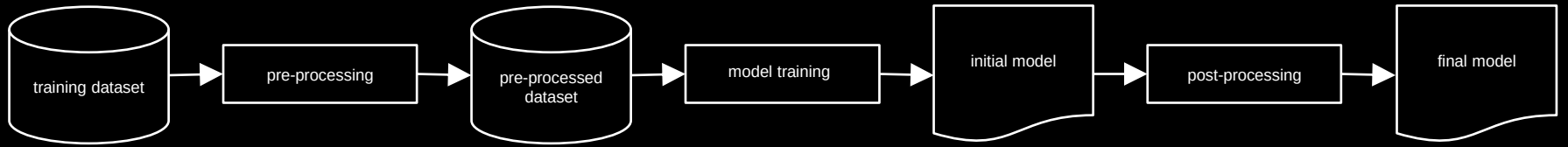


No shortcuts in modeling

				
Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

No shortcuts in modeling

distributional robustness
adversarial robustness
fairness
explainability
uncertainty quantification



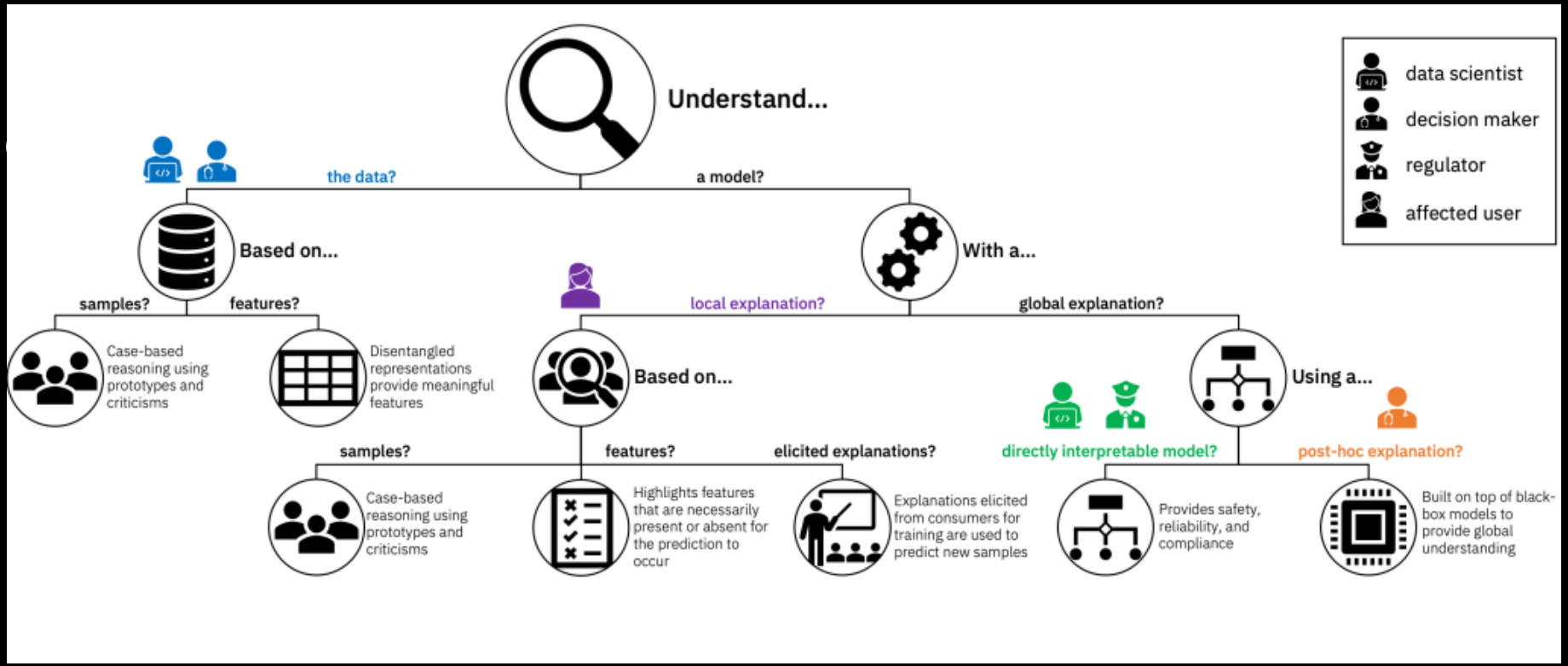
domain adaptation
data sanitization
bias mitigation pre-processing
disentangled representations
data uncertainty

domain robustness
gradient shaping/adversarial training
bias mitigation in-processing
directly interpretable models
model uncertainty

patching
bias mitigation post-processing
post hoc explanations
total uncertainty

Explanation: A justification for a machine learning prediction

Explainability



Unwanted bias

places privileged groups at systematic advantage

and unprivileged groups at systematic disadvantage.

Where does unwanted bias come from?

Problem misspecification.

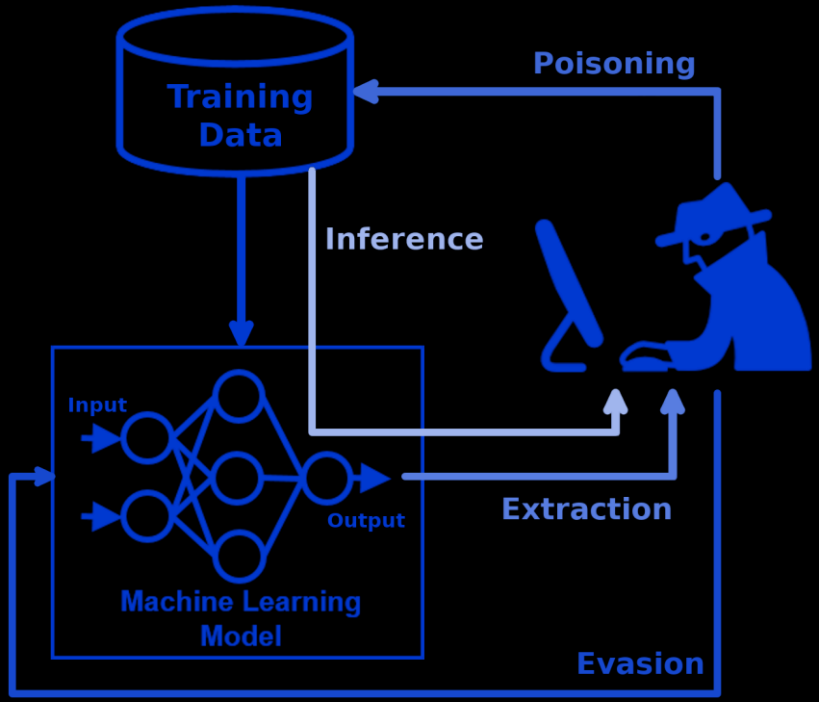
Data engineering.

Prejudice in historical data.

Undersampling.

Adversary: Malicious actors trying to meet their own goals

Adversarial robustness

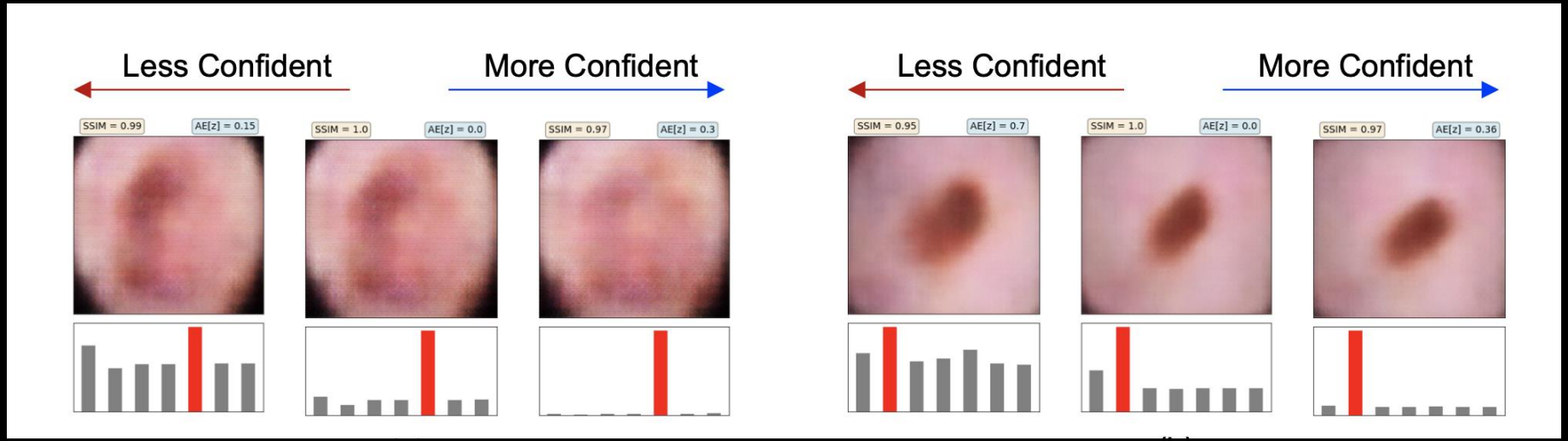


Detecting, preventing, and certifying against attacks by malicious adversaries.

Pushing AI to its limits as a test.

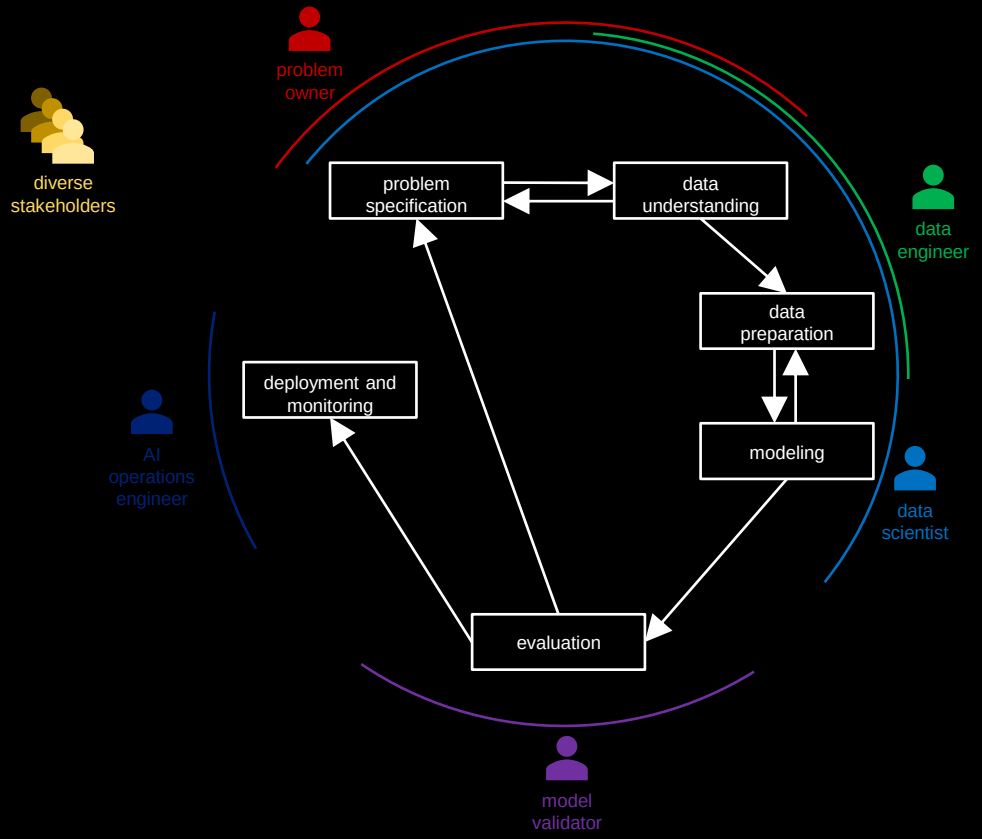
UQ: Does the model know when it doesn’t know?

Uncertainty quantification in skin disease diagnosis



No shortcuts in evaluation

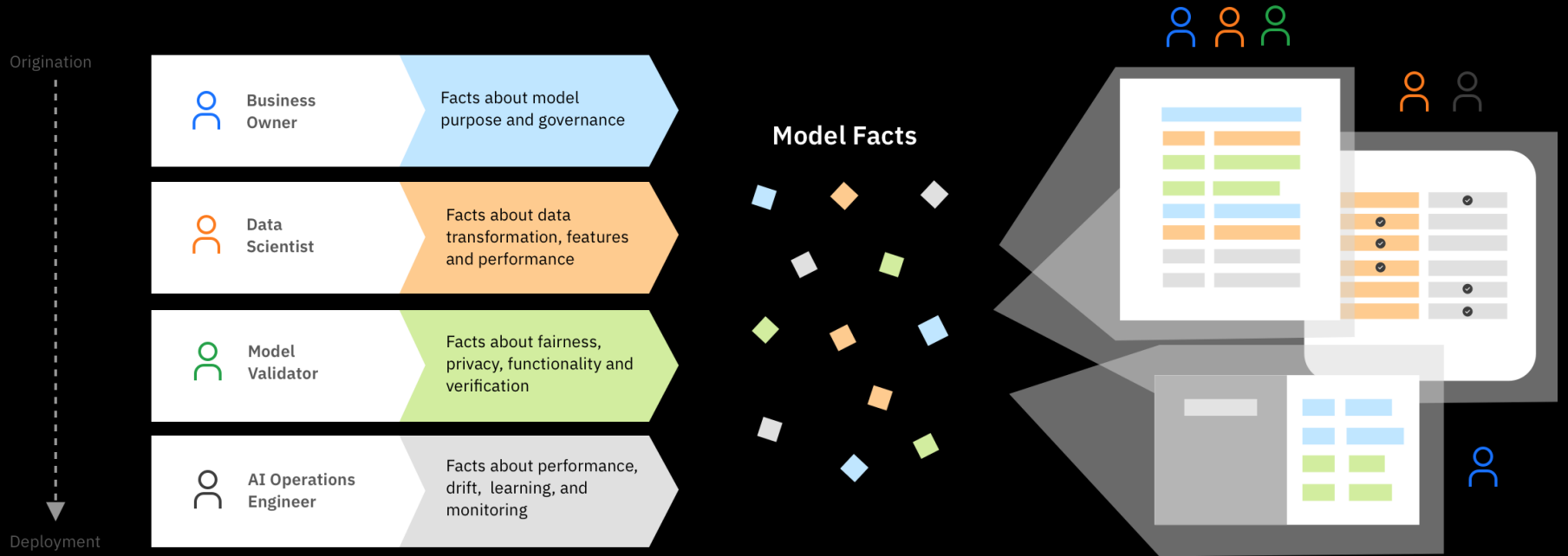
Take advice from a panel of diverse voices



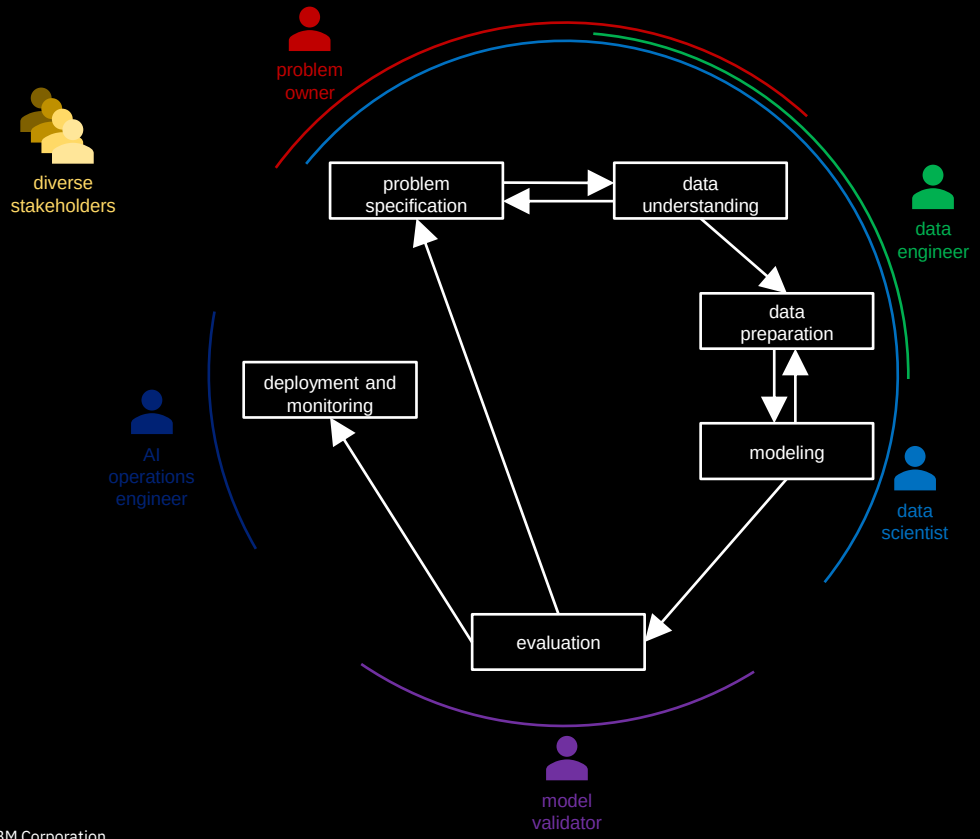
No shortcuts in monitoring

Predictive Performance	Data Scientist's Dataset	Validator's 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
Explainability			
Faithfulness Mean	0.31	0.36	0.35

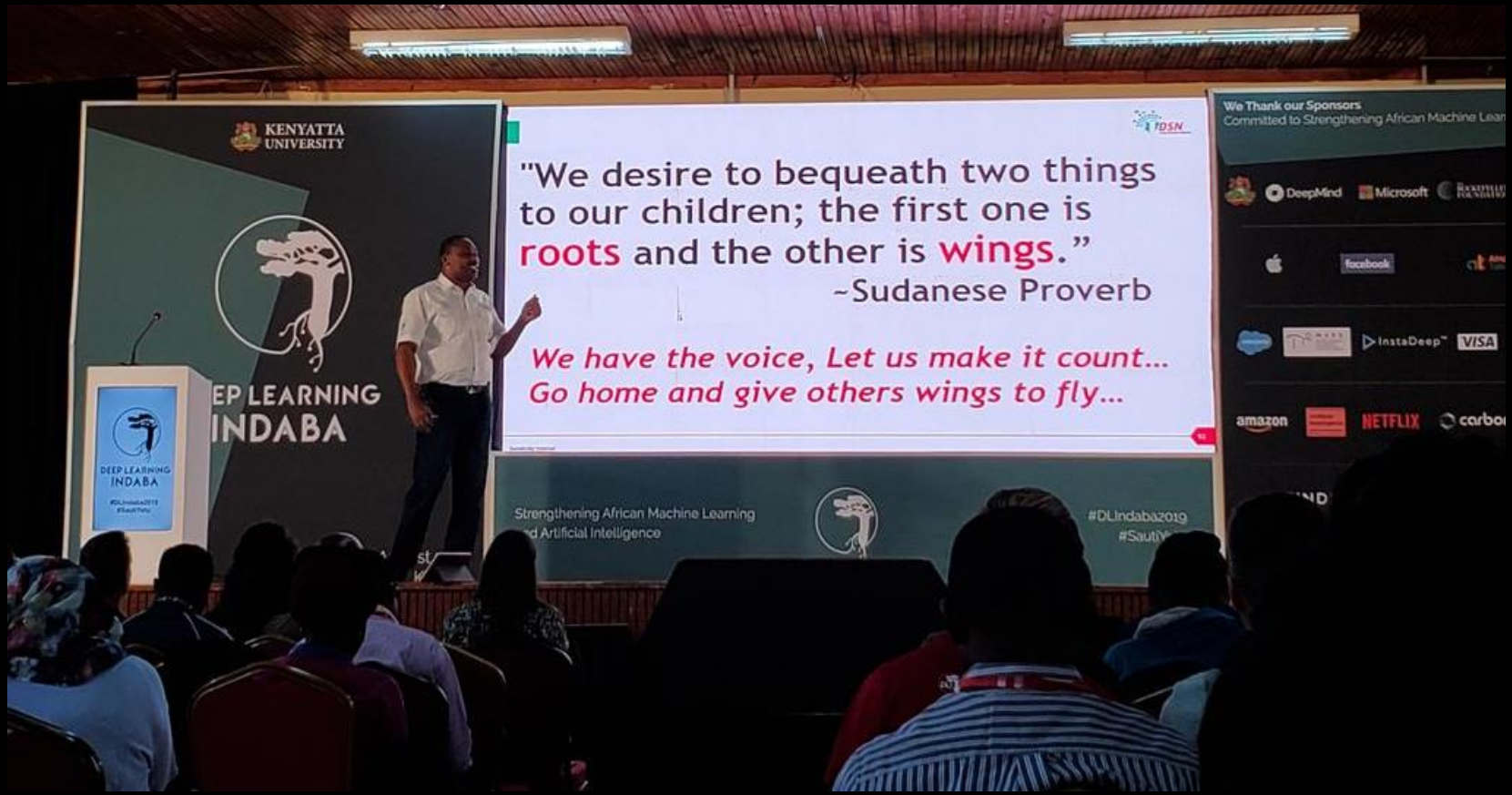
AI FactSheets for transparency throughout development



Don’t take shortcuts anywhere in the AI lifecycle



Roots and wings



Our wings to you: open-source toolkits

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

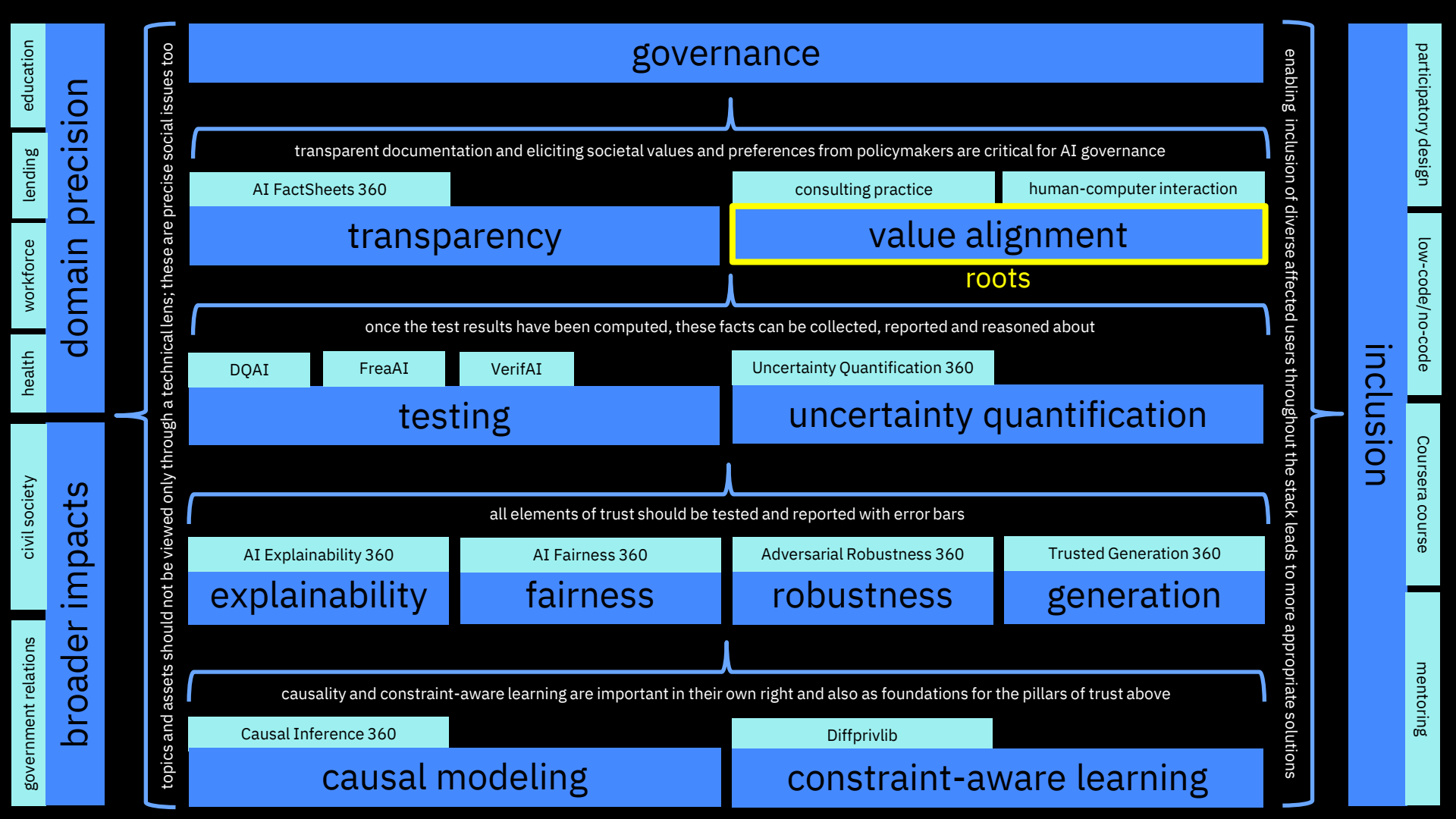
AI Explainability 360 http://aix360.mybluemix.net/

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

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

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

Causal Inference 360 https://cif360-dev.mybluemix.net/


education
lending
workforce
health
domain precision
civil society
government relations
broader impacts
topics and assets should not be viewed only through a technical lens; these are precise social issues too
governance
transparent documentation and eliciting societal values and preferences from policymakers are critical for AI governance
AI FactSheets 360
transparency
consulting practice
human-computer interaction
value alignment
roots
once the test results have been computed, these facts can be collected, reported and reasoned about
DQAI
FreaAI
VerifAI
testing
Uncertainty Quantification 360
uncertainty quantification
all elements of trust should be tested and reported with error bars
AI Explainability 360
explainability
AI Fairness 360
fairness
Adversarial Robustness 360
robustness
Trusted Generation 360
generation
causality and constraint-aware learning are important in their own right and also as foundations for the pillars of trust above
Causal Inference 360
causal modeling
Diffprivlib
constraint-aware learning
enabling inclusion of diverse affected users throughout the stack leads to more appropriate solutions
inclusion
participatory design
low-code/no-code
Coursera course
mentoring

Thank you

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