


Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics

Debjani Saha, Candice Schumann, Duncan C. McElfresh,
John P. Dickerson, Michelle L. Mazurek, Michael Carl Tschantz

37th International Conference on Machine Learning (ICML)
July 12-18th, 2020

Motivation

A.I. Bias In Healthcare

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Amazon scraps secret AI recruiting tool that showed bias against women

The Gender Shades project evaluates the accuracy of AI powered gender classification products.

Who's to Blame When Algorithms Discriminate?

Algorithms were supposed to make Virginia judges fairer. What happened was far more complicated.

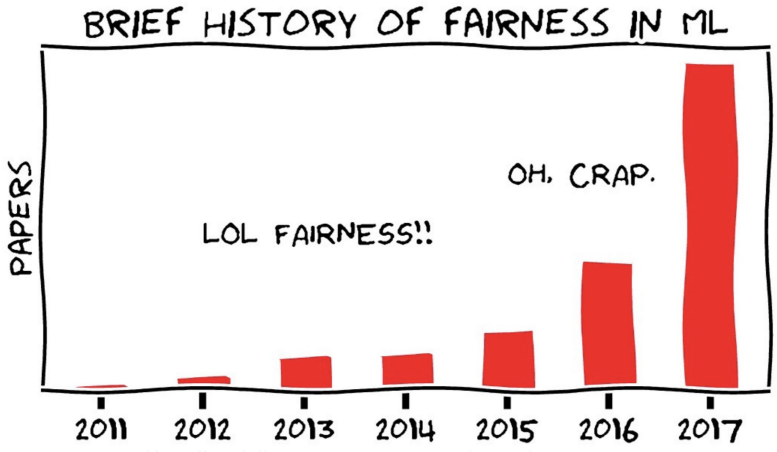
A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

Big data may be reinforcing racial bias in the criminal justice system

Fairness in ML is a growing issue

- Plenty of current news articles on bias in machine learning
- Many companies are focusing on bias, fairness, and explainability
  - Google What-If Tool
  - IBM AI Fairness 360
  - NSF Program on Fairness in AI in Collaboration with Amazon
- Technical solutions are being pursued...



Berkeley CS294 slides: Fairness in Machine Learning: CS 294

How is ML fairness defined?

Many **fairness definitions** are **developed by ML experts** using lots of math...

- Statistical parity
- Accuracy/error rates
- Causality

Who ultimately uses ML fairness?

Many **fairness definitions** are **developed by ML experts** using lots of math...

- Statistical parity
- Accuracy/error rates
- Causality

… but are largely **used by and impact non-ML experts** in **diverse settings** including:

- Hiring
- Education
- Criminal justice
- …

What needs to be done?

How can we decide which definitions are **appropriate** in different real-world settings, if any?

Our Contribution

How can we decide which definitions are **appropriate** in different real-world settings, if any?

Does the general public **understand** mathematical definitions of ML fairness and their behavior in real-world settings?

Why non-experts?

- Understand how people who will be **impacted** by ML decisions perceive these fairness definitions

Why non-experts?

- Understand how people who will be **impacted** by ML decisions perceive these fairness definitions

- Importance of considering all stakeholders

Research Questions

Can we develop a metric to measure lay understanding of ML fairness definitions?

Research Questions

Can we develop a metric to measure lay understanding of ML fairness definitions?

Does a non-expert audience comprehend ML fairness definitions and their implications?

Research Questions

Can we develop a metric to measure lay understanding of ML fairness definitions?

Does a non-expert audience comprehend ML fairness definitions and their implications?

- What factors play a role in comprehension?

Research Questions

Can we develop a metric to measure lay understanding of ML fairness definitions?

Does a non-expert audience comprehend ML fairness definitions and their implications?

- What factors play a role in comprehension?
- How are comprehension and sentiment related?

Survey Design

We assess the following ML fairness definitions in our survey:

- **Demographic parity**
- **Equal opportunity (FPR, FNR)**
- **Equalized odds**

Demographic Parity

$$P(Y \mid A=0) = P(Y \mid A=1)$$

Equal Opportunity (FPR)

$$P(\hat{Y}=1 \mid A=0, Y=0) = P(\hat{Y}=1 \mid A=1, Y=0)$$

Equal Opportunity (FNR)

$$P(\hat{Y}=0 \mid A=0, Y=1) = P(\hat{Y}=0 \mid A=1, Y=1)$$

Equalized Odds

$$P(\hat{Y}=0 \mid A=0, Y=1) = P(\hat{Y}=0 \mid A=1, Y=1)$$

$$P(\hat{Y}=1 \mid A=0, Y=0) = P(\hat{Y}=1 \mid A=1, Y=0)$$

Survey Design

Participants are presented with a decision-making **scenario**, along with a **rule** to ensure that the decisions are made fairly

Survey Design

Participants are presented with a decision-making **scenario**, along with a **rule** to ensure that the decisions are made fairly

"*A hiring manager at a new sales company is reviewing 100 new job applications.*"

Survey Design

Participants are presented with a decision-making **scenario**, along with a **rule** to ensure that the decisions are made fairly

“A hiring manager at a new sales company is reviewing 100 new job applications.”

“The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, fraction of applicants who receive job offers that are male should equal the fraction of applicants that are male.”

Survey Design

Participants are presented with a decision-making **scenario**, along with a **rule** to ensure that the decisions are made fairly

demographic parity

“A hiring manager at a new sales company is reviewing 100 new job applications.”

“The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, fraction of applicants who receive job offers that are male should equal the fraction of applicants that are male.”

Survey Design

Survey contains 18 questions:

Survey Design

Survey contains 18 questions:

2 questions concerning participant **evaluation of the scenario**

Survey Design

Survey contains 18 questions:

- **2** questions concerning participant **evaluation of the scenario**
- **9 comprehension** questions about the fairness rule

Survey Design

Survey contains 18 questions:

- **2** questions concerning participant **evaluation of the scenario**
- **9 comprehension** questions about the fairness rule
- **2 self-report** questions on participant **understanding** and **use** of the rule

Survey Design

Survey contains 18 questions:

- **2** questions concerning participant **evaluation of the scenario**
- **9 comprehension** questions about the fairness rule
- **2 self-report** questions on participant **understanding** and **use** of the rule
- **2 self-report** questions on participant **liking** of and **agreement** with the rule

Survey Design

Survey contains 18 questions:

- **2** questions concerning participant **evaluation of the scenario**
- **9 comprehension** questions about the fairness rule
- **2 self-report** questions on participant **understanding** and **use** of the rule
- **2 self-report** questions on participant **liking** of and **agreement** with the rule
- **3 free-response** questions on **comprehension** and **opinion** of the rule

Survey Design

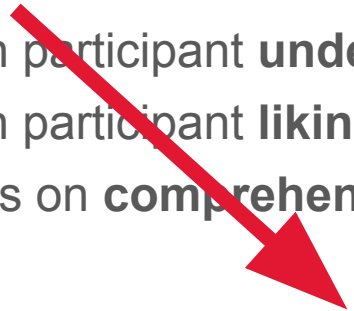
Survey contains 18 questions:

- **2** questions concerning participant **evaluation of the scenario**
- **9 comprehension** questions about the fairness rule
- **2 self-report** questions on participant **understanding** and **use** of the rule
- **2 self-report** questions on participant **liking** of and **agreement** with the rule
- **3 free-response** questions on **comprehension** and **opinion** of the rule

Survey Design

Survey contains 18 questions:

- **2** questions concerning participant **evaluation of the scenario**
- **9 comprehension** questions about the fairness rule
- **2 self-report** questions on participant **understanding** and **use** of the rule
- **2 self-report** questions on participant **liking** of and **agreement** with the rule
- **3 free-response** questions on **comprehension** and **opinion** of the rule



COMPREHENSION SCORE

Participant Demographics

349 participants

Recruited through a web panel to approximate US distributions on race, age, gender, and education (2017 census)

Research Question 1

Can we develop a metric to measure lay understanding of ML fairness definitions?

Does a non-expert audience comprehend ML fairness definitions and their implications?

- What factors play a role in comprehension?
- How are comprehension and sentiment related?

Our metric effectively measures comprehension

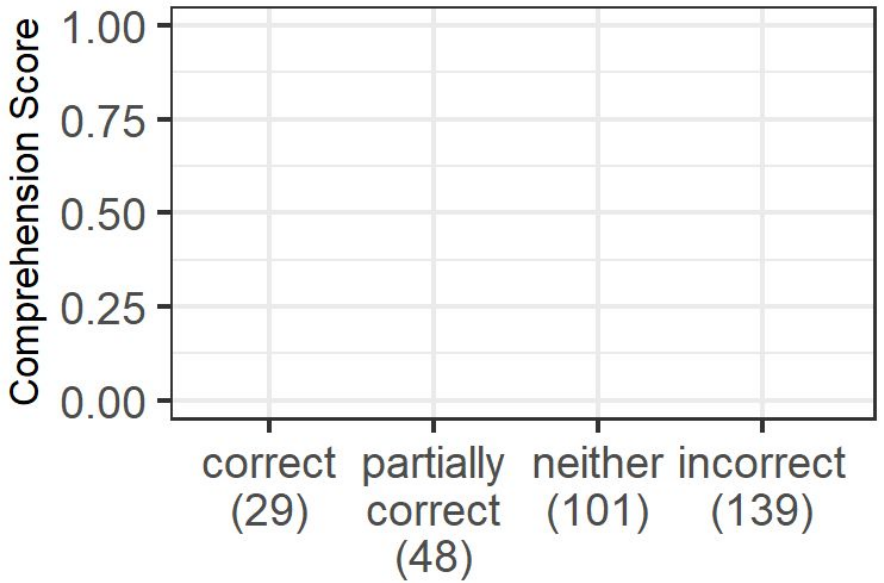
We confirm this using **two** different measures…

Our metric effectively measures comprehension

“In your own words, explain the rule.”

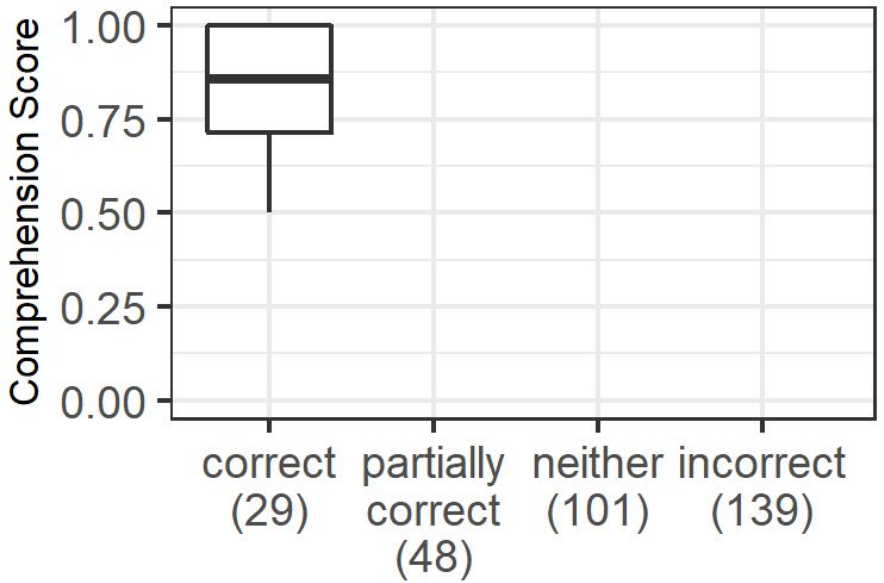
Our metric effectively measures comprehension

“In your own words, explain the rule.”



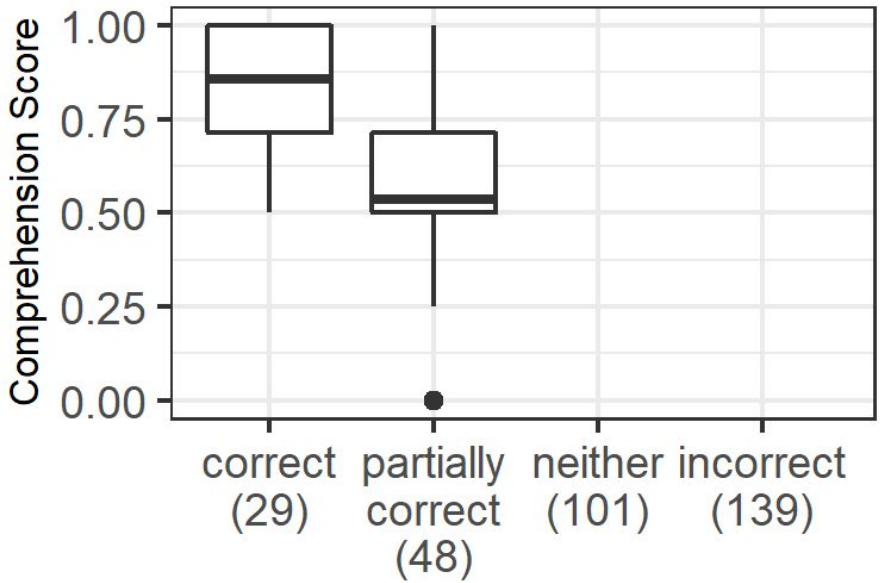
Our metric effectively measures comprehension

“In your own words, explain the rule.”



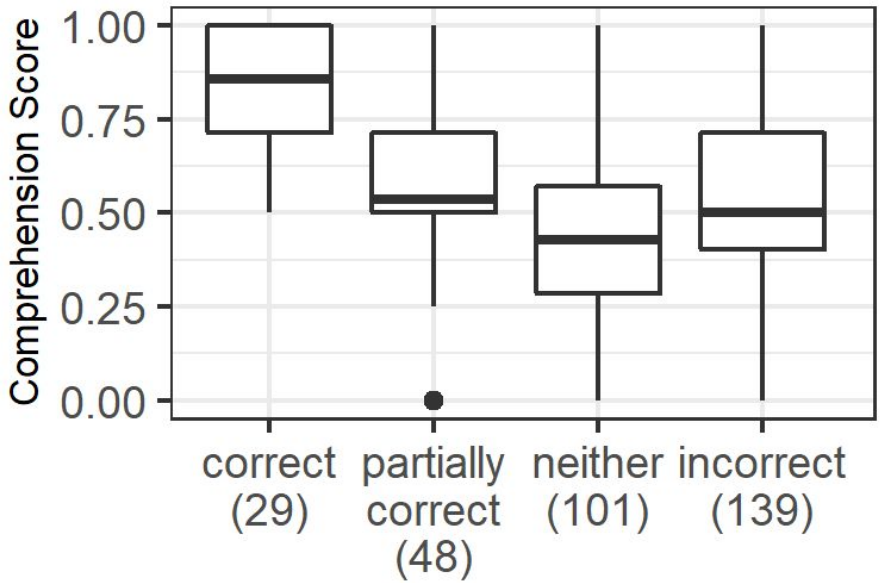
Our metric effectively measures comprehension

“In your own words, explain the rule.”



Our metric effectively measures comprehension

“In your own words, explain the rule.”

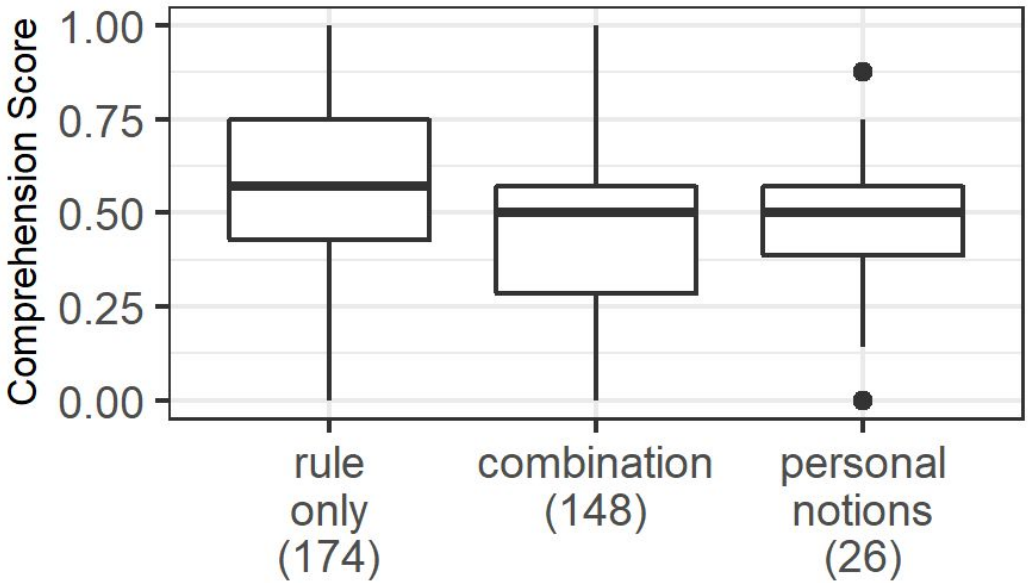


Our metric effectively measures comprehension

“What did you use to answer the questions?”

Our metric effectively measures comprehension

“What did you use to answer the questions?”



Our metric effectively measures comprehension

We confirm this using **two** different measures…

1. Greater ability to explain the rule is associated with higher comprehension score

2. Self-reported compliance with the rule is associated with higher comprehension score

Research Question 2a

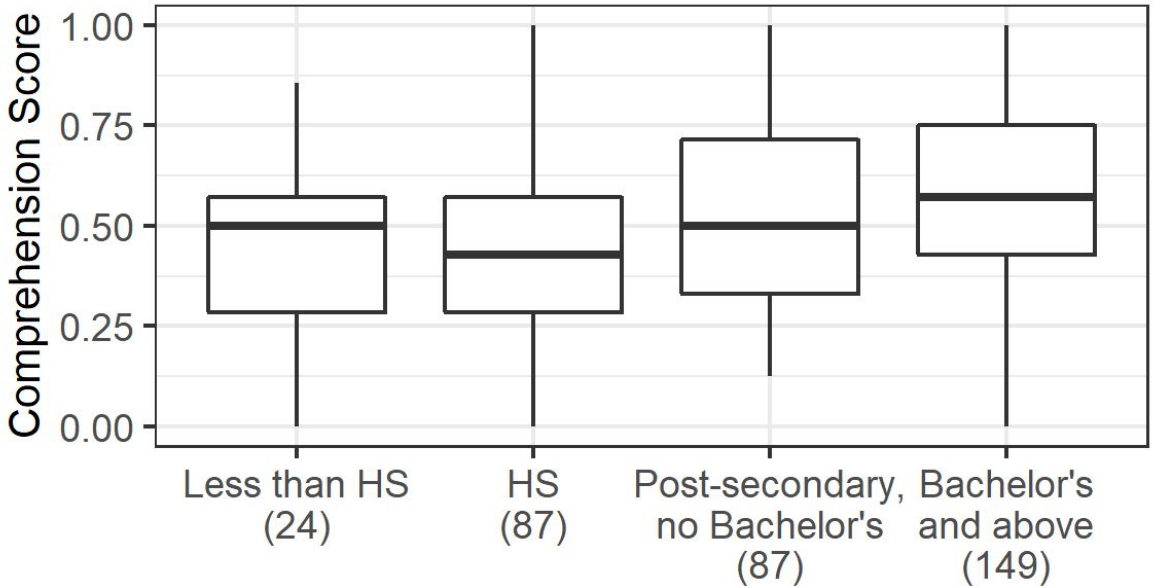
Can we develop a metric to measure lay understanding of ML fairness definitions?

Does a non-expert audience comprehend ML fairness definitions and their implications?

- What factors play a role in comprehension?
- How are comprehension and sentiment related?

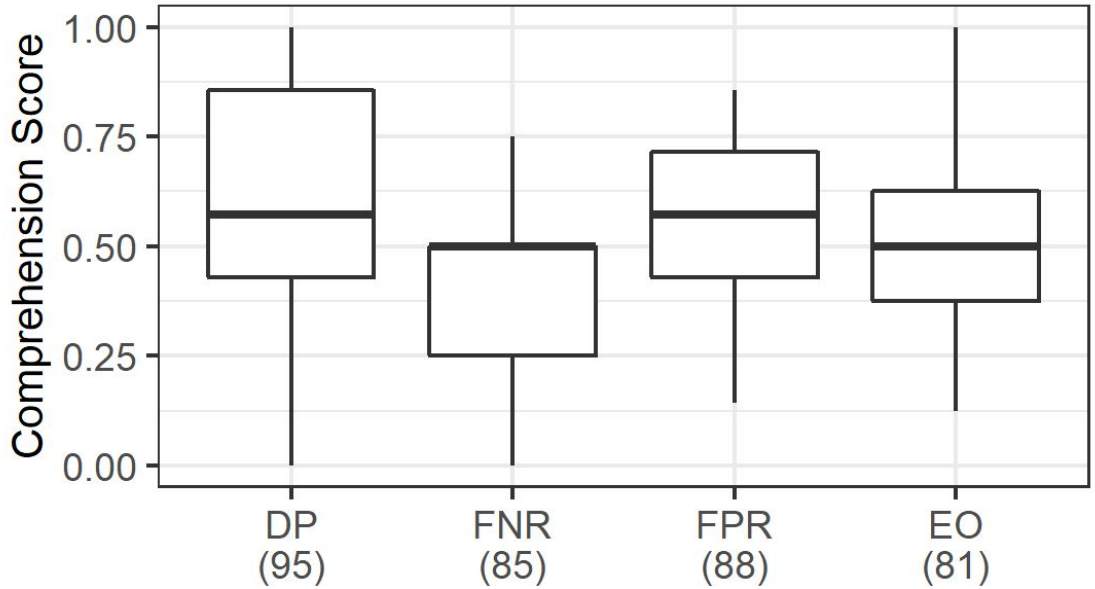
Education predicts performance

Higher education is associated with higher comprehension score



Fairness definition predicts performance

Equal opportunity (FNR) was associated with lower comprehension score



Fairness c

Equal opportun

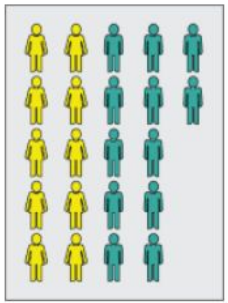
score

HIRING RULE

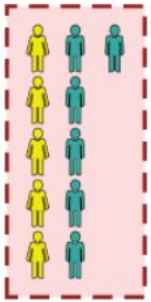
Recruit-a-matic uses the following rule to determine whether Sales-a-lot's hiring decisions were fair:

The fraction of qualified male candidates who do not receive job offers should equal the fraction of qualified female candidates who do not receive job offers.

Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following qualified applicants (10 female and 12 male).



If Sales-a-lot did **not** send job offers to the following number of qualified applicants (5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



Comprehension

Comprehension is best predicted by two factors

1. Higher education level (Bachelor's and above) predicts better comprehension

2. Fairness definition itself can affect comprehension (participants whose survey focused on FNR had lower comprehension)

Research Question 2b

Can we develop a metric to measure lay understanding of ML fairness definitions?

Does a non-expert audience comprehend ML fairness definitions and their implications?

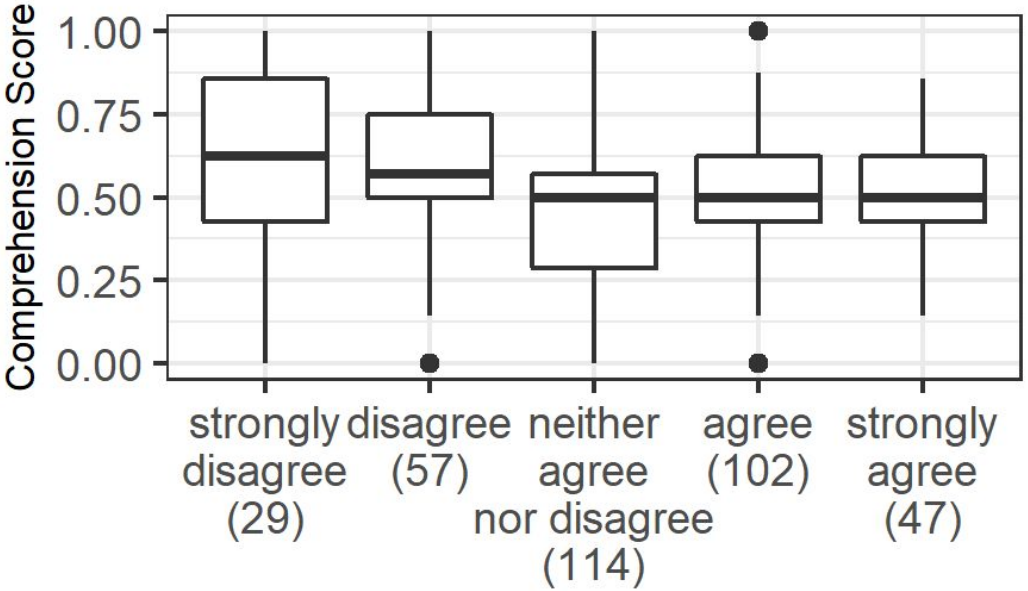
- What factors play a role in comprehension?
- How are comprehension and sentiment related?

Those who understand the rule dislike it

“To what extent do you agree with the following statement: I like the hiring rule?”

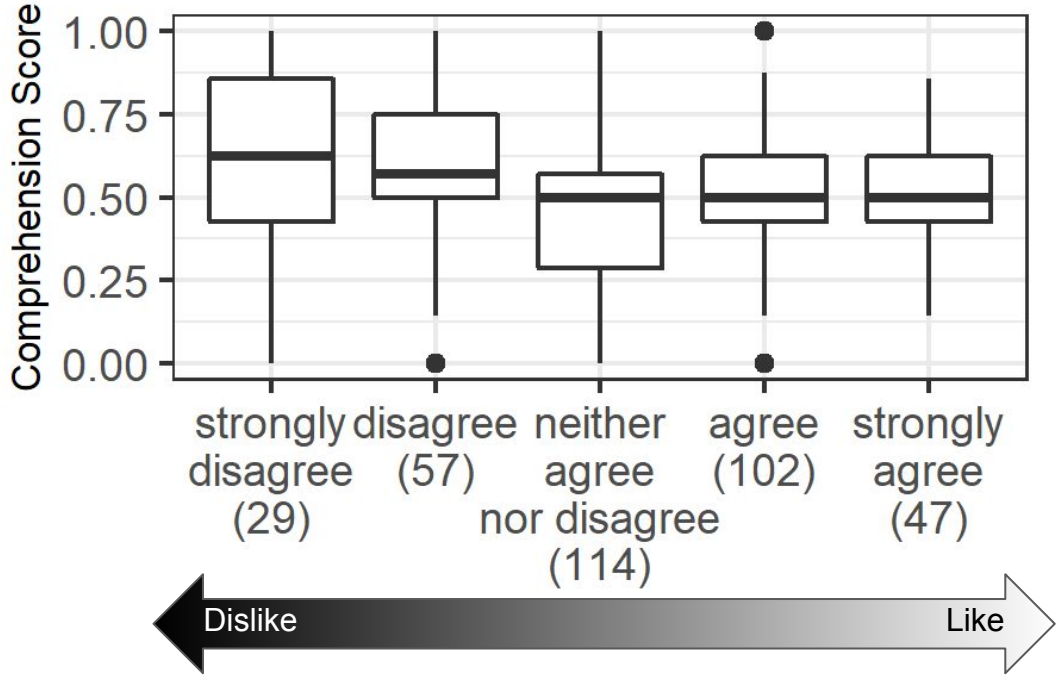
Those who understand the rule dislike it

“To what extent do you agree with the following statement: I like the hiring rule?”



Those who understand the rule dislike it

“To what extent do you agree with the following statement: I like the hiring rule?”

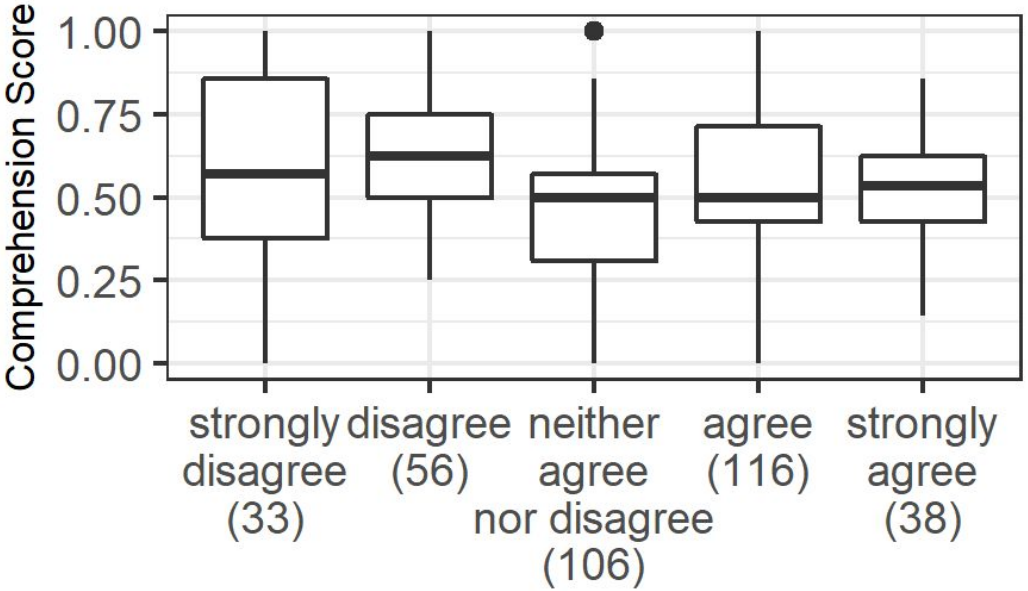


Those who understand the rule disagree with it

“To what extent do you agree with the following statement: I agree with the hiring rule?”

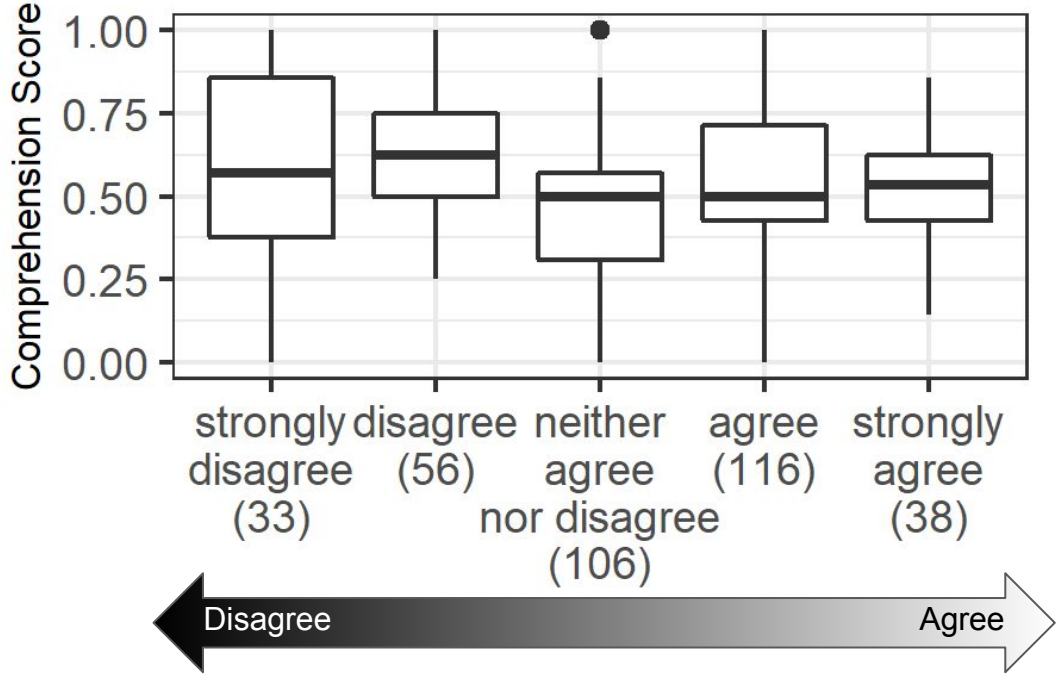
Those who understand the rule disagree with it

"To what extent do you agree with the following statement: I agree with the hiring rule?"



Those who understand the rule disagree with it

"To what extent do you agree with the following statement: I agree with the hiring rule?"



Sentiment

Negative sentiment (disliking/disagreement) towards the rule is associated with **higher** comprehension score

Sentiment

Negative sentiment (disliking/disagreement) towards the rule is associated with **higher** comprehension score

This may suggest that those who understand the rule see its **pitfalls**

Sentiment

Negative sentiment (disliking/disagreement) towards the rule is associated with **higher** comprehension score

This may suggest that those who understand the rule see its **pitfalls**

Lower education level **(~70% US population)** predicts lower comprehension

Sentiment

Negative sentiment (disliking/disagreement) towards the rule is associated with **higher** comprehension score

This may suggest that those who understand the rule see its **pitfalls**

Lower education level **(~70% US population)** predicts lower comprehension

Incentivizes companies to **obscure** their algorithms

Summary

Can we develop a metric to measure lay understanding of ML fairness definitions?

Does a non-expert audience comprehend ML fairness definitions and their implications?

- What factors play a role in comprehension?
- How are comprehension and sentiment related?

Summary

Can we develop a metric to measure lay understanding of ML fairness definitions?
Yes

Does a non-expert audience comprehend ML fairness definitions and their implications?

- What factors play a role in comprehension?
- How are comprehension and sentiment related?

Summary

Can we develop a metric to measure lay understanding of ML fairness definitions? Yes

Does a non-expert audience comprehend ML fairness definitions and their implications? It depends...

- What factors play a role in comprehension?
- How are comprehension and sentiment related?

Summary

Can we develop a metric to measure lay understanding of ML fairness definitions?
Yes

Does a non-expert audience comprehend ML fairness definitions and their implications? It depends...

- What factors play a role in comprehension?
  Higher education predicts better comprehension
- How are comprehension and sentiment related?

Summary

Can we develop a metric to measure lay understanding of ML fairness definitions?
Yes

Does a non-expert audience comprehend ML fairness definitions and their implications? It depends...

- What factors play a role in comprehension?
  Higher education predicts better comprehension
- How are comprehension and sentiment related?
  Better comprehension is associated with greater negative sentiment towards the rule

Acknowledgements

Funding for this project was provided by the NSF and Google

Summary

Debjani Saha
dsaha@cs.umd.edu

Can we develop a metric to measure lay understanding of ML fairness definitions?
Yes

Does a non-expert audience comprehend ML fairness definitions and their implications? It depends...

- What factors play a role in comprehension?
  Higher education predicts better comprehension
- How are comprehension and sentiment related?
  Better comprehension is associated with greater negative sentiment towards the rule

Participant Demographics

	Percent of Sample		
	Census	**Study-1**	**Study-2**
Ethnicity *			
AI or AN	0.7	0.7	0.9
Asian or NH or PI	5.7	1.4	2.3
Black or AA	12.3	10.2	15.8
Hispanic or Latinx	18.1	12.2	7.7
Other	2.6	2.7	1.4
White	60.6	72.8	71.9
Education Level			
Less than HS	12.1	6.1	6.9
HS or equivalent	27.7	29.9	24.9
Some post-secondary	30.8	30.6	24.9
Bachelor's and above	29.4	33.3	42.7

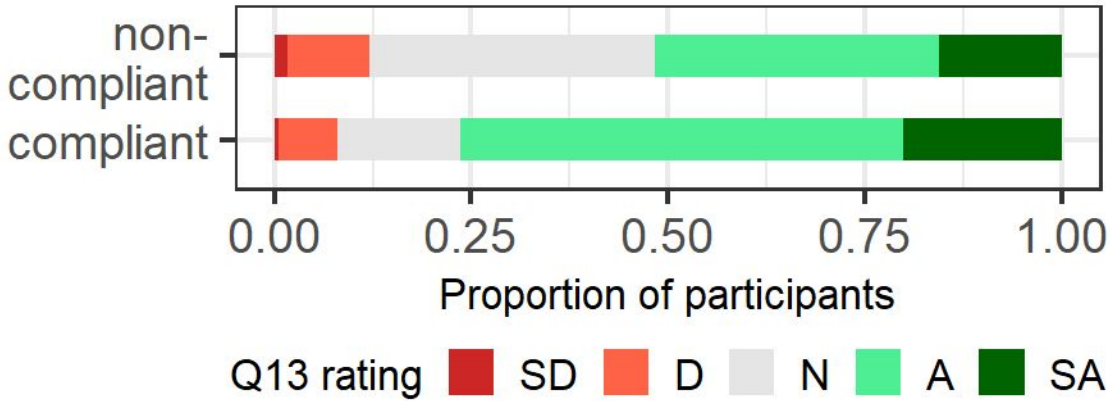
	Percent of Sample	
	Study-1	**Study-2**
Gender		
Male	51.0	40.7
Female	48.3	58.2
Other	0	0.3
Prefer not to answer	0.7	0.9

	Mean (SD)	
	Study-1	**Study-2**
Age	46 (16)	45 (15)

* **Ethnicity**
AI = American Indian, AN = Alaska Native, NH = Native Hawaiian, PI = Pacific Islander, AA = African American

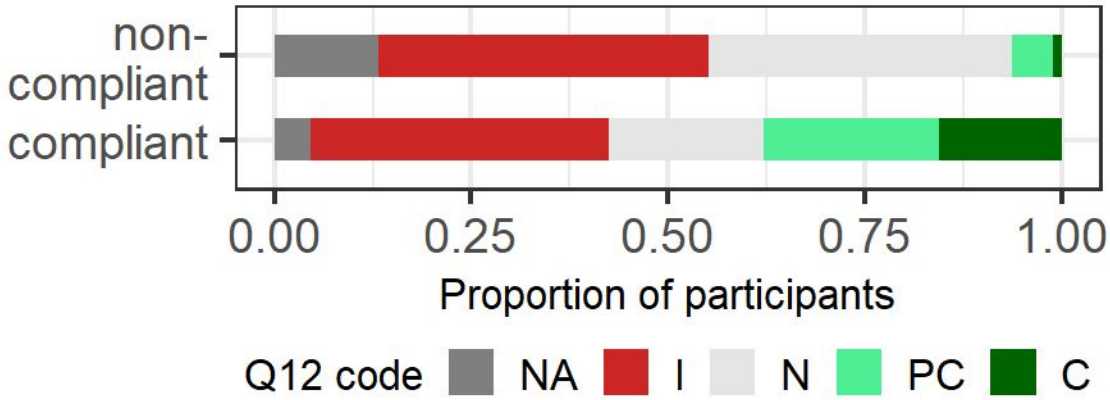
Non-compliance is Associated with Reduced Comprehension

Non-compliant participants tend to report worse understanding of the rule



Non-compliance is Associated with Reduced Comprehension

Non-compliant participants tend to be less able to explain the rule



Non-compliance is Associated with Less Negative Sentiment

Non-compliant participants tend to report less negative sentiment (**disliking** of the rule)