accenturedigital

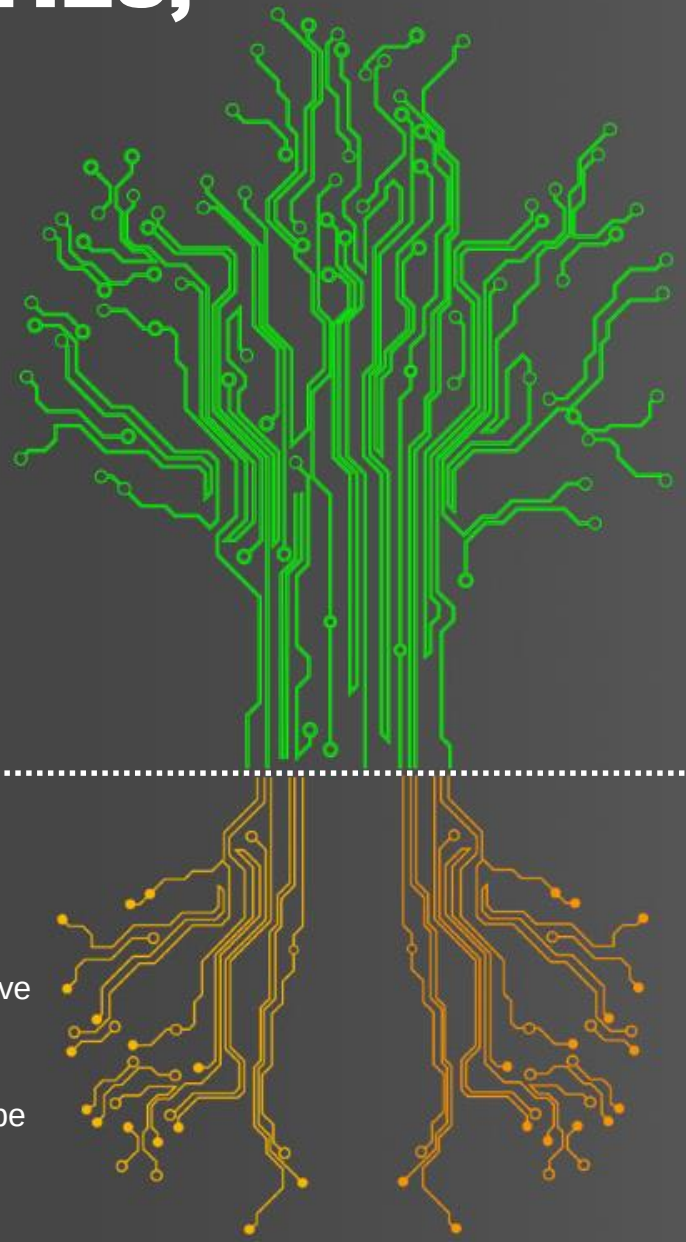
RESPONSIBLE AI

DR. RUMMAN CHOWDHURY
GLOBAL LEAD, RESPONSIBLE AI
@RUCHOWDH
WWW.RUMMANCHOWDHURY.COM

The **Responsible AI team at Accenture** is dedicated to creating human-centric Artificial Intelligence. Our goal is to understand and address the social, regulatory and economic impact of this technology from development to deployment and beyond. The team also serves as the starting point for governance internally at Accenture.

The Responsible Artificial Intelligence Team led by Dr. Rumman Chowdhury, data scientist and social scientist, and Deborah Santiago, senior leadership in Accenture Legal.

AI OPPORTUNITIES, RISKS AND CHALLENGES



OPPORTUNITY TO UNLOCK TRAPPED VALUE

AI affords a tremendous opportunity not only to increase efficiencies and reduce costs, but has the capacity to help rethink businesses and solve critical problems.

UNINTENDED CONSEQUENCES

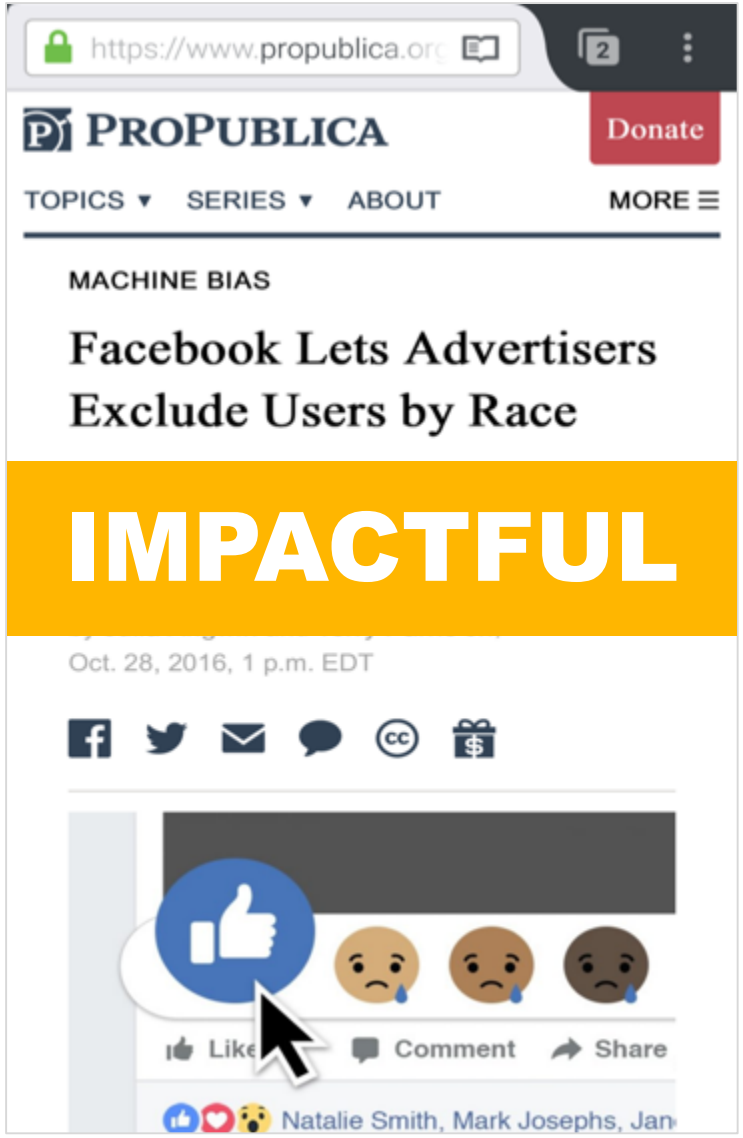
Unexpected, but harmful, outcomes have led to consumer backlash and legal problems. Launching AI without an understanding of its social impact can be risky.

COMPLIANCE CHALLENGE

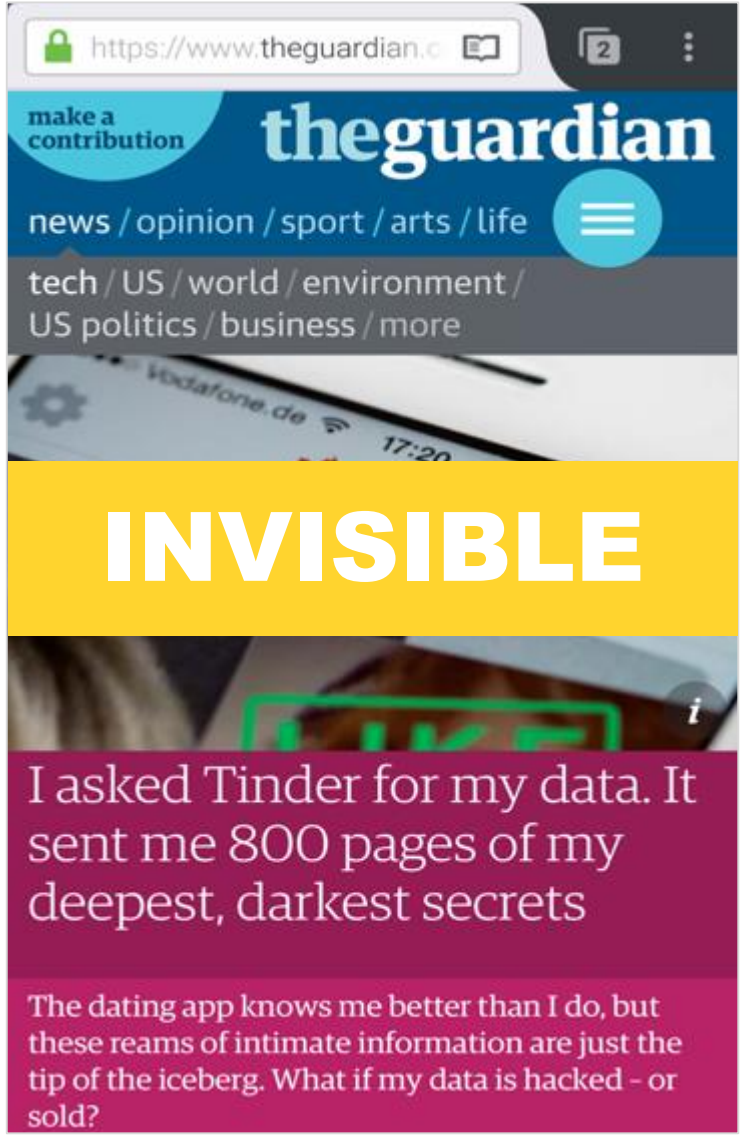
Deploying AI without anchoring to robust compliance and core values may expose a business to significant risks including employment/HR, data privacy, health and safety issues. The potential fines and sanctions can be business threatening.

EROSION OF TRUST: THE THREE I'S

IMMEDIATE

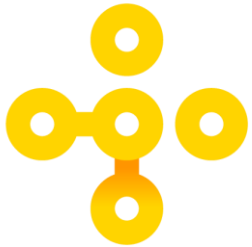


IMPACTFUL



INVISIBLE

WHAT ARE CLIENTS ASKING?



OPERATIONS

How do I guard against data breaches and hacking?

What architecture or internal frameworks do I have to build to be in compliance with GDPR and other regulations?

How can I institute ethical design practices at all levels of the AI project lifecycle?

How do I maintain and monitor my AI solutions to understand how my AI evolves over time?

HUMAN AGENCY

How do I create a personalized consumer experience without violating a customer's sense of privacy?

How do I create AI solutions that enable good decision-making by my employees and keep the human in the loop?

How can I be transparent about my AI solutions to build trust?

FUTURE OF WORK

How do I address a situation where a significant number of my employees lose their jobs due to automation?

How do I transition my employees to an AI workforce?

What are the skills needed for the future of work? How can I identify them?

How does leadership need to change or evolve in the AI economy?

HOW DO WE INTERACT WITH AI?

AI HELPS INFORM DECISIONS HUMANS MAKE

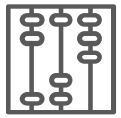
- Doctors making diagnoses
- Hiring processes
- Parole decisions

AI MAKES DECISIONS ON BEHALF OF HUMANS

- Credit risk decision
- Flagging fraudulent transactions
- Image identification or facial recognition

HOW MIGHT BIAS CREEP IN?

EXPERIMENTAL BIAS INTRODUCED BY DESIGN OR BY DATA



DATA BIAS

- Selection or sampling bias: Is your data representative of the population the model will be used on?
- Measurement bias: Both measurement instrument and operationalization can be faulty.

RESPONSE /REPORTING BIAS

- How is the data being picked up, and might that introduce bias?
- Is the data sensitive in nature; is there reason to misrepresent the truth? Will people have the same metrics of reporting (e.g., yelp effect)?

DESIGN BIAS

- What assumptions are you making about your model and its applicability to the question?
- Are you able to determine counterfactuals?

HOW MIGHT BIAS CREEP IN?
SOCIETAL BIAS

Data is not an objective truth.
It is reflective of pre-existing institutional, cultural, and social biases.

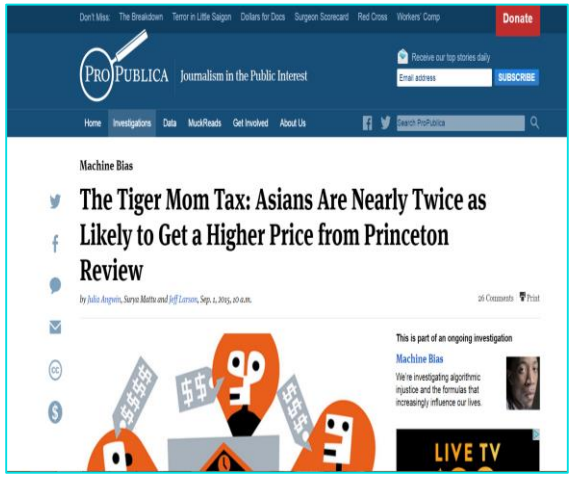
LOSS OF OPPORTUNITY

Does our system create different access outcomes for different groups relating to jobs, insurance, benefits, housing and education?

Are there reasonable alternatives that have less of a disparate impact on those groups?

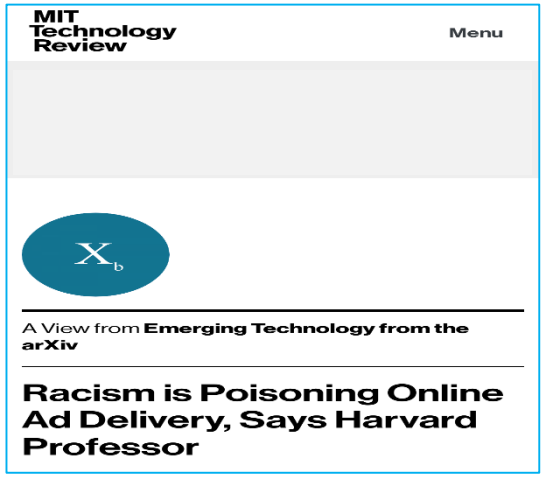
ECONOMIC LOSS

Does our system create different access outcomes for different groups relating to credit, goods and services or narrowing of choice in general?



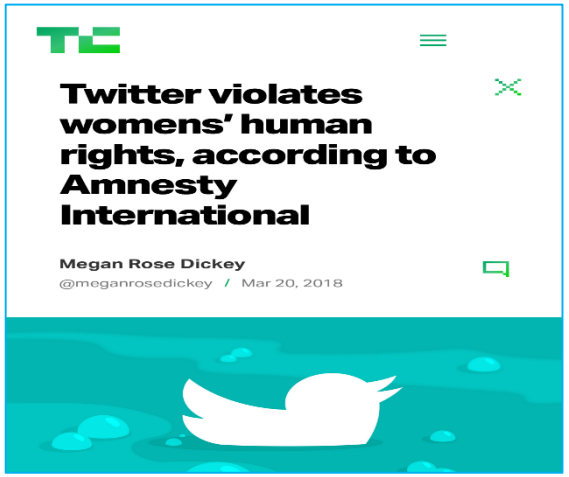
SOCIAL DETRIMENT

Does our system create filter bubbles, reinforce stereotypes or create confirmation bias?



LOSS OF LIBERTY

Does our system balance the concerns around surveillance, human rights, free speech, etc. in a way that reflects our values?



APPLIED RESPONSIBLE AI

PRINCIPLES OF ‘RESPONSIBILITY’

Explaining vs. Understanding

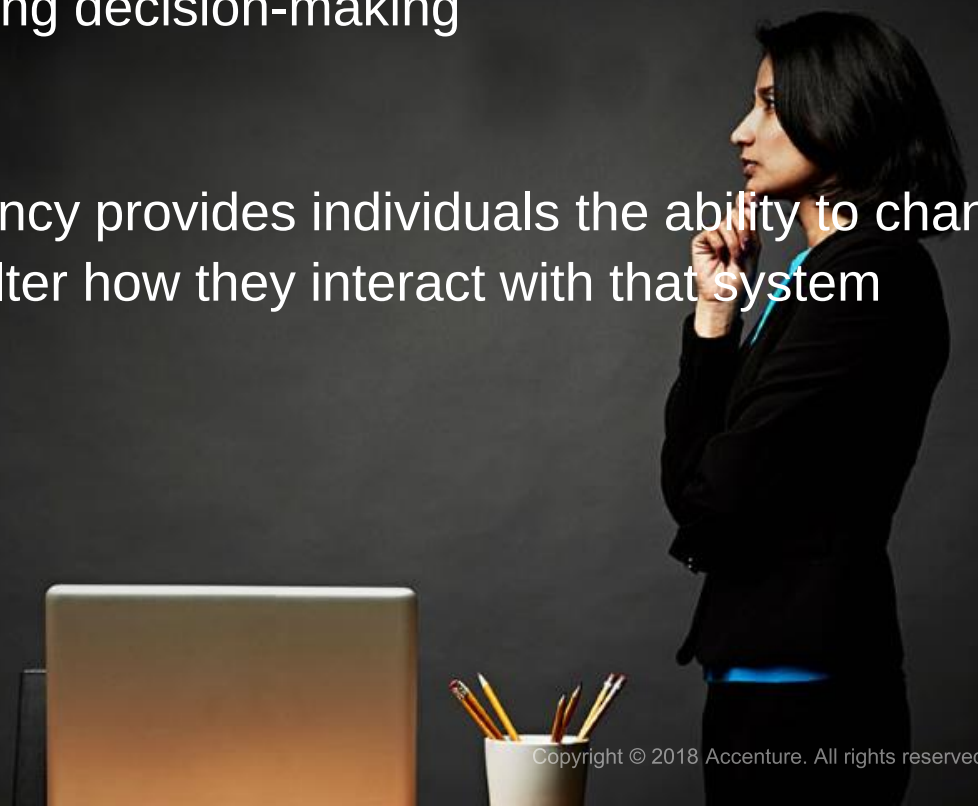
Explaining is a top-down method of providing reasoning for actions

Understanding addresses the needs of the user to provide relevant and actionable outcomes

Transparency vs. Agency

Transparency provides insights into the systems driving decision-making

Agency provides individuals the ability to change or alter how they interact with that system



Copyright © 2018 Accenture. All rights reserved.

HOW MIGHT WE ALGORITHMICALLY COMBAT UNINTENDED CONSEQUENCES?

Improving Explanation

LIME: Local Interpretable Model-Agnostic Explanations

For any classifier system, how might we provide an explanation by judiciously selecting examples to explain the algorithm's 'thought process'?

Visual Descriptions

How might visual classifiers be improved to provide human-understandable and valuable text explanations using NLP?

Identifying or Removing Bias

Path-Specific Counterfactual Fairness

If we identify sensitive variables, can we eliminate the bias introduced by them by comparing all potential causal pathways and their alternatives?

Transparent Model Distillation

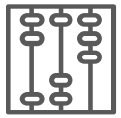
Can we 'unpack the black box' by training two new models on a black box models, and compare the outcome to identify bias?

LIME

GLOBAL UNDERSTANDING VIA LOCAL EXPLANATION

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Ribiero, Singh, Guestrin

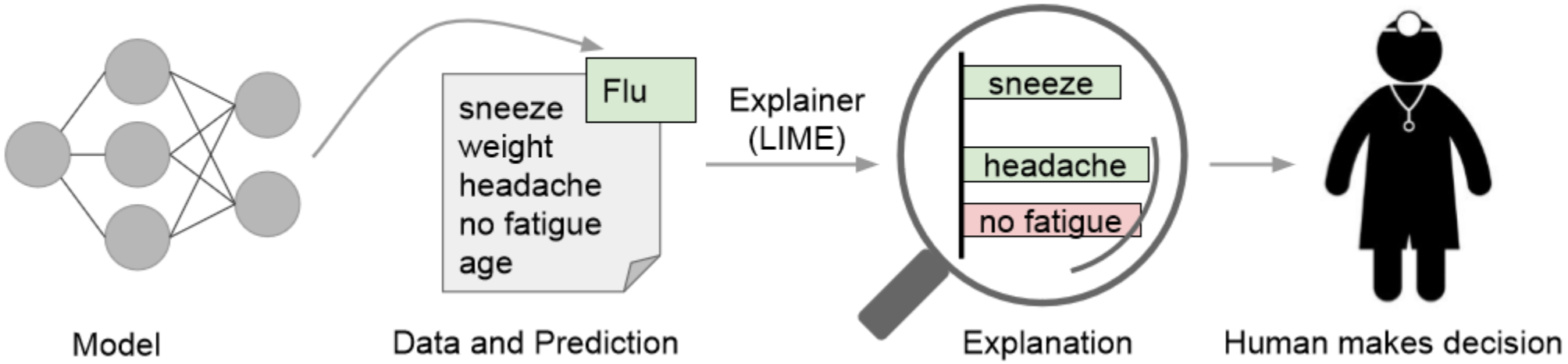


BUILDING TRUST

- People want to know that they can trust a model’s output
- “Transparency” doesn’t mean “understanding” – visualizing decision trees is transparent but not understandable

CASE SELECTION

- Specify a ‘pick step’ which prioritizes which cases to explain, given which variables have the most value
- An ‘explanation matrix’ compares instances and feature importance, so the most valuable feature is the most represented, and the subset consists of as many features as possibly by priority rank.



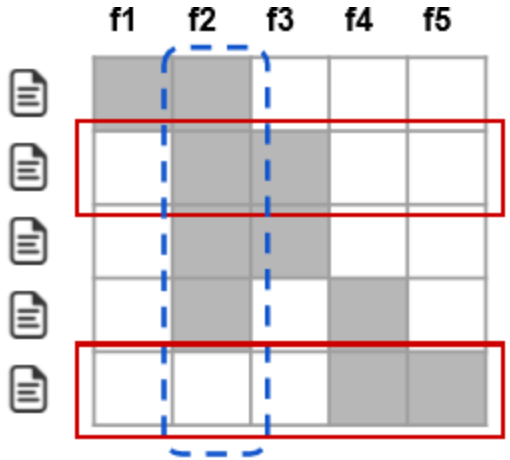


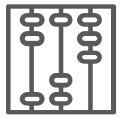
Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

VISUAL DESCRIPTION

VISUAL EXPLANATIONS USING NATURAL LANGUAGE

Generating Visual Explanations

Hendricks, et al

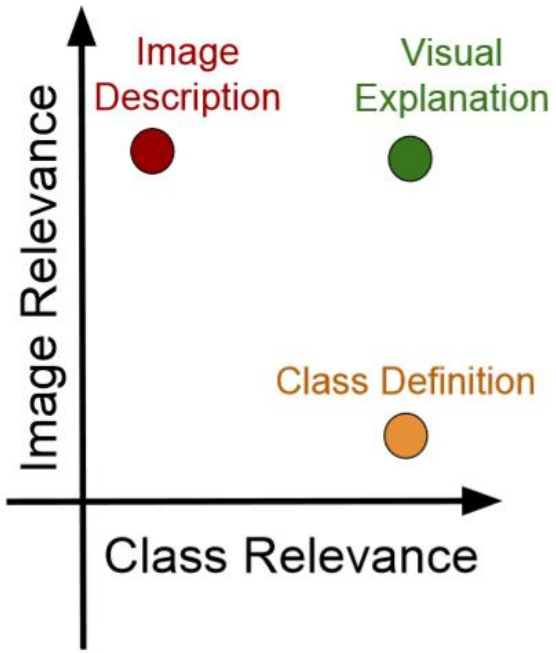


WHAT IS ‘USEFUL’?

- A useful explanation is one that is class-specific and provides discriminative features rather than simply explanatory.
- Ex: Not useful to say “This is a Barn Swallow because it has wings” – all birds have wings. What is specific to a Barn Swallow?

ALGORITHMIC USEFUL SENTENCE GENERATION

- Convolutional neural network extracts visual features + 2 LSTMs generate text output
- TF-IDF prioritizes sentences that use less common words, thus finding novel explanations



Western Grebe

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Western Grebe* is a waterbird with a yellow pointy beak, white neck and belly, and black back.
Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Laysan Albatross

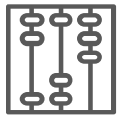
Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

PATH-SPECIFIC COUNTERFACTUAL FAIRNESS

‘UNFAIRNESS’ AS A CAUSAL EFFECT OF SENSITIVE VARIABLES



CAUSAL INFERENCE

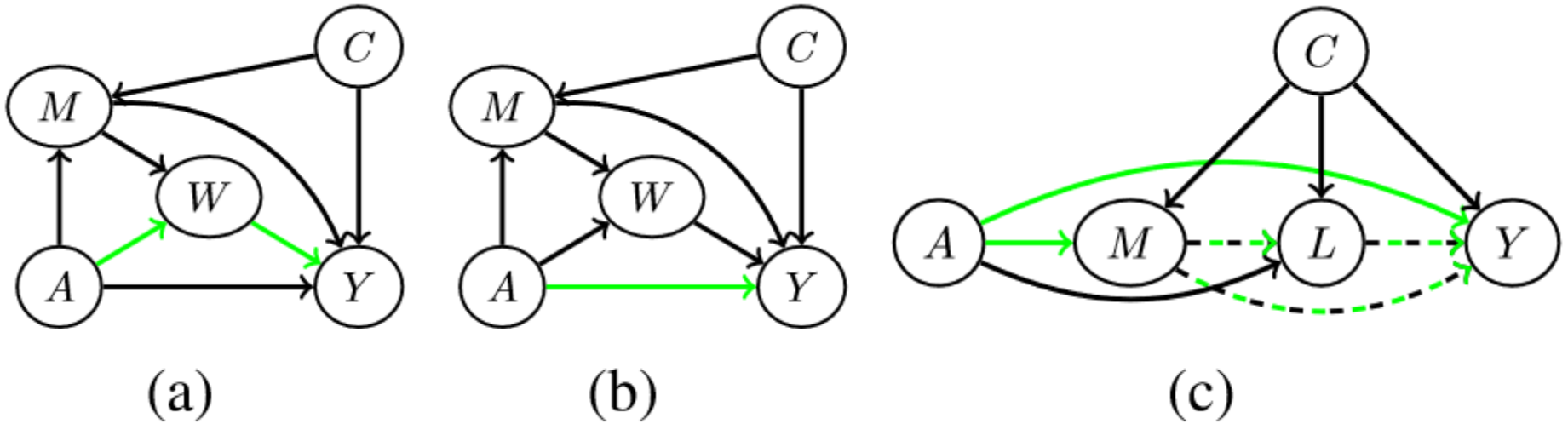
- Causality means that a particular variable has downstream impacts on ‘descendent’ variables, directly or indirectly. We can map GCM’s (graphical causal models)
- A decision is ‘fair’ if the same outcome would have occurred in expectation for the counterfactual outcome.. We limit to pathways that result from sensitive variables.

REMOVING BIAS BY ERASING UNFAIRNESS

- We calculate Maximum Mean Discrepancy (MMD) between sensitive pathways v counterfactuals.
- Introduce a latent variable to remove discrepancy by accounting for the difference.

Path-Specific Counterfactual Fairness

Chiappa, Gillam



WE ISOLATE OUR GCM TO ACCOUNT FOR THE DOWNSTREAM IMPACTS OF SENSITIVE VARIABLE A.
A) IS INDIRECT IMPACT
B) IS DIRECT IMPACT
C) ILLUSTRATES BOTH

Table 1. In order columns represent: unfair test accuracy, fair test accuracy, and MMD values for H_m, H_l, and H_r ($\times$ 10,000) for the UCI Adult dataset. Rows represent values after 5,000, 8,000, 15,000, and 20,000 training steps.

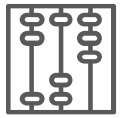
82.88%	81.66%	610.85	13.31	3.73	3.10	3.12
82.85%	80.21%	6.73	2.80	3.75	2.88	3.10
82.71%	79.41%	2.97	3.45	0.25	0.07	0.49
80.60%	73.98%	3.19	6.31	0.22	0.10	0.47

TRANSPARENT MODEL DISTILLATION

THE STUDENT BECOMES THE TEACHER

Auditing Black-Box Models Using Transparent Model Distillation With Side Information

Tan, et al

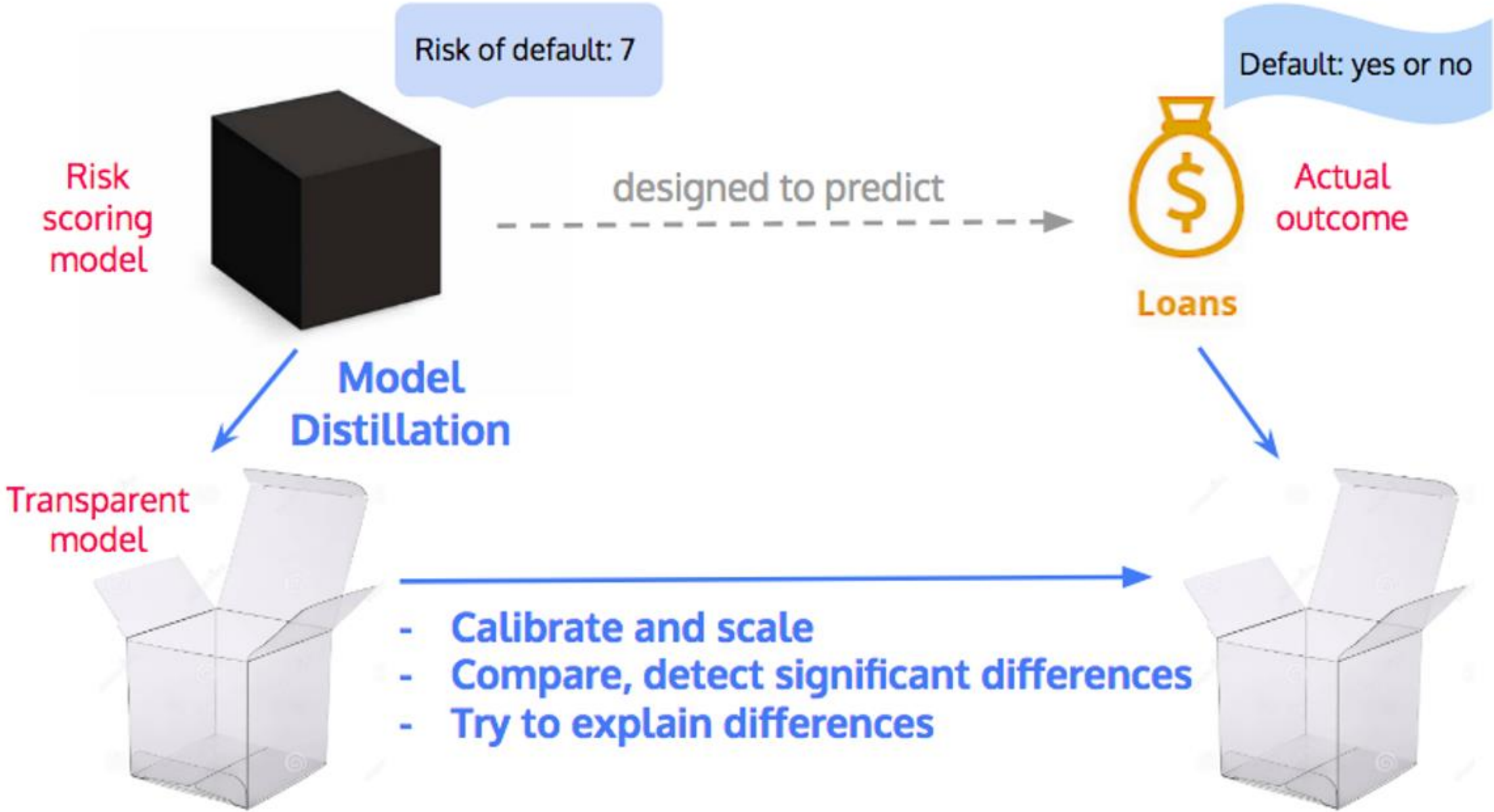


STUDENT MODEL A TRAINS ON BLACK BOX MODEL OUTPUT

- Proprietary or black box models allow no insight into their risk function. Outcomes may be biased, but we don't know what led to that bias.
- We train an explainable student model on the outcome of a black box model, so we have an estimate of the original model's risk function.

STUDENT MODEL B TRAINS ON ACTUAL OUTCOMES

- A second explainable model is trained on the true outcomes.
- Any difference between model A and model B is bias in the original algorithm.


Risk of default: 7
Default: yes or no
Risk scoring model
designed to predict
Actual outcome
Loans
Model Distillation
Transparent model
- Calibrate and scale
- Compare, detect significant differences
- Try to explain differences

Transparent Model Distillation, With Side Information

Student model of black-box risk score teacher r_S, trained on teacher's outputs y^S:

Input: label y^S, features $\boldsymbol{x}$

Output: prediction $\hat{y}^S$

Model functional form: $\hat{y}^S = f^S(\boldsymbol{x})$

Model of actual outcome y_O:

Input: label y^O, features $\boldsymbol{x}$ (same as student model), same data points as student model

Output: prediction $\hat{y}^O$

Model functional form: $\text{logit}(\hat{y}^O) = f^O(\boldsymbol{x})$, since $\hat{y}^O$ is binary. This model is not a student model, as actual outcomes are not labeled by another model.

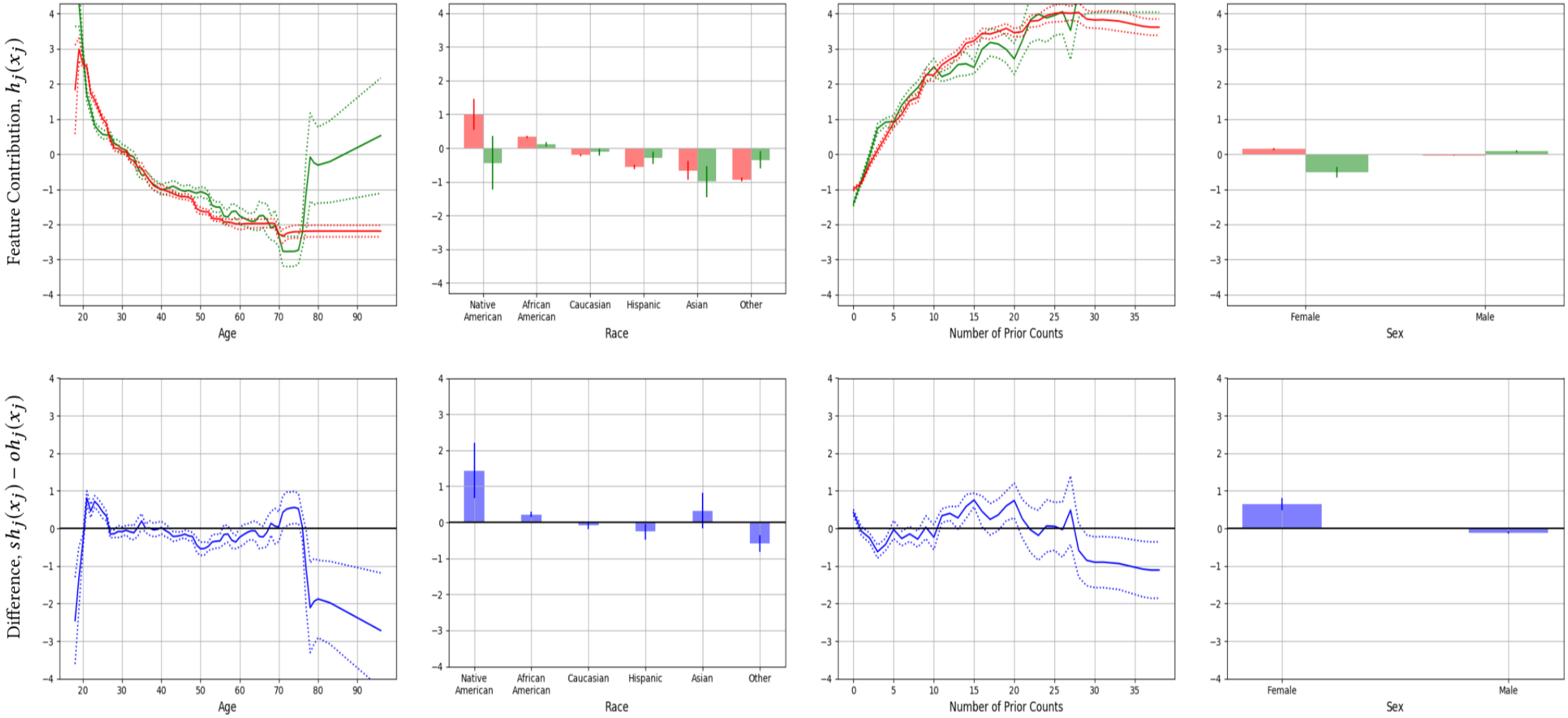


Figure 4: GA2M shaped feature contributions for four features to the COMPAS risk score student model (in red) and the actual recidivism outcome model (in green). For categorical features such as race and sex, categories are ordered in decreasing importance to the score student model. The blue line captures the difference between the two models (score student - outcome). All plots mean-centered on the vertical axes. See Appendix for additional features and interactions between *pairs* of features.

CONCLUSIONS

- **AI applied in the real world can lead to unintended biased outcomes.**
- **Not as simple as removing a sensitive variable:**
  - Sometimes you can't without sacrificing significant explanatory value
  - Sometimes correlations pick up the variable as a proxy (COMPAS)
- **Algorithmic solutions for identifying and possibly neutralizing bias are feasible.**
- **Need a combination of good design, good explanations, and algorithmic solutioning.**

THANK YOU

DR. RUMMAN CHOWDHURY
GLOBAL LEAD, RESPONSIBLE AI
ACCENTURE APPLIED INTELLIGENCE
@RUCHOWDH
WWW.RUMMANCHOWDHURY.COM