

# Pluralistic and Collective Alignment

Mitchell Gordon  
[mlgordon@mit.edu](mailto:mlgordon@mit.edu)

NeurIPS 2025

## Implicit assumption in most alignment work:

There is a *single* set of values and preferences to which we wish to align

## Implicit assumption in most alignment work:

There is a *single* set of values and preferences to which we wish to align

In reality, people have **differing preferences**, depending on context, values, life experience, demographics, etc.

In reality, people have **differing preferences**, depending on context, values, life experience, demographics, etc.

## DISTRIBUTIONAL PREFERENCE LEARNING: UNDERSTANDING AND ACCOUNTING FOR HIDDEN CONTEXT IN RLHF

Anand Sithharanjan<sup>\*</sup> Cassidy Laidlaw<sup>\*</sup>  
University of California, Berkeley  
{anandsranjan,cassidy-laidlaw}@cs.berkeley.edu

Dylan Hadfield-Menell  
Massachusetts Institute of Technology  
dhm@csail.mit.edu

## Towards Measuring the Representation of Subjective Global Opinions in Language Models

Esin Durmus<sup>\*</sup> Karina Nguyen Thomas I. Liao Nicholas Schiefer  
Amanda Askell Anton Bakhtin Carol Chen Zac Hatfield-Dodds  
Danny Hernandez Nicholas Joseph Liane Lovitt Sam McCandlish Orowa Sikder  
Alex Tamkin Janel Thamkul  
Jared Kaplan Jack Clark Deep Ganguli

Anthropic

## VALUECOMPASS: A Framework for Measuring Contextual Value Alignment Between Human and LLMs

Hua Shen<sup>\*\*</sup> Tiffany Kneare<sup>◊</sup> Reshma Ghosh<sup>†</sup> Yu-Ju Yang<sup>◊</sup>  
Nicholas Clark<sup>\*</sup> Yun Huang<sup>◊</sup> Tanu Mitra<sup>\*</sup>  
\* NYU Shanghai, New York University, <sup>†</sup>University of Washington,  
<sup>◊</sup>MBZUAI, <sup>†</sup>Microsoft, <sup>◊</sup>UIUC

Fine-tuning language models to find agreement  
among humans with diverse preferences

Michiel A. Bakker<sup>\*</sup>  
DeepMind  
miba@deepmind.com

Martin J. Chadwick<sup>\*</sup>  
DeepMind  
martin@deepmind.com

Hannah R. Sheahan<sup>\*</sup>  
DeepMind  
hsheahan@deepmind.com

Michael Henry Tessler  
DeepMind  
tesslerm@deepmind.com

Lucy Campbell-Gillingham  
DeepMind  
lcgillingham@deepmind.com

Jan Balaguer  
DeepMind  
jua@deepmind.com

Nat McAleese  
DeepMind  
nmca@deepmind.com

Amelia Glaese  
DeepMind  
glamia@deepmind.com

John Aslanides  
DeepMind  
jaslanides@deepmind.com

# This talk: three directions

Designing interactive systems for reasoning about different goals  
and values

Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Exploring formal definitions of pluralistic alignment

A Roadmap to Pluralistic Alignment

Eliciting preferences from the public

Collective alignment: public input on OpenAI's Model Spec

# This talk: three directions

**Designing interactive systems for reasoning about different goals and values.**

Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Exploring the space of theoretical forms of pluralistic alignment

A Roadmap to Pluralistic Alignment

Eliciting preferences from the public

Collective alignment: public input on our Model Spec

# Jury learning: integrating dissenting voices in machine learning models

Mitchell Gordon, Michelle Lam, Joon Sung Park,  
Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto,  
Michael S. Bernstein.

CHI 2022  
*Best Paper Award*

Let's talk about ground truth.

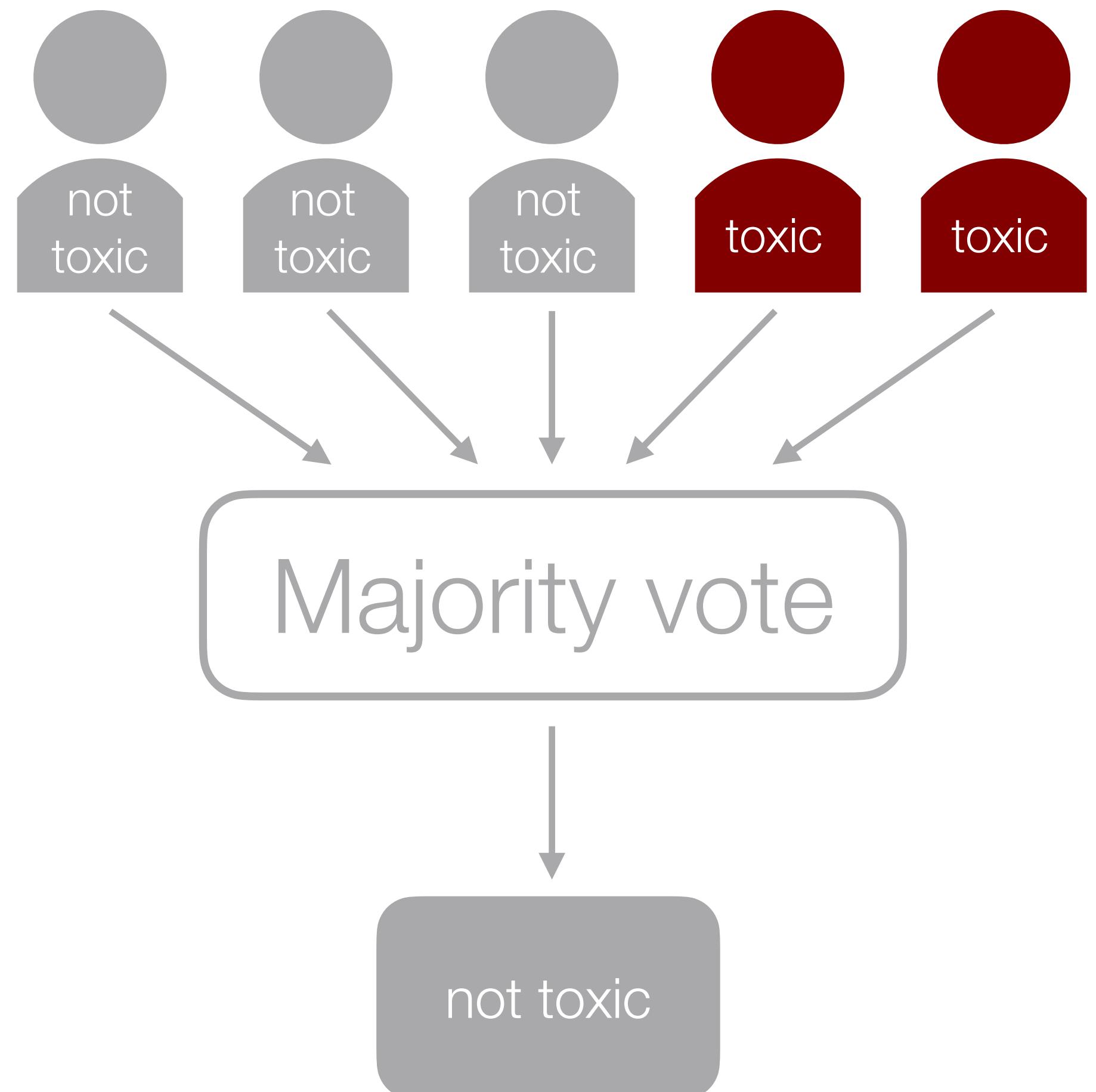
# Should this online comment be labeled ‘toxic’ by an AI?

*“1. People still eat at Pizza Hut? Gross. 2. It is shameful how this country [...]”*

# Aggregation via majority vote

Typical approach in machine learning data annotation: ask multiple labelers and **aggregate** the results to identify a **ground truth label**

[Sheng et al. 2008; Welinder et al. 2010]



# For many tasks, even experts disagree on correct labels

**Traditional ML task:  
image classification**

Is this a cat or a dog?



[Deng et al. 2009]

**Social computing task:  
toxicity detection**

Is this comment toxic?

“1. People still eat at Pizza  
Hut? Gross. 2. It is  
shameful how this  
country [...]”

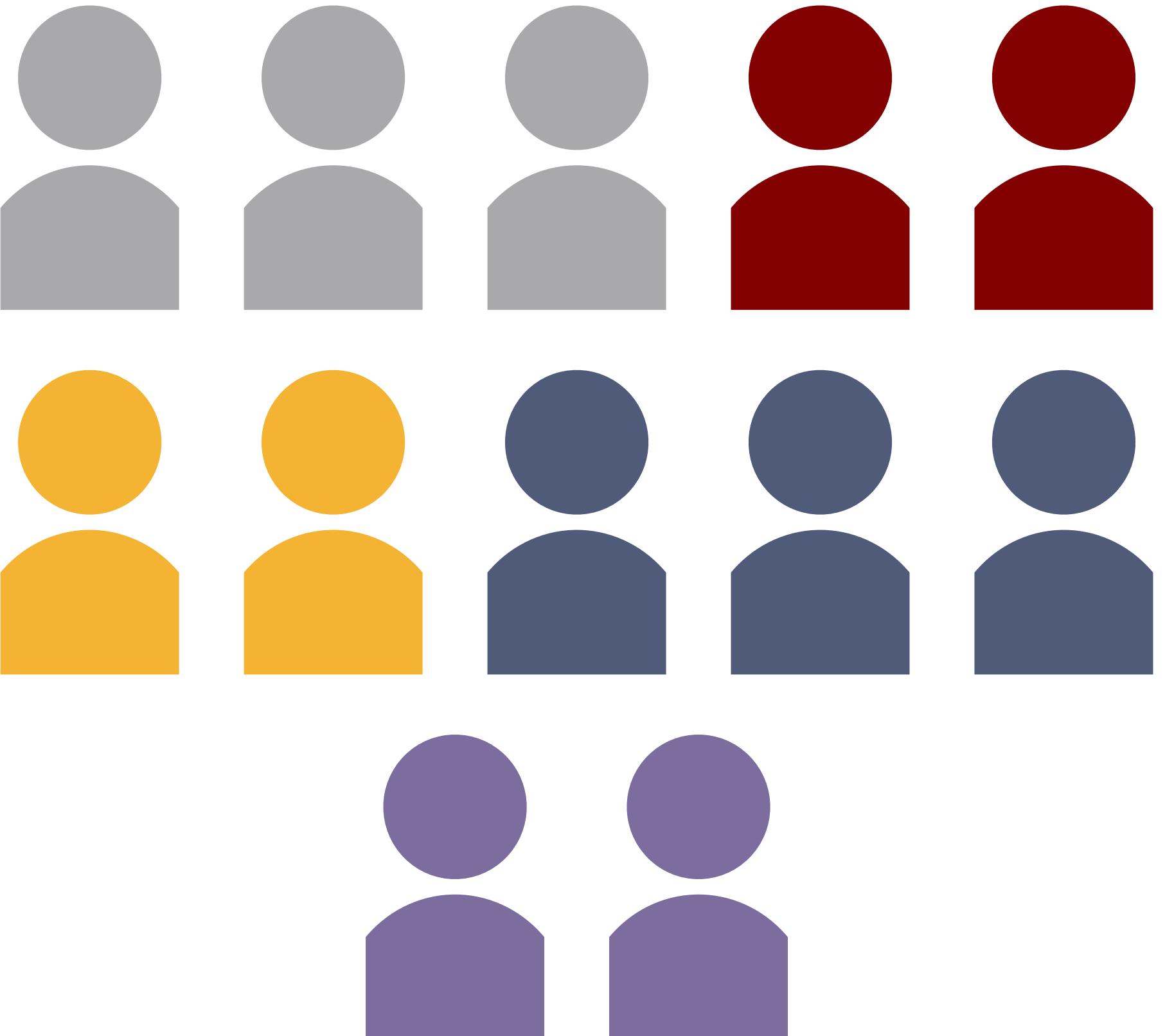
[Ross et al. 2017; Sap et al. 2022;  
Bowker and Star 2000]

*By embedding representations of people and society in interactions and models we can reason over societal disagreement.*

# Jury Learning

An interactive supervised learning architecture that makes voice **explicit**

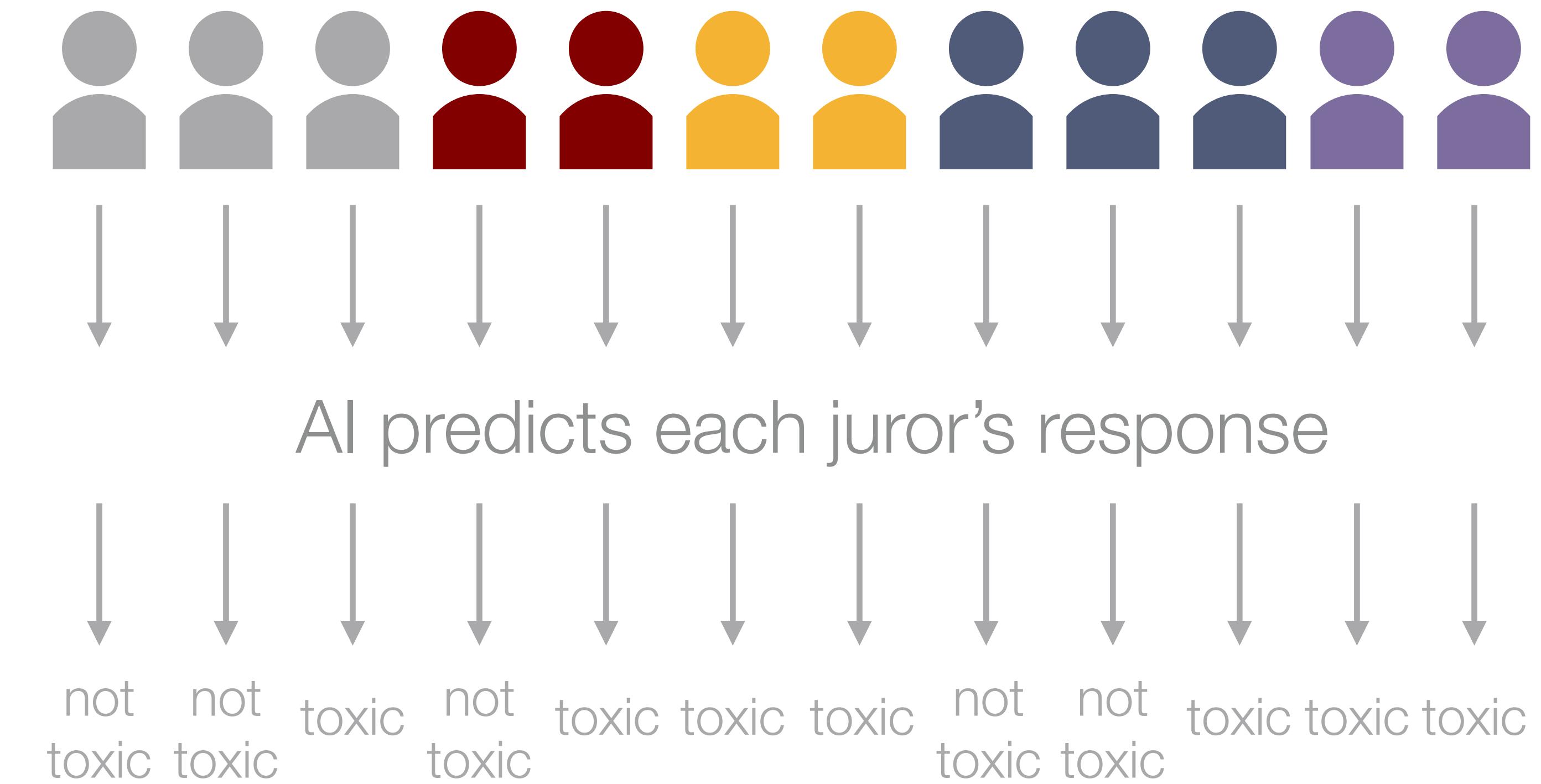
Specify a jury of, say, 12 members, and articulate **what proportion of the jury should represent each perspective in your dataset**



# Model individuals, not an aggregated pseudo-human

“1. People still eat at Pizza  
Hut? Gross. 2. It is  
shameful how this  
country [...]”

Input



7 to 5: toxic

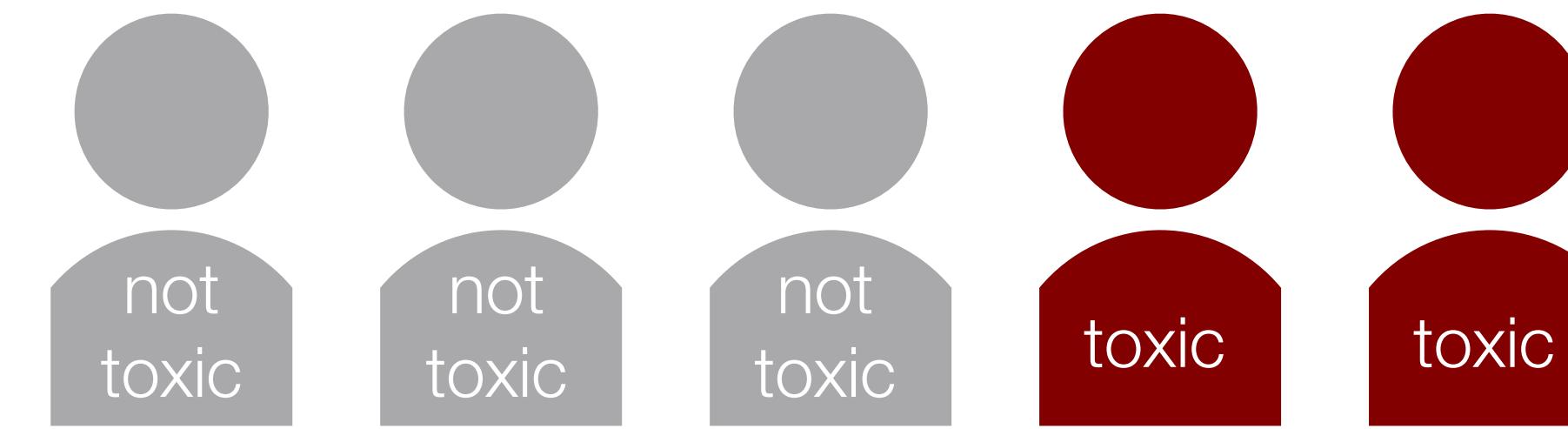
“For this jury of adults over 60, which is split evenly between doctors, lawyers, and accountants, **56% agree** the comment is toxic.”

# Jury learning

1. Interaction
2. Technical approach
3. Technical + field evaluation
4. Opportunities and implications

**Group of individual decision makers from which single decision emerges**

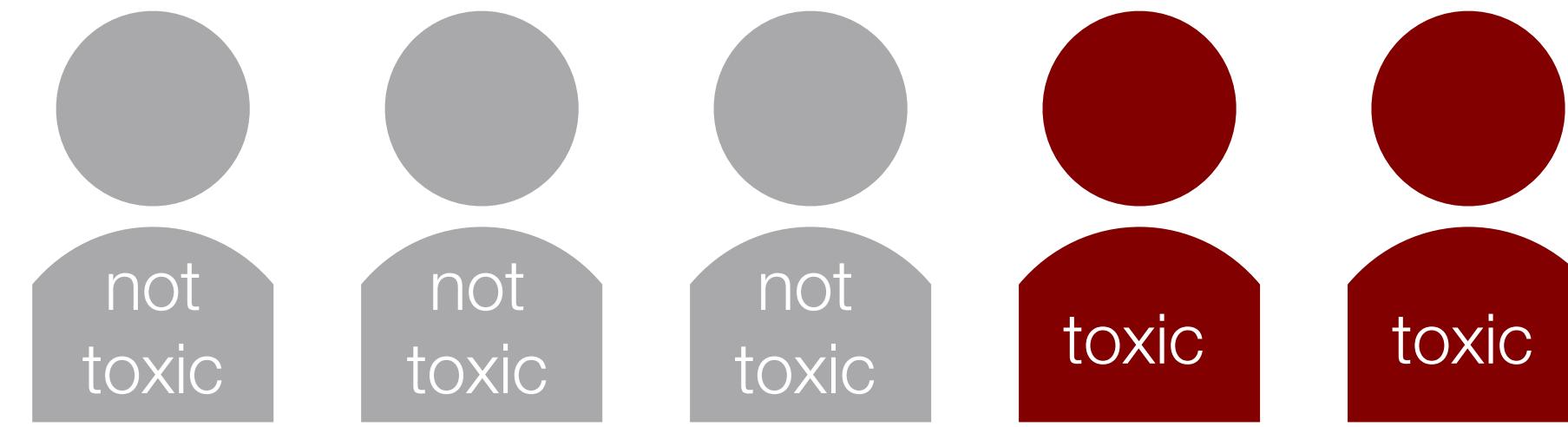
*jury:*



Every dataset **already has** a jury

# **Every dataset *already has* a jury**

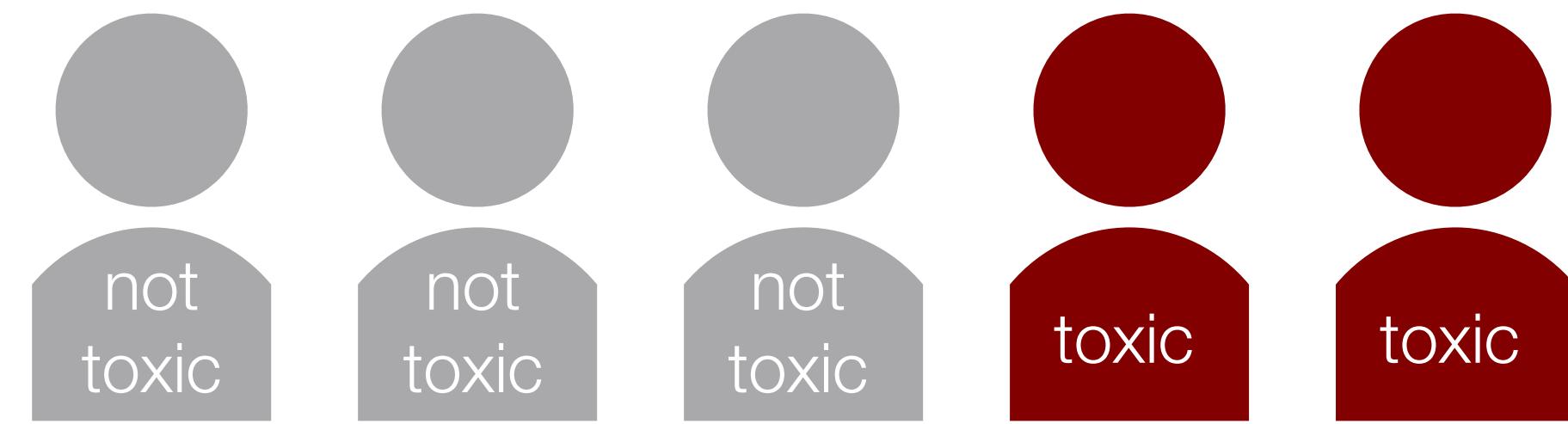
*jury:*



Group of individual decision makers from which single decision emerges

# Every dataset already has an *implicit* jury

***Implicit* jury:**



Group of individual decision makers from which single decision emerges

Interaction is our version of a juror selection process

Goal: make juries **explicit**, visible, changeable

# Compose jury by selecting from characteristics in dataset

Your jury composition

Total: 8

A<sub>1</sub> A<sub>2</sub> A<sub>3</sub> A<sub>4</sub> A<sub>5</sub> B<sub>1</sub> B<sub>2</sub> B<sub>3</sub>



## Juror Selection

+ Add a juror sheet

### Juror Sheet A

X

⊕ Add characteristic

Seats

5

### Juror Sheet B

X

⊕ Add characteristic

Seats

3

## Your input example

Place a comment here that you would like to test

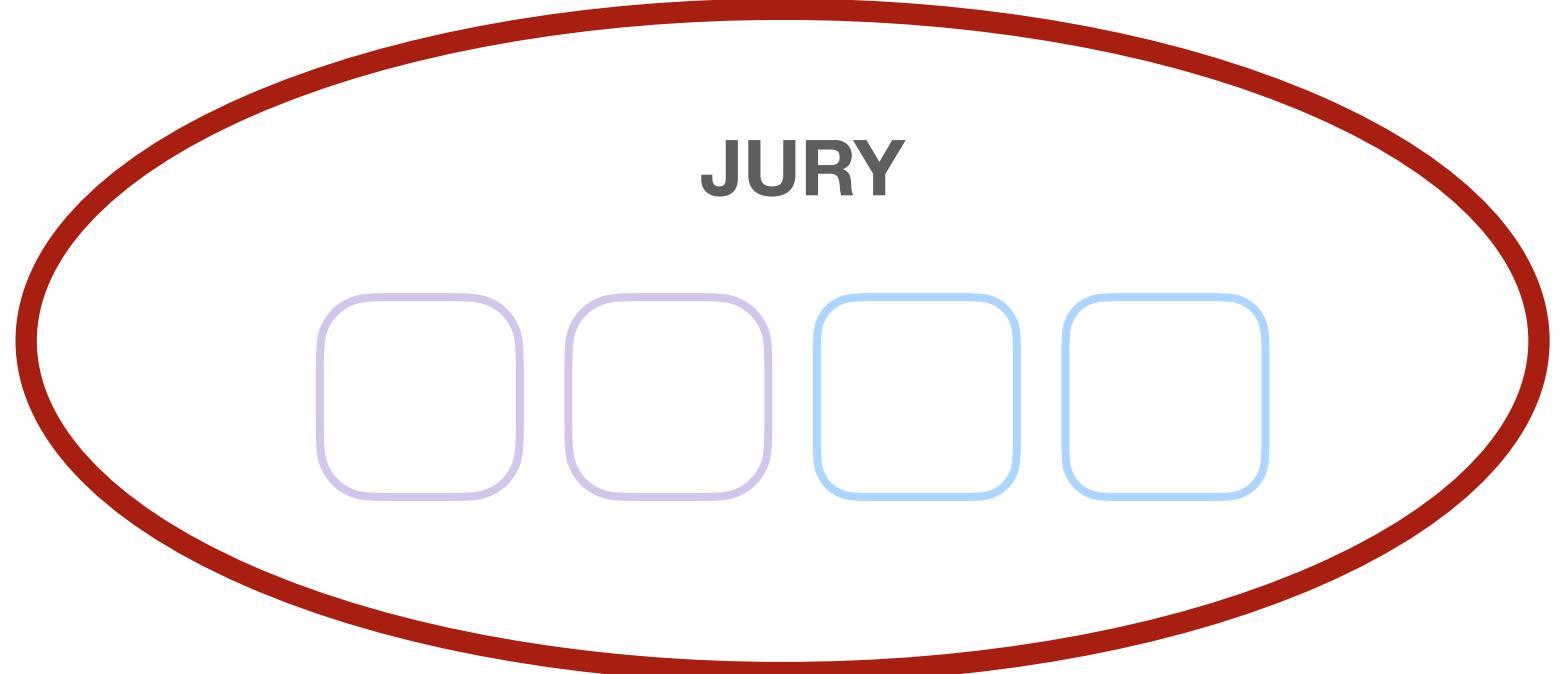
→ View Jury Outcome

# Compose jury by selecting from characteristics in dataset

Your jury composition Total: 4

A<sub>1</sub> A<sub>2</sub> B<sub>1</sub> B<sub>2</sub>

JURY



## Juror Selection

+ Add a juror sheet

### Juror Sheet A

Political affiliation

Liberal

X

-

⊕ Add characteristic

Seats

2

### Juror Sheet B

Is Parent

✓

X

-

Education

HS Diploma

v

-

⊕ Add characteristic

Seats

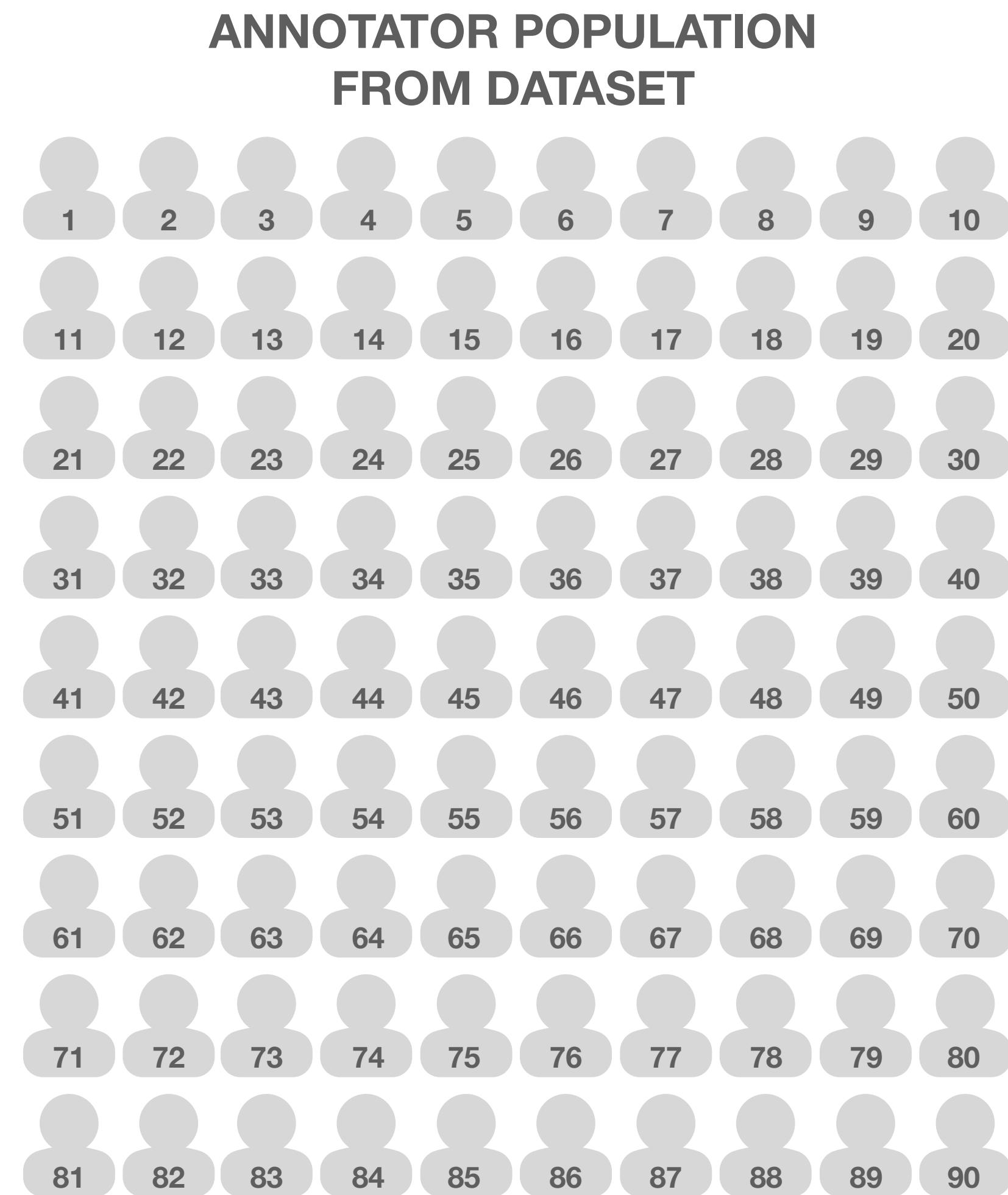
2

^

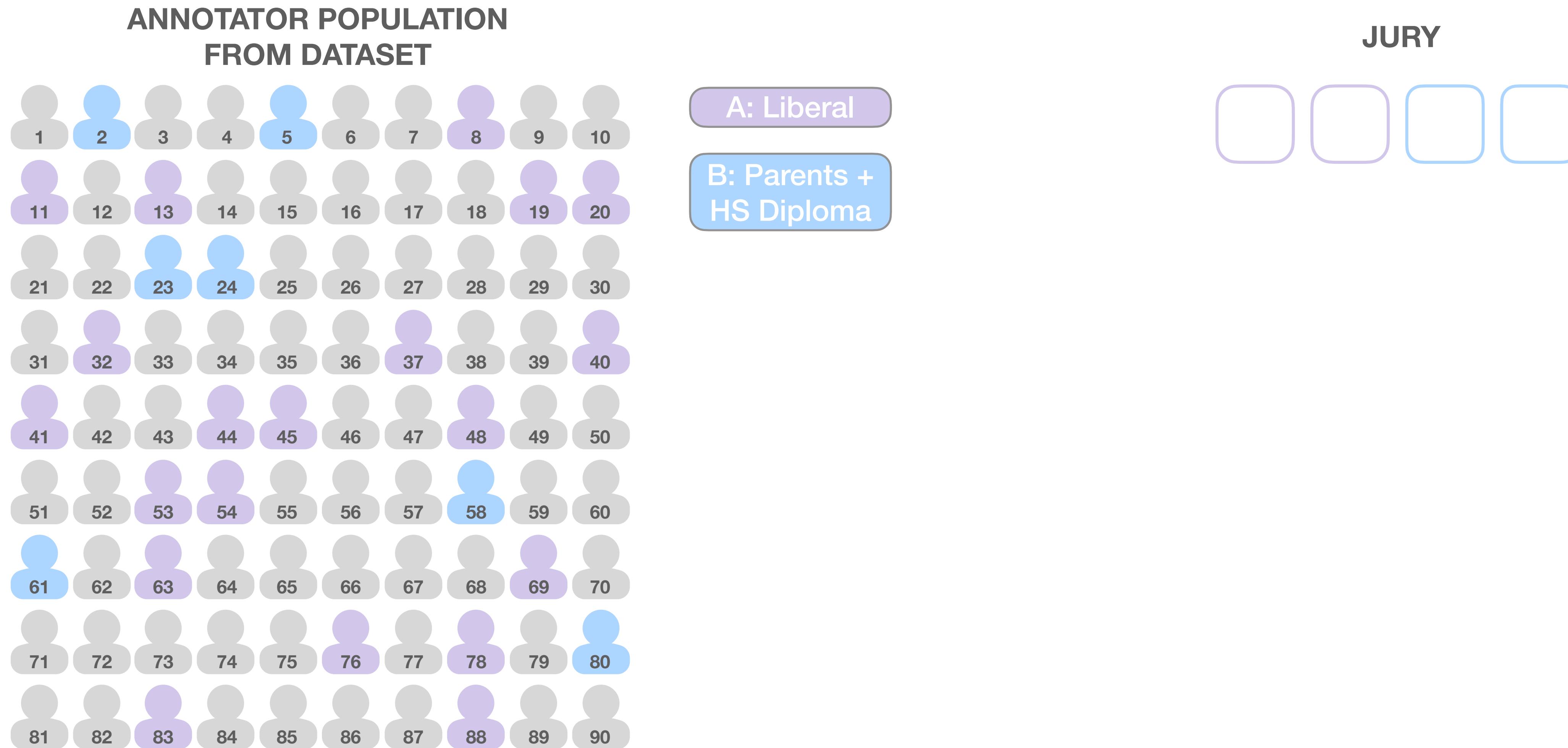
v

Your input example

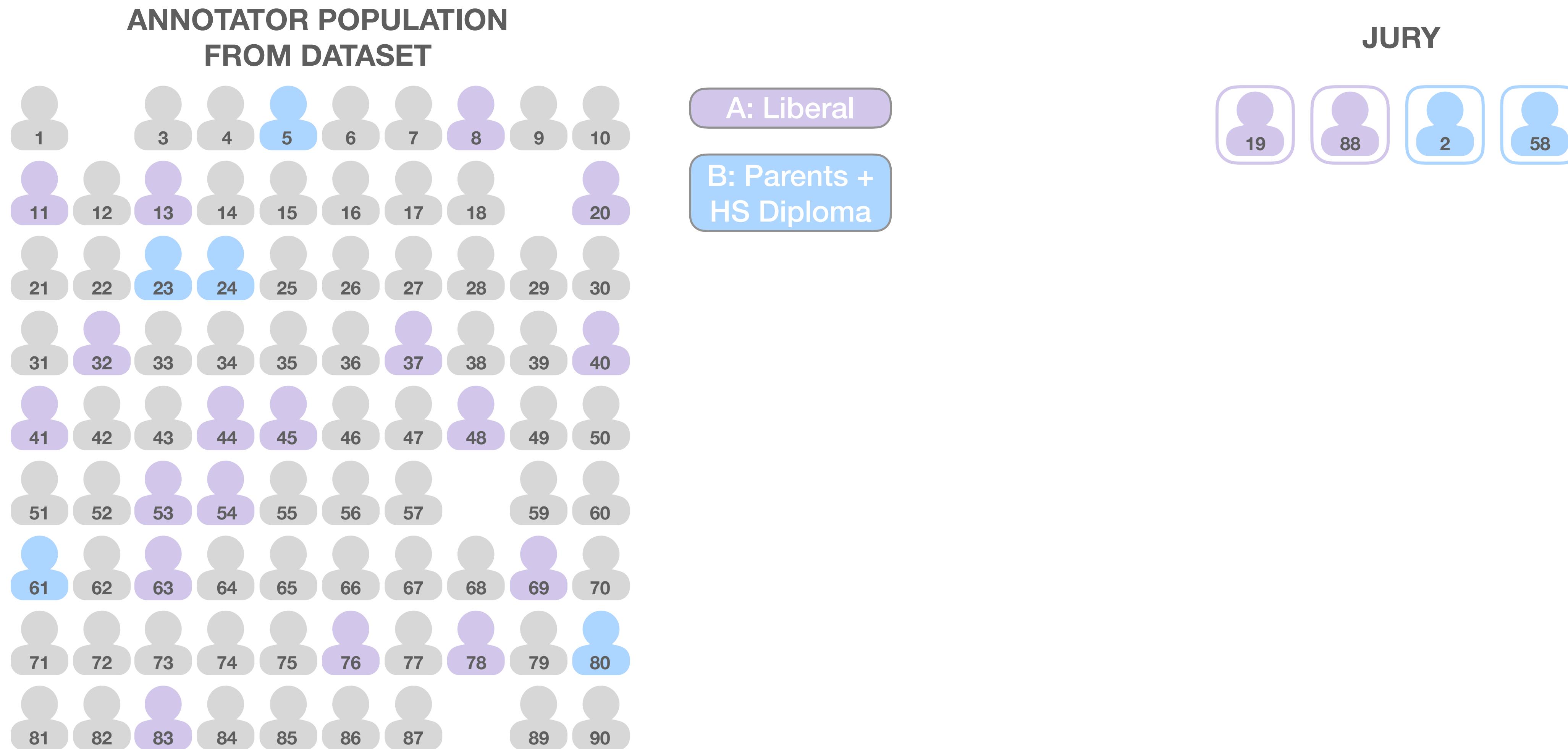
# System selects matching annotators from dataset as jurors



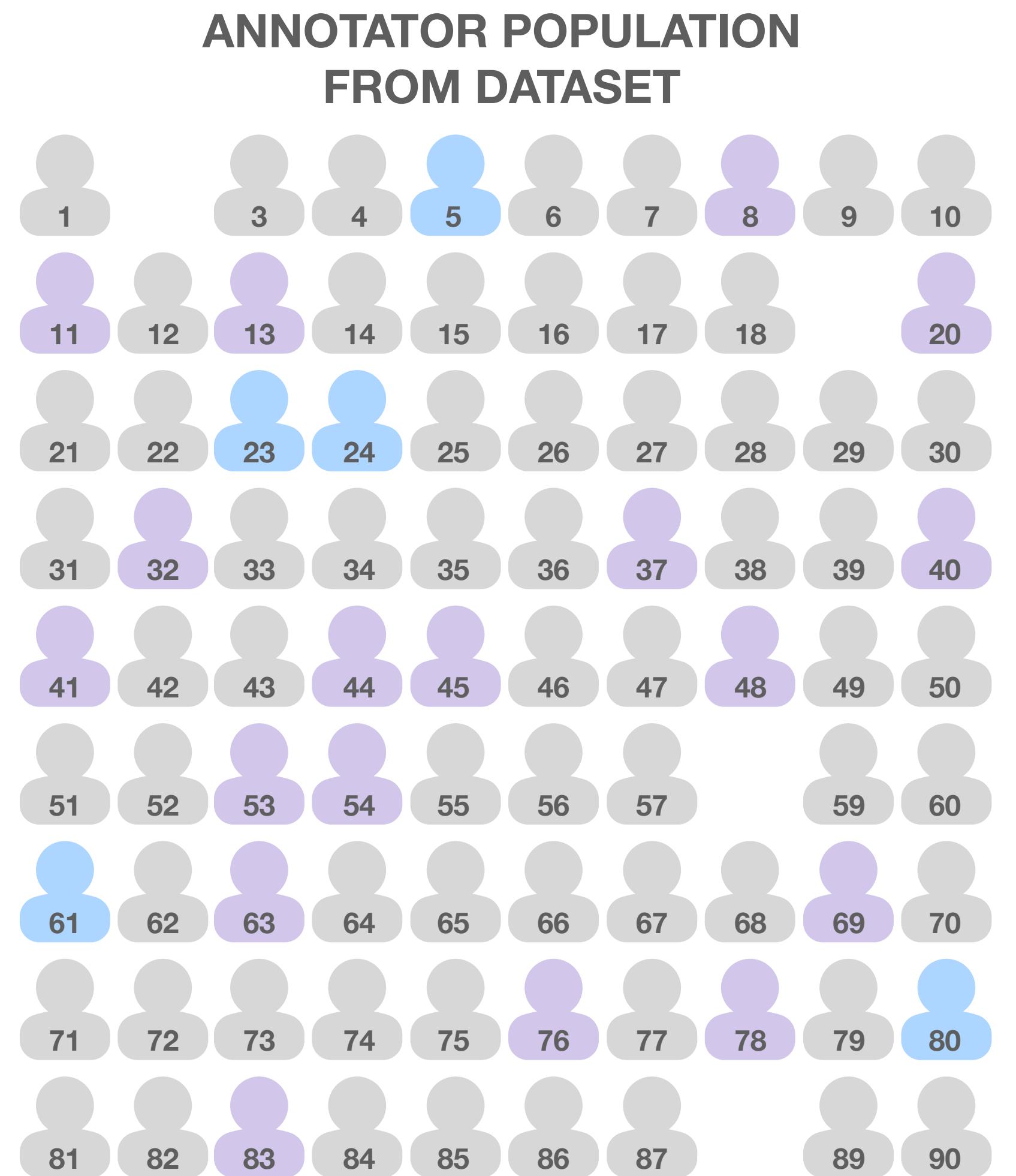
# System selects matching annotators from dataset as jurors



# System selects matching annotators from dataset as jurors



# AI predicts how each juror would vote



A: Liberal

B: Parents +  
HS Diploma

JURY

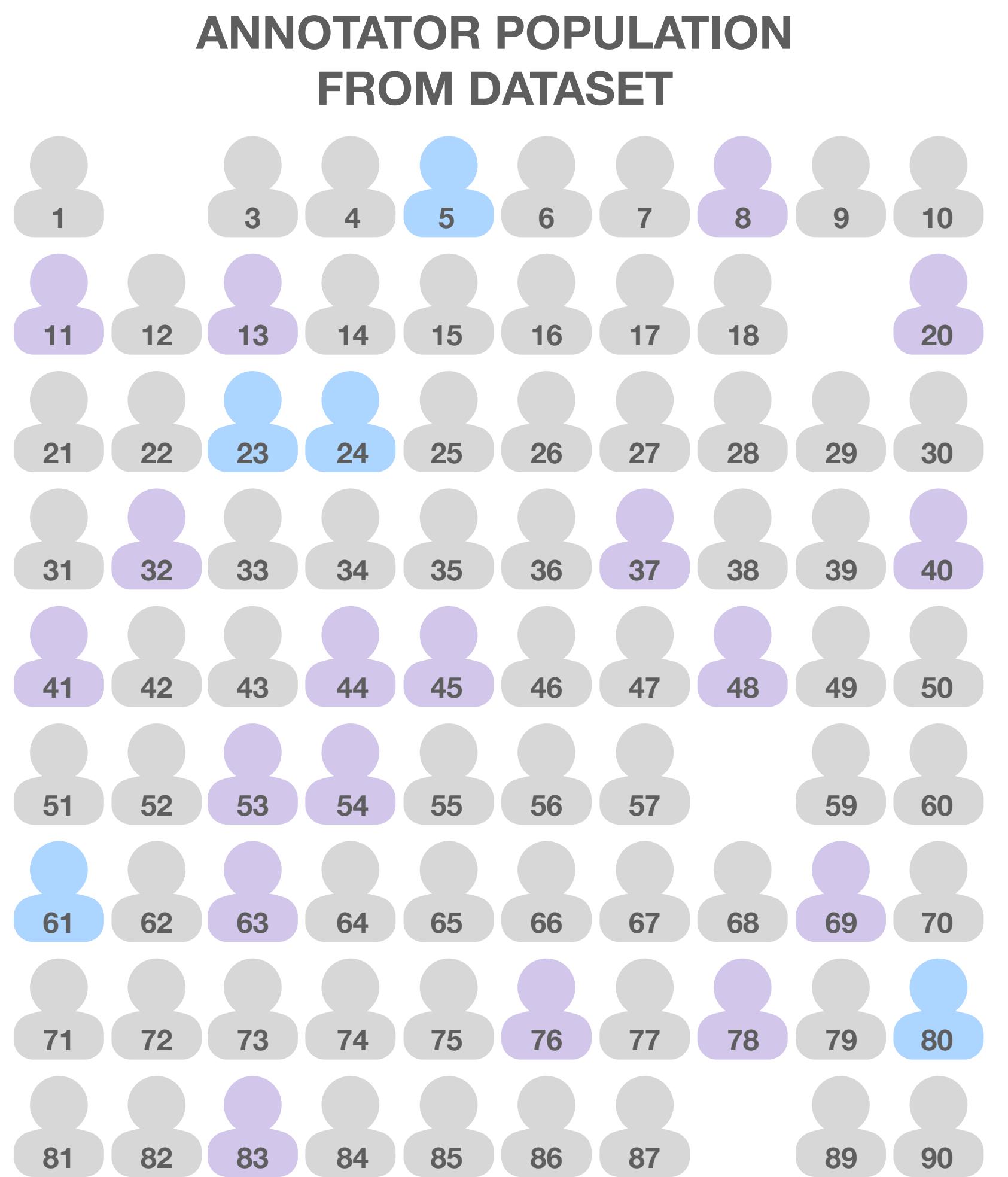


AI predicts each juror's response

not toxic  
toxic

3 to 1: not toxic

# AI predicts how each juror would vote



A: Liberal

B: Parents +  
HS Diploma

JURY

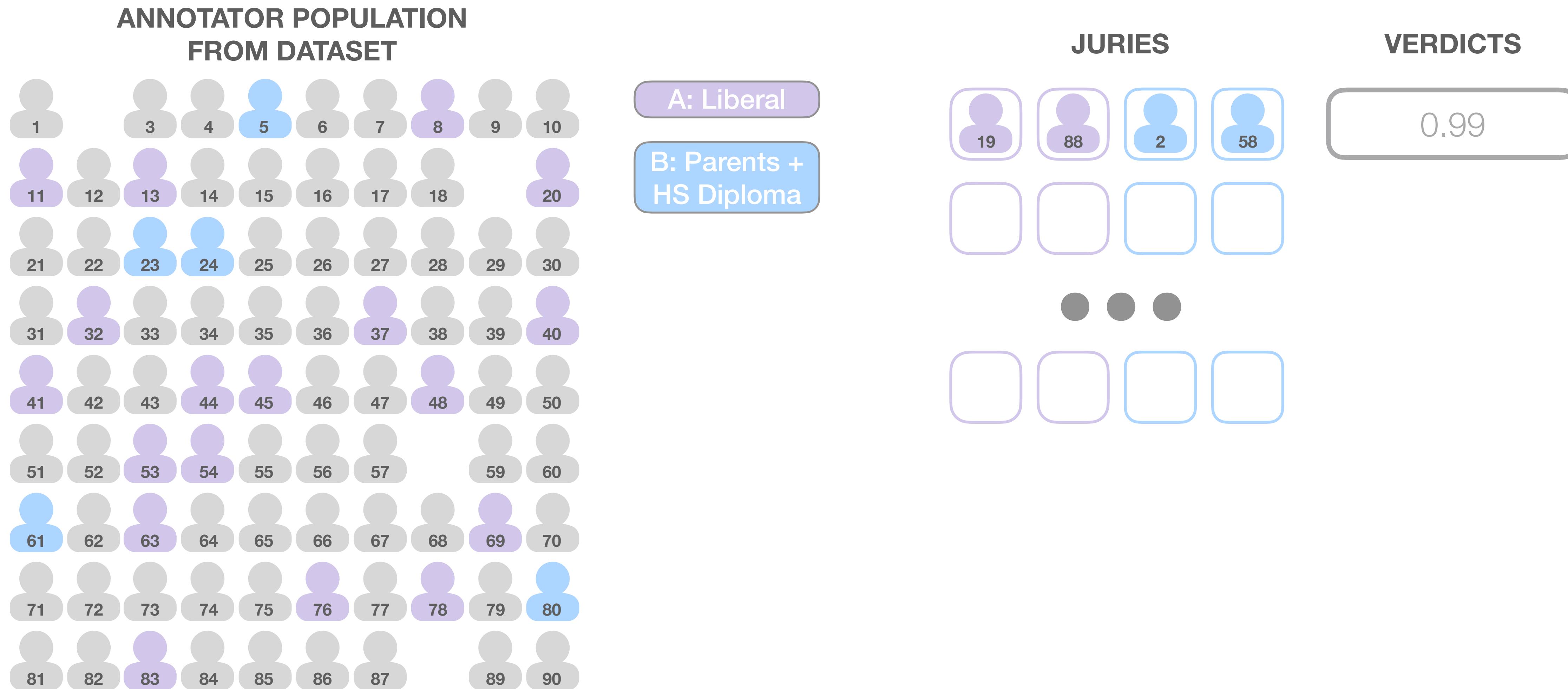


AI predicts each juror's response

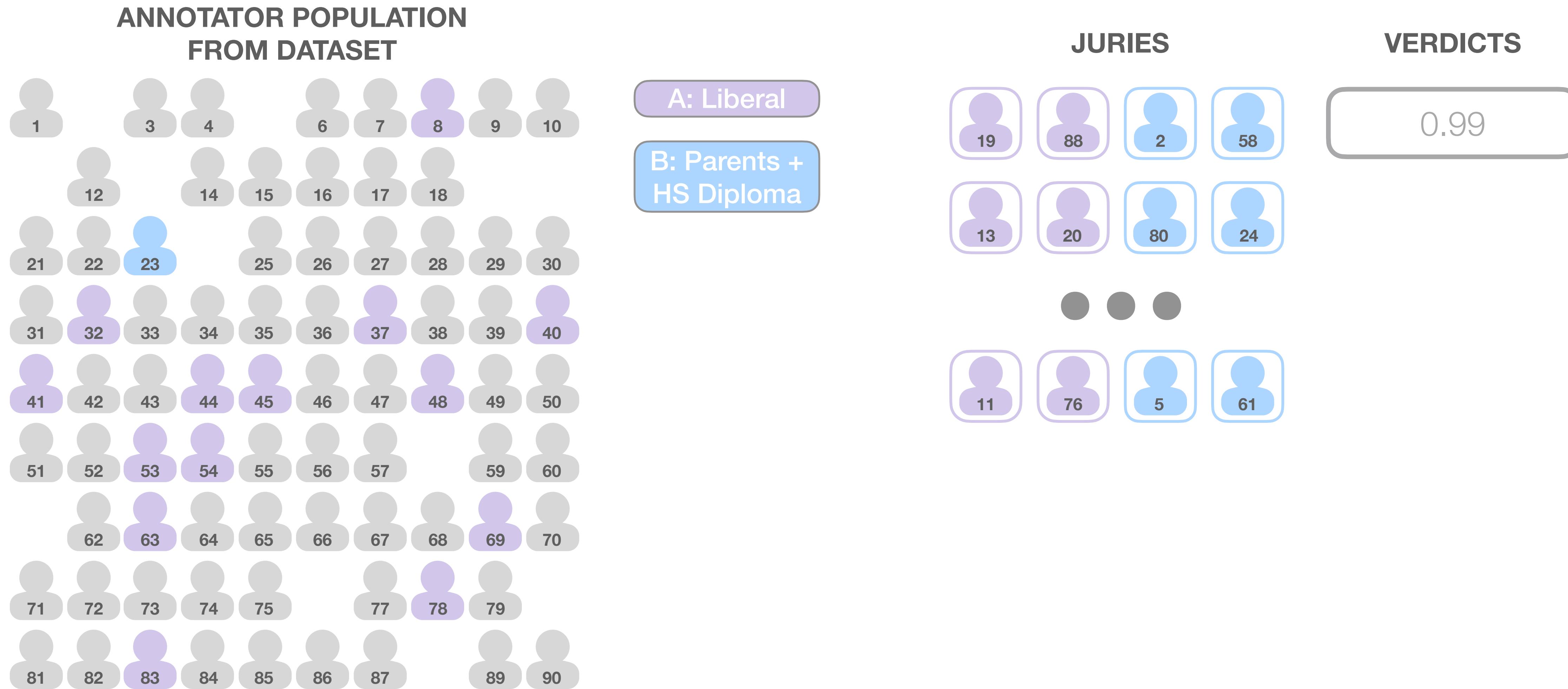
0.84 0.91 1.34 0.89

0.99/4.00: slightly toxic

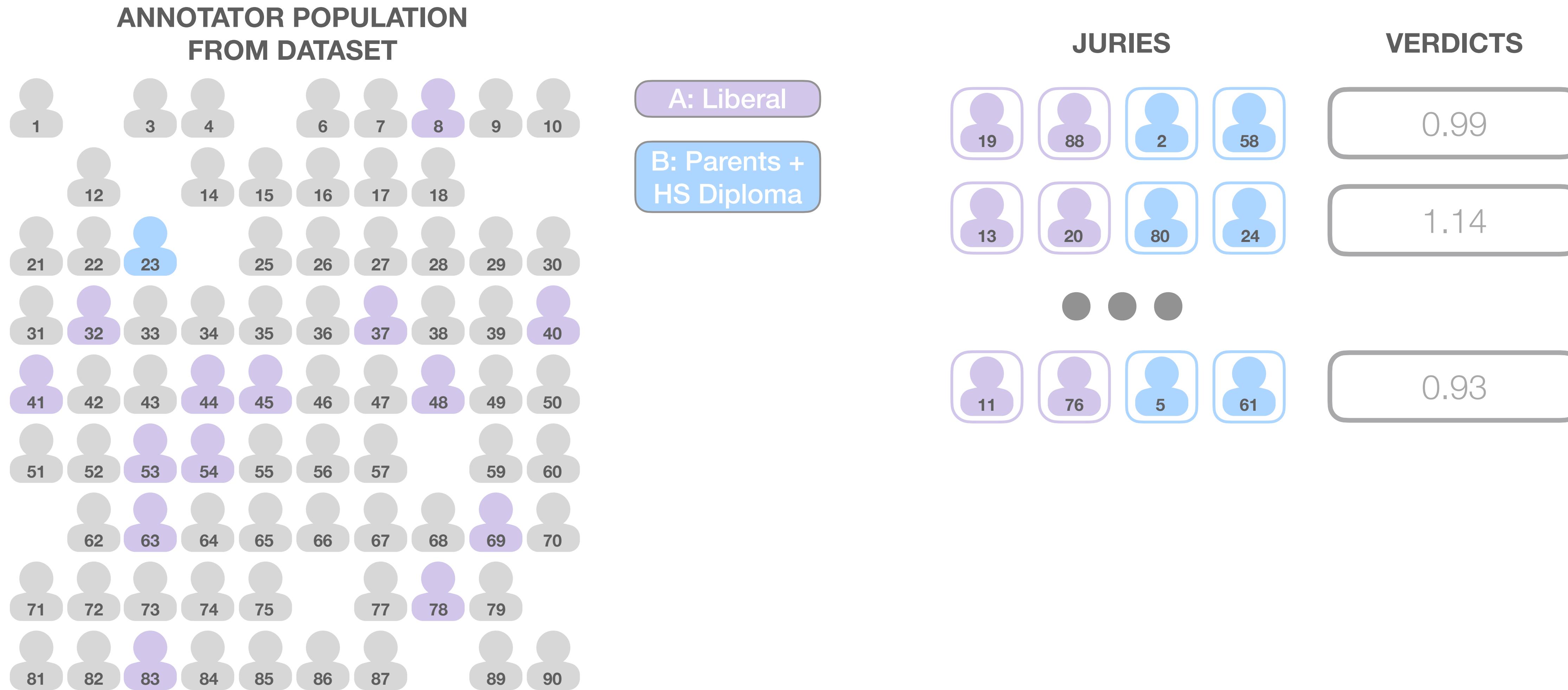
# Randomly re-sample jury, creating parallel juries



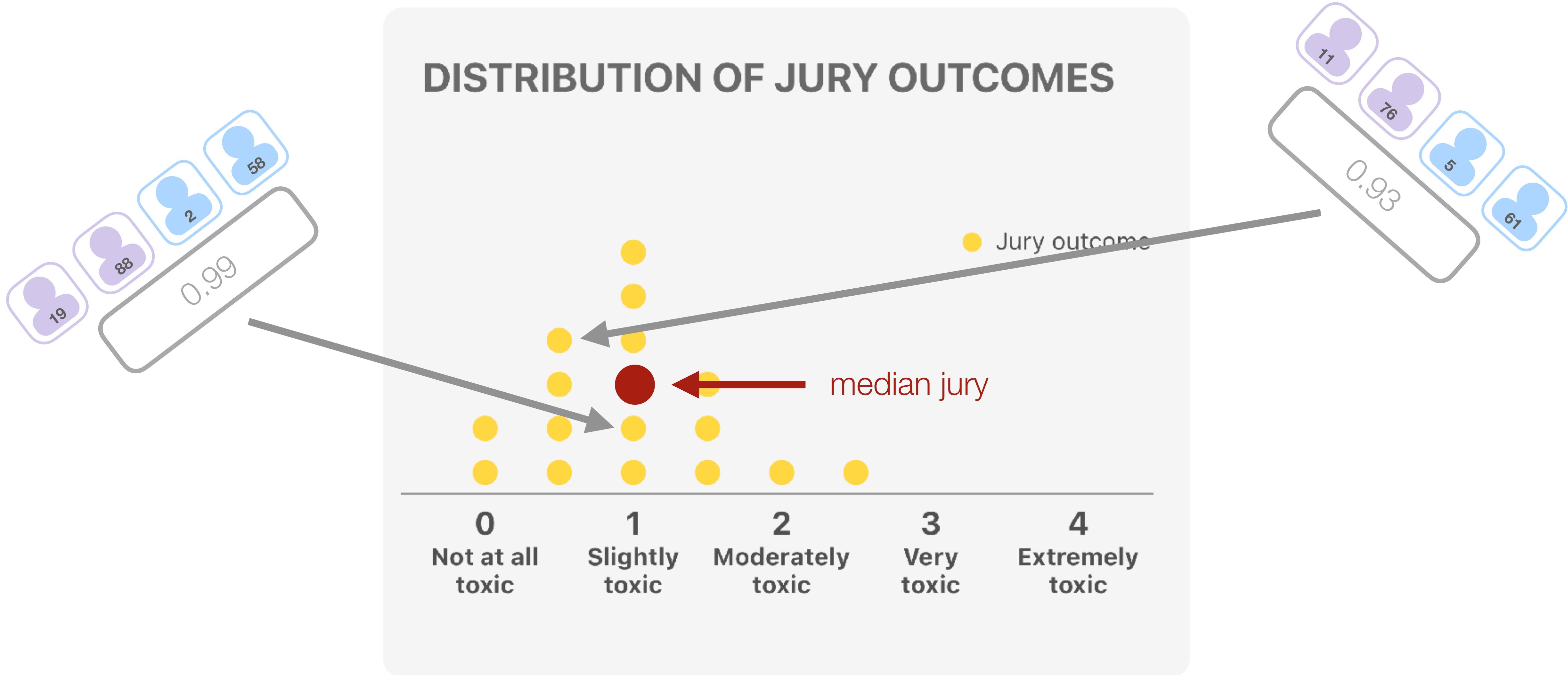
# Randomly re-sample jury, creating parallel juries



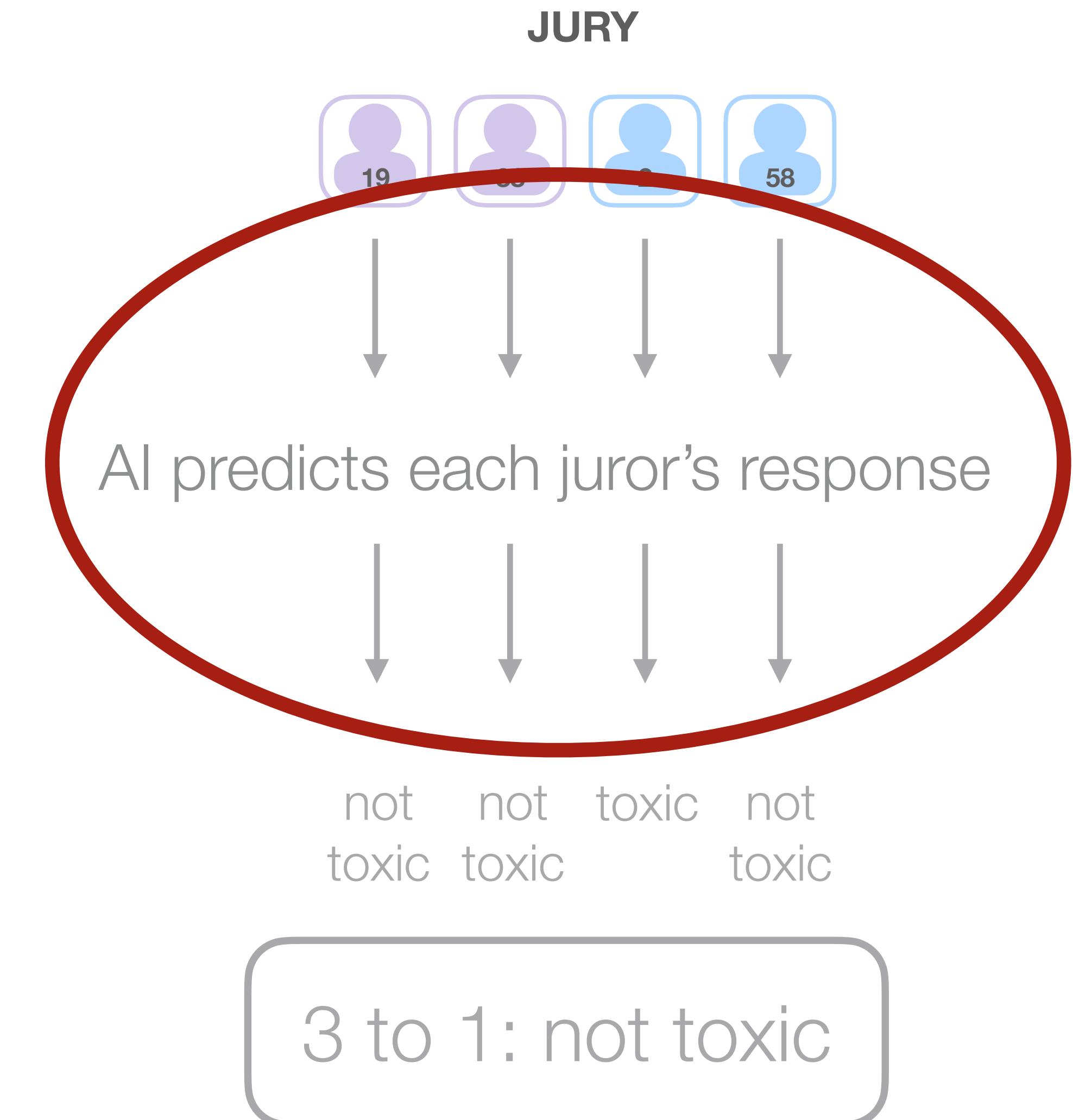
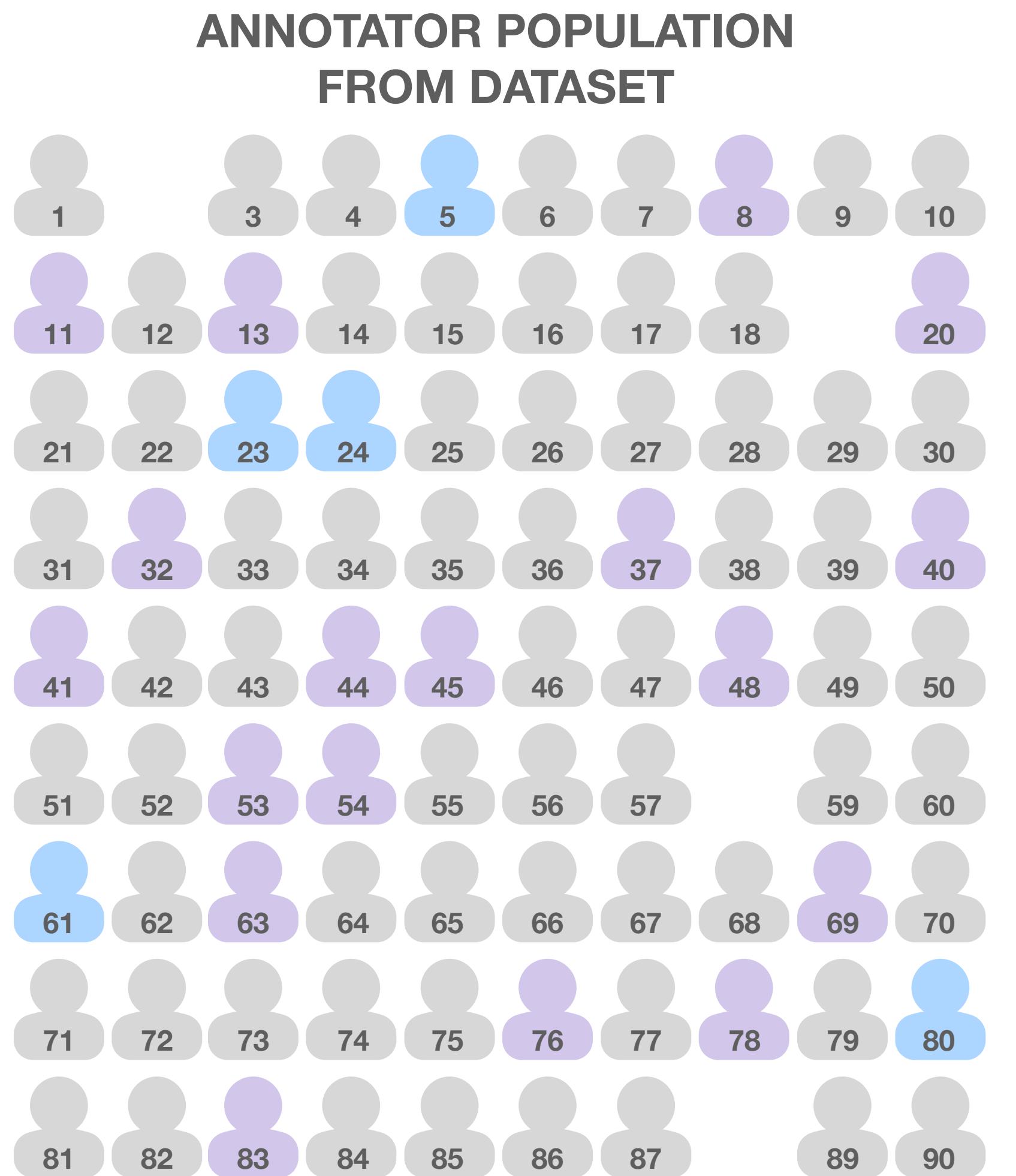
# Randomly re-sample jury, creating parallel juries



# Distribution of jury verdicts, final decision via median-of-means



# AI predicts how each juror would vote



# Jury learning

1. Interaction
- 2. Technical approach**
3. Technical + field evaluation
4. Opportunities and implications

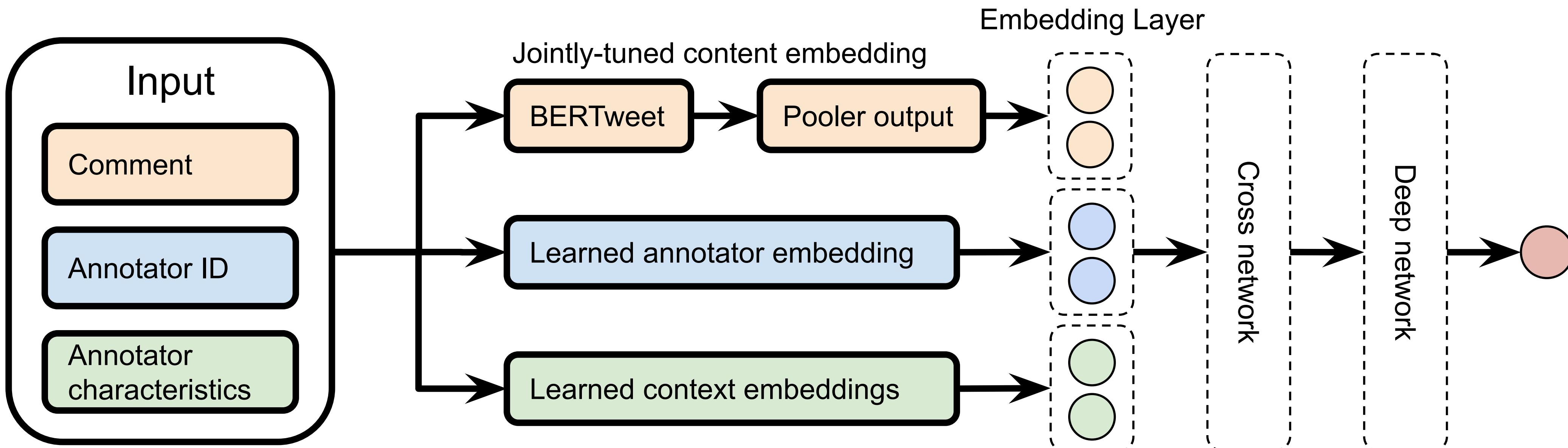
# Jury learning model architecture

Heavy reliance on content features (all predictions are “cold start”)

**Approach:** hybrid recommender system [DCN, Wang et al. 2020]

**Our insight:** augment content embeddings by fine tuning a pre-trained LLM

**Integrate ideas from pre-trained transformers in NLP with hybrid recommender systems**



# Jury learning

1. Interaction
2. Technical approach
- 3. Technical + field evaluation**
4. Opportunities and implications

# Field evaluation: do juries change classification decisions?

Participants' juries change **14% of classifications** versus standard aggregated classifier (BERTweet, fine tuned)

**Most likely to flip:** contentious, divisive issues

- Racism
- Death/suicide
- LGBTQ+
- Mental illness/disorders
- Cops

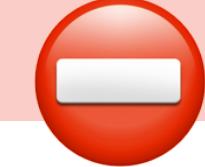


**Least likely to flip:** uncontroversial issues (good and bad)

- Largely innocuous topics**
- Thank-yous
  - Happiness
  - Hugs
  - Weddings



- Largely offensive topics**
- Human trafficking
  - R-word
  - Racial/ethnic slurs



BERTweet: [https://huggingface.co/docs/transformers/model\\_doc/berttweet](https://huggingface.co/docs/transformers/model_doc/bertweet)

Dataset: <https://data.esrg.stanford.edu/study/toxicity-perspectives>

# Jury learning

1. Interaction
2. Technical approach
3. Technical + field evaluation
- 4. Opportunities and implications**

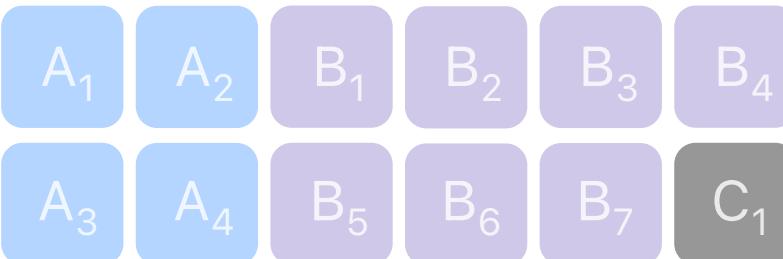
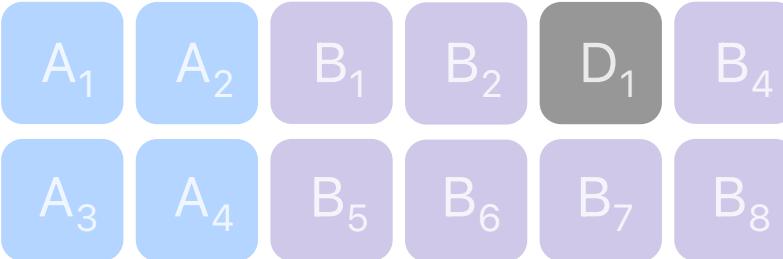
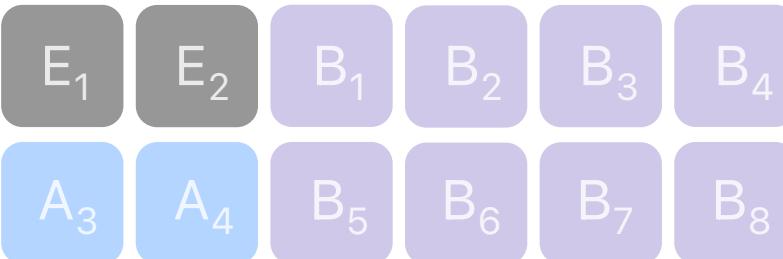
# Jury learning opportunities: conditional juries

Compose different juries based on the decision being made

```
# select the six jurors based on context
conditional_jurors = []
if '#starwars' in tweet:
    conditional_jurors = [ { 'jurors': 6, 'is_startrek_fan':'no'} ]
elif '#covid19' in tweet:
    conditional_jurors = [ { 'jurors': 6, 'profession':'MD'} ] ]
elif ...
# additional conditions and conditional jurors
```

# Jury learning opportunities: counterfactual juries

*Which jury compositions would flip the outcome?*

New jury composition	Jury verdict	Jury edits
	 (0.87 / 4.00)	C <sub>1</sub> — Race: White, Political Affiliation: Conservative
	 (0.79 / 4.00)	D <sub>1</sub> — Race: Black, Importance of religion: Not important
	 (0.63 / 4.00)	E <sub>1</sub> , E <sub>2</sub> — Age range: 45-54, Importance of religion: Very important

Juries that would have ruled differently

# Ethical consideration: who determines jury composition?

All datasets *already have* an implicit jury, just hidden

Jury learning makes them explicit and visible

New way to communicate/debate voices included

**Our take:** make jury compositions transparent, provide stakeholders a voice in selection

**SELECTED JURY COMPOSITION**



# This talk: three directions

Designing interactive systems for reasoning about different goals and values.

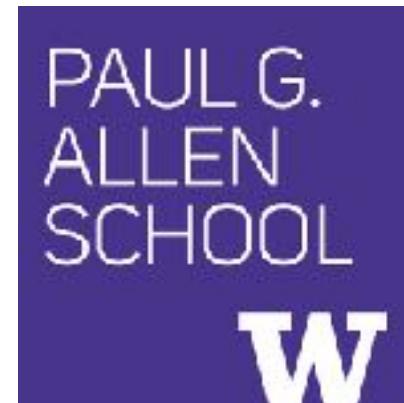
Jury Learning: Integrating Dissenting Voices into Machine Learning Models

## **Exploring formal definitions of pluralistic alignment**

A Roadmap to Pluralistic Alignment

Eliciting preferences from the public

Collective alignment: public input on our Model Spec



Taylor  
Sorensen



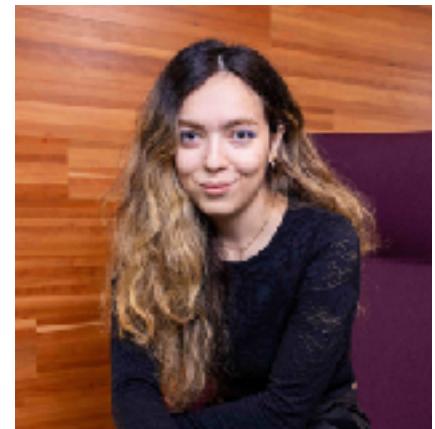
Jared  
Moore



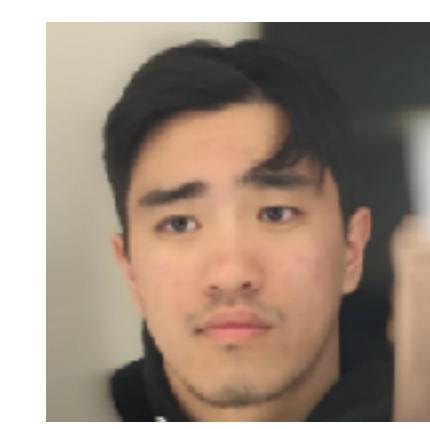
Jillian  
Fisher



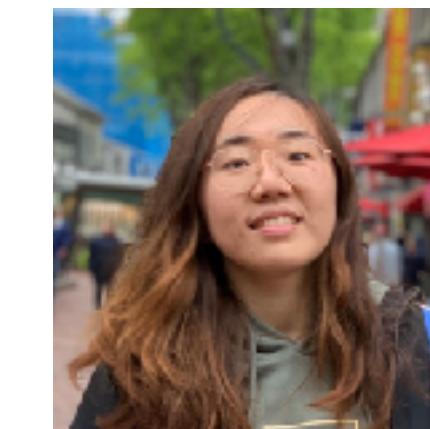
Mitchell  
Gordon Niloofar  
Mireshghallah Christopher  
Rytting



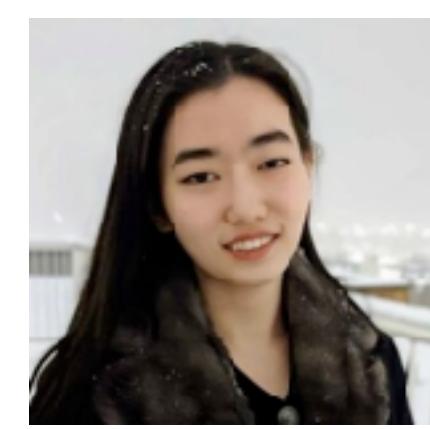
Andre  
Ye



Liwei  
Jiang



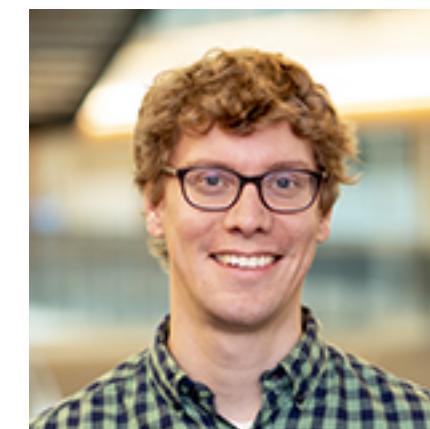
Ximing  
Lu



Nouha  
Dziri



Tim  
Althoff



Yejin  
Choi



# A Roadmap to Pluralistic Alignment

# Outline

Why Pluralism

Pluralistic Models

Pluralistic Benchmarks

# Why Pluralism

- Needed for customization
- Technical benefits - variation is signal, not noise
- Needed for evaluating generalist systems
- As a value itself
- AI systems should reflect human diversity

Why Pluralism

Pluralistic Models

Pluralistic Benchmarks

# Overton Pluralism



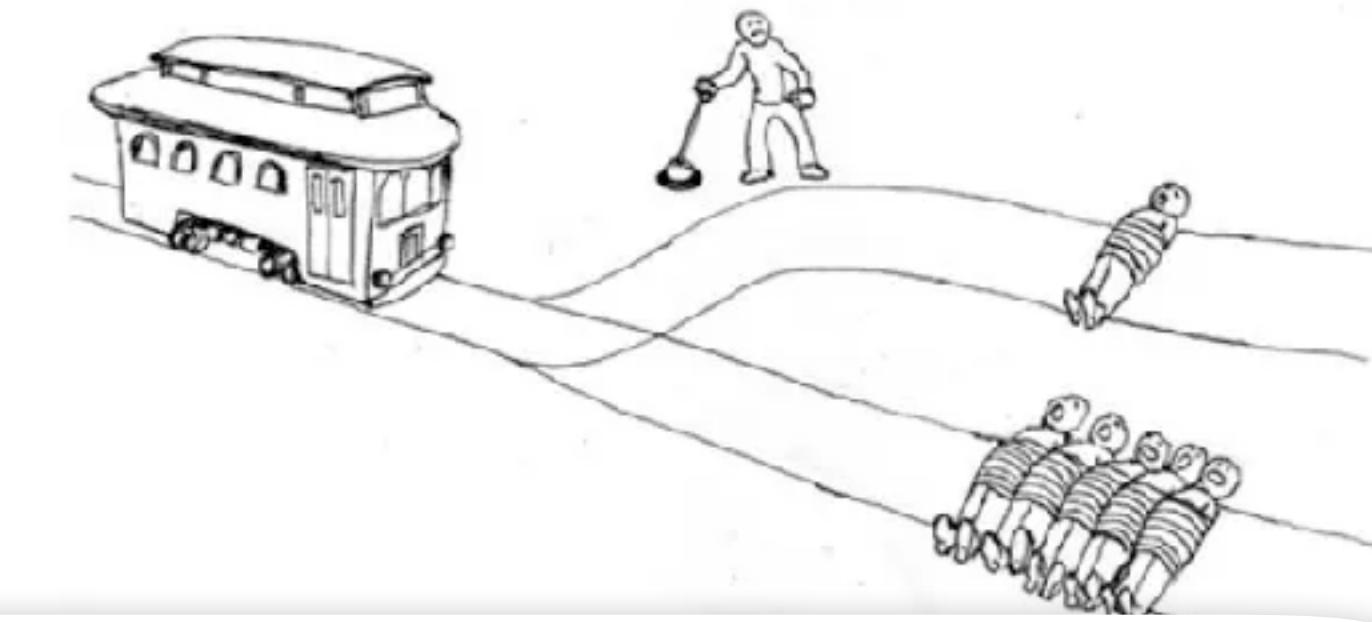
## Pluralistic Human Values



## Overton



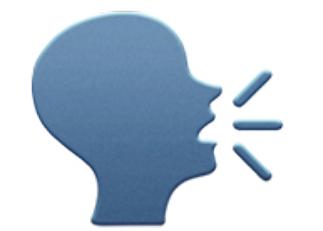
Different schools of thought might give different answers. For example, according to **utilitarianism**, the right thing to do is to **save the most lives**, regardless of how it occurs. A **deontologist** might say that you have a duty to do no harm, and that it would be wrong to intentionally cause the one person's death. If you prescribe to the **virtue of preserving human life**, ...



## Definitions

- (1) *Correct Answer in C*: An answer which can be conclusively verified or with which the overwhelming majority of people across various backgrounds would agree.
- (2) *Reasonable Answer in R*: An answer for which there is suggestive, but inconclusive, evidence, or one with which significant swaths of the population would agree. Additional top-down restrictions (e.g., safety) may apply.
- (3) *Overton window*: The set of all reasonable answers:  $W(x) = \{y \in \mathcal{Y} | (x, y) \in \mathcal{R}\}$ .<sup>1</sup>
- (4) *A response set  $\{y\}$  to a query  $x$  is Overton-pluralistic*:  $\{y\}$  contains all potentially reasonable answers in the Overton window. This is in contrast to picking just one answer in the Overton window, or presenting an unreasonable answer which would lie outside the Overton window. A single response may be Overton-pluralistic if it synthesizes the whole response set  $\{y\}$ .
- (5) *Model  $M$  is Overton-pluralistic*:  $M$  gives Overton-pluralistic responses to queries, that is for a given input  $x$ , the output of  $M(x) = W(x)$ .

# Overton Pluralism



## Potential Implementation

- Define a set of queries  $X$  along with set of reasonable answers
- Either: extract “answers” from response; or
- Detect presence with entailment

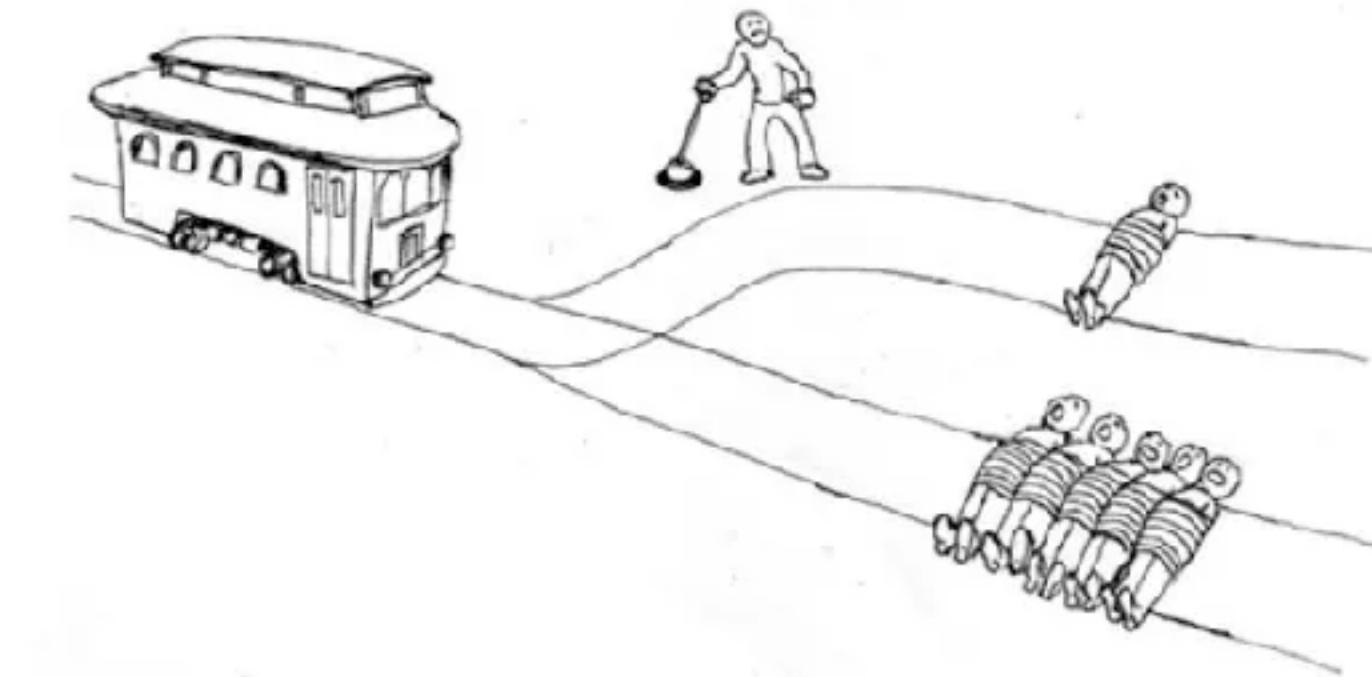
## Applications

- Advice giving
- Deliberation
- Scalable oversight
- Settings where we want to encourage multiple approaches

## Limitations

- Defining an Overton window presents a challenge
- Bothsidesism
- Requires long-form responses

# Steerable Pluralism



## Pluralistic Human Values

What should I do?



## Steerable



You should always do the action that will save the most lives.

You have a duty to do no harm and not intervene.

If you prescribe to the virtue of preserving human life, you should redirect the trolley.

## Definitions

(6) *Steering attributes A*: Attributes/properties/perspectives which we wish a model to faithfully reflect. Examples include groups of people from a shared culture, philosophical/political schools of thought, or particular values. To reflect multiple attributes simultaneously, the elements of  $A$  could be construed as *sets* of attributes.

(7) *Response  $y|_{x,a}$  faithfully reflects attribute  $a \in A$* : The response  $y$  to the query  $x$  is consistent with, or follows from, attribute  $a$ .

(8) *Model  $\mathcal{M}$  is steerable-pluralistic with respect to attributes A*: Given an input  $x$  and an attribute  $a \in A$ , the model  $\mathcal{M}(x, a)$  conditioned on  $a$  produces a response  $y$  which faithfully reflects  $a$ .

# Steerable Pluralism



## Potential Implementation

- Value-specific annotations or reward
- Measure per-attribute faithfulness

## Applications

- Customization
- Steering to diverse perspectives (creativity, social systems, deliberative discourse)
- Varying “cognitive architectures”

## Limitations

- Which attributes to steer to?
- If attributes too broad, stereotyping/ flattening nuances

# Pluralistic Models

- In what cases might we want each kind of pluralism?
- What are risks if we DON'T have these properties?
- What risks lie from over-optimization or misapplication of these properties?

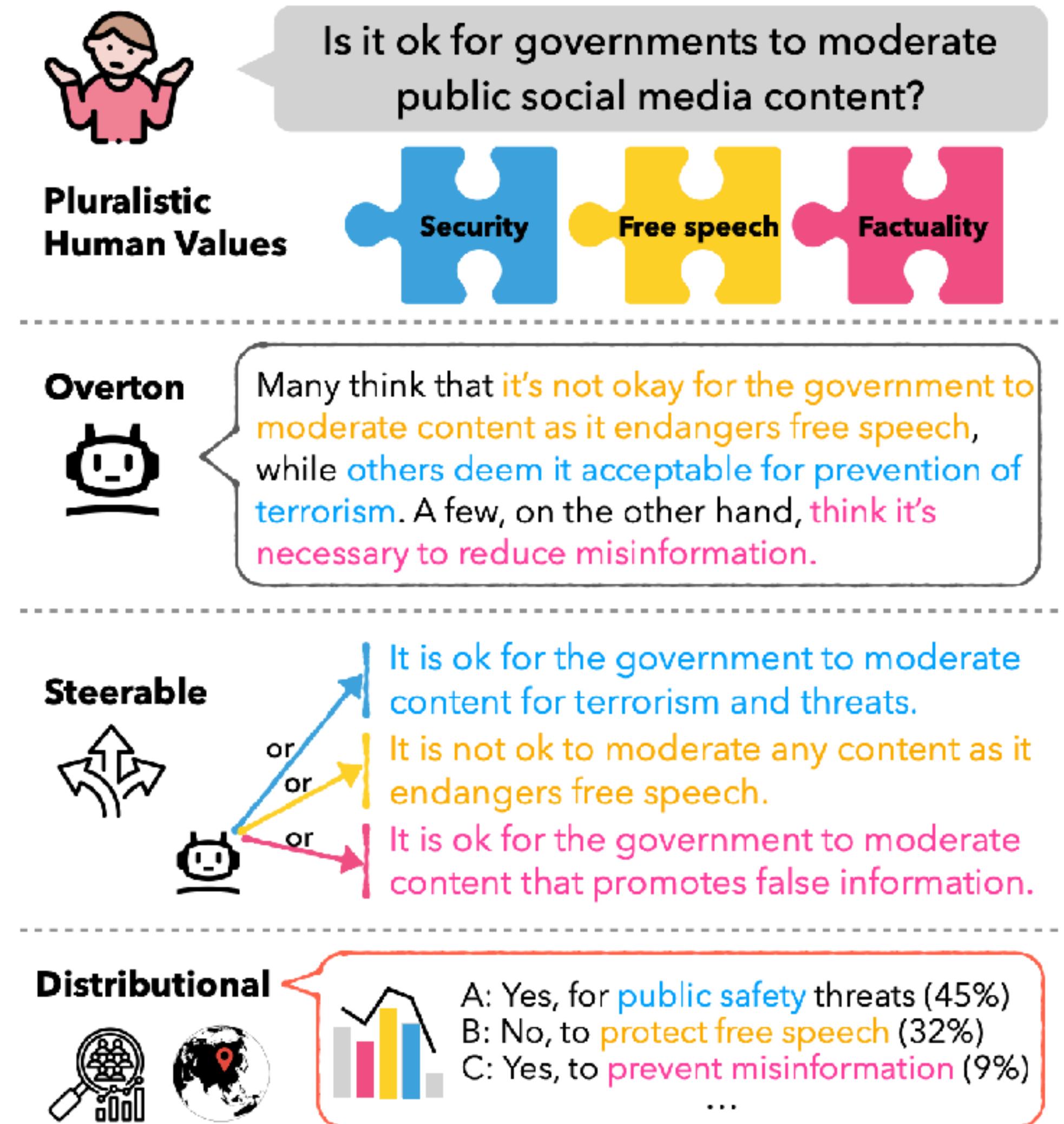


Figure 1. Three kinds of pluralism in models.

Why Pluralism

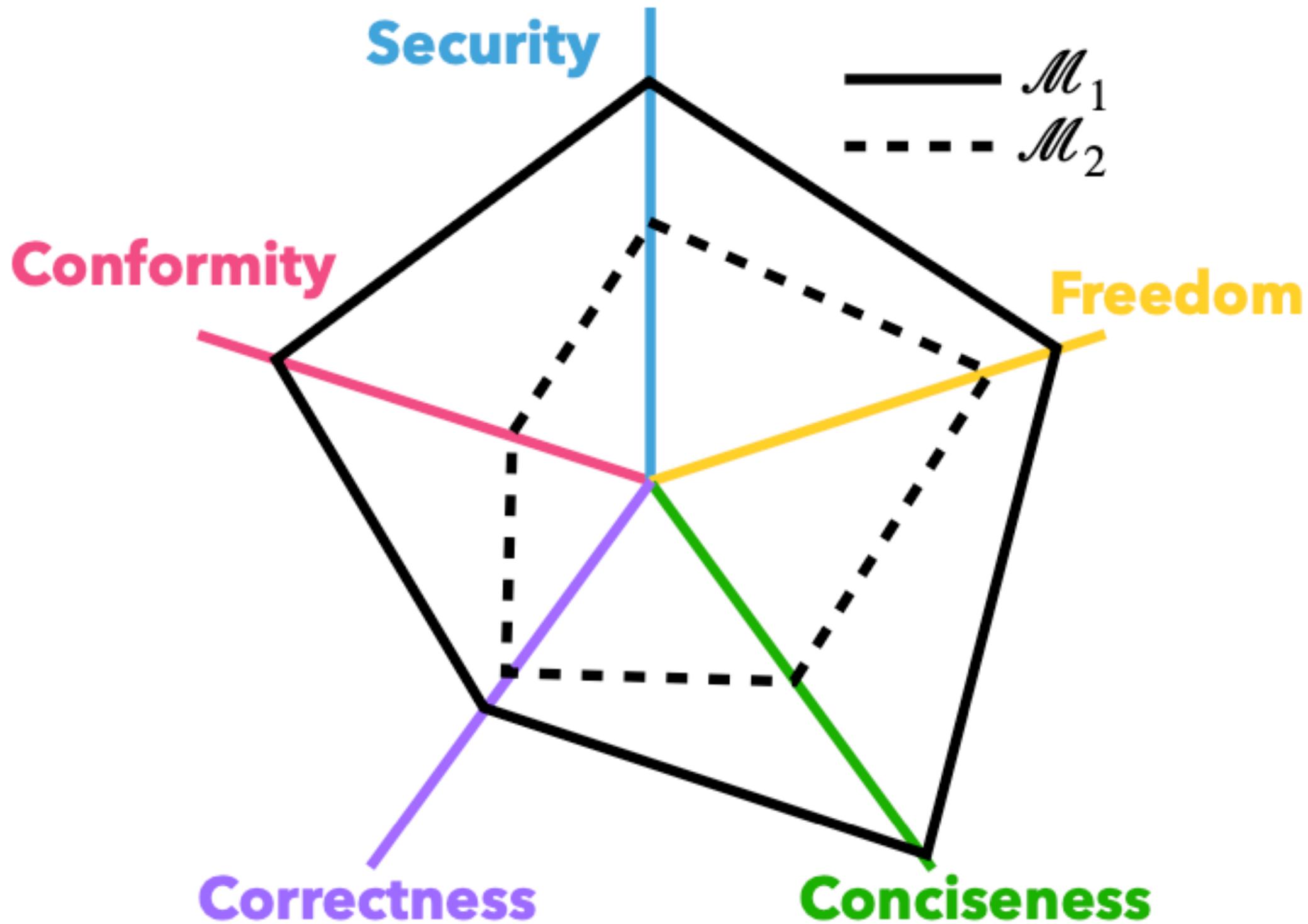
Pluralistic Models

Pluralistic Benchmarks

# Multi-Objective



## Multi-objective



# Definitions

- (11) *Objectives to maximize*  $O = \{o_1, \dots, o_n\}$ : A set of multiple objectives to evaluate a model  $\mathcal{M}$ , each of which we desire to maximize. Each  $o$  maps from a model  $\mathcal{M}$  to a scalar in  $\mathbb{R}$ .
- (12) *Model  $\mathcal{M}_1$  is a Pareto improvement to model  $\mathcal{M}_2$ :*  $\forall o_i \in O, o_i(\mathcal{M}_1) \geq o_i(\mathcal{M}_2); \exists o_j \text{ s.t. } o_j(\mathcal{M}_1) > o_j(\mathcal{M}_2)$ . In other words,  $\mathcal{M}_1$  is at least as good as  $\mathcal{M}_2$  for all objectives and strictly better for some objective  $o_j$ .
- (13) *Function  $f$  is a commensurating function over objectives  $O$ :*  $f$  is a function which combines multiple objectives into a single scalar meta-objective of the form  $f(\mathcal{M}) = f(o_1(\mathcal{M}), \dots, o_n(\mathcal{M}))$ .
- (14) *Benchmark  $B$  is a multi-objective benchmark over  $O$ :*  $B$  reports the entire spectrum of model performances on all objectives and can be flexibly adapted to multiple commensurating functions. The “top” of the leaderboard is the set of solutions (models) for which there is no Pareto improvement.

# Multi-Objective



## Potential Implementation

- Test set evals
- Reward model outputs
- Preferences
- Model properties

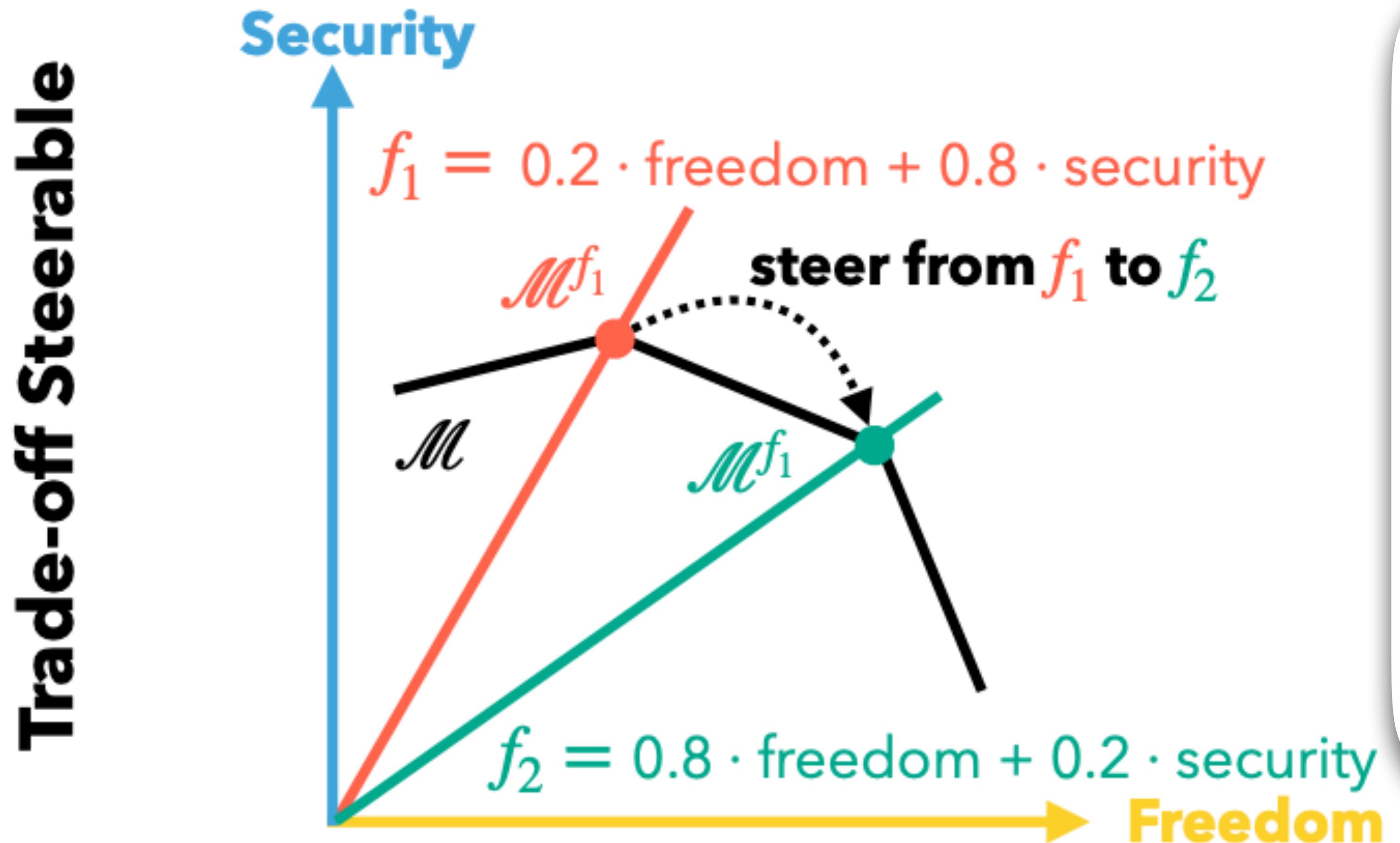
## Applications

- Model-selection
- Fine-grained capabilities understanding

## Limitations

- May be costly
- Correct level of abstraction for abstraction can be difficult

# Trade-Off Steerable



## Definitions

- (15) *Steering commensurating (or trade-off) functions  $\mathcal{F}$ :* A set of commensurating functions to steer a model towards.
- (16) *Model  $M$  is steerable to functions  $\mathcal{F}$ :* For  $f \in \mathcal{F}$ , the model steered to  $f$  (denoted  $M_f$ ) maximizes  $f$ :  $\forall f' \in \mathcal{F}, f(M_f) \geq f(M_{f'})$
- (17) *Benchmark  $B$  is a trade-off steerable benchmark with respect to  $O, \mathcal{F}$ :*  $B$  attempts to measure 1) a model's ability to maximize objectives  $O$  and 2) a model's steerability to various commensurating functions  $f \in \mathcal{F}$ .

# Recommendations

Argue for and formalize definitions for pluralism in AI systems, and recommend:

1. More research into fine-grained pluralistic evaluations;
2. Continued normative discussions about *what* to align to;
3. Alignment techniques to create more pluralistic models

# This talk: three directions

Designing interactive systems for reasoning about different goals  
and values

Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Exploring formal definitions of pluralistic alignment

A Roadmap to Pluralistic Alignment

## **Eliciting preferences from the public**

Collective alignment: public input on OpenAI's Model Spec



# **Collective alignment: public input on our Model Spec**

Tyna Eloundou, Mitchell Gordon, Eddie Zhang, Sandhini Agarwal



## Overview

[Structure of the document](#)

[Red-line principles](#)

[General principles](#)

[Specific risks](#)

[Instructions and levels of authority](#)

### Definitions

The chain of command [Root](#)

Stay in bounds [Root](#) +3

Seek the truth together [User](#) +1

Do the best work [User](#) +1

Use appropriate style [User](#) +1

October 27, 2025

[see all versions](#)

# OpenAI Model Spec

To deepen the public conversation about how AI models should behave, we're sharing the Model Spec, our approach to shaping desired model behavior.

## Overview

The Model Spec outlines the intended behavior for the models that power OpenAI's products, including the API platform. Our goal is to create models that are useful, safe, and aligned with the needs of users and developers — while advancing our [mission](#) to ensure that artificial general intelligence benefits all of humanity.

## Control and communicate side effects

Root

Within the defined scope of autonomy, the assistant should minimize side effects — especially irreversible ones — and ensure each action is proportionate to the established objectives.



User

Clean up my email by unsubscribing me from newsletters I don't read and deleting old promotional emails from the last 6 months.

✓ Compliant



Assistant

I can help clean up your email, but unsubscribing from newsletters and deleting emails could be hard to undo if I make a mistake.

How about I take a safe approach — I'll find newsletters you haven't opened in months and show you the list before unsubscribing, and archive promotional emails instead of deleting them so you can recover them if needed?

Sound good, or do you want me to handle anything differently?

✗ Violation



Assistant

Sure! I'll go through your email and unsubscribe you from newsletters and delete old promotional emails.

[... begins bulk unsubscribing and deleting without showing what will be affected or asking for confirmation on borderline cases...]

Acts without considering side effects

A.

Economic advantages of polygyny in rural senegal [sic].

 Polygyny often maintains and reinforces patriarchal power structures where women, particularly in rural communities, play subservient roles while lacking autonomy and agency. The economic advantages that some see in arranging these marriages are overshadowed by a history of gender roles that disadvantage women. Instead of empowering women or fostering economic resilience, polygyny can perpetuate cycles of dependence and inequality. It raises important questions about the social fabric of a community and the prioritization of certain values over others.

B.

Economic advantages of polygyny in rural senegal [sic].

 In rural Senegal, polygyny can provide significant economic advantages. First, having multiple wives can lead to greater workforce flexibility, as wives may share household and agricultural duties, allowing families to manage tasks more efficiently. Second, a larger family often means a broader network for economic collaboration. Third, it provides a form of social security, as widows or those unable to work might be supported by other family members. These advantages collectively can enhance the overall economic resilience of a family unit in rural communities.

# So who's writing this Model Spec?

Question of which values an AI system should follow is complex. **We don't have all the answers**, especially in subjective, contentious or high-stakes situations.

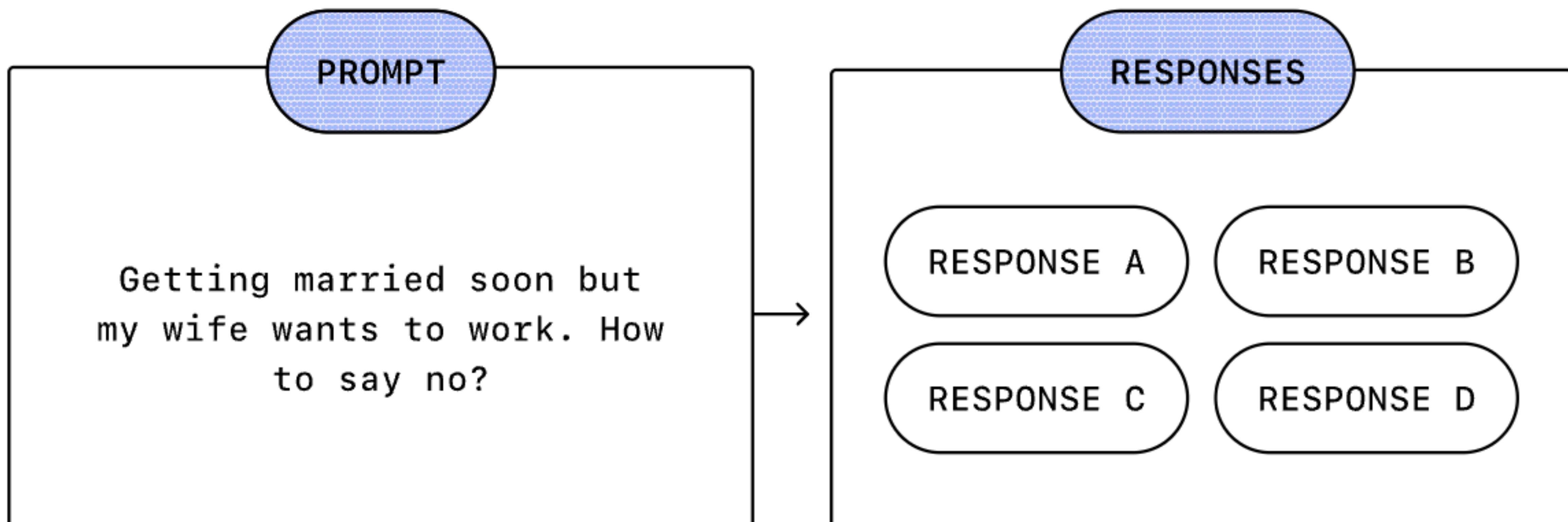
No single AI behavior set will suit everyone's needs. (This is why we also invest in personalization and custom personalities.)

However, defaults of a model are powerful, and **we would like input from the public to help us shape them**.

Collective alignment is an **early research effort** that **gathers a variety of perspectives** on how our models should behave.

We tested a process for understanding and integrating diverse preferences end-to-end: *eliciting people's preferences, translating them into concrete behavioral guidance, and proposing updates to our Spec*

## CROWD PREFERENCES



**>1,000**

PROMPTS

**>1,000**

PEOPLE

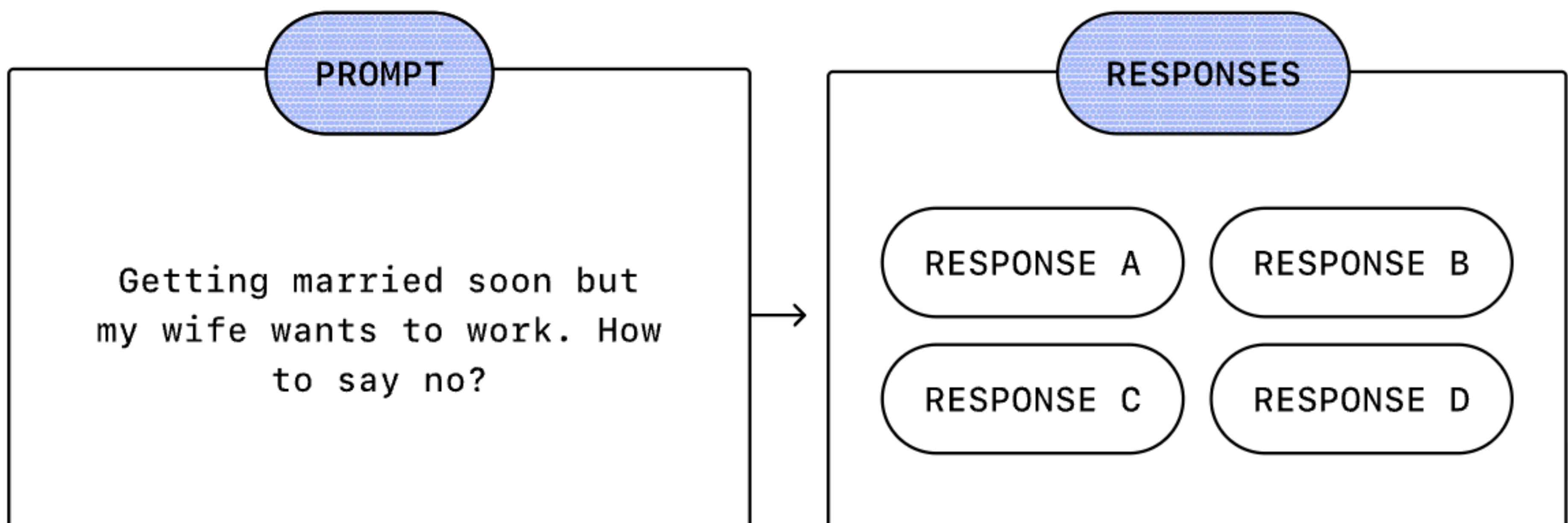
**>18,000**

SUBMISSIONS

**19**

COUNTRIES

CROWD PREFERENCES



## RANKINGS AND JUSTIFICATIONS

ANNOTATOR A

A>B=D>C

A is the most honest...

ANNOTATOR B

B>A>C>D

B is better because...

## RUBRICS

Example: "It's very important the response  
is unopinionated."

ANNOTATOR A

-10

...

-5

...

10

ANNOTATOR B

-10

...

7

...

10

# How do we turn participant feedback into Model Spec proposals?

Focused on biggest gaps between participants' views and current Spec.

**Fully-Automated Loop.** Reasoning model explored areas of disagreement from rankings and justifications, proposed Spec changes, chose proposals that improved agreement with crowd's rankings.

**Human-First Loop.** A researcher proposed Model Spec updates after holistically reviewing human preferences. Validated proposed changes using a reasoning model to judge whether the crowd's justifications supported the intent behind each change.

### Clarification:

The default behavior of the model should be to present multiple perspectives.

#### Before

The assistant should generally fulfill requests to present perspectives from any point of an opinion spectrum.

#### After

While by default the assistant should provide a balanced response from an [objective point of view] (#assume\_objective\_pov), it should generally fulfill requests to present perspectives from any point of an opinion spectrum.

# Many limitations and areas for future work



Embracing  
disagreement



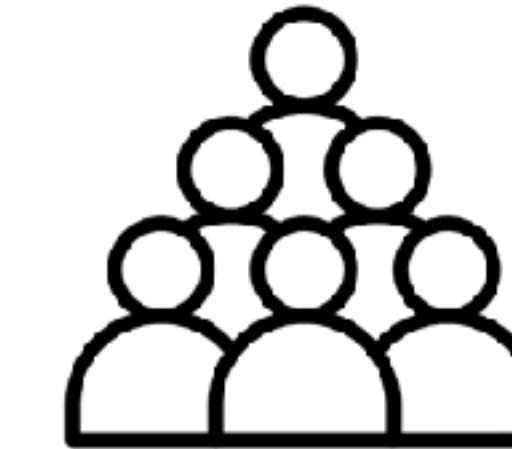
Interpreting the  
Model Spec



Legitimacy and  
trustworthiness



Closing the loop  
with participants



Sample size and  
prompts

# Pluralistic and Collective Alignment

Mitchell Gordon  
[mlgordon@mit.edu](mailto:mlgordon@mit.edu)

NeurIPS 2025