ILTACON

Fair and Ethical AI: A Closer Look At the Legal Domain

Corey Ouellette
Customer Center Lead, TR Labs

John Hudzina
Lead Research Engineer, TR Labs

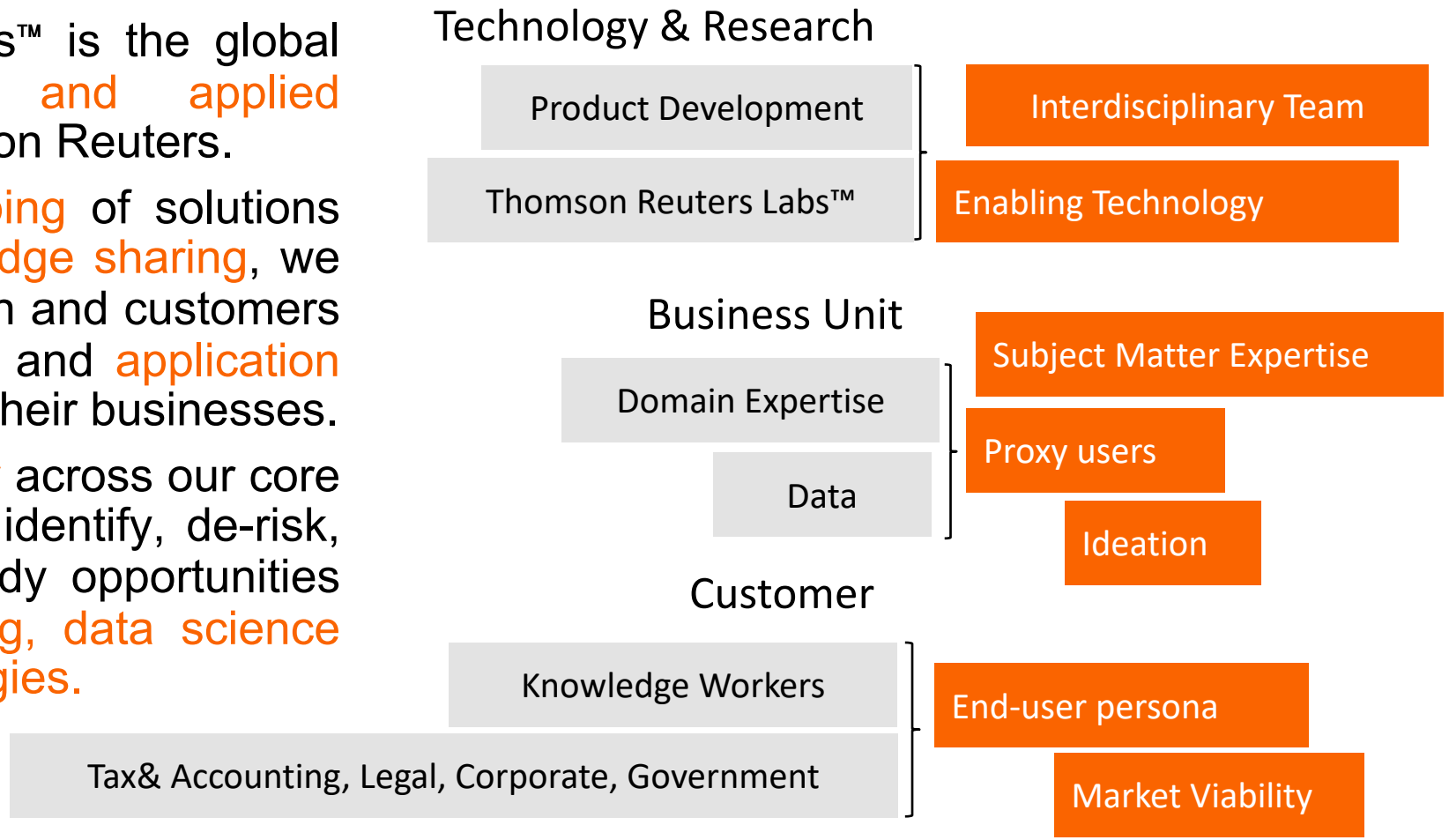
Dhivya Chinnappa
Research Scientist, TR Labs

Dawn Sepehr
Research Scientist, TR Labs

Aug 24, 2021

Thomson Reuters Labs

- Thomson Reuters Labs™ is the global reaching innovation and applied research arm of Thomson Reuters.
- Through rapid prototyping of solutions and continuous knowledge sharing, we support our organization and customers with the understanding and application of new technologies to their businesses.
- We work collaboratively across our core customer segments to identify, de-risk, and activate future-ready opportunities in AI, machine learning, data science and emerging technologies.



Thomson Reuters Labs: Skills and Locations

Technical Functions

Data Science & Research

- Natural Language Processing
- Machine Learning & Neural Networks
- Search & Text Similarity

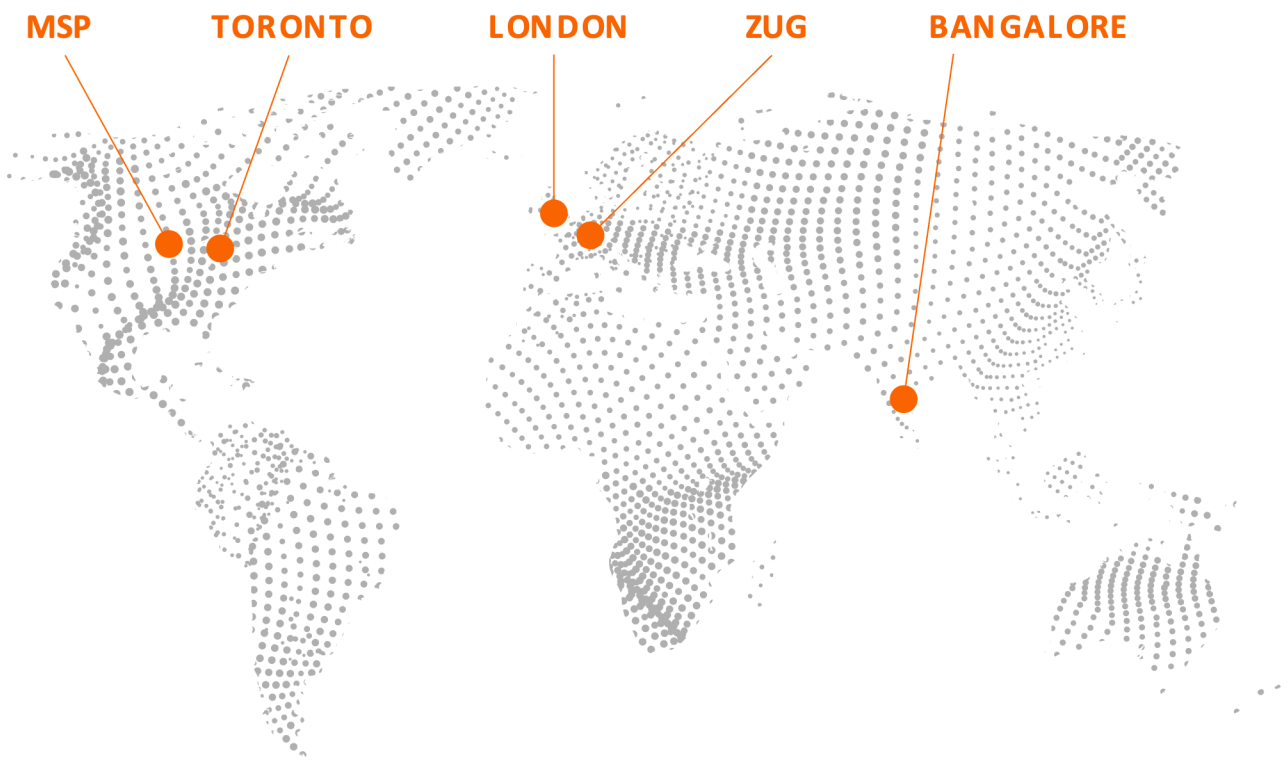
Engineering

- Cloud
- Machine Learning
- Data Engineering

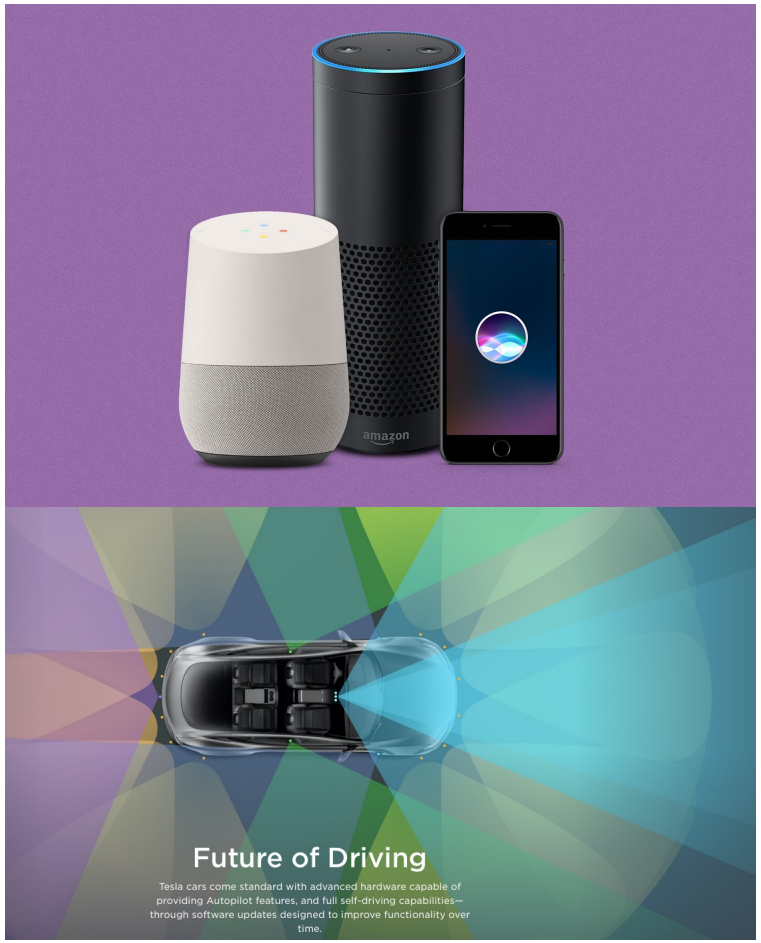
Design & UX

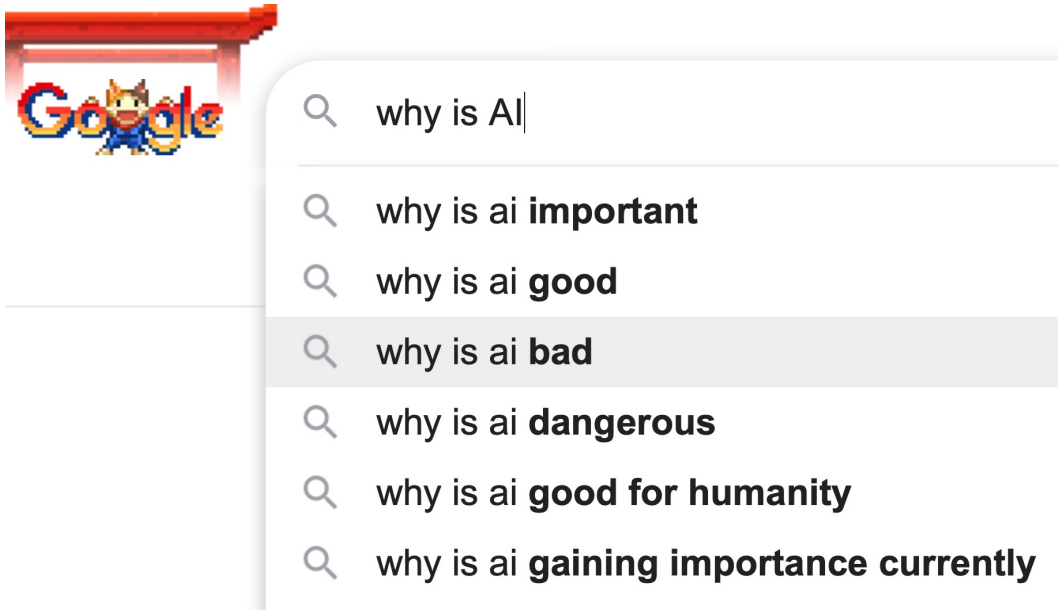
- Human+AI Interaction
- Customer Co-creation Workshops
- Data Visualization

100+ talented colleagues operating globally



AI Everywhere





Fairness and Trustworthiness of AI

REUTERS

World Business Markets Breakingviews

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

BIGOTRY ENCODED: RACIAL BIAS IN TECHNOLOGY

by Taylor Synclair Goethe | published Mar. 2nd, 2019

illustration by Aria Dines

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Anja Lambrecht and Catherine Tucker*

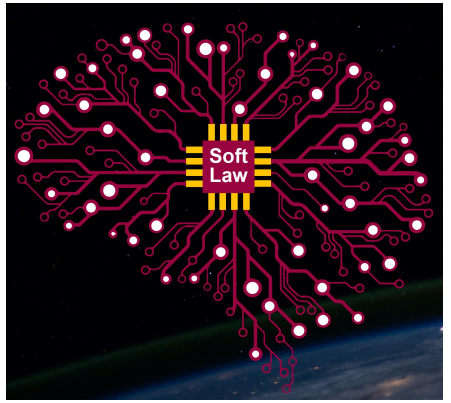
March 9, 2018

Legal Requirements

- EU: General Data Protection Regulation (GDPR)
  - ✓ Art. 13: Information to be provided where personal data are collected from the data subject
  - ✓ Art. 14: Information to be provided where personal data have not been obtained from the data subject
    - Providing further information necessary to ensure fair and transparent processing
    - Meaningful information about the logic involved in the case of existence of automated decision-making
- US:
  - ✓ Equal Credit Opportunity Act: prohibits creditors from discriminating in any aspect of a credit transaction on the basis of an applicant's race, color, religion, national origin, sex, marital status or age
  - ✓ California Consumer Privacy Act (CCPA): The right to know about the personal information a business collects about them and how it is used and shared

[https://gdpr-info.eu]
[https://www.ftc.gov/enforcement/statutes/equal-credit-opportunity-act]
[https://www.oag.ca.gov/system/files/attachments/press_releases/CCPA%20Fact%20Sheet%20%2800000002%29.pdf]

Soft Law Governance of AI



- Study by Arizona State University (ASU) published in 2021*
  - "Soft Law: a program that sets substantive expectations, but is not directly enforceable by government"
  - Not bounded by a geographic jurisdiction and can be developed, amended, and adopted by any entity
- Goal: inform decision-makers with evidence, practices, and recommendations that can be harnessed to enhance soft law programs
- Identified 634 soft law AI programs
  - 90% were published between 2016-2019
  - Limited geographical diversity

  (mostly countries in Europe and NA)

Table 11 - Top Five Themes and Sub-Themes

Theme	# of labels
1 Education – displacement of labor	815
2 Transparency and explainability	805
3 Ethics	776
4 Security	591
5 Bias	506

Sub-theme	% of database
1 General transparency	43.38%
2 General mention of discrimination or bias	38.33%
3 AI literacy	38.33%
4 Acting in favor of AI ethics	29.34%
5 Human control and involvement in AI decision-making	27.92%

- National Institute of Standards and Technology+
  - ✓ a step towards consensus standards and a risk-based framework for trustworthy and responsible AI

*[Gutierrez, Carlos Ignacio and Marchant, Gary E., A Global Perspective of Soft Law Programs for the Governance of Artificial Intelligence, 2021]
+[A Proposal for Identifying and Managing Bias within Artificial Intelligence, 2021]

Thomson Reuters AI Principles

At Thomson Reuters, trust is one of our most important values.

Thomson Reuters has drafted these AI principles to promote trustworthiness in our continuous design, development, and deployment of AI and they will evolve as the field of AI matures:

1. That Thomson Reuters will prioritize **safety, security, and privacy** throughout the design, development and deployment of our AI products and services.
2. That Thomson Reuters will strive to maintain a **human-centric approach**, and will strive to design, develop and deploy AI products and services that **treat people fairly**.

Complete list of our AI principles can be found here or scanned here:

Agenda

- How Bias Finds its Way into AI Models
- Bias and the Legal Industry
- Practical Example of Bias

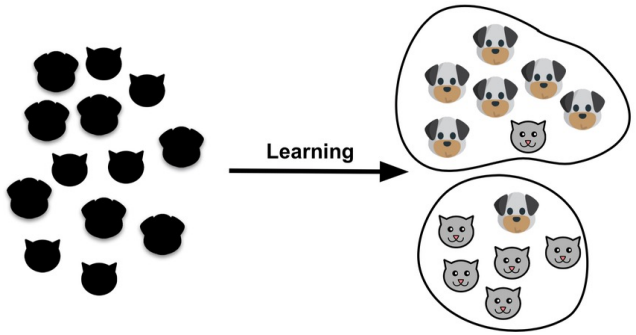
Where are you in your implementation of AI based systems for your firm/legal department ?

1. Just starting to take a look
2. Have attended one or more presentations and/or met with our advisors
3. We are using AI to automate some of our repetitive tasks
4. We have several AI systems in place and are moving forward towards implementing compliance or planning solutions
5. Doesn’t apply to our company

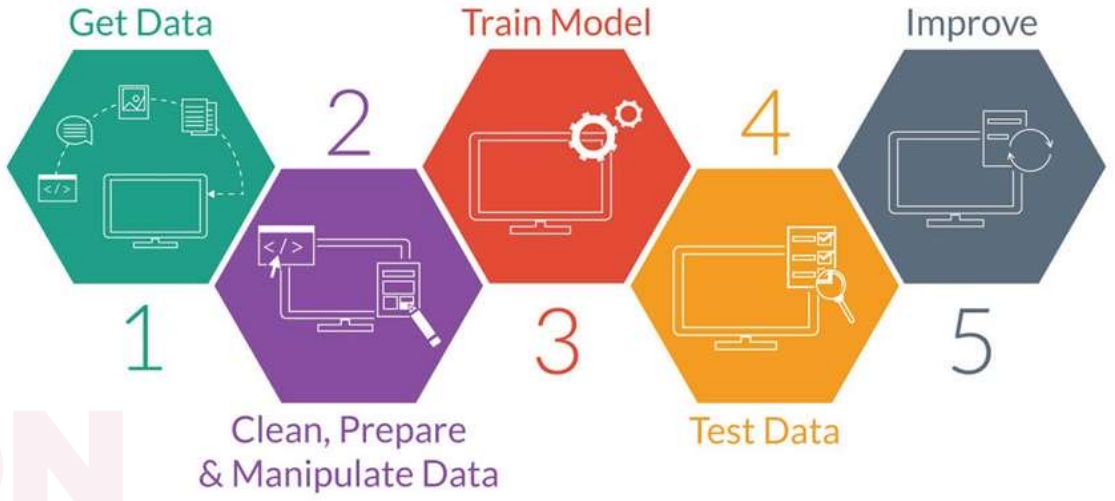
How Bias Finds its Way to AI Models

Bias vs Fairness

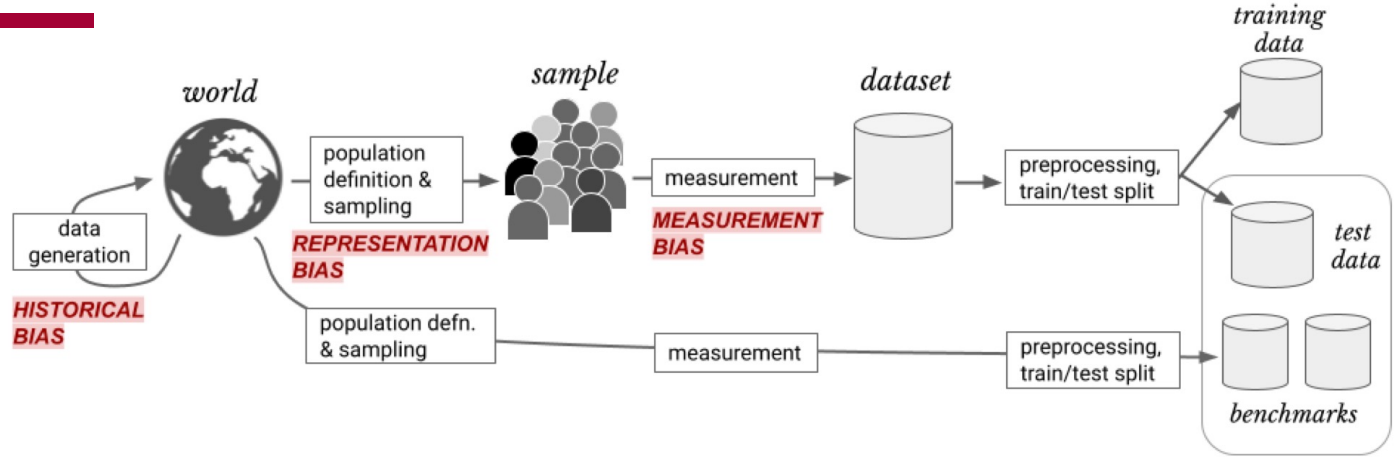
- Not all AI systems are unfair



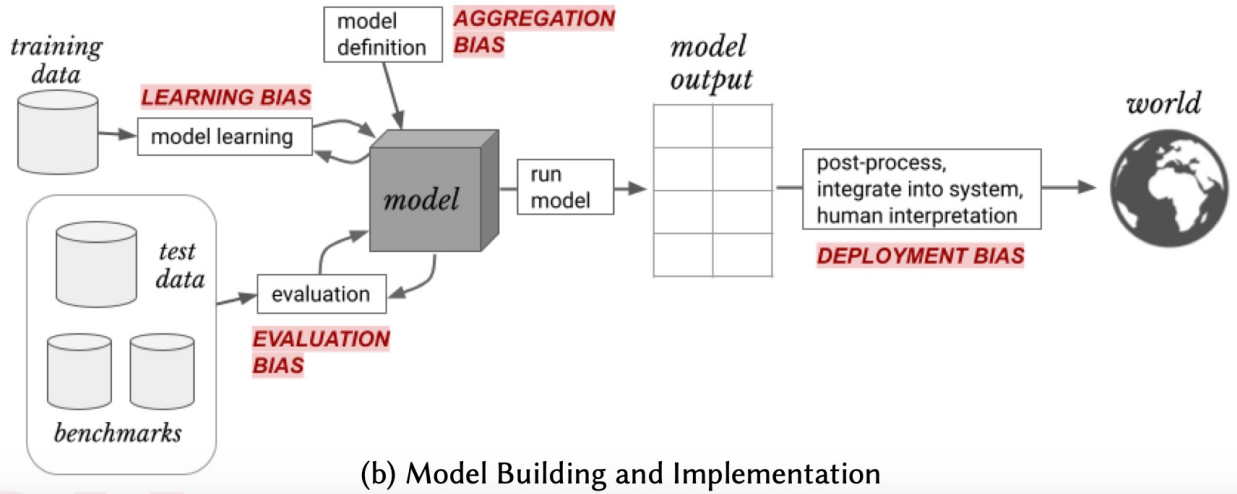
- Data vs Model



Different Types of Bias in the Machine Learning (ML) Pipeline



(a) Data Generation



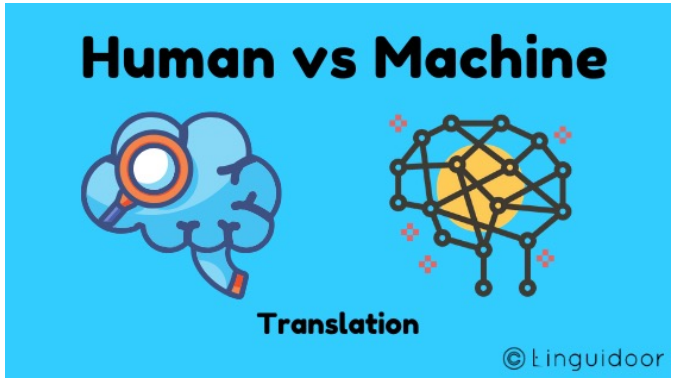
(b) Model Building and Implementation

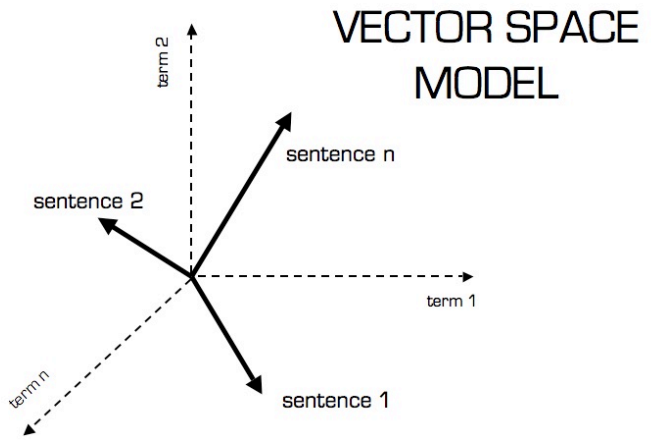
[Harini Suresh and John V. Guttag , A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, arXiv, 1901.10002, 2021]

Data Generation: Historical Bias

1. Historical Bias

- ❑ Different Types of Embeddings (e.g., word2vec, GloVe, BERT)
- ❑ Learned from large corpus of text → nurse, women and engineer, men





Window Size	Text	Skip-grams
2	[The **wide** road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road **shimmered** in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot **sun**].	sun, the sun, hot

[https://www.tensorflow.org/tutorials/text/word2vec]

Data Generation: Representation Bias

2. Representation Bias

- Happens when the development sample under-represents some part of the population:
  - ✓ Training data distribution do not reflect the application data distribution
  - ✓ Training data contains under-represented groups
  - ✓ Sampling bias
- Geographic Diversity in Image Datasets:
  - ✓ ImageNet: 45% of the images from the United States, the remaining is mostly from North America and Western Europe

Data Generation: Measurement Bias

3. Measurement Bias

- ❑ Happens when choosing, collecting, or computing **features and labels** to use in a prediction problem
- ❑ Feature or label is a proxy (a concrete measurement) for some construct (an idea or concept)
    - ✓ Different qualities of the proxy for different groups in the dataset
- ❑ Racial bias detected in predictive algorithms to help patients with complex health needs: At a given risk score, Black patients are considerably sicker than White patients
    - ✓ Unequal access to health care
    - ✓ Health care costs used as a proxy for risk

[Ziad Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations, 2019]

Model Building: Aggregation Bias

5. Aggregation Bias

- ❑ Happens when a one-size-fits-all model is used
  - ✓ Data contains underlying groups or types of examples that should be considered differently (e.g., people or groups with different backgrounds, cultures or norms)
- ❑ Can lead to a model that is not optimal for any group, or a model that is fit to the dominant population
- ❑ Social Media Analysis: Analyzing Twitter posts of gang-involved youth in Chicago
  - ✓ Harmful misclassification of the tweets

	Loss			Other			Aggression			
Labels	p	r	f	p	r	f	p	r	f	Macro F1
Gold	77.08	56.92	65.49	88.04	95.76	91.74	50	27.59	35.56	64.26
Distant	50.00	48.46	49.22	85.63	84.50	85.06	19.72	24.14	21.71	52.00

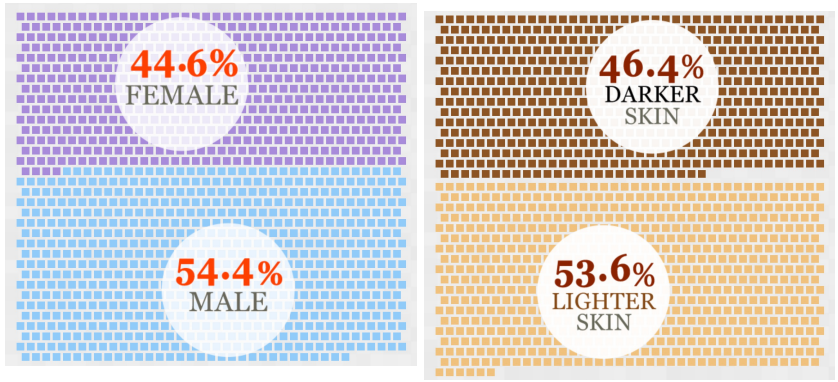
Table 2: SVM performance trained on hand-labeled vs distantly-labeled data. The difference between F1 scores is statistically significant with p=0.001.

[Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing , 2020]

Model Building: Evaluation Bias

6. Evaluation Bias

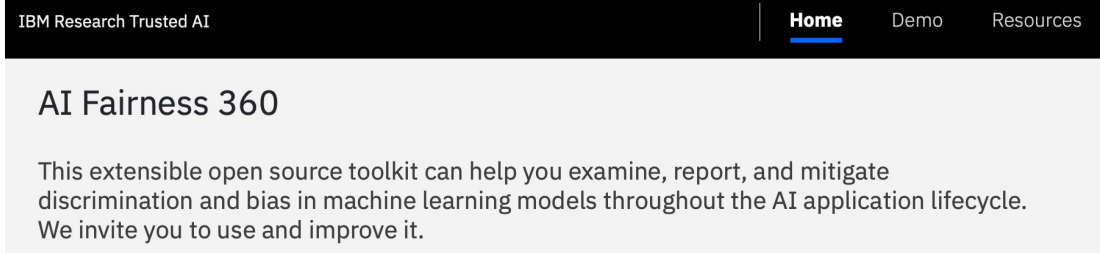
- Happens when evaluating a model, if the benchmark data does not represent the population that the model will serve
- Commercial gender classification AI systems



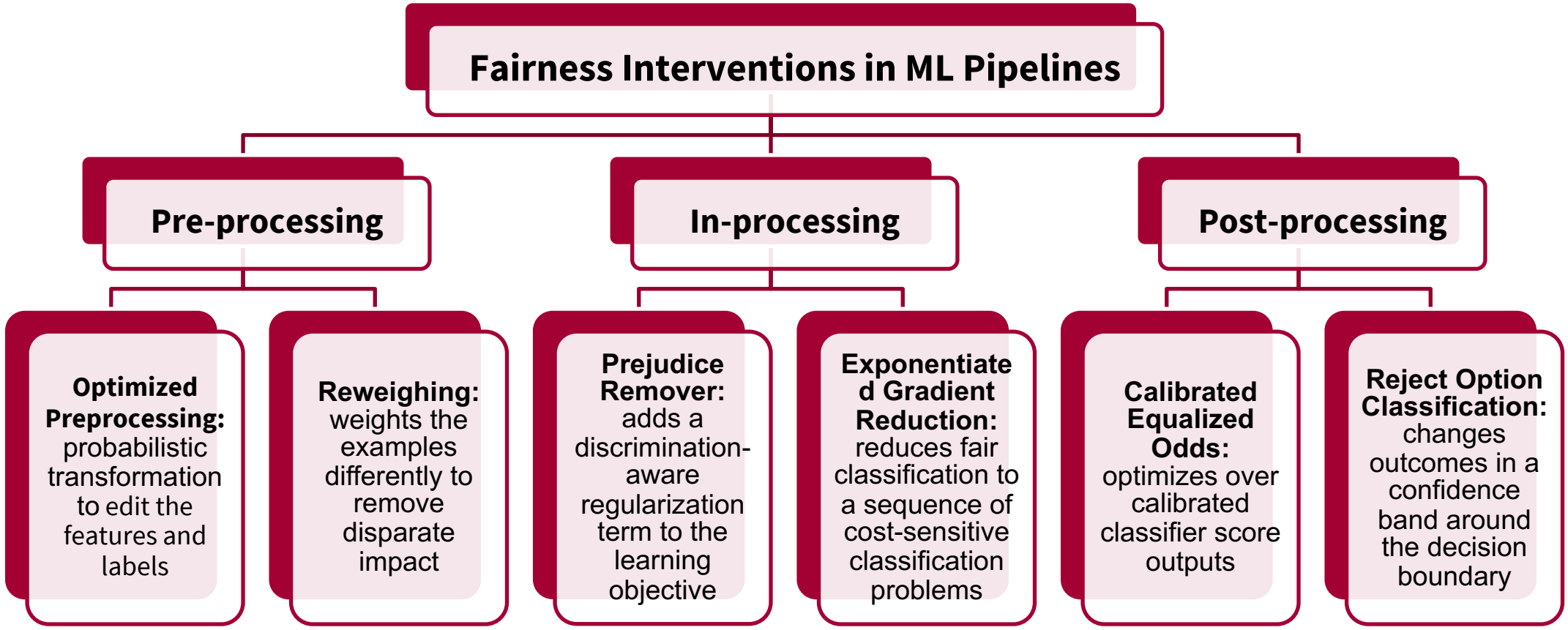
Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

[Gender Shades Project, http://gendershades.org/overview.html]



Detection and Mitigation Strategies



[Practical Fairness, Aileen Nielsen, O'Reilly Media, 2020]
[AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, Rachel K. E. Bellamy et al., 2018]

What are the most significant barriers your organization is experiencing with adopting AI? (Please select all that apply)

1. Concerns over job elimination
2. Concerns over the privacy and security of data
3. Concerns over AI fairness and trustworthiness
4. Lack of organizational commitment or strategy to support AI
5. Lack of technological infrastructure to support AI
6. Lack of talent and appropriate skill sets to do AI work

Bias and the Legal Industry

AI Systems in the Legal Domain

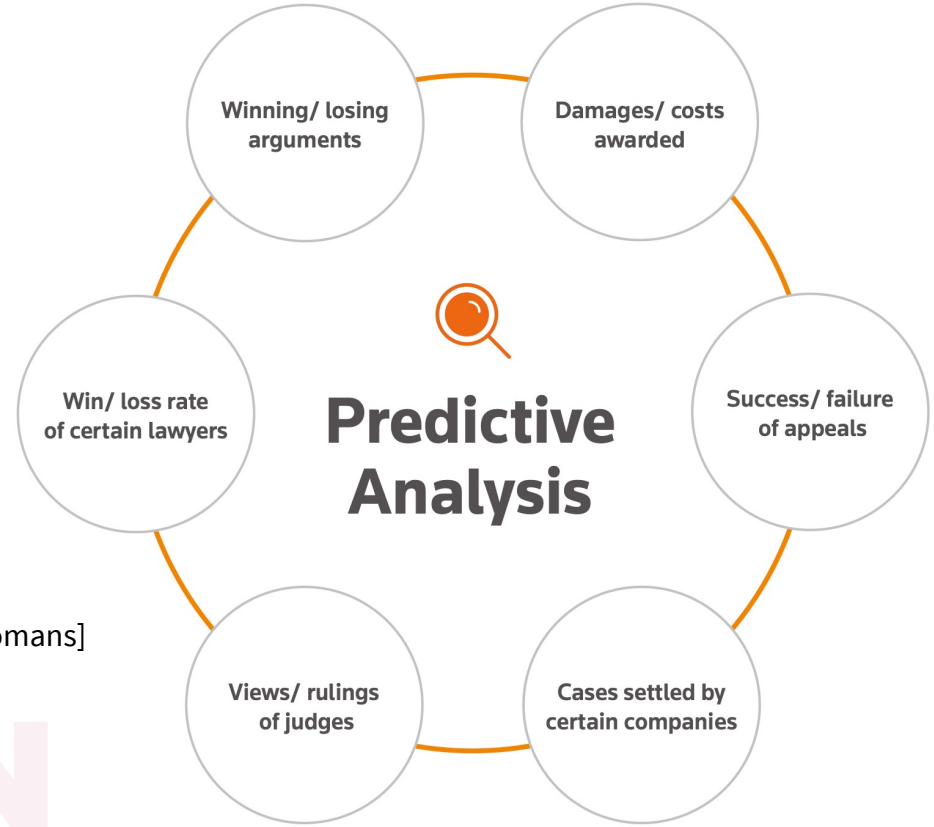
Three main branches of legal AI

[Legal AI, A beginner's guide, By Richard Tromans]

AI Systems in the Legal Domain (cont'd)

Legal Data Research:

- ❑ Knowledge systems: Legal research along practice lines
- ❑ Predictive systems: Case outcome prediction based on specific matters and/or litigation trends based on court outcomes



[Legal AI, A beginner's guide, By Richard Tromans]

Bias in the Legal System

- Biases in the law:
  - explicit biases, conscious expressions of prejudice
  - implicit biases, subconscious prejudices
- Demonstrated in:
  - criminal sentencing: “black defendants are sentenced to almost two months more in prison compared to their white counterparts”[1]
  - stop-and-frisk policy: “persons of African and Hispanic descent were stopped more frequently than whites”[2]
  - application of police force: “racial disparities in police use of force persist even when controlling for racial distribution of local arrest rates”[3]

[1][Crystal Yang, Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing, 2015]
[2][Andrew Gelman et al., An Analysis of the New York City Police Department’s “Stop-and-Frisk” Policy in the Context of Claims of Racial Bias, 2007]
[3][Goff P, Lloyd T, Geller A, Raphael S and Glaser J, The science of justice: Race, arrests, and policy use of force, 2016]

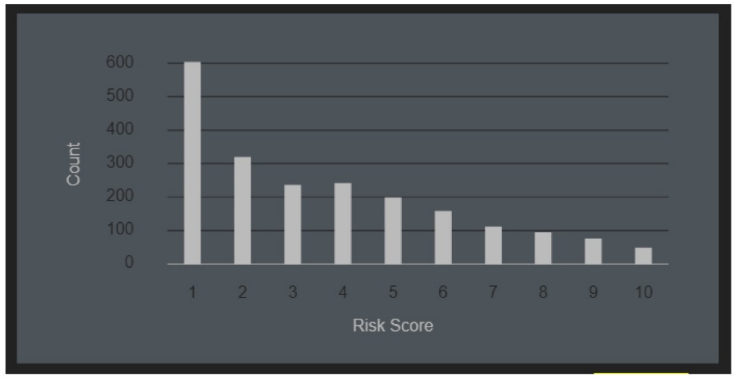
Biased AI Systems in the Legal Domain

- Risk Assessments in the Criminal Justice System
  - COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism system to predict the likelihood that a defendant will re-offend

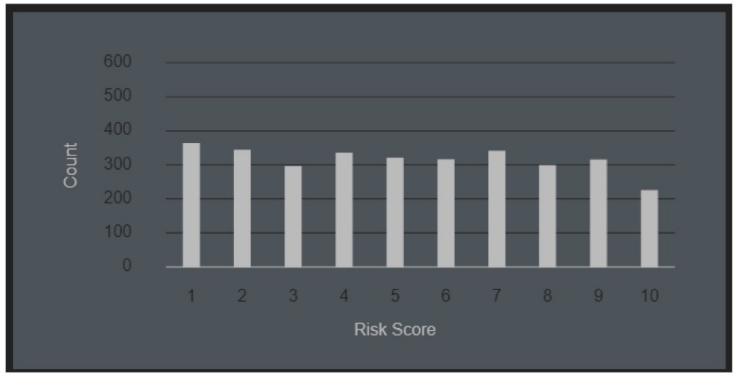
These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not.

[ProPublica analysis of data from Broward County, Fla.]

White Defendants' Risk Scores



Black Defendants' Risk Scores



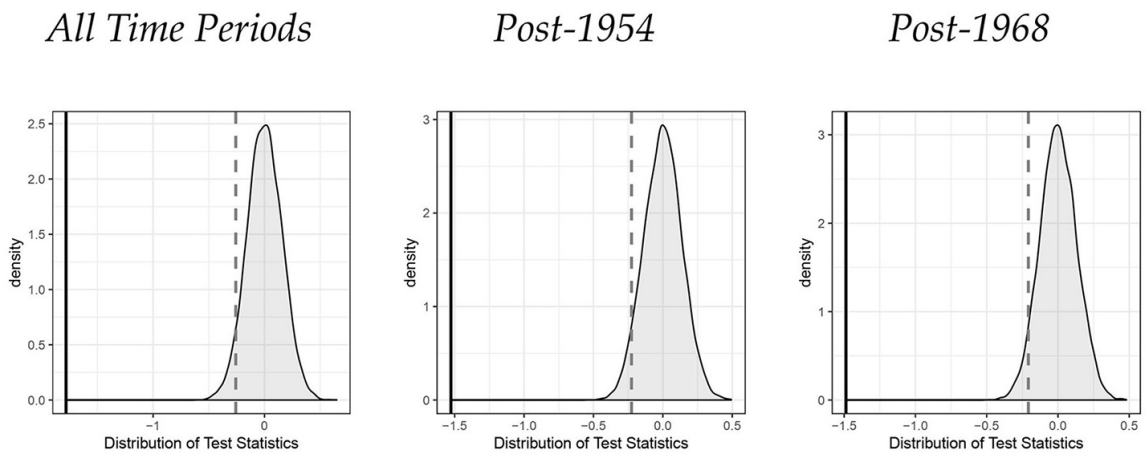
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Biased AI Systems in the Legal Domain (cont'd)

- Implicit racial bias found in judicial opinions
  - ✓ Dataset: more than 1 million appellate court opinions from US state and federal courts
  - ✓ Method: Word Embedding Association Test (WEAT)
  - ✓ Observation: African-American names are more frequently associated with unpleasant/negative concepts, European-American names are more frequently associated with pleasant/positive concepts
  - ✓ Similar observations for US supreme court, US courts of appeal, and state courts of last resort data

Figure 1. Distributions of test statistics from 10,000 iterations. These plots provide the distribution of test statistics estimated across random samples of the target characteristic (pleasant / unpleasant). The vertical grey dashed line indicates a one-sided 5% significance test, and the vertical black line indicates the observed test statistics for indicated opinions.



[Douglas Rice, et al., Racial bias in legal language, 2019]

Challenges of Bias Detection in Legal Systems

- Adapting the tests for legal language
  - ✓ Legal specific terminology (e.g., pro hac vice means for this time only and isn't the combination of pro, hac, & vice)
  - ✓ First names vs pronouns
  - ✓ Positive/negative words in legal vs social media
  - ✓ Normative text vs legal text (e.g., murder vs breathing)

Practical Example

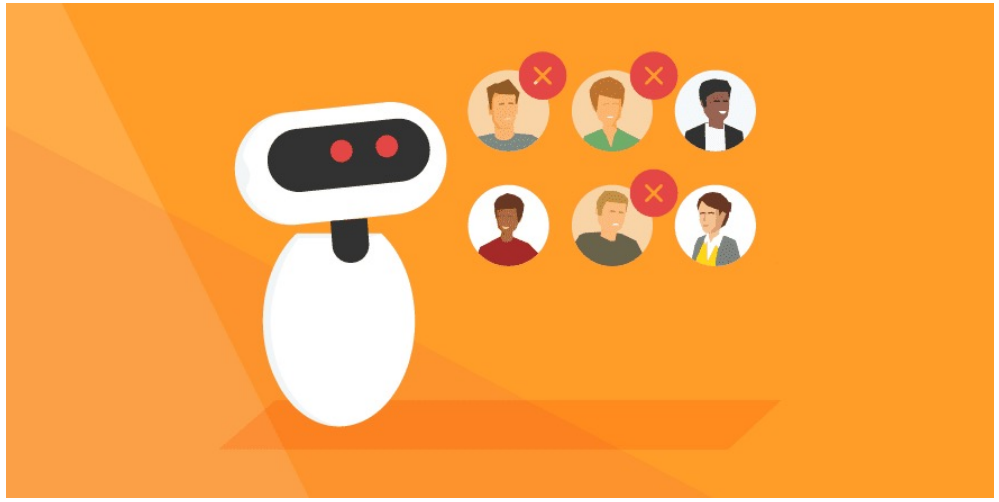
Practical Example

You can follow along with us here or scanned here:

Final version - public

Concluding Remarks

- ❑ Consequences of unchecked AI systems
- ❑ Legal regulations and soft law governance for AI
- ❑ Sources of bias (data and model) and mitigation strategies
- ❑ Legal domain specific concerns



What are your future concerns related to AI ?
(Please select all that apply)

1. Who is ultimately accountable for AI based decisions?
2. Ethical issues that may arise as AI becomes more powerful.
3. Our competitors will learn how to use AI more effectively than we do.
4. None – We expect AI to have only a positive impact on our organization.

Thank you!