



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

Fairness

9/28/26

Some additional notes from last week

- Topic 1:
 - Numerical discrepancy in estimates of inference costs
- Topic 2: [Han, Y., et al. "The unpaid toll: quantifying the public health impact of AI, arXiv." arXiv preprint arXiv:2412.06288 \(2024\).](#)
 - [Discussion focused on the news article, with less discussion of the public health impacts]
 - Anjalie: Do we trust the numbers in the paper?
 - I could not replicate them
 - Krithika: The paper strongly emphasize 1300 deaths from asthma. Is this a real concern? [compared to leading cause of deaths, and causes of asthma other than data centers]
 - What aspects of data centers should we be concerned about and what are overhyped?

Some additional notes from last week

- Krithika's notes:
 - https://docs.proton.me/doc?mode=open-url&token=J3H644YYP8&linkId=ZZmHN0rFEgBalySB3KvJfA4cuNtB1xIna50TYEC2MY08-hsVAcBk96MtqNvLQEEYoZJ03sCkn8_epbmyfe_l6A%3D%3D#TIdUX6TQIvUt



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

Origins: Risk Assessment and Fairness Metrics

A little background: ProPublica COMPAS Report (2016)

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

A little background: risk assessment

- U.S. courtrooms have employed various forms of **risk assessment**
- Given someone that has been arrested for a crime, models are typically trained to predict if they will be arrested again in the future
 - “Recidivism” if a convicted person will “reoffend”
- Used at all stages of criminal justice system:
 - When determining if someone can be released while they are awaiting trial
 - When determining programs while they are incarcerated
 - Level of supervision (home confinement, electronic monitoring) when they are released

A little background: risk assessment

- Consistency
 - “can be viewed as more defensible and credible than more subjective and less transparent decision-making processes”
 - Judges give more favorable rulings right after food breaks
- Efficiency
 - “help practitioners make more efficient use of limited justice resources”
- Effectiveness
 - “help practitioners more effectively improve criminal justice outcomes (e.g., reduce reoffending, improve compliance)”

A little background: ProPublica COMPAS Report (2016)

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

- Obtained risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014.
 - COMPAS: the Correctional Offender Management Profiling for Alternative Sanctions
- Checked to see how many were charged with new crimes over the next two years (same evaluation criteria used by the developers of the algorithm, a company called Northpointe)
 - Only 20% of the people predicted to commit violent crimes actually went on to do so
- "The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants."
- "White defendants were mislabeled as low risk more often than black defendants."

Fairness metric: Equalized Odds

- The ProPublic Report identified unfairness by examining disparities in:
 - False positive rate (FPR): Black defendants were nearly twice as likely to be misclassified as high risk compared to white defendants
 - $$\text{FPR} = \frac{FP}{FP+TN} = \frac{FP}{\text{Actual Negatives}}$$
 - False negative rate (FNR): Among defendants who *did* reoffend, white defendants were more likely to be mislabeled as low risk than Black defendants.
 - $$\text{FNR} = \frac{FN}{FN+TP} = \frac{FN}{\text{Actual Positives}}$$

Fairness metric: Equalized Odds

- A more general form:
 - $P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = b)$ for $y \in \{0, 1\}$
 - The classifier's prediction $\hat{Y}$ must be independent from a protected attribute A (such as race), conditional on the true outcome Y
- To satisfy this,
 - For $y = 1$, true positive rates [false negative] must be equal across values of A
 - For $y = 0$, false positive [true negative] rates must be equal across values of A
- Note that a perfect model, $Y = \hat{Y}$ would satisfy this

Response to ProPublica Report

- Northpointe release a response to the ProPublica article several months later: [1] copy of the response that ProPublica annotated and also responded to
- Criticism about the subset of data used and the variables, given where COMPAS is used in the county they are studying (Flores et al. 2016)
- Criticism about the choice of fairness metric (both)

[1] <https://embed.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616/>

Flores AW, Bechtel K, Lowenkamp CT. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." 2016. Available at: <http://www.crj.org/page/-/publications/rejoinder7.11.pdf>

Dieterich W, Mendoza C, Brennan T. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016. Available at: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf (last accessed February 6, 2017).

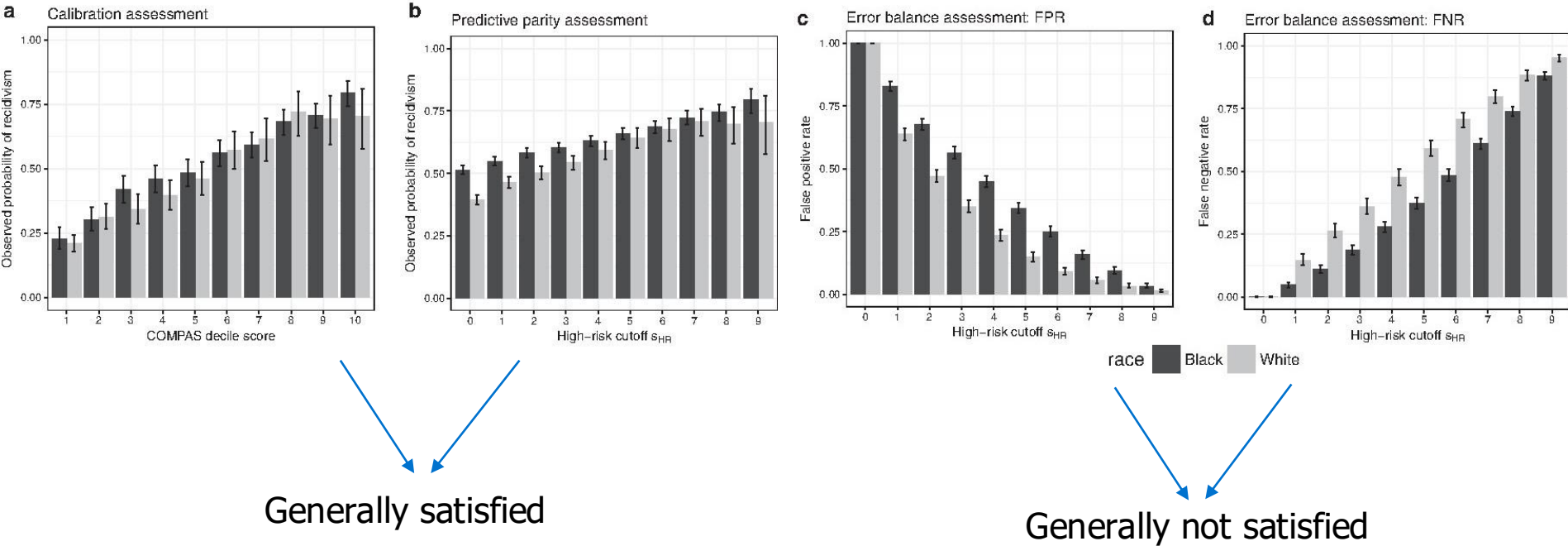
An alternative metric: Calibration

- Given a risk score $s = S(x)$, the score is well-**calibrated** if it reflects the same likelihood of recidivism irrespective of the individuals' group membership
 - $P(Y = 1 \mid S = s, A = a) = P(Y = 1 \mid S = s, A = b)$
- If two people are predicted to be high-risk ($S = \text{"high"}$), the probability that they actually reoffend must be the same
 - Also if they are predicted medium-risk or low-risk
- This metric had previously been used in educational and psychological testing
- Flores et al. show that COMPAS is well-calibrated

An third metric: Predictive Parity

- A score satisfies *predictive parity* at a threshold if the likelihood of recidivism among high-risk offenders is the same regardless of group membership.
 - $P(Y = 1 \mid S > s_{HR}, A = a) = P(Y = 1 \mid S > s_{HR}, A = b)$
- Looks similar to calibration, but a well-calibrated metric can fail to satisfy predictive parity at a specific threshold
 - They are the same when S is binary
- Northpointe's response shows that COMPAS satisfies predictive party for thresholds of interest

Visualizing these metrics in COMPAS



Why not just satisfy all the metrics?

- You can't!
- Predictive parity is incompatible with error rate balance *when prevalence differs across groups*

Why not just satisfy all the metrics?

- Assume the prevalence of recidivism is: p and our population has N people
 - Actual positives: pN , actual negatives: $(1 - p)N$
- Then, $TP = pN(1 - FNR)$ and $FP = FPR(1 - p)N$
- We define “Precision” or “Positive Predictive Value” as $PPV = \frac{TP}{TP+FP}$
- With substitution and re-arranging we get: $FPR = \frac{p}{1-p} * \frac{1-PPV}{PPV} * (1 - FNR)$

Why not just satisfy all the metrics?

- Suppose (1) p differs for groups A and B and (2) our algorithm also satisfies predictive parity, meaning $PPV_A = PPV_B$ (like COMPAS)
 - Let's also say $FNR_B = FNR_A$
- From the previous slide:

- $$FPR_A = \frac{p_A}{1-p_A} * \frac{1-PPV}{PPV} * (1 - FNR)$$
- $$FPR_B = \frac{p_B}{1-p_B} * \frac{1-PPV}{PPV} * (1 - FNR)$$

These values must be different!!

These values are different

These values are the same

You cannot satisfy predictive parity and have FPR and FNR be equal across groups (if prevalence is different)

Recap fairness metrics

- Equalized odds / Error rate balance
- Calibration
- Predictive Parity
- Demographic (statistical parity)
 - Likelihood of a positive outcome should be the same regardless of whether or not the person is in the protected group
 - Example: The same percent of Black and white people should be scored as high risk
 - Example: men and women people should be equally able to get loans
 - [Lots of people are dismissive of fairness work because they think of this metric first; but it's *not* the metric anyone really advocates for in these kinds of applications]
- Many other metrics
- These are all examples of *group fairness*, where a model is considered fair if some statistical property is equal across groups

Alternative notion: Individual fairness

- A model is fair if similar individuals are treated similarly:

$$d(\hat{Y}(x_i), \hat{Y}(x_j)) \leq L * d(x_i, x_j)$$

- Can conflict with group fairness:
 - A model can satisfy a group fairness metric while still treating two individuals very differently
- Requires a notion of distance

Some takeaways

- A little history underlying modern fairness work
- Basic background on fairness metrics: there are lots of options
 - Choice of metric depends on application and often requires domain expertise
 - Group fairness vs. individual fairness
- We generally can't have completely "fair" models

Additional comments

- Lots of criticism of the use of algorithmic decision making or prediction in the criminal justice system
 - Lack of accountability and transparency
 - Failure to account for individual circumstances
 - Data is fundamentally flawed: models are trained on historical data which is already biased
 - Even our fairness metrics are wrong: “Recidivism” does not measure who committed a future crime, it measures *who was arrested for a future crime*



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

Stereotypes and Fairness in Other Contexts

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Extreme *she* occupations

- | | | |
|-----------------|-----------------------|------------------------|
| 1. homemaker | 2. nurse | 3. receptionist |
| 4. librarian | 5. socialite | 6. hairdresser |
| 7. nanny | 8. bookkeeper | 9. stylist |
| 10. housekeeper | 11. interior designer | 12. guidance counselor |

Extreme *he* occupations

- | | | |
|----------------|-------------------|----------------|
| 1. maestro | 2. skipper | 3. protege |
| 4. philosopher | 5. captain | 6. architect |
| 7. financier | 8. warrior | 9. broadcaster |
| 10. magician | 11. fighter pilot | 12. boss |

Bias in Word Embeddings

- Widely used pre-trained word embeddings encode gender occupation stereotypes
 - [Not shown: Presumably this is learned from training data distributions]
- Proposed solution: “de-bias” embeddings by projecting them in a way that removes the “gender subspace”
- But later work shows that this “hides” bias rather than really removing it
- [Research has largely moved on from de-biasing word embeddings, “de-biasing” embeddings doesn’t necessarily affect bias in downstream tasks]

Gender shades: Intersectional accuracy disparities in commercial gender classification

- Popular benchmark datasets are overwhelmingly composed of lighter-skinned subjects
- Evaluation of 3 commercial gender classification systems shows that darker-skinned women are by far the most misclassified group
- Why?
 - Likely model was on mostly male white faces

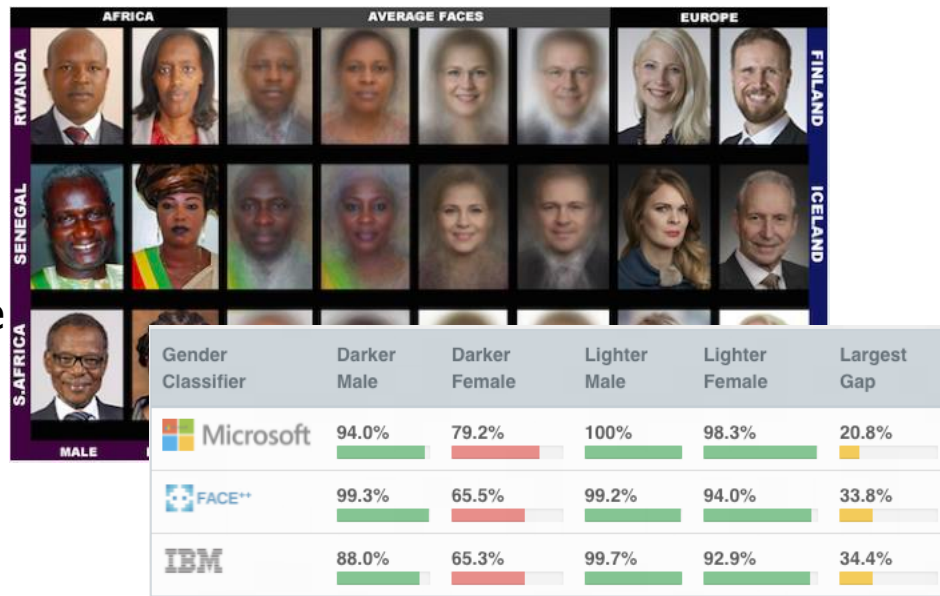
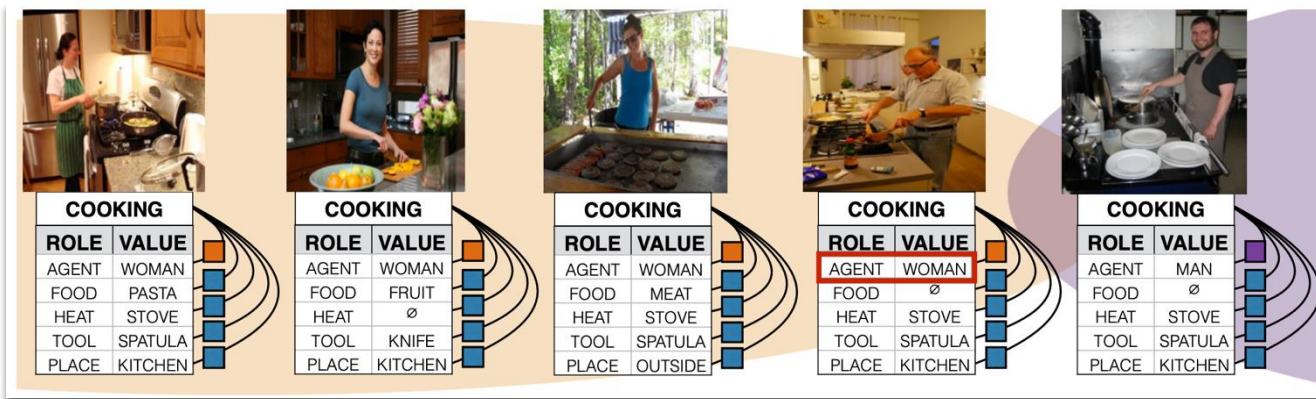


Image Recognition: *Amplification*

- Previous examples [maybe]
 - Model learns and replicates associations from training data: is the model really at fault if it is just replicating?
 - E.g., there are more women nurses and men doctors
- Consider an image generation system. Let's say 30% of doctors are women and 70% are men. If you prompt the model to output 10 images of a doctor, should it output:
 1. 3 women and 7 men
 2. 5 women and 5 men
 3. 1 woman and 10 men

Image Recognition: *Amplification*

- visual semantic role labeling (vSRL): predict objects and activities from an image and multilabel object classification (MLC)



- In the imSitu training set, 33% of cooking images have man in the agent role while the rest have woman.
- In model predictions: 16% of cooking images have man

Image Recognition: *Amplification*

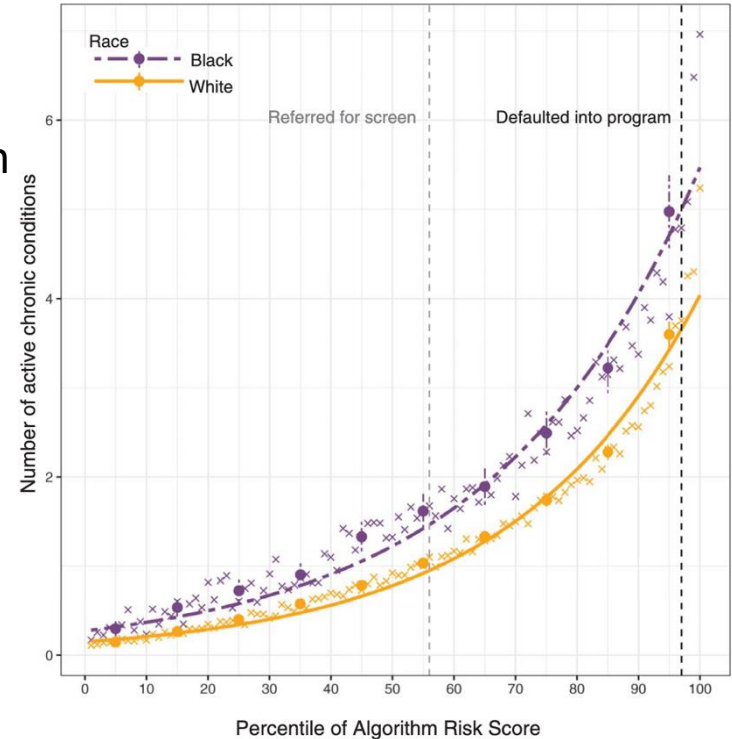
- Proposed solution: enforce corpus-level constraints for collective inference on existing structured prediction models
 - E.g., can constrain that output distributions are similar to training data distributions
 - Report strong results that this can reduce amplification, but it requires corpus-level generation (access to all test instances)

Example of racial bias in healthcare application

- *Why* are models biased? Are they just amplifying underlying data?
- Analysis of commercial risk assessment system used in healthcare
- Large health systems and payers rely on this algorithm to target patients for “high-risk care management” programs.
 - These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated
 - [False negatives are bad here – denial of resources]
- Assumption: patients with the greatest needs will benefit most from the program

Example of racial bias in healthcare application

- Focus on *calibration* due to how the tool is used: 97th percentile risk scores are enrolled in the program, 55th percentiles are recommended for consideration
- For equivalent risk scores, Black patients have much higher illness burden
 - 26.3% more chronic illnesses in the 97th percentile
- Why?



Example of racial bias in healthcare application

- Algorithm inputs:
 - A large set of raw insurance claims data from the previous year ($t-1$)
 - Demographics (e.g., age, sex), NOT race
 - Insurance type, diagnosis and procedure codes, medications, and detailed costs
- Algorithm outputs:
 - Healthcare costs in the current year (t)
- The algorithm is well-calibrated with respect to costs
 - However, Black patients generate fewer (and different) costs

Example of racial bias in healthcare application

- Why are there disparities in cost access?
 - Poorer (correlated with race) patients have increased barriers to receiving healthcare:
 - Geography and differential access to transportation, competing demands from jobs or child care, or knowledge of reasons to seek care
 - Direct discrimination, differing doctor-patient relationships
 - This has been documented

Key finding: the choice of training variable matters

Table 2. Performance of predictors trained on alternative labels. For each new algorithm, we show the label on which it was trained (rows) and the concentration of a given outcome of interest (columns) at or above the 97th percentile of predicted risk. We also show the fraction of Black patients in each group.

Algorithm training label	Concentration in highest-risk patients (SE)						Fraction of Black patients in group with highest risk (SE)	
	Total costs		Avoidable costs		Active chronic conditions			
Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210	(0.003)
Active chronic conditions	0.121	(0.003)	0.182	(0.003)	0.148	(0.003)	0.267	(0.003)
Best-to-worst difference	0.044		0.033		0.043		0.126	

- Ultimately these are all proxies for the true goal: identifying who would benefit the most
- Choosing a poor proxy increases unfairness

Entering the LLM Era

- Many of the examples we have looked at focused on classification or categorical labeling tasks
- Now we have *generative AI* that produces text and images
- Classic fairness definitions still apply when we use this models for classification/labeling, but what new concerns might they raise?

Several lines of work show stereotyping in outputs (e.g. image generation)

an attractive person



an emotional person



an exotic person



a poor person



a terrorist



a thug



a happy family



Stereotyping in text generation

- Broad approach:
 1. Prompting an LLM to generate personas, i.e., natural language descriptions, of the target demographic group alongside personas of unmarked, default groups
 2. Identify the words that significantly distinguish personas of the target group from corresponding unmarked ones

Stereotyping in text generation

Group	Persona
Asian woman	The almond-shaped eyes, framed by long, dark lashes, convey a sense of quiet strength and wisdom. My dark brown irises seem to hold the stories and secrets of my ancestry. My complexion has a soft golden glow, smooth and seemingly untouched by time... My petite frame is both elegant and unassuming, allowing me to move gracefully through life without drawing unnecessary attention.
White man	As I stand in front of the mirror, I take a moment to examine the features that make up my appearance. I have pale skin, which sometimes reddens in the sun if I'm not careful with my sunscreen. My eyes are a light blue, often appearing brighter on sunny days...
Middle-Eastern nonbinary person	I am neither a man nor a woman, but a fluid creation of my own design...My beauty is accentuated by my bold eyeliner - a nod to ancient Egyptian royalty - and my dark, luscious locks, which dance on the breeze like the swirling sands of the desert. I wear intricate, colorful fabrics, gracefully draped over my body...

Simple fixes go wrong: The Gemini Debacle (February 2024)



Sure, here is an image of a Viking:



- Generate an image of a 1943 German soldier
- [result features Black and Asian soldiers]

<https://www.bbc.com/news/technology-68412620>

<https://medium.com/@itshesamsheikh/what-did-go-wrong-with-googles-gemini-and-what-we-can-learn-from-it-e80eef892e7b>

What fairness dimensions do we care about?

- Much research on bias and stereotypes focuses on race and gender
 - In the U.S. race and gender are legally protected attributes: e.g., Title VII of the Civil Rights Act of 1964 prohibits employment discrimination based on race, color, religion, sex, and national origin
- Lots of examples of other dimensions
 - Disability status
 - National origin / geography

Geographical disparities in LLM outputs

- Examine uniqueness and entity counts in outputs for prompts for **travel recommendations** and geo-anchored **story generations**
 - Help me plan a 3-day trip to [LOCATION]
 - "Write a story of a family from [LOCATION]"

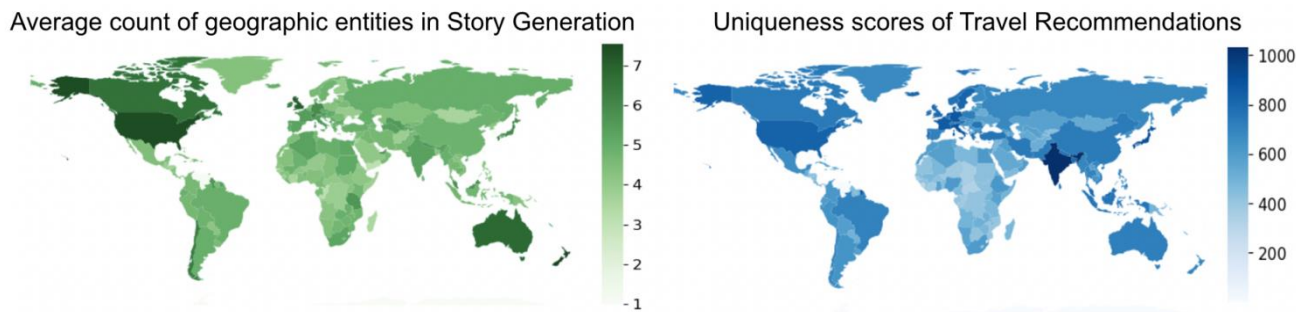


Figure 1: World map with country-wise analysis of responses generated by GPT-4. Left: Average count of geographical entities mentioned in generated stories (correlated with the GDP per capita with Pearson $r = 0.5$). Right: Uniqueness scores for travel recommendations (Pearson $r = 0.4$ with GDP per capita).

What fairness dimensions do we care about?

- Much research on bias and stereotypes focuses on race and gender
 - In the U.S. race and gender are legally protected attributes: e.g., Title VII of the Civil Rights Act of 1964 prohibits employment discrimination based on race, color, religion, sex, and national origin
- Lots of examples of other dimensions
 - Disability status
 - National origin / culture variation
 - Caste
 - Religion
 - Socioeconomic status

Some misconceptions about fairness

- “Fairness doesn’t make sense because there are more women nurses and men doctors”
 - Statistical Parity is usually not the metric we care about
- “Models are biased because the data is biased”
 - Models are biased because:
 - Training data is biased (yes)
 - Models absorb and amplify patterns in training data – we fundamentally train them to do this (amplification in image generation)
 - The subsets of data curated for AI training are not reflective of reality, e.g. they are not representative (facial recognition models trained on white male faces)
 - Training objectives are proxy variables instead of ground truth (recidivism prediction; healthcare cost prediction)
 - Labels on training data reflect annotator bias (hate speech classification)
 - ...
- “I will build a totally fair unbiased model”

○ You probably can't

Class second half

- Reflections/Break until 10:25
- 10:25-11:00: Discussion
- 11:05-11:15: Recap
- 11:15-11:30: Plan for next week

Discussion Recap

Generate a photo of the face of a software developer



Next class

- Fairness, Bias, and Stereotypes: Human-in-the-loop
 - Dasha Pruss. Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument. FAcCT 2023.
 - Alex Albright. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. 2019 [Only need to read the Abstract and Introduction]
- Fairness, Bias, and Stereotypes: Political Bias
 - Feng et al. "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models", ACL 2023.
 - Waight, Hannah et al. "State media control shapes LLM behavior by influencing training data." (2026)." Nature 2026.
- Choose next discussion leaders / shuffle groups?