



# Towards Participatory Machine Learning: Selecting Jurors for Content Moderation AIs

Joyce Chen, Mitchell L Gordon, James Landay, Tatsunori B Hashimoto, Michael S Bernstein



## Motivation

State-of-the-art machine learning classifiers evaluate content by **implicit “majority vote.”**

Decisions around whether content is toxic therefore often **override the voices of minorities**, whom are frequently the **targets of such toxicity.**

Jury Learning allows practitioners, such as **content moderators**, to explicitly choose the **voices** they want their **classifier to listen to.**

## Research Questions

We want to study how real community-centered content moderators use Jury Learning:

RQ0 Does Jury Learning enable an **effective deliberative process** about whose voices a ML classifier should emulate?

RQ1 What **negotiations and trade-offs** do moderators make when constructing a jury moderator? How do those **deliberative discussions differ** from those around thresholding Perspective API?

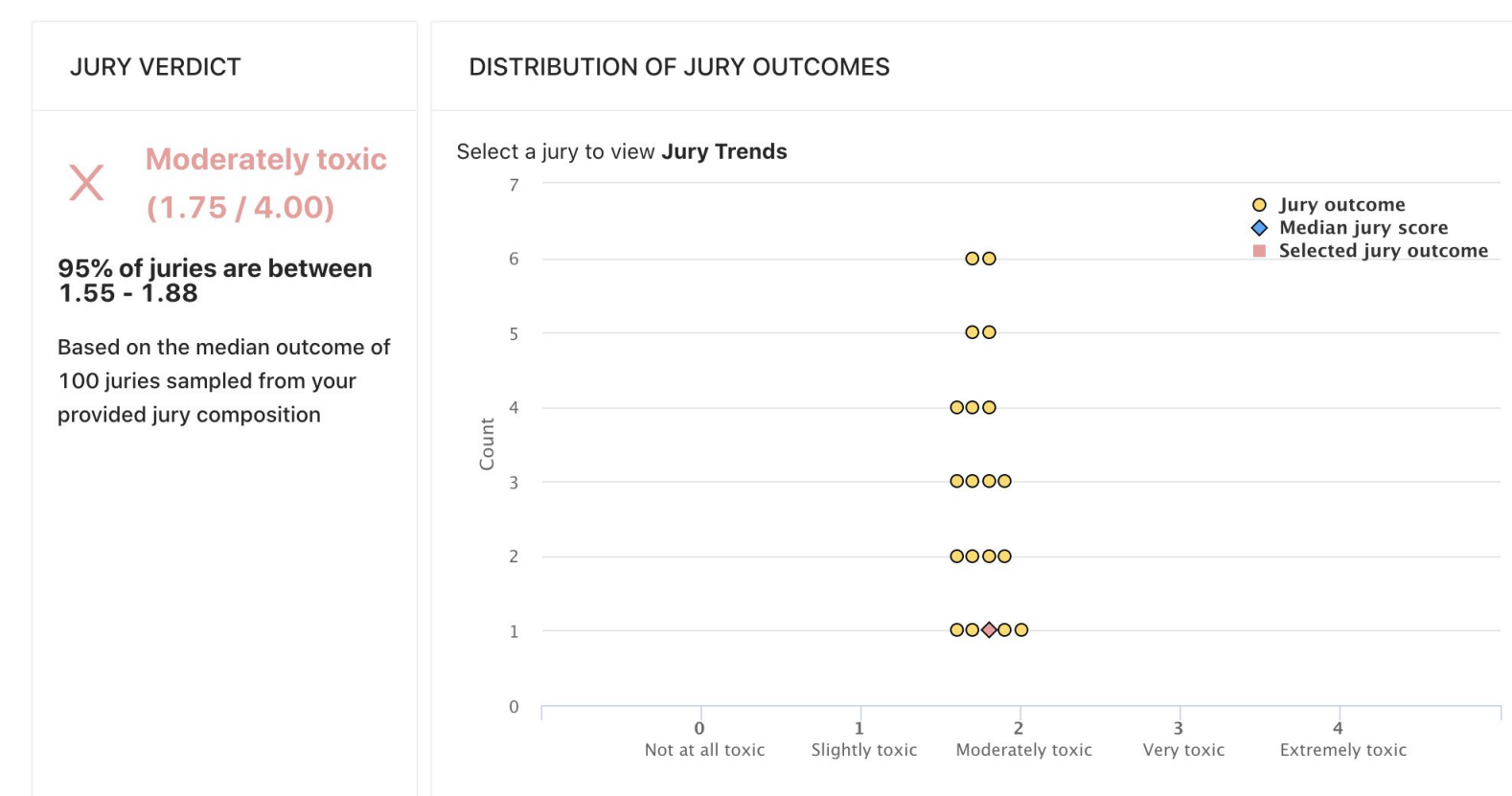
RQ2 How does the **composition of deliberatively-constructed juries** compare to individually-constructed juries? Is the deliberative jury perceived as **more legitimate and trustworthy** in making content moderation decisions than Perspective API?

## Study

Reddit moderators choose settings of AI content moderator **individually**  
Evaluate AI via 7-point Likert survey: **satisfaction** with AI’s decisions, **trust** in implementing the AI in their subreddit etc.  
Choose same settings via **group deliberation** (team of 3)  
Evaluate AI again through survey and live interview

## Jury Learning and Perspective API Systems

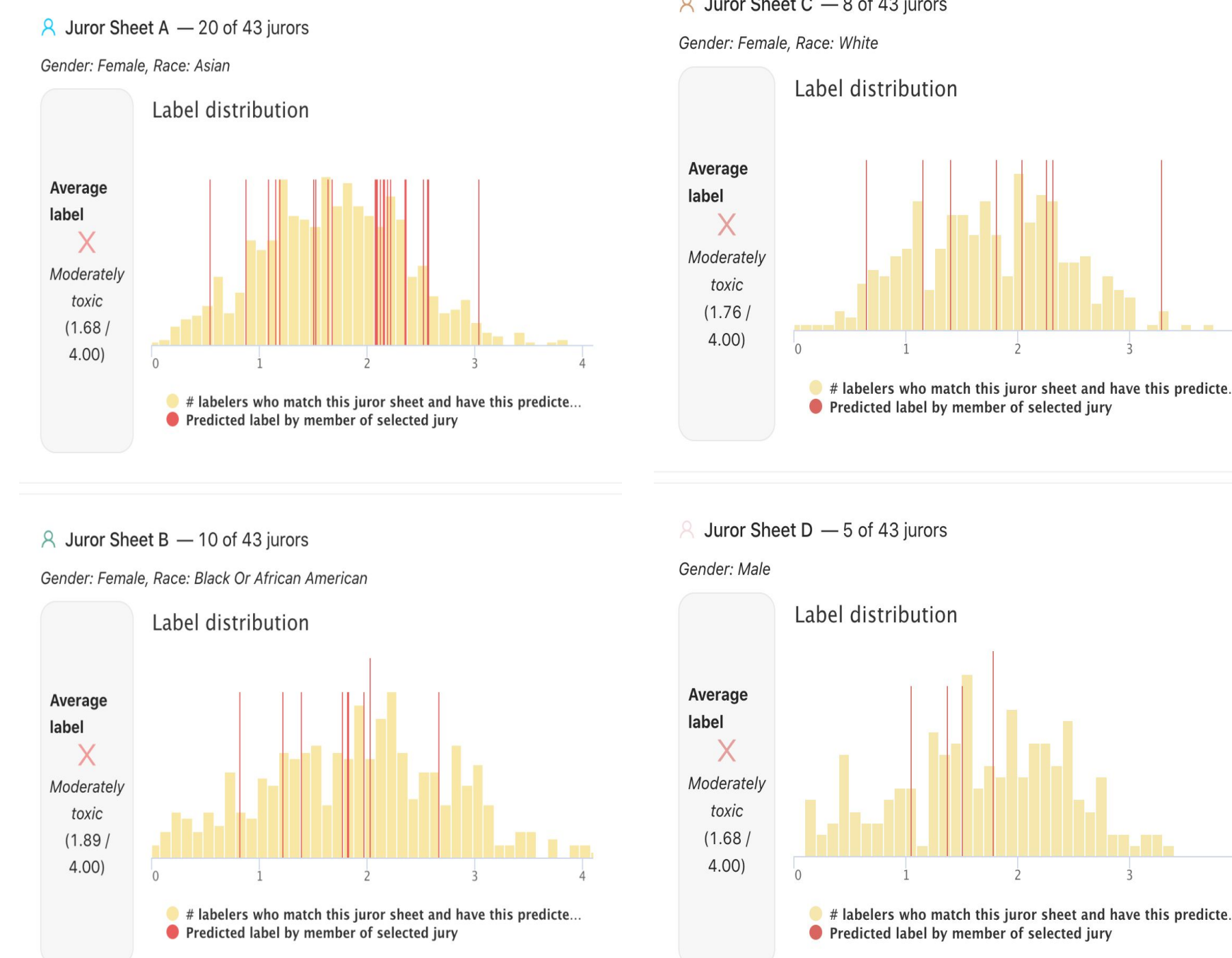
The interface shows a 'Test a New Comment' section with a text input field containing 'what an awful guest' and a 'Submit' button. Below this, 'Toxicity Results' are displayed for three comments: 'what an awful guest' (55.36% likely to be toxic), 'what an awful movie' (44.81% likely to be toxic), and 'your taste is awful' (59.59% likely to be toxic).



The 'Juror Selection' interface allows users to create four juror sheets (A, B, C, D) with specific characteristics:

- Juror Sheet A:** Race: Asian, Gender: Female, Seats: 20
- Juror Sheet B:** Race: Black Or African American, Gender: Female, Seats: 10
- Juror Sheet C:** Race: White, Gender: Female, Seats: 8
- Juror Sheet D:** Gender: Male, Seats: 5

An input example is provided: 'Is skincare a tool of the patriarchy to oppress women or do you wash your face so often because you're like super vapid'.



## Initial results

**Jury Learning provides new mental frameworks**  
“What are the *comments* on our sub?” → “Who are the *people affected?*”

“What does our subreddit *currently* look like?” → “What do we want our sub to *ideally* look like?”

N = 40 (Reddit moderators, >20 unique subreddits)  
15 sessions (10 mod teams from the same subreddit, 5 from different subreddits)

**Survey scores** (1: strongly disagree, 7: strongly agree)

	Jury Learning	Perspective
Perceived Legitimacy	$\mu = 2.589$ $\sigma = 1.066$	$\mu = 3.568$ $\sigma = 1.278$
Process	$\mu = 4.544$ $\sigma = 1.366$	$\mu = 5.316$ $\sigma = 1.209$
Representation	$\mu = 4.553$ $\sigma = 1.492$	$\mu = 4.079$ $\sigma = 1.570$

Not statistically significant  
Does not separate **technical/usability difficulties** from evaluation of the AI itself

**High-level themes** (from transcripts)

Reasoning about **AI & its errors**  
Reasoning about **users & demographics**  
**Ideal community** design & norms  
**Representation friction**

## Next steps

1. Continue iterating on our codes and synthesizing evidence for our high-level themes
2. Analyze the types of comments tested by moderators in each condition

## Acknowledgements

Many thanks to my fantastic mentor Mitchell Gordon, Michael Bernstein, James Landay, Tatsu Hashimoto, and the CURIS coordinators!