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?

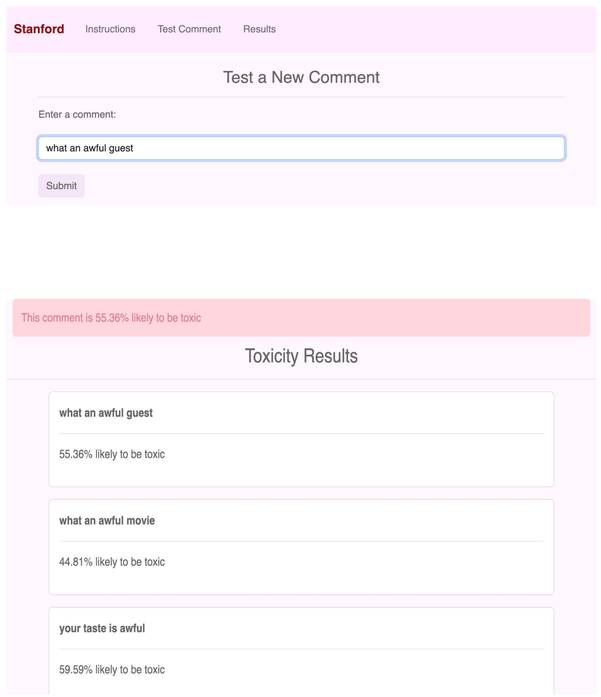
RQ1 What **negotiations and trade-offs** do moderators make when constructing a jury moderator? How do those **deliberative discussions differ** from those around thresholding Perspective API?

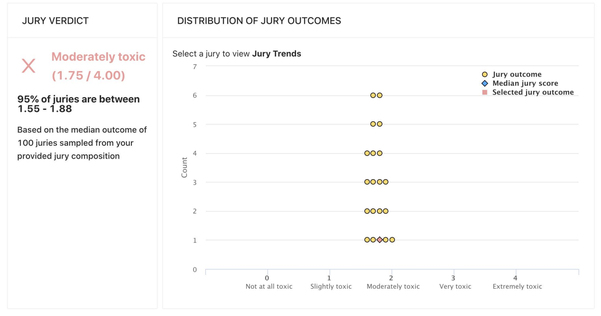
RQ2 How does the **composition of deliberatively-constructed juries** compare to individually-constructed juries? Is the deliberative jury perceived as **more legitimate and trustworthy** in making content moderation decisions than Perspective API?

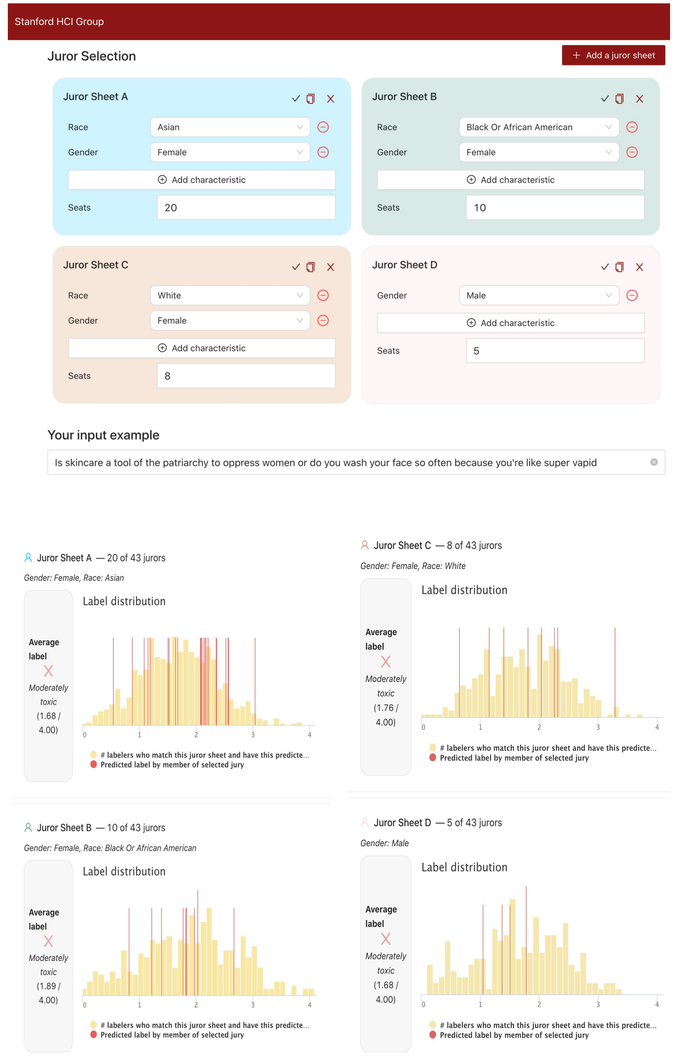
Study

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







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	μ = 2.589 σ = 1.066	μ = 3.568 σ = 1.278
Process	μ = 4.544 σ = 1.366	μ = 5.316 σ = 1.209
Representation	μ = 4.553 σ = 1.492	μ = 4.079 σ = 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!