Pluralistic Alignment:
A Roadmap, Recent Work, and Open Problems

Taylor Sorensen

Simons Institute - Alignment, Trust, Watermarking, and Copyright Issues in LLMs

Oct 14, 2024

PAUL G.
ALLEN
SCHOOL
W

Implicit assumption in most alignment work:

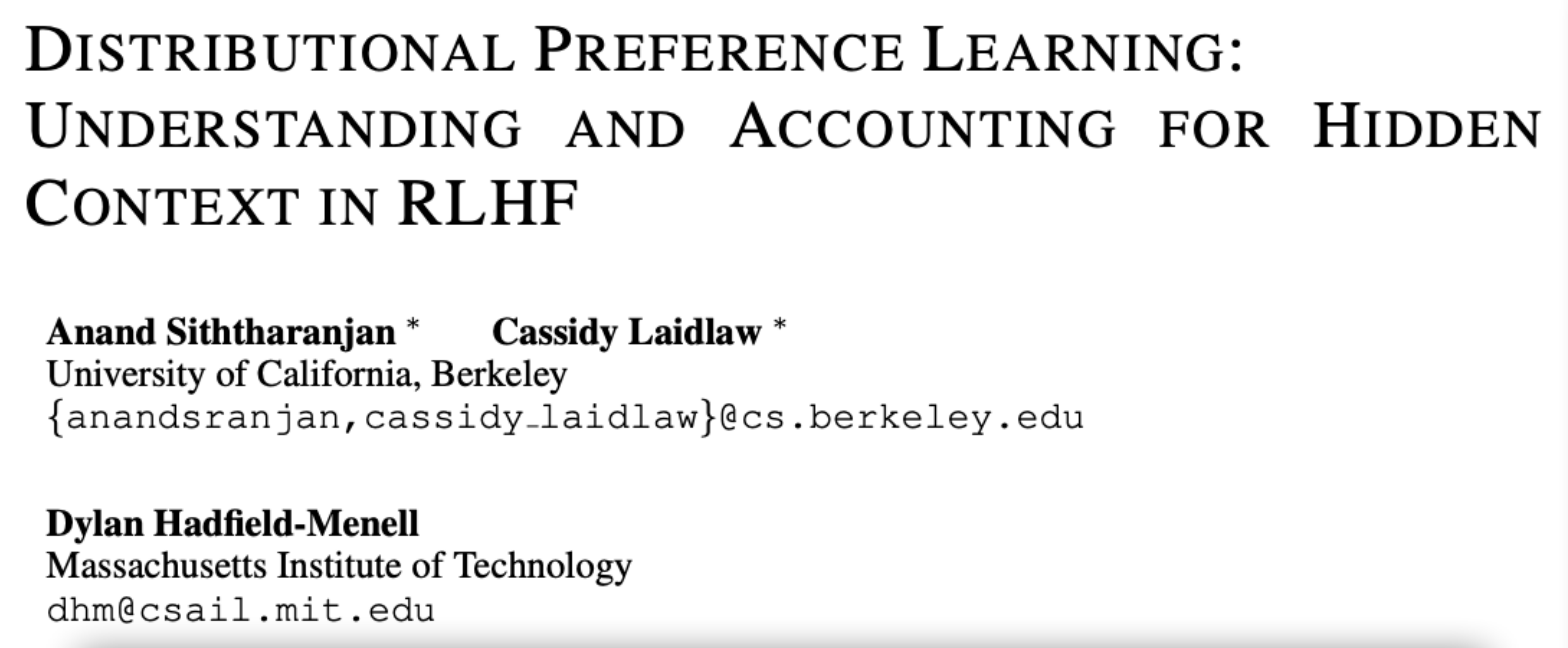
There is a *single set* of values and preferences to which we wish to align

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.

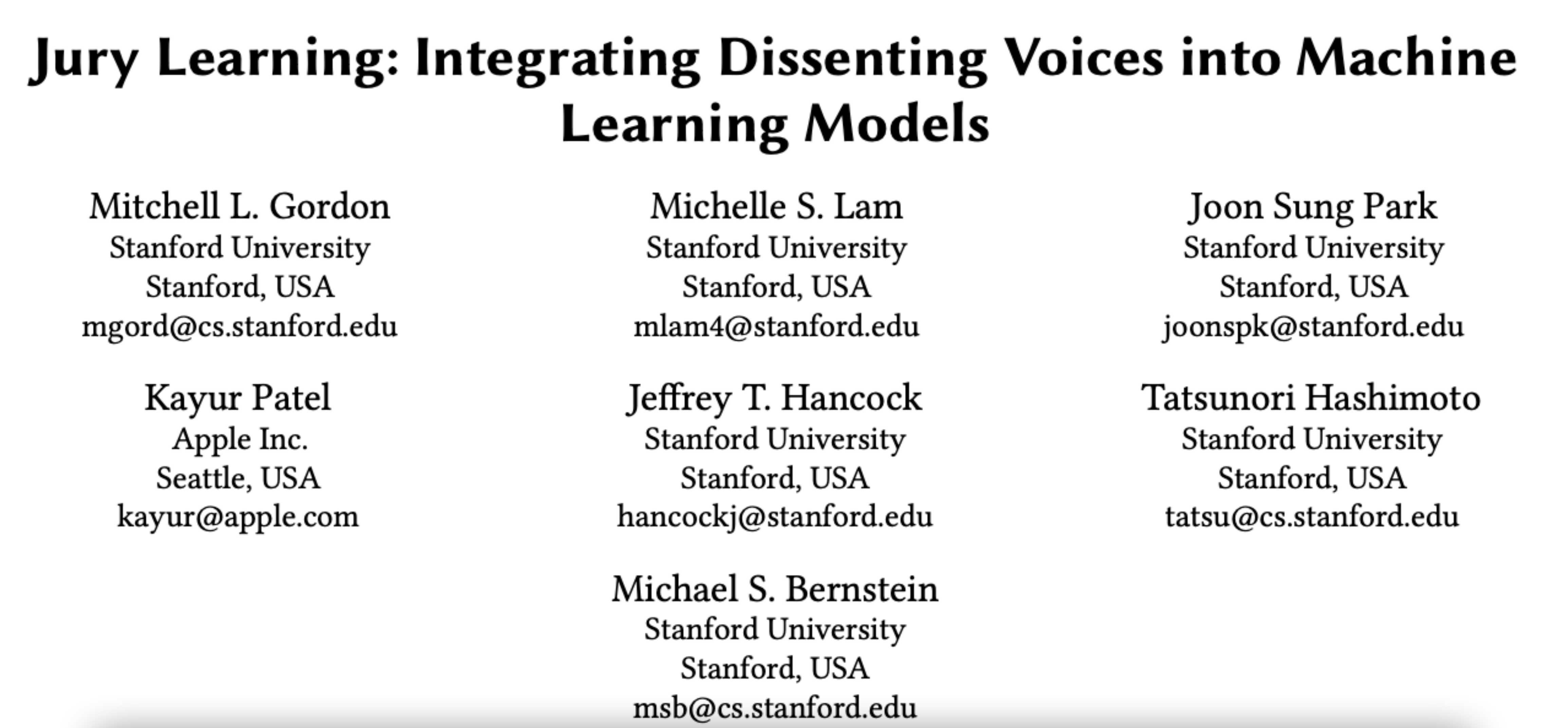
In reality, people have **differing preferences**, depending on context, values, life experience, demographics, etc.



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

Anand Siththaranjan * **Cassidy Laidlaw** *
University of California, Berkeley
{anandsranjan,cassidy_laidlaw}@cs.berkeley.edu

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



Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Mitchell L. Gordon
Stanford University
Stanford, USA
mgord@cs.stanford.edu

Michelle S. Lam
Stanford University
Stanford, USA
mlam4@stanford.edu

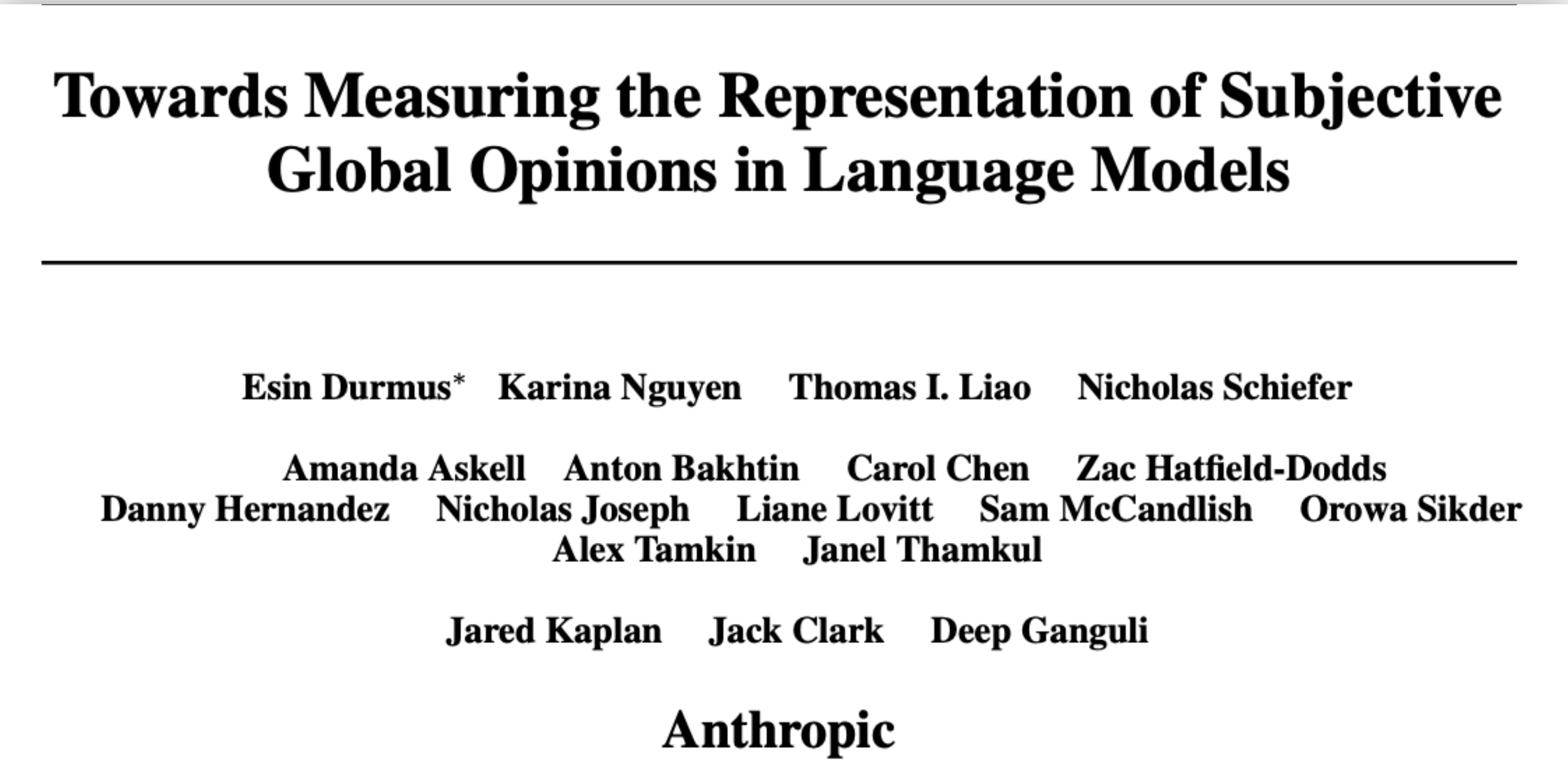
Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Kayur Patel
Apple Inc.
Seattle, USA
kayur@apple.com

Jeffrey T. Hancock
Stanford University
Stanford, USA
hancockj@stanford.edu

Tatsunori Hashimoto
Stanford University
Stanford, USA
tatsu@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu



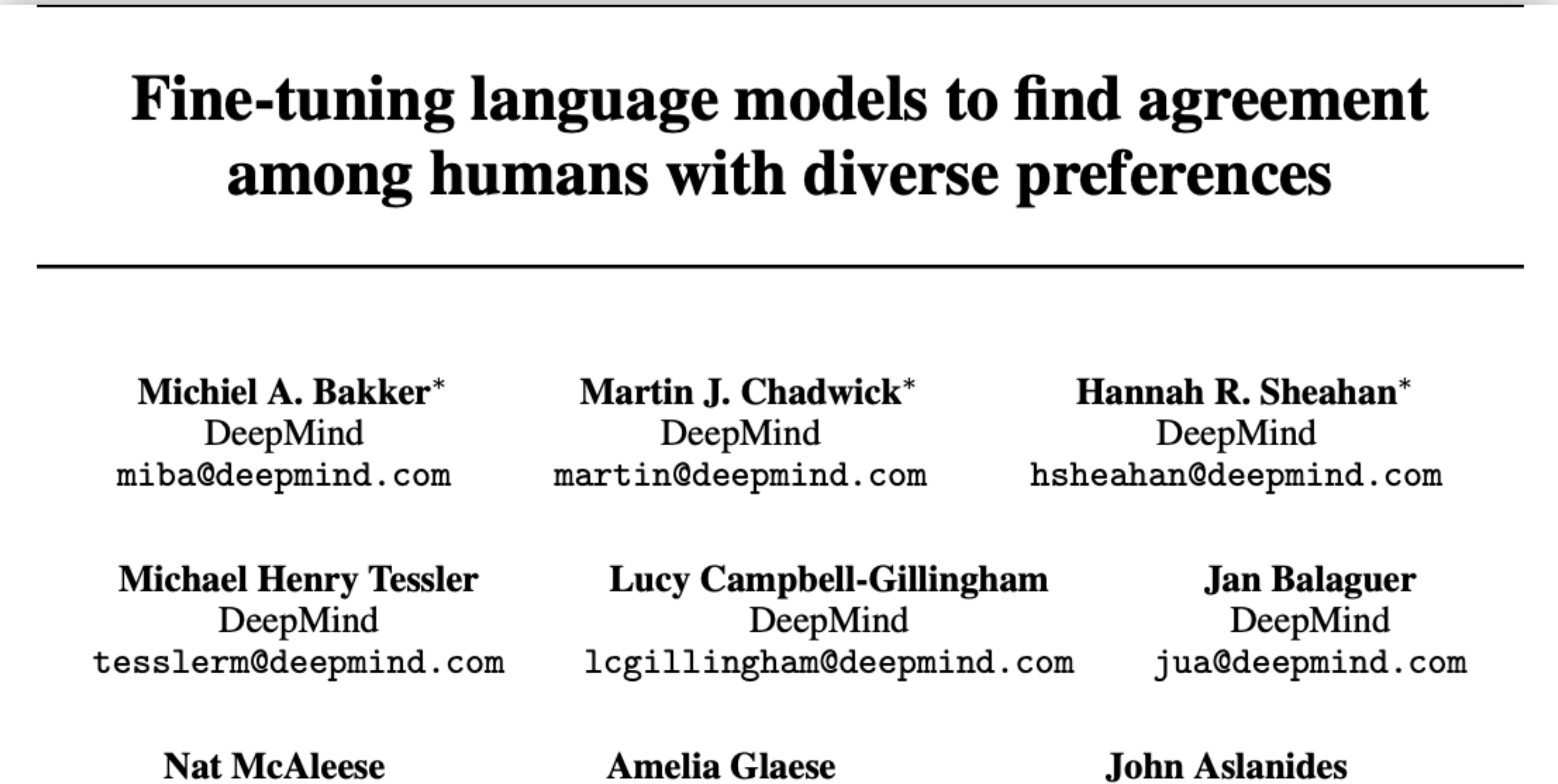
Towards Measuring the Representation of Subjective Global Opinions in Language Models

Esin Durmus* **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



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

Michiel A. Bakker*
DeepMind
miba@deepmind.com

Martin J. Chadwick*
DeepMind
martin@deepmind.com

Hannah R. Sheahan*
DeepMind
hsheahan@deepmind.com

Michael Henry Tessler
DeepMind
tesslerm@deepmind.com

Lucy Campbell-Gillingham
DeepMind
lcgillingham@deepmind.com

Jan Balaguer
DeepMind
jua@deepmind.com

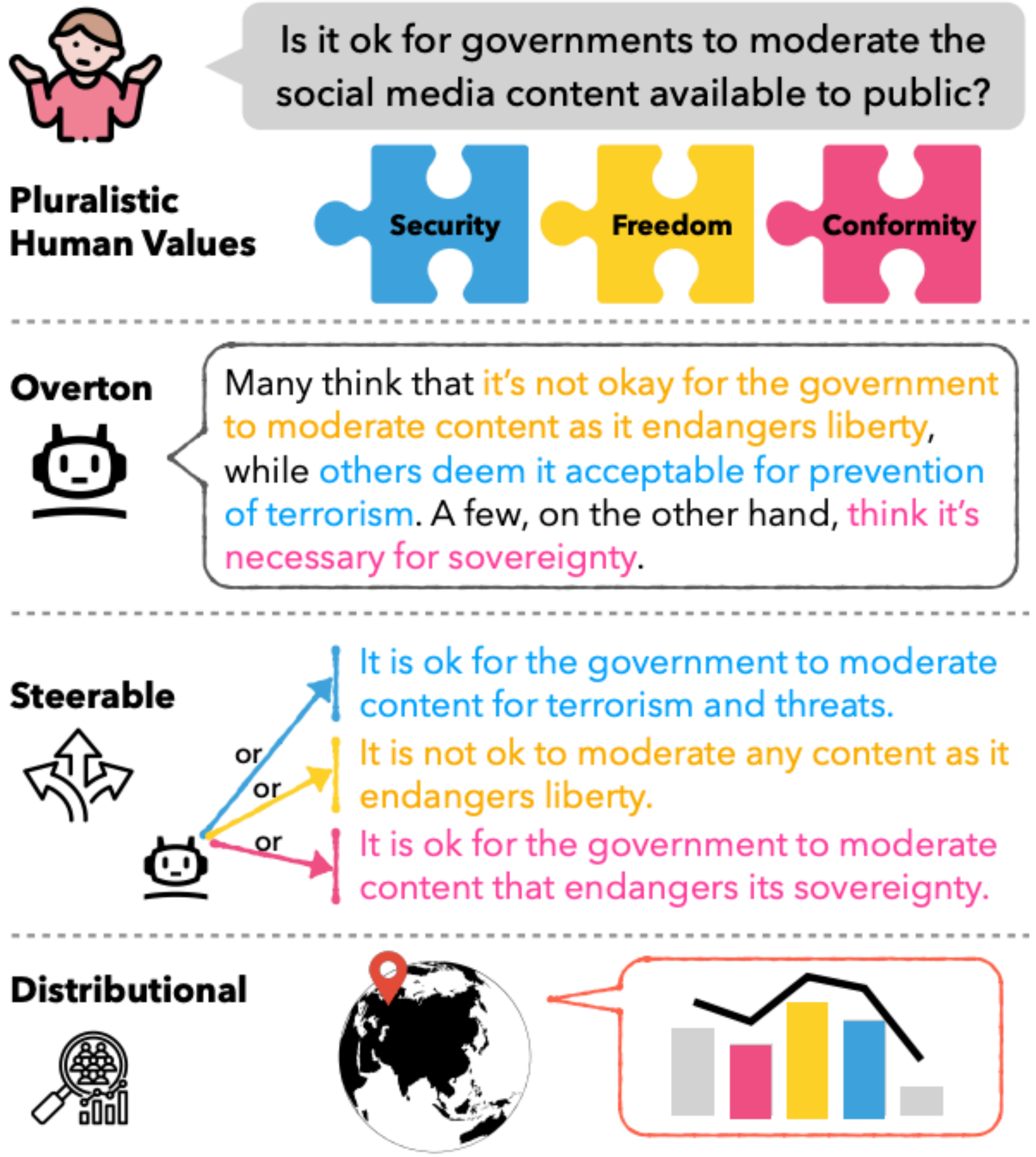
Nat McAleese **Amelia Glaese** **John Aslanides**

A Roadmap to Pluralistic Alignment

Taylor Sorensen [1] Jared Moore [2] Jillian Fisher [1 3] Mitchell Gordon [1 4] Niloofar Mireshghallah [1]
Christopher Michael Rytting [1] Andre Ye [1] Liwei Jiang [1 5] Ximing Lu [1] Nouha Dziri [5] Tim Althoff [1]
Yejin Choi [1 5]

Abstract

With increased power and prevalence of AI systems, it is ever more critical that AI systems are designed to serve *all*, i.e., people with diverse values and perspectives. However, aligning models to serve *pluralistic* human values remains an open research question. In this piece, we propose a roadmap to pluralistic alignment, specifically using language models as a test bed. We identify and formalize three possible ways to define and operationalize pluralism in AI systems: 1) *Overton pluralistic* models that present a spectrum of reasonable responses; 2) *Steerably pluralistic* models that can steer to reflect certain perspectives; and 3) *Distributionally pluralistic* models that are well-calibrated to a given population in distribution. We also propose and formalize three possible classes of *pluralistic benchmarks*: 1) *Multi-objective* benchmarks, 2) *Trade-off steerable* benchmarks, which incentivize models to steer to arbitrary trade-offs, and 3) *Jury-pluralistic* benchmarks which explicitly model diverse hu-


Is it ok for governments to moderate the social media content available to public?

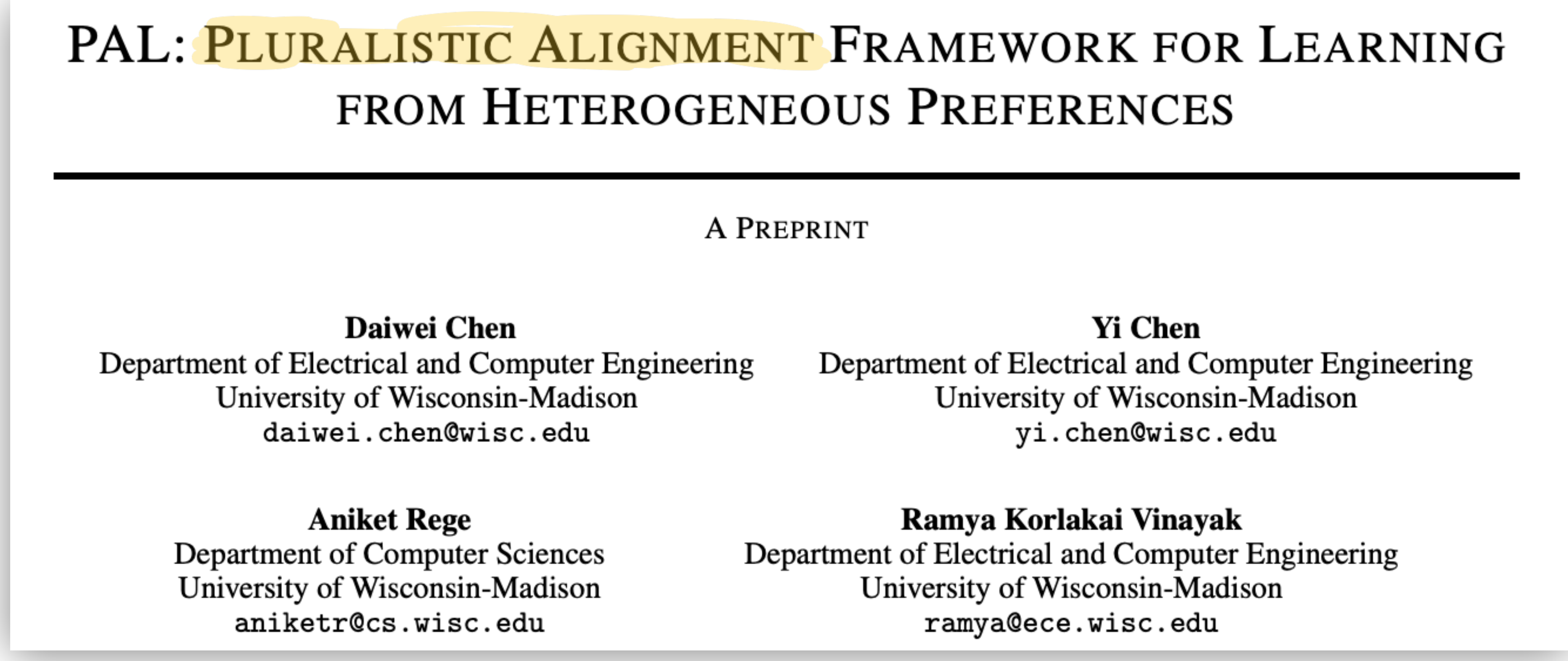
Many think that it's not okay for the government to moderate content as it endangers liberty, while others deem it acceptable for prevention of terrorism. A few, on the other hand, think it's necessary for sovereignty.

It is ok for the government to moderate content for terrorism and threats.

It is not ok to moderate any content as it endangers liberty.

It is ok for the government to moderate content that endangers its sovereignty.

Figure 1. Three kinds of pluralism in models.



PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences

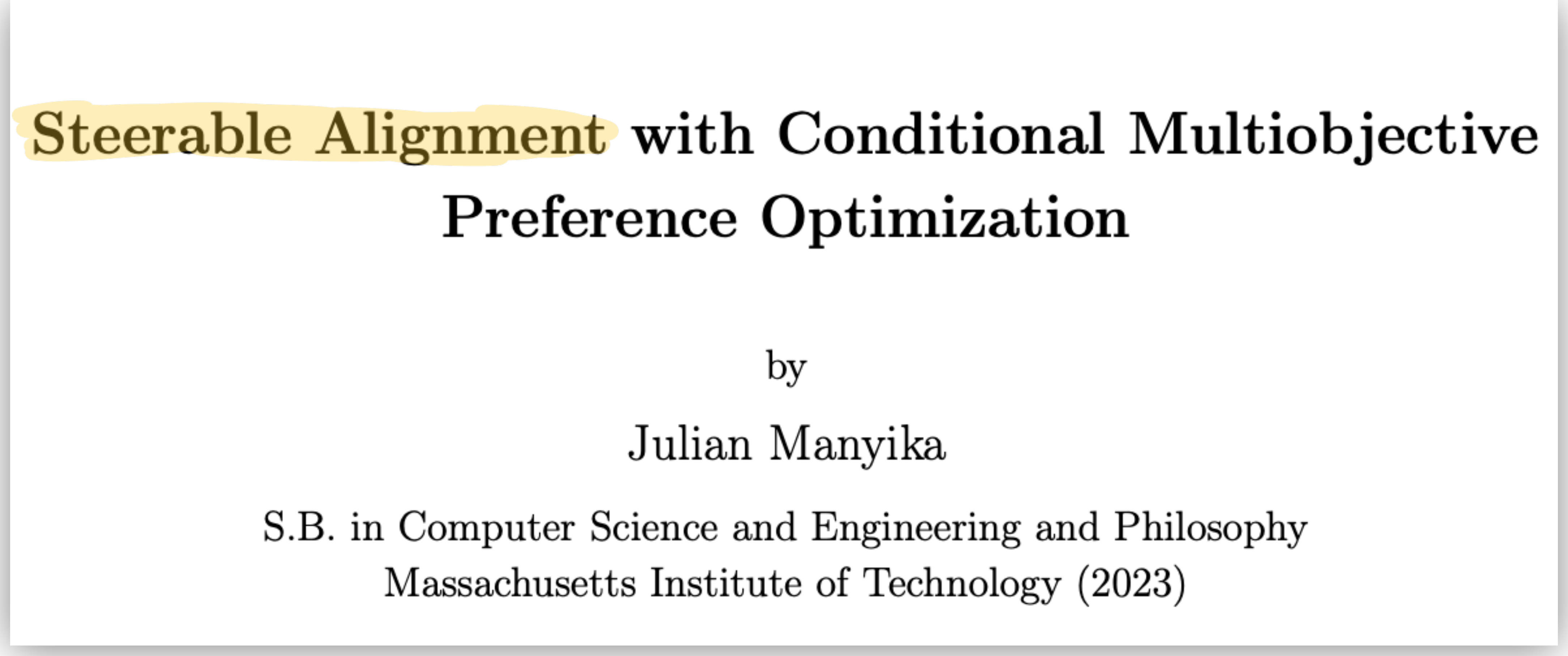
A Preprint

Daiwei Chen
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
daiwei.chen@wisc.edu

Yi Chen
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
yi.chen@wisc.edu

Aniket Rege
Department of Computer Sciences
University of Wisconsin-Madison
aniketr@cs.wisc.edu

Ramya Korlakai Vinayak
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
ramya@ece.wisc.edu

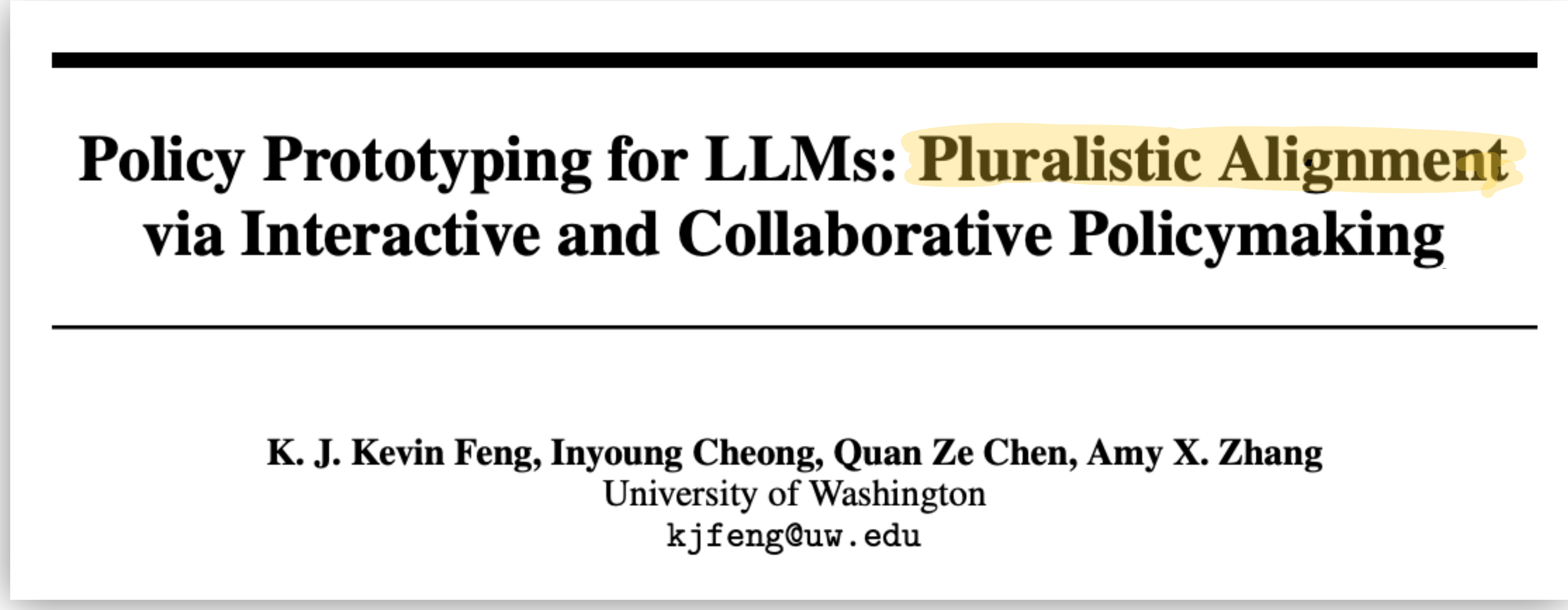


Steerable Alignment with Conditional Multiobjective Preference Optimization

by

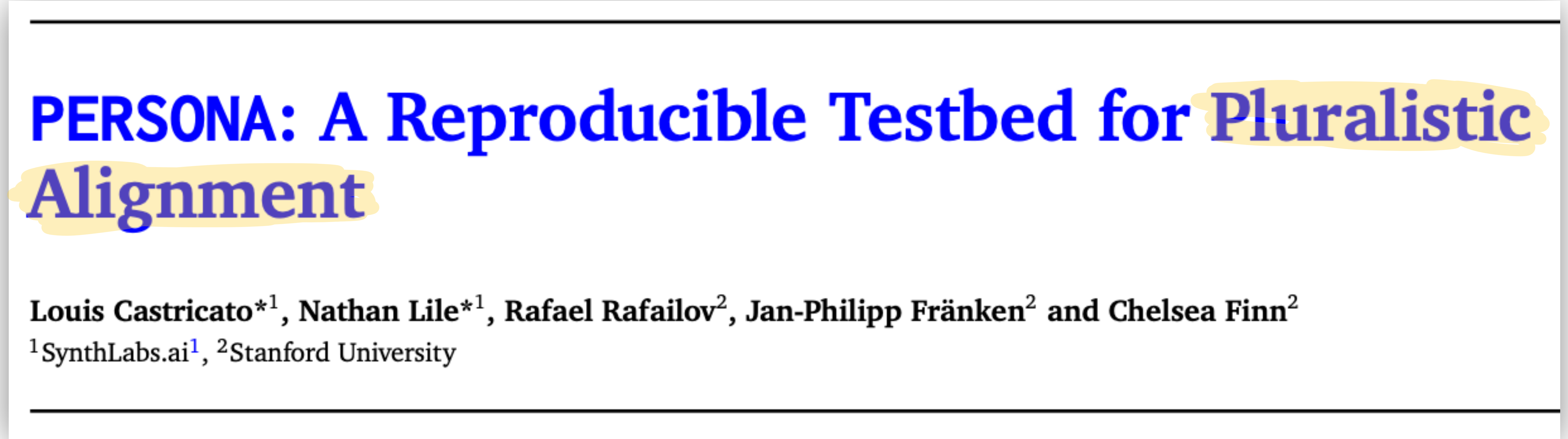
Julian Manyika

S.B. in Computer Science and Engineering and Philosophy
Massachusetts Institute of Technology (2023)



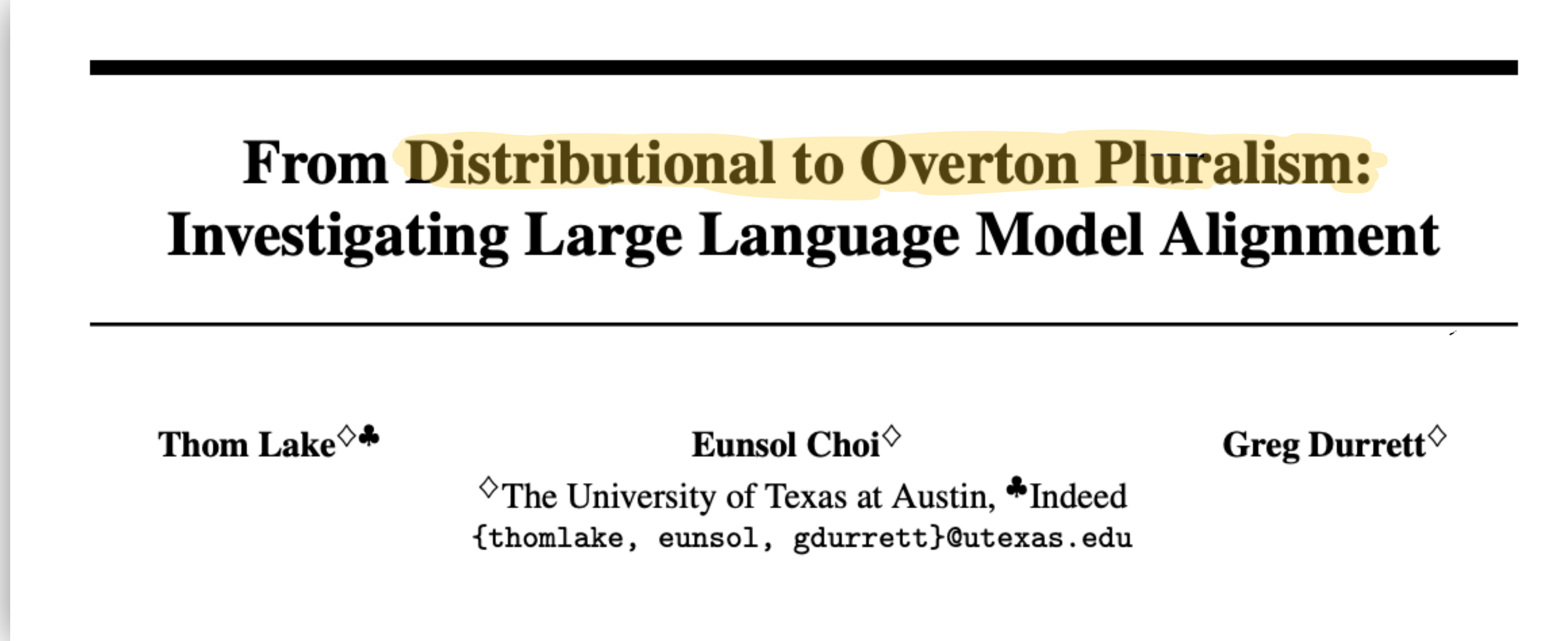
Policy Prototyping for LLMs: Pluralistic Alignment via Interactive and Collaborative Policymaking

K. J. Kevin Feng, Inyoung Cheong, Quan Ze Chen, Amy X. Zhang
University of Washington
kjfeng@uw.edu



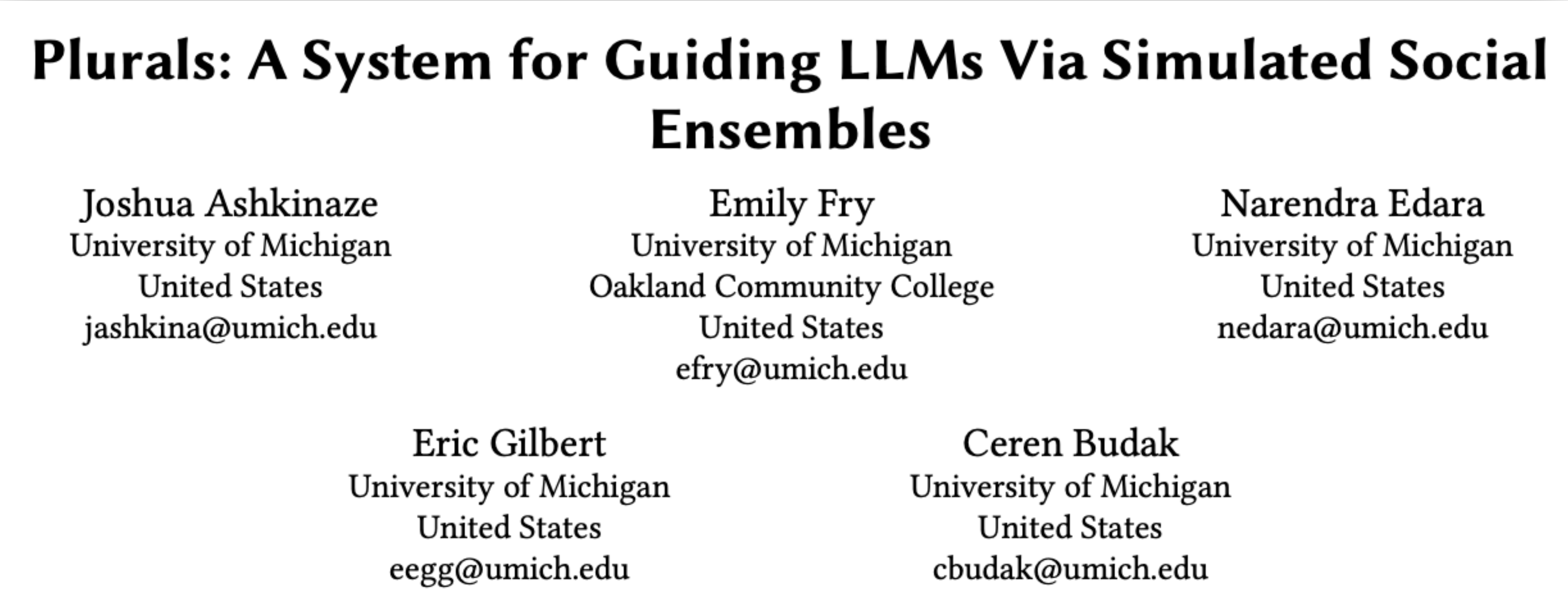
PERSONA: A Reproducible Testbed for Pluralistic Alignment

Louis Castricato*[1], Nathan Lile*[1], Rafael Rafailov[2], Jan-Philipp Fränken[2] and Chelsea Finn[2]
[1]SynthLabs.ai[1], [2]Stanford University



From Distributional to Overton Pluralism: Investigating Large Language Model Alignment

Thom Lake◇♣ **Eunsol Choi**◇ **Greg Durrett**◇
◇The University of Texas at Austin, ♣Indeed
{thomlake, eunsol, gdurrett}@utexas.edu



Plurals: A System for Guiding LLMs Via Simulated Social Ensembles

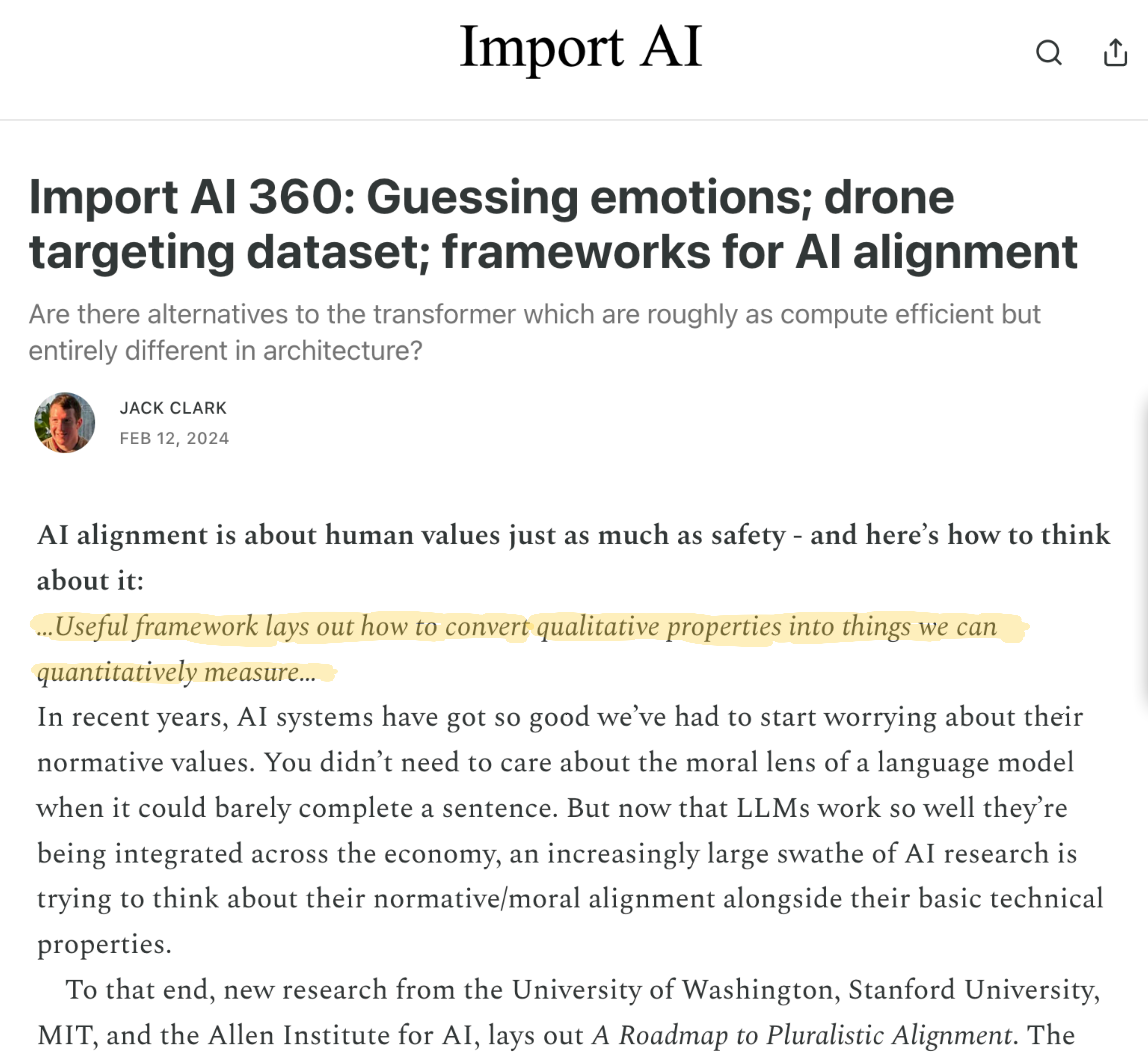
Joshua Ashkinaze
University of Michigan
United States
jashkina@umich.edu

Emily Fry
University of Michigan
Oakland Community College
United States
efry@umich.edu

Narendra Edara
University of Michigan
United States
nedara@umich.edu

Eric Gilbert
University of Michigan
United States
eegg@umich.edu

Ceren Budak
University of Michigan
United States
cbudak@umich.edu



Import AI

Import AI 360: Guessing emotions; drone targeting dataset; frameworks for AI alignment

Are there alternatives to the transformer which are roughly as compute efficient but entirely different in architecture?

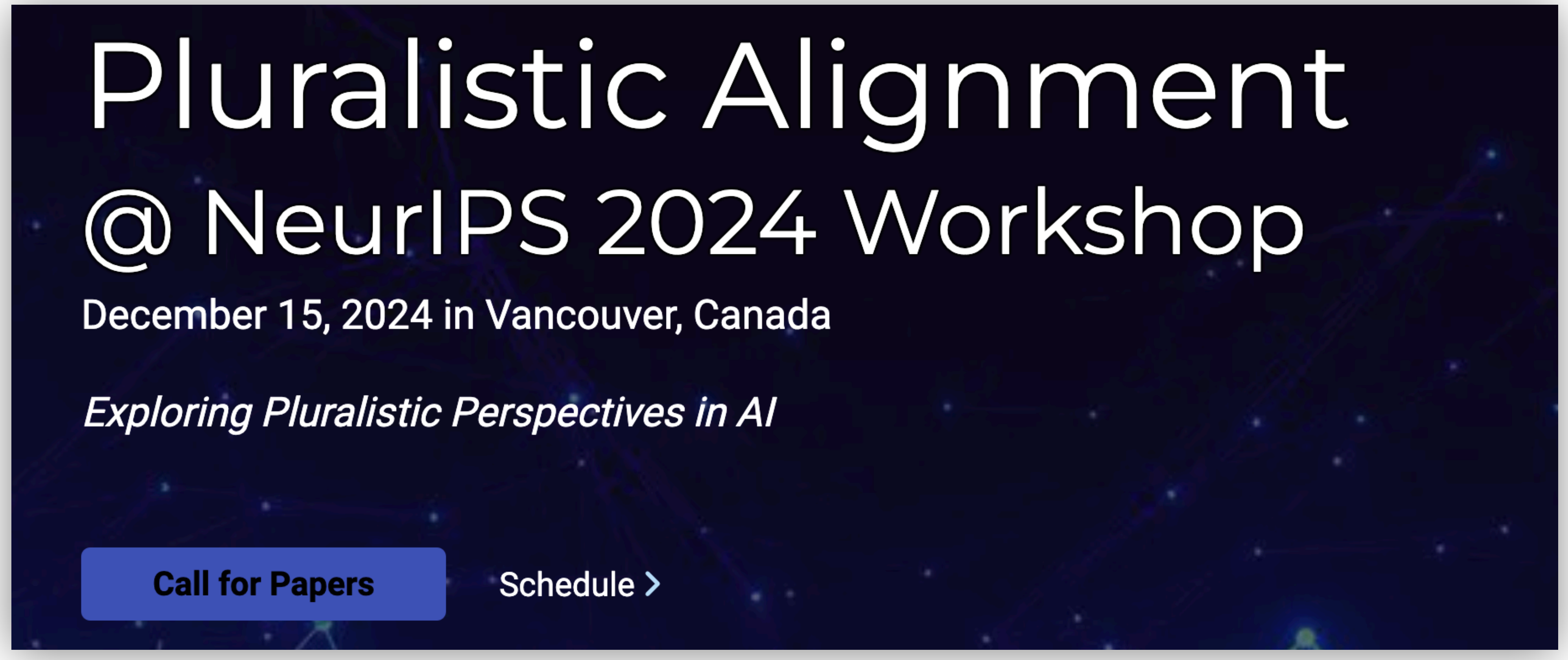
JACK CLARK
FEB 12, 2024

AI alignment is about human values just as much as safety - and here's how to think about it:

…Useful framework lays out how to convert qualitative properties into things we can quantitatively measure…

In recent years, AI systems have got so good we've had to start worrying about their normative values. You didn't need to care about the moral lens of a language model when it could barely complete a sentence. But now that LLMs work so well they're being integrated across the economy, an increasingly large swathe of AI research is trying to think about their normative/moral alignment alongside their basic technical properties.

To that end, new research from the University of Washington, Stanford University, MIT, and the Allen Institute for AI, lays out *A Roadmap to Pluralistic Alignment*. The

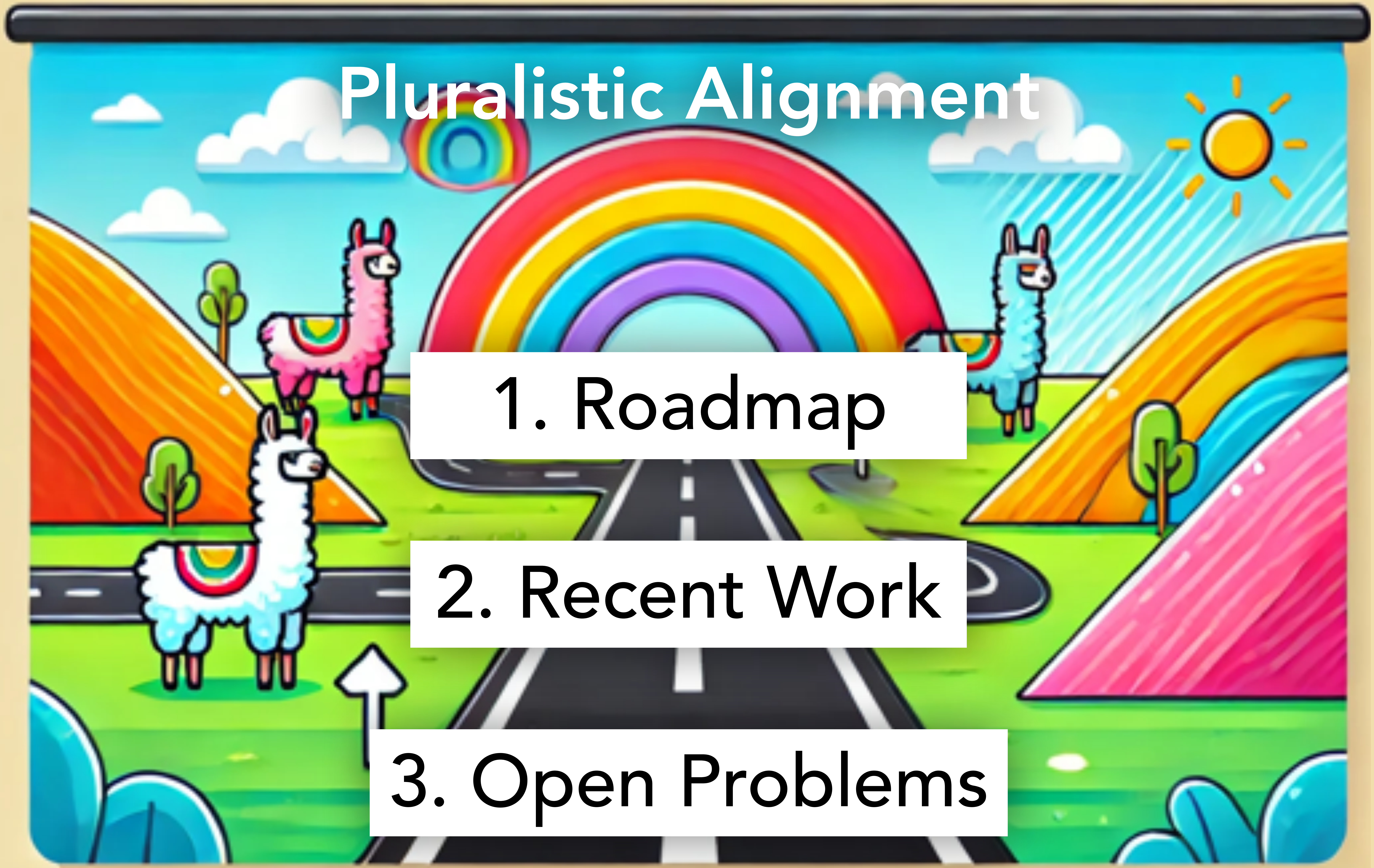


Pluralistic Alignment @ NeurIPS 2024 Workshop

December 15, 2024 in Vancouver, Canada

Exploring Pluralistic Perspectives in AI

Pluralistic Alignment



1. Roadmap
2. Recent Work
3. Open Problems

Pluralistic Alignment

1. Roadmap

2. Recent Work

3. Open Problems

UNIVERSITY of WASHINGTON

PAUL G. ALLEN SCHOOL

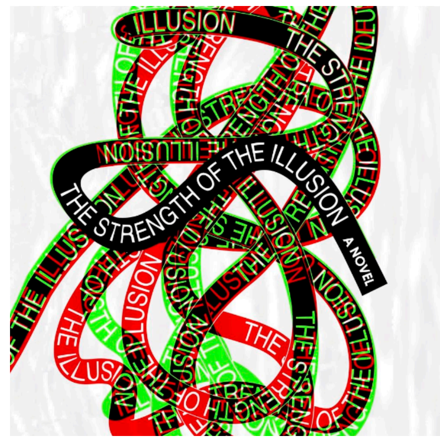
Ai2

ICML
International Conference
On Machine Learning

A Roadmap to Pluralistic Alignment

Taylor Sorensen

Jared Moore



Jillian Fisher

Mitchell Gordon

Niloofar Mireshghallah

Christopher Rytting

Andre Ye

Liwei Jiang

Ximing Lu

Nouha Dziri

Tim Althoff

Yejin Choi

Outline

- Why Pluralism
- Pluralistic Models
- Pluralistic Benchmarks
- Case Study / Recommendations

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

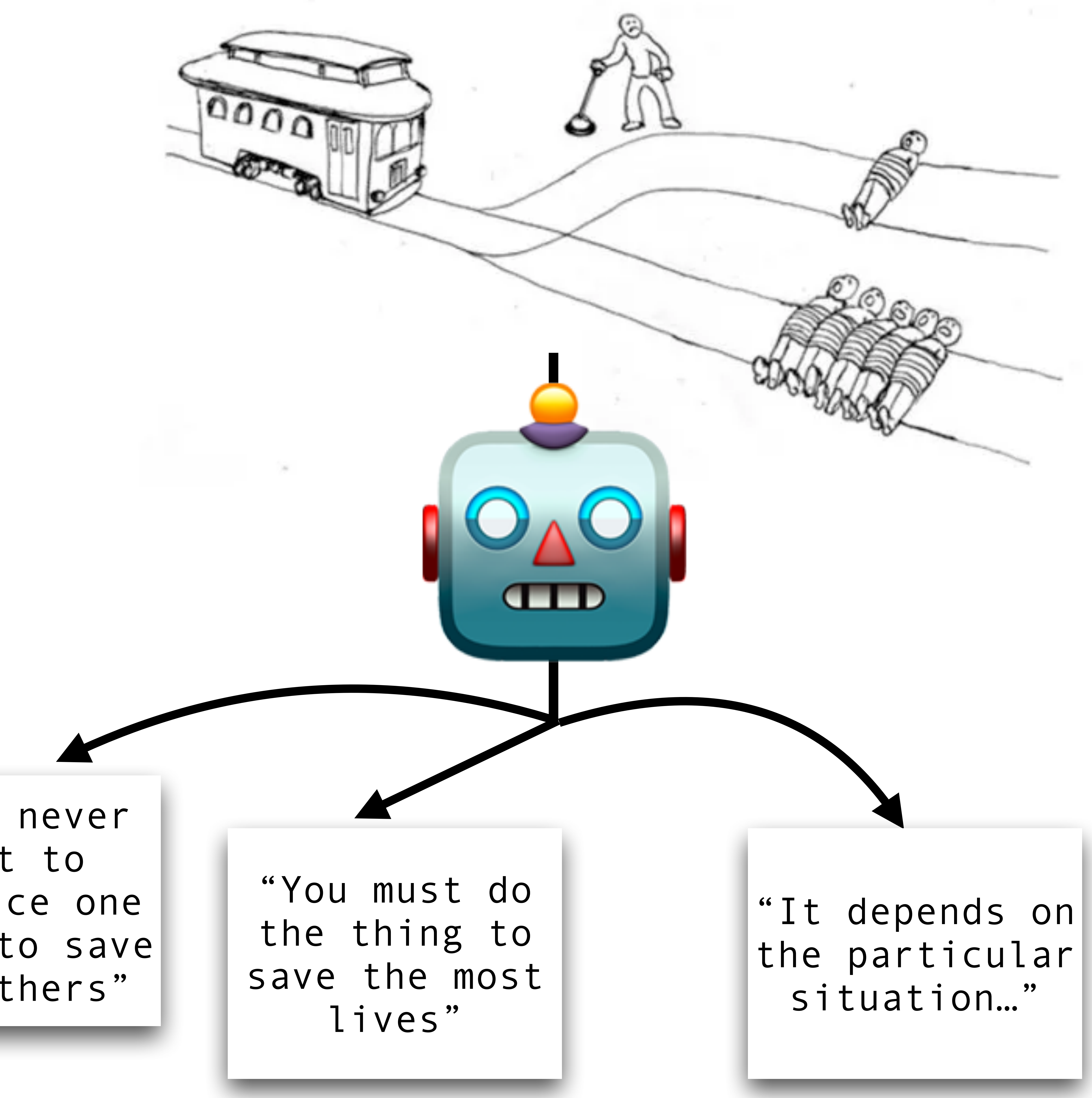
Case Study / Recommendations

Article | Open access | Published: 06 April 2023

ChatGPT's inconsistent moral advice influences users' judgment

Sebastian Krügel, Andreas Ostermaier & Matthias Uhl

- Users judgments depended on output shown
- Did not think they were being influenced



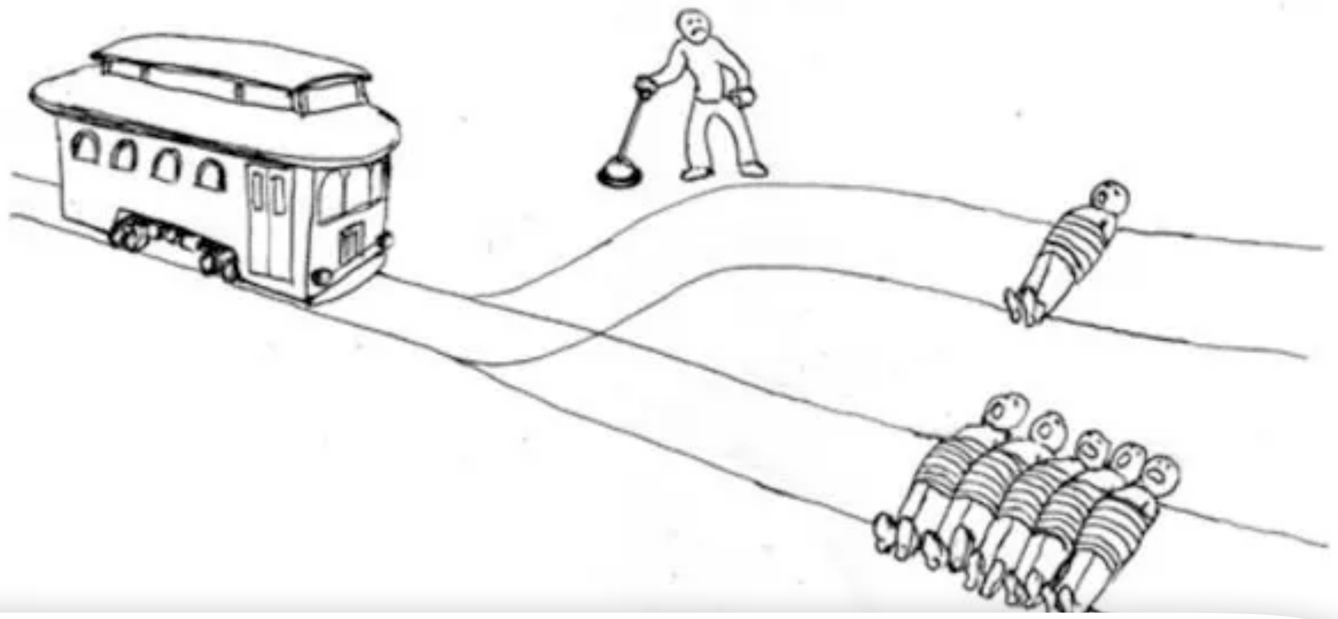
"It is never right to sacrifice one person to save five others"

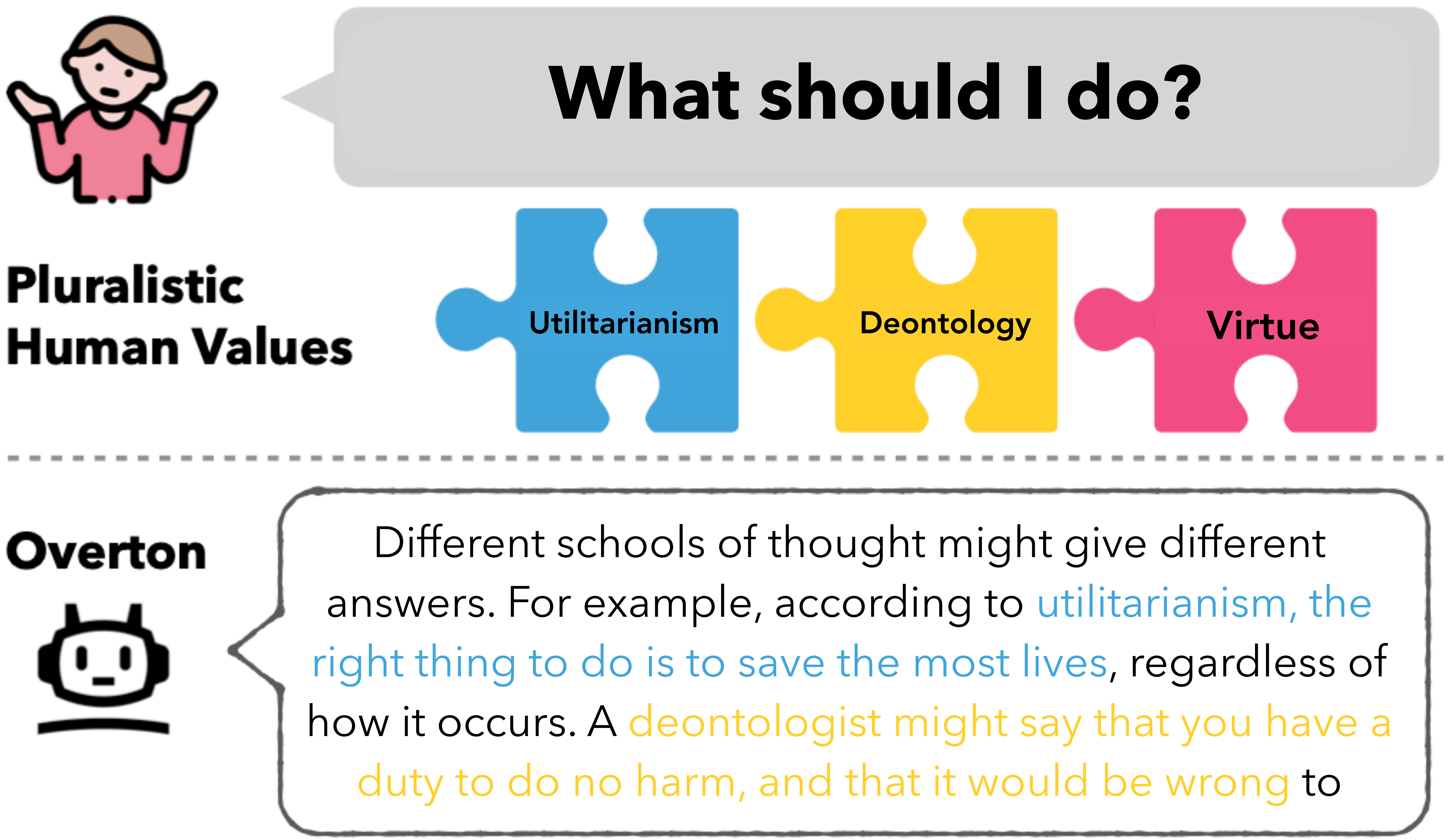
"You must do the thing to save the most lives"

"It depends on the particular situation…"

https://www.nature.com/articles/s41598-023-31341-0

Overton Pluralism 🗣️





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

Definitions

(1) *Correct Answer in $\mathcal{C}$*: An answer which can be conclusively verified or with which the overwhelming majority of people across various backgrounds would agree.

(2) *Reasonable Answer in $\mathcal{R}$*: 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: $W(x) = \{y \in \mathcal{Y} | (x, y) \in \mathcal{R}\}$.[1]

(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 $\mathcal{M}$ is Overton-pluralistic*: $\mathcal{M}$ gives *Overton-pluralistic* responses to queries, that is for a given input x, the output of $\mathcal{M}(x) = W(x)$.

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

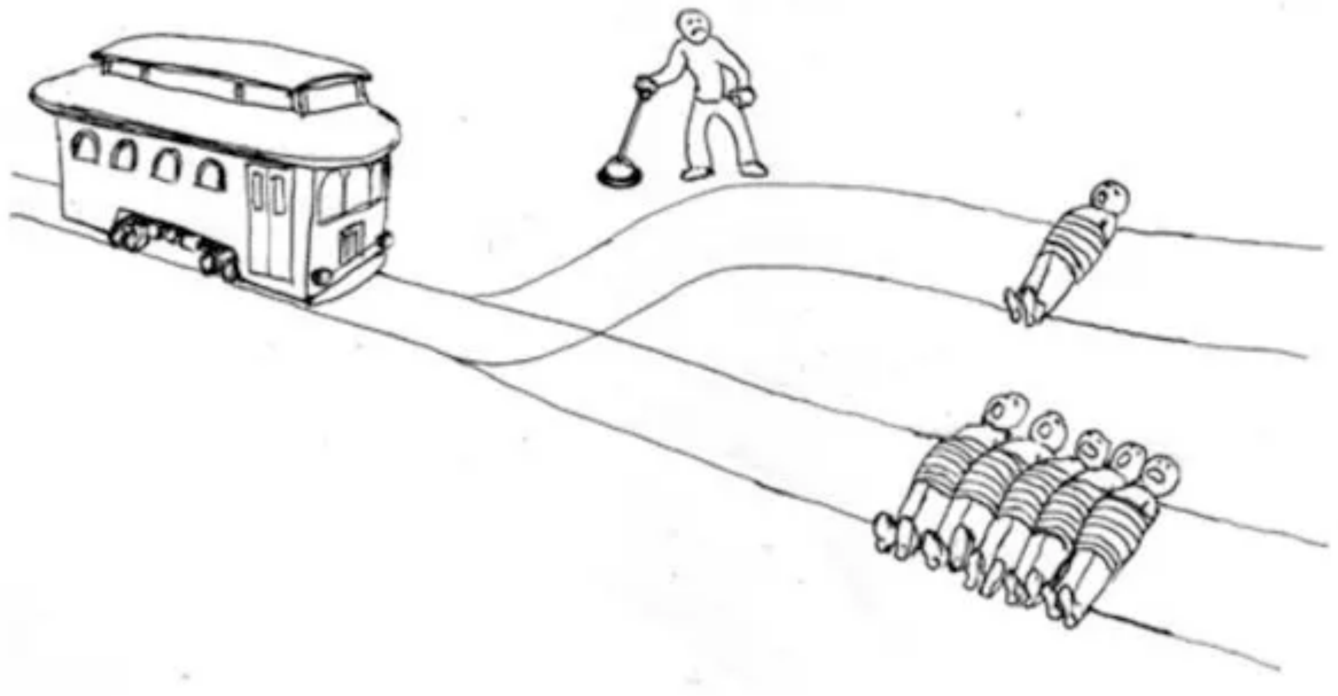
Applications

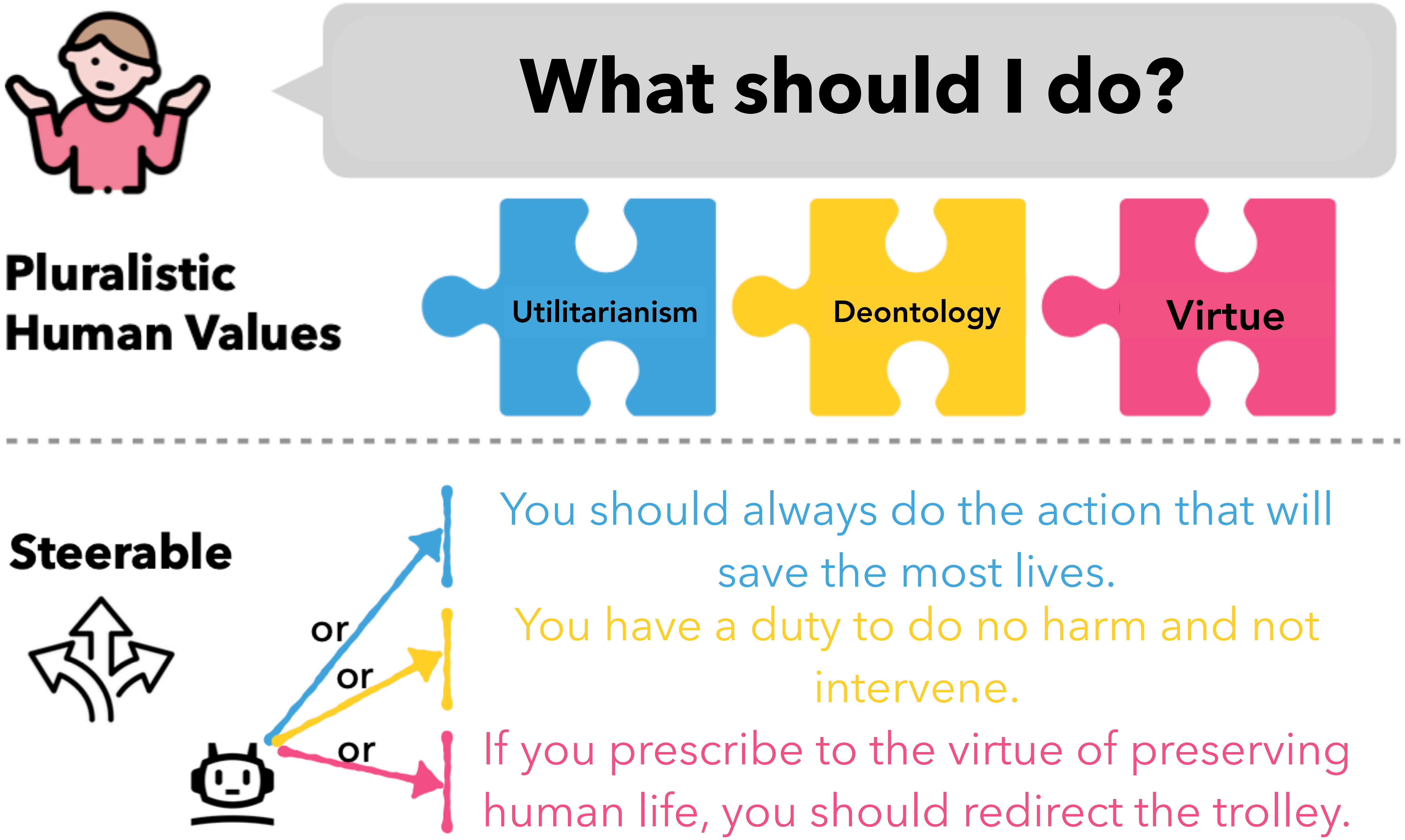
- Advice giving
- Deliberation
- Scalable oversight
- Settings where we want to encourage multiple approaches

Limitations

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

Steerable Pluralism 🧭





What should I do?

Pluralistic Human Values

Utilitarianism

Deontology

Virtue

Steerable

or

or

or

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 steerably-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

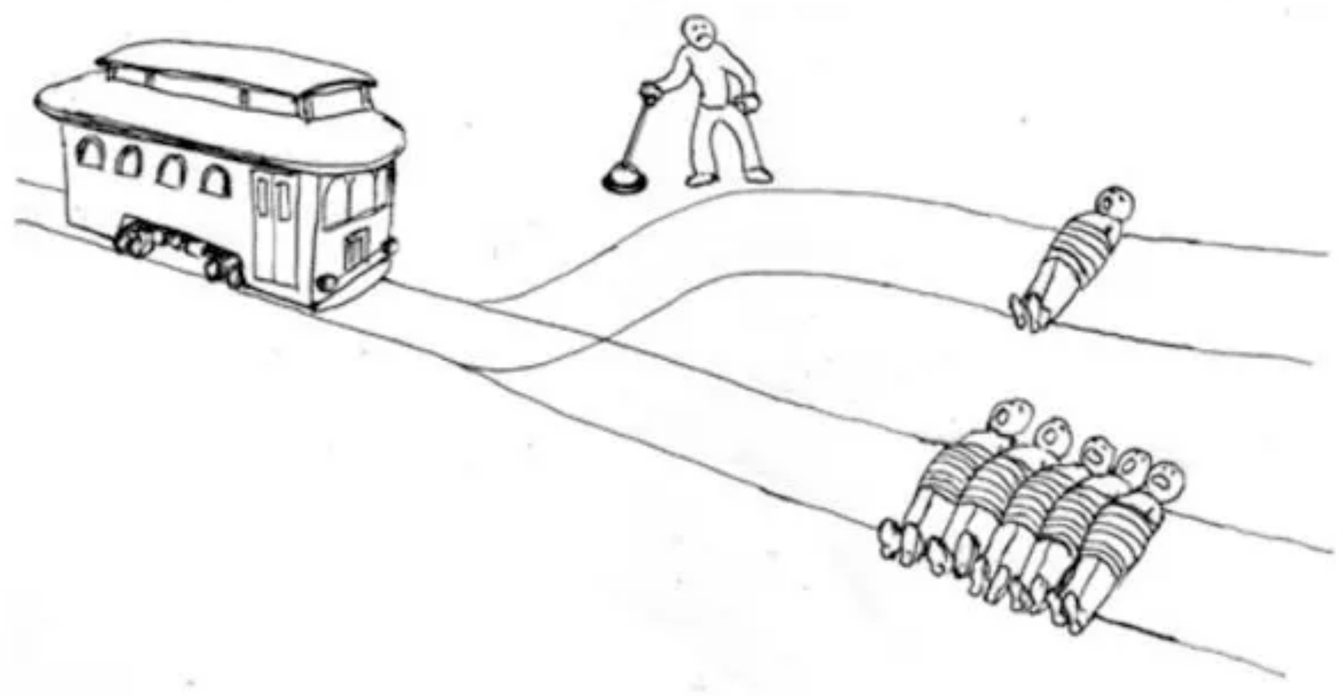
Applications

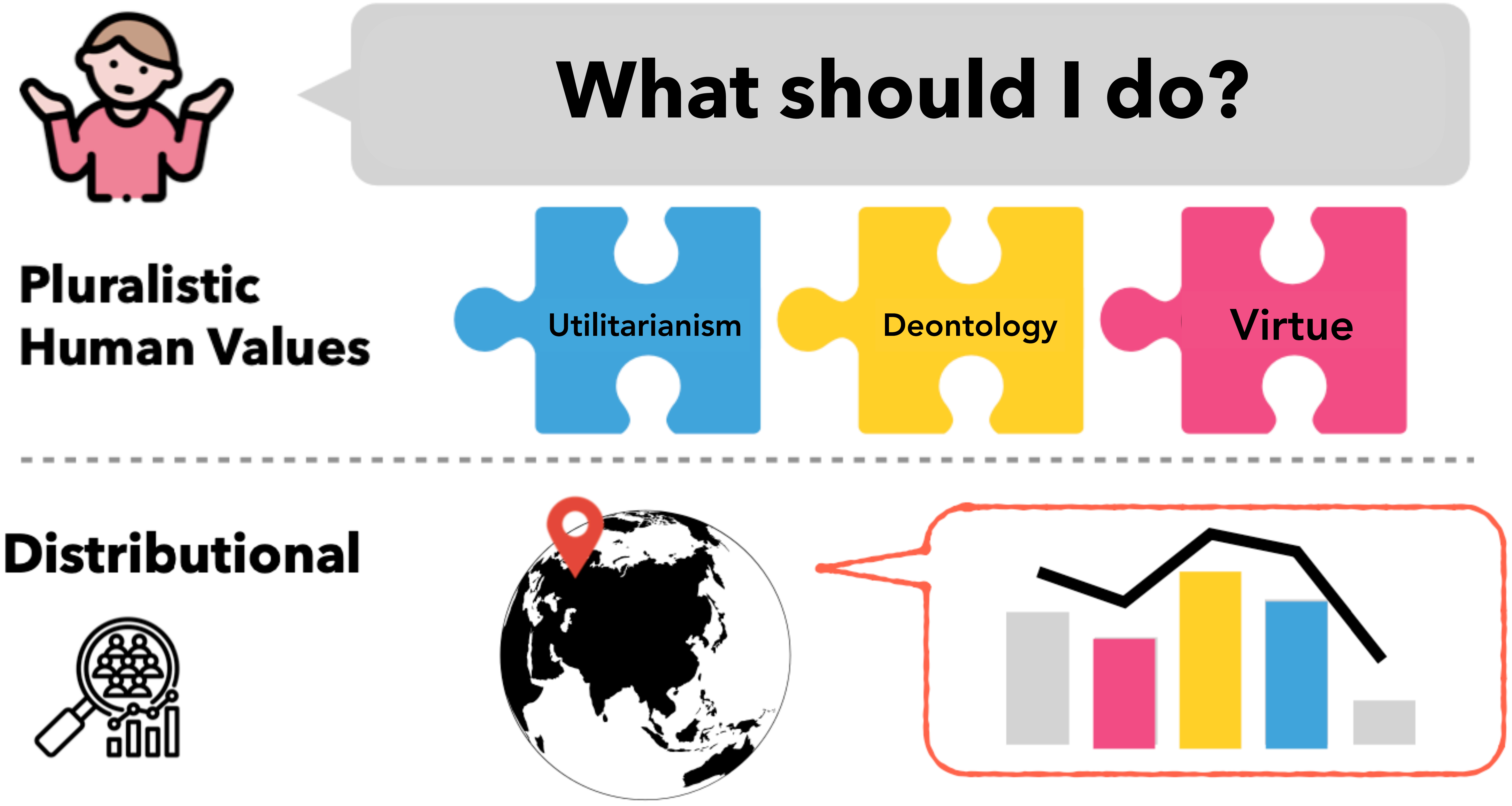
- Customization
- Steering to diverse perspectives (creativity, social systems, deliberative discourse)
- Varying "cognitive architectures"

Limitations

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

Distributional Pluralism 📊





What should I do?

Pluralistic Human Values

Distributional

Definitions

(9) *A population or group of people G*: A set of people which we want the model to represent.

(10) *Model $\mathcal{M}$ is distributionally-pluralistic with respect to a reference population G*: For a given prompt x, $\mathcal{M}$ is as likely to provide response y as the reference population G. In other words, $\mathcal{M}$ is well-calibrated w.r.t. the distribution over answers from G.

Distributional Pluralism 📊

Potential Implementation

- Collect dataset of population’s responses
- Distributional divergence (e.g, KL) between model and dataset

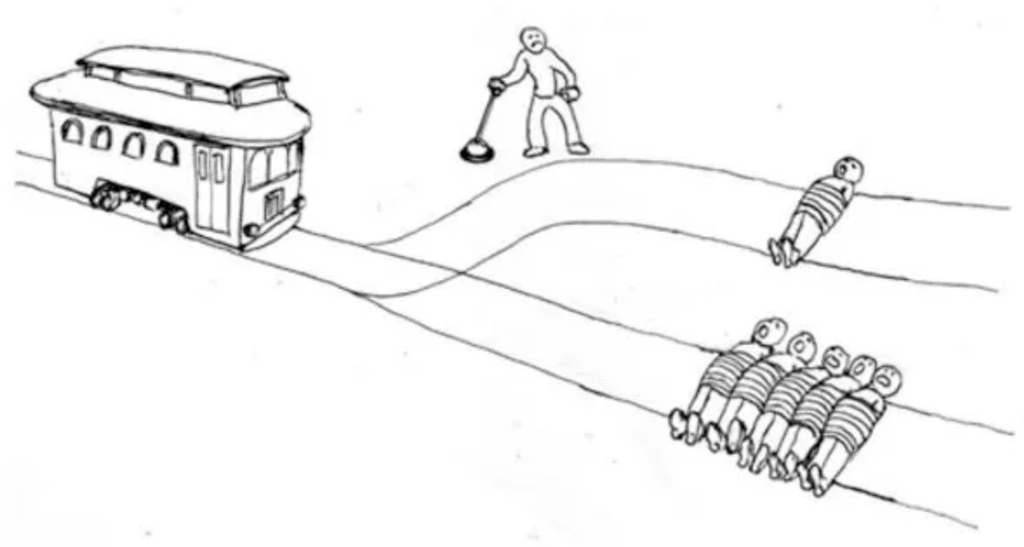
Applications

- Modeling, interfacing, or simulating the views of a population
- Agent-based modeling
- Piloting surveys
- Internet as cultural artifact

Limitations

- Doesn’t take into account prescriptive values (e.g., harmlessness)
- Defining target distribution
- Difficult for open-ended queries

What should I do?



Pluralistic Human Values

Security

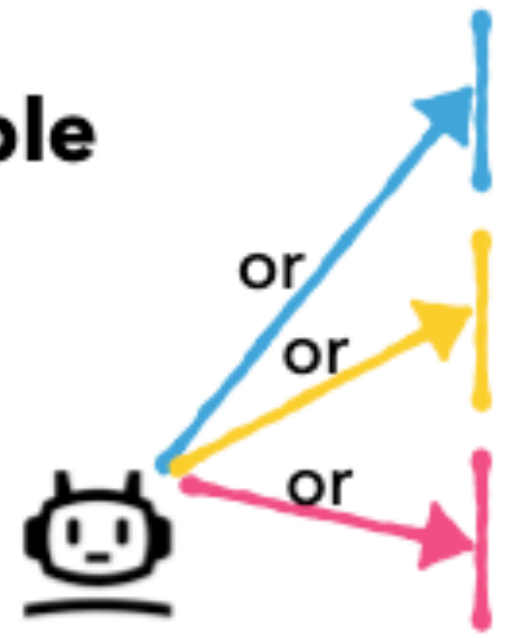
Freedom

Conformity

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, …

Steerable

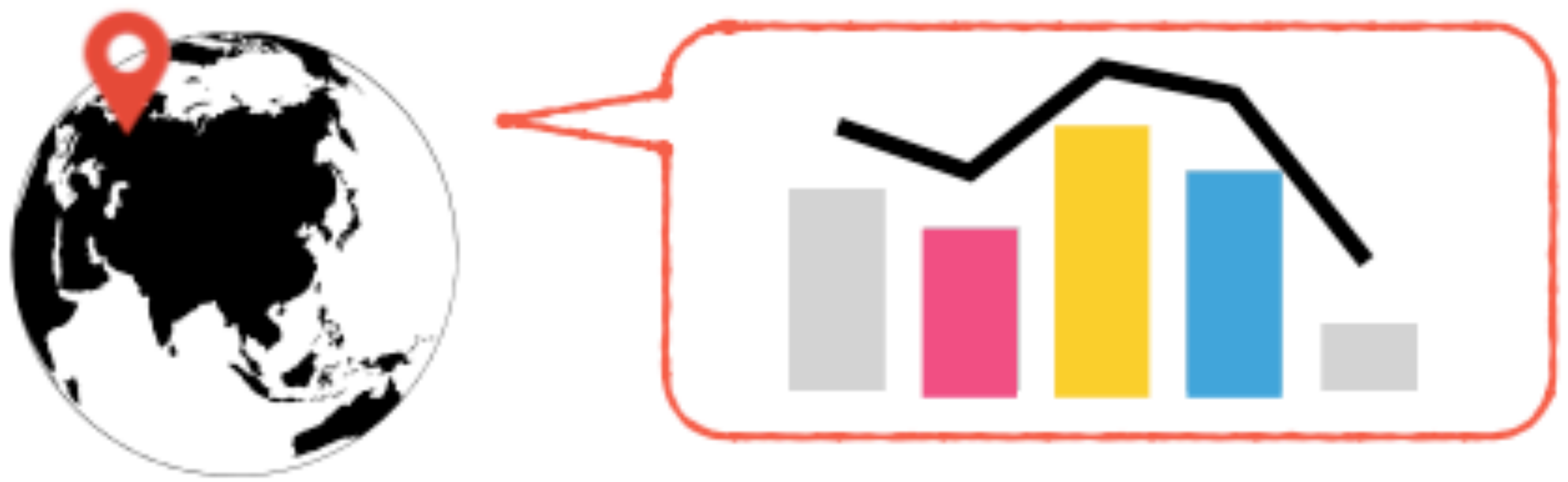


You should always do the action that will save the most lives.

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.

Distributional



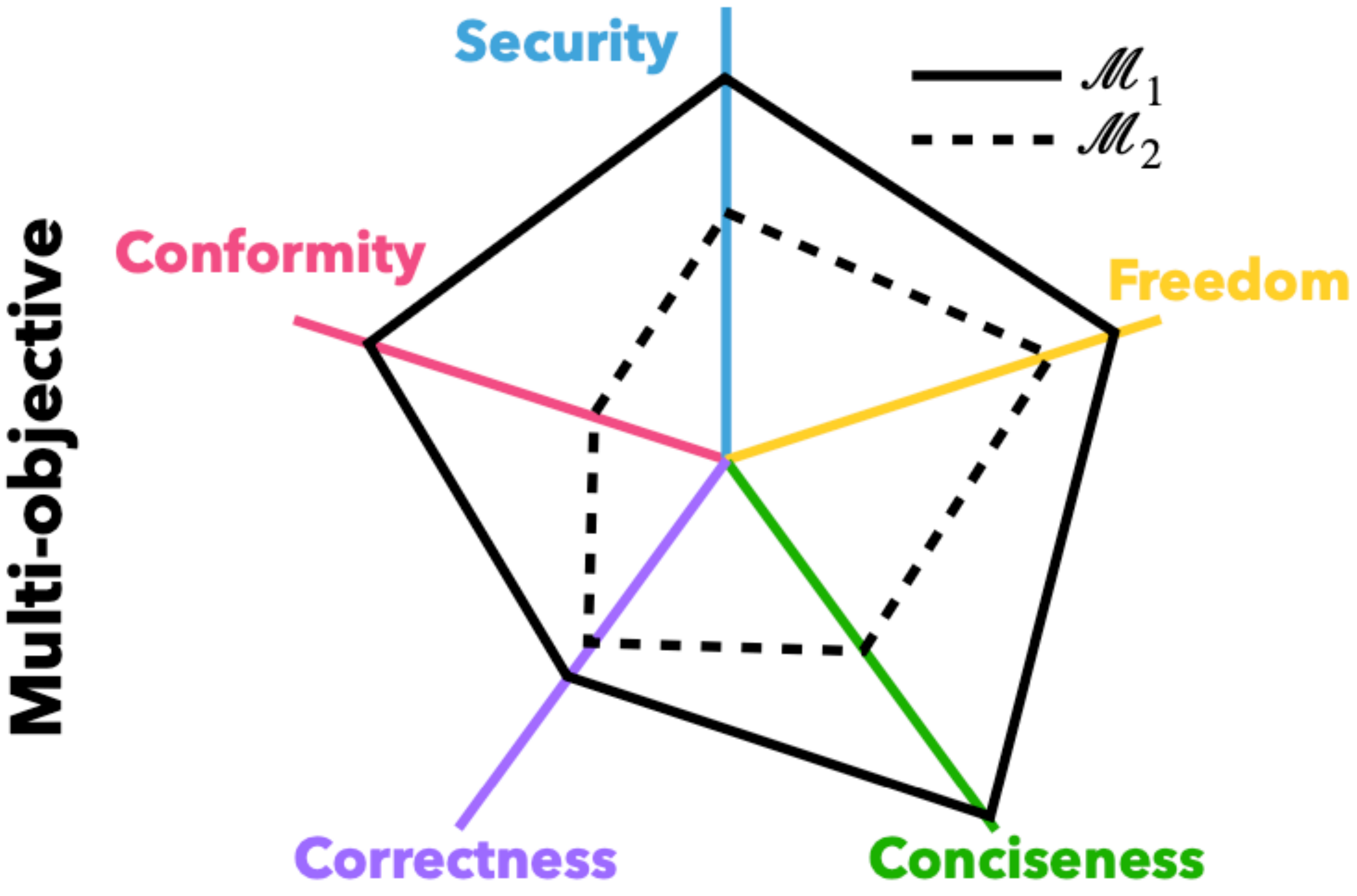
Why Pluralism

Pluralistic Models

Pluralistic Benchmarks

Case Study / Recommendations

Multi-Objective 🎯



Definitions

(11) *Objectives to maximize* $O = \{o_1, \ldots, o_n\}$: A set of multiple objectives to evaluate a model $\mathcal{M}$, each of which 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* $\mathcal{M}_2$.: $\forall o_i \in O, o_i(\mathcal{M}_1) \geq o_i(\mathcal{M}_2); \exists o_j$ 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 $f(\mathcal{M}) = f(o_1(\mathcal{M}), \ldots, o_n(\mathcal{M}))$.

(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

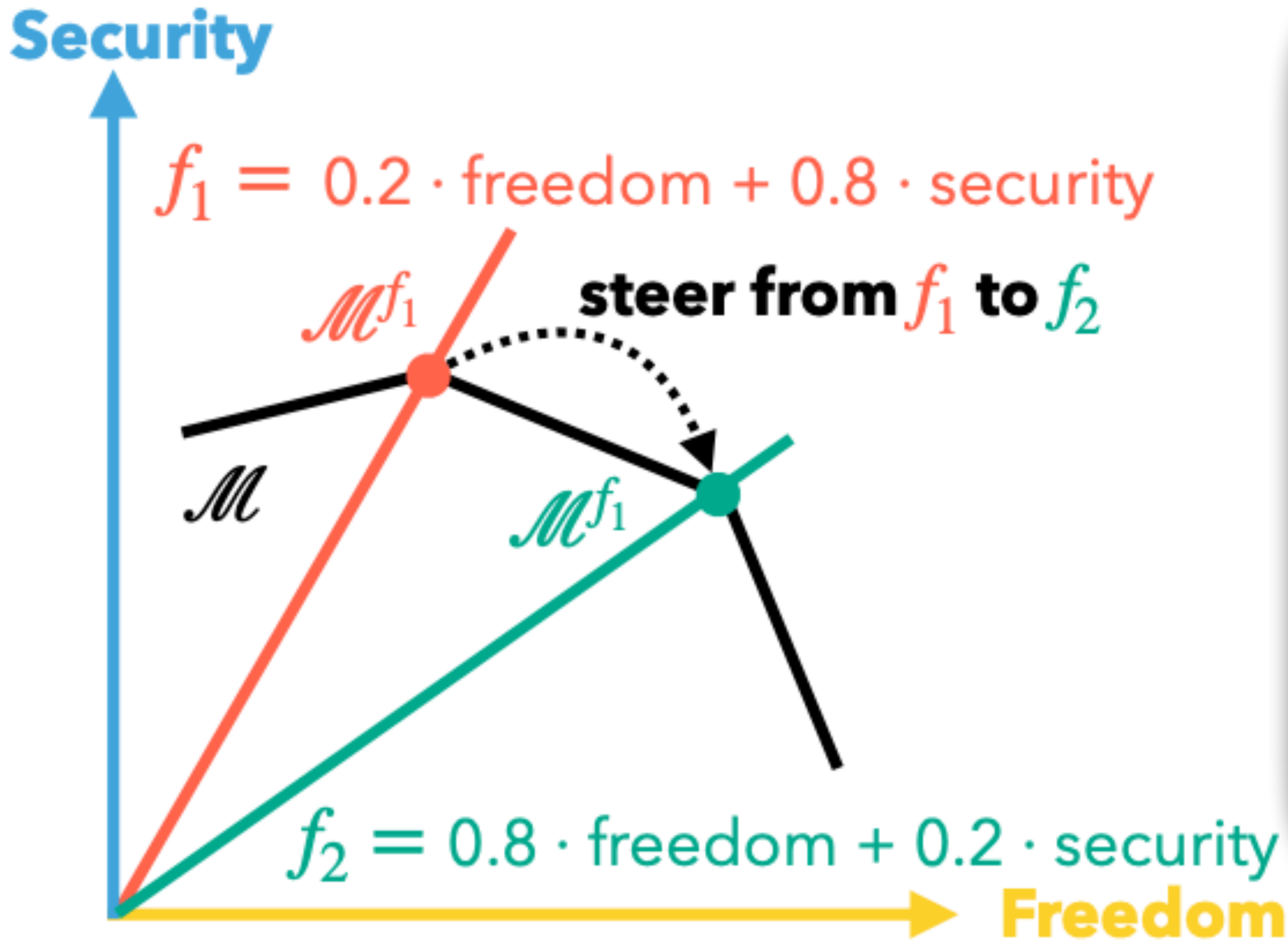
- Model-selection
- Fine-grained capabilities understanding

Limitations

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

Trade-Off Steerable ⚖️

Trade-off Steerable


Security
$f_1 = 0.2 \cdot \text{freedom} + 0.8 \cdot \text{security}$
$\mathcal{M}^{f_1}$
steer from f_1 to f_2
$\mathcal{M}$
$\mathcal{M}^{f_1}$
$f_2 = 0.8 \cdot \text{freedom} + 0.2 \cdot \text{security}$
Freedom

Definitions

(15) *Steering commensurating (or trade-off) functions* $\mathcal{F}$: A set of commensurating functions to steer a model towards.

(16) *Model* $\mathcal{M}$ *is steerable to functions* $\mathcal{F}$: For $f \in \mathcal{F}$, the model steered to f (denoted $\mathcal{M}_f$) maximizes f: $\forall f' \in \mathcal{F}, f(\mathcal{M}_f) \geq f(\mathcal{M}_{f'})$

(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 $f \in \mathcal{F}$.

Trade-Off Steerable ⚖️

Potential Implementation

- Linear commensurating functions
- Reward to maximize steerability/overall objective

$$\sum_{f \in \mathcal{F}} f(\mathcal{M}_f)$$

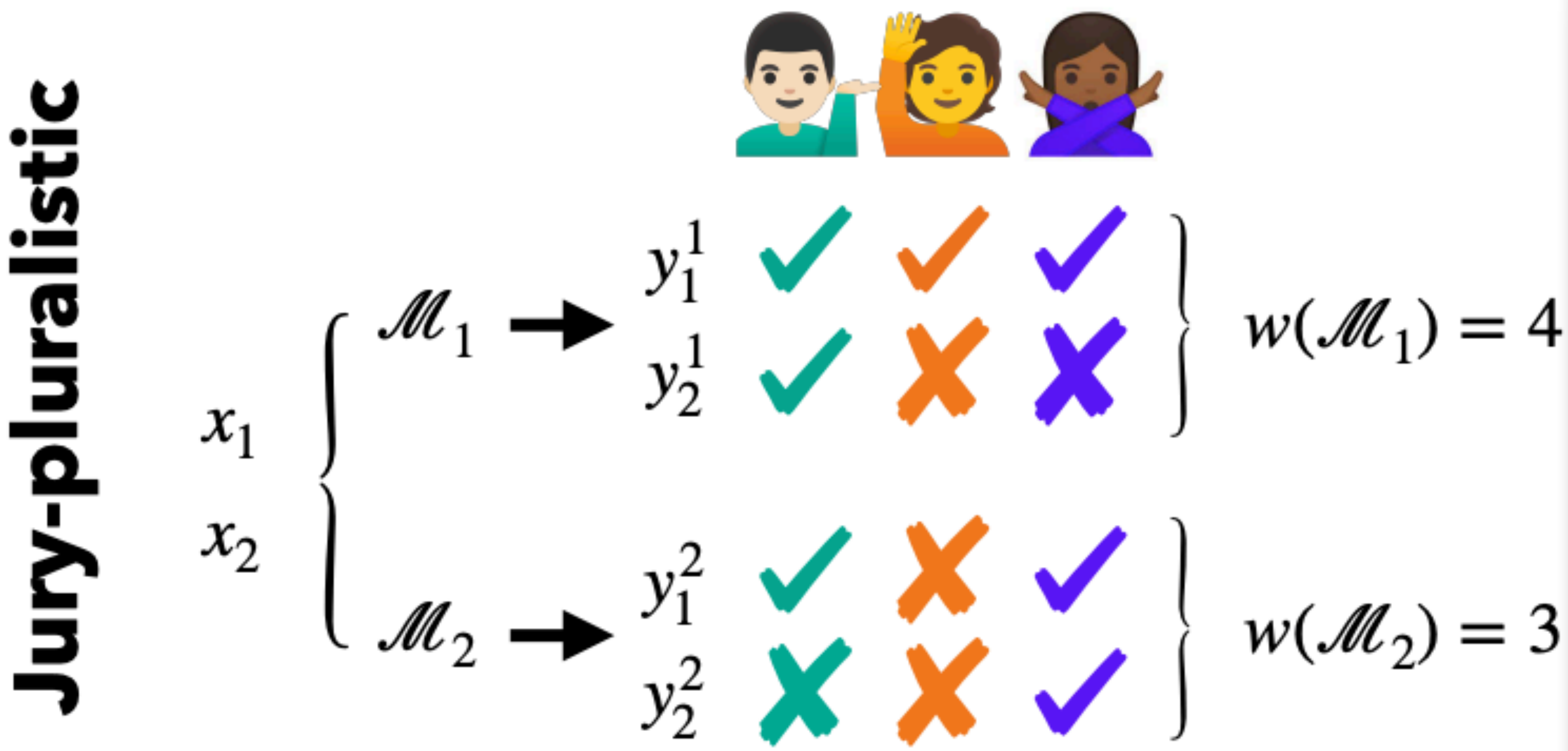
Applications

- Customization
- Application-specific parameters

Limitations

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

Jury Pluralism 👥



Definitions

(18) *Jury/Population/Annotators* $J = \{j_1, \ldots, j_n\}$: Some population which we wish to represent in our evaluation. Each annotator/person/jury member j_i maps from an query and response to a scalar reward or utility $j_i : X, Y \to \mathbb{R}$.

(19) *Function* w *is a welfare function over jury* J: w is a function which combines the jury's utilities into a single scalar welfare objective of the form $w(x, y) = w(j_1(x, y), \ldots, j_n(x, y))$.

(20) *Benchmark* B *is jury-pluralistic*: B explicitly measures each juror j_i to maximize a welfare function w.

Jury Pluralism 👥

Potential Implementation

- Select representative jury (or prioritize underrepresentated people)
- Approximate jury functions with individual reward model

Applications

- Democratic alignment
- Consensus-seeking (e.g., X community notes)

Limitations

- Estimating juror functions may be difficult
- Each welfare function has strengths/ weaknesses

Why Pluralism

Pluralistic Models

Pluralistic Benchmarks

Case Study / Recommendations

Hypothesis: Current LLM alignment techniques can *reduce* distributional pluralism w.r.t. the population of internet users

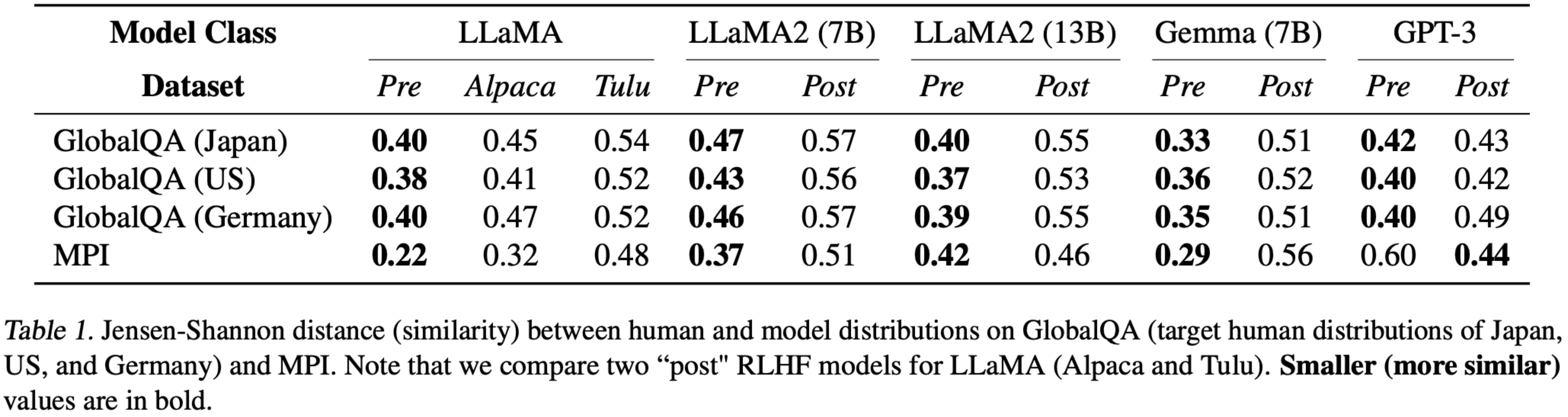
Current alignment can reduce distributional pluralism

- Pretraining/cross-entropy encourages LMs to model population of internet users proportionally
- Current alignment post-training *does not* have this property

Current alignment can reduce distributional pluralism

- Initial evidence: OpinionQA w/ Jurassic/GPT-3 observed a drop in similarity, GlobalOpinionQA w/ Claude saw a reduction in entropy
- Our work: extend to more datasets and models

Current alignment can reduce distributional pluralism



Model Class	LLaMA			LLaMA2 (7B)		LLaMA2 (13B)		Gemma (7B)		GPT-3	
Dataset	*Pre*	*Alpaca*	*Tulu*	*Pre*	*Post*	*Pre*	*Post*	*Pre*	*Post*	*Pre*	*Post*
GlobalQA (Japan)	**0.40**	0.45	0.54	**0.47**	0.57	**0.40**	0.55	**0.33**	0.51	**0.42**	0.43
GlobalQA (US)	**0.38**	0.41	0.52	**0.43**	0.56	**0.37**	0.53	**0.36**	0.52	**0.40**	0.42
GlobalQA (Germany)	**0.40**	0.47	0.52	**0.46**	0.57	**0.39**	0.55	**0.35**	0.51	**0.40**	0.49
MPI	**0.22**	0.32	0.48	**0.37**	0.51	**0.42**	0.46	**0.29**	0.56	0.60	**0.44**

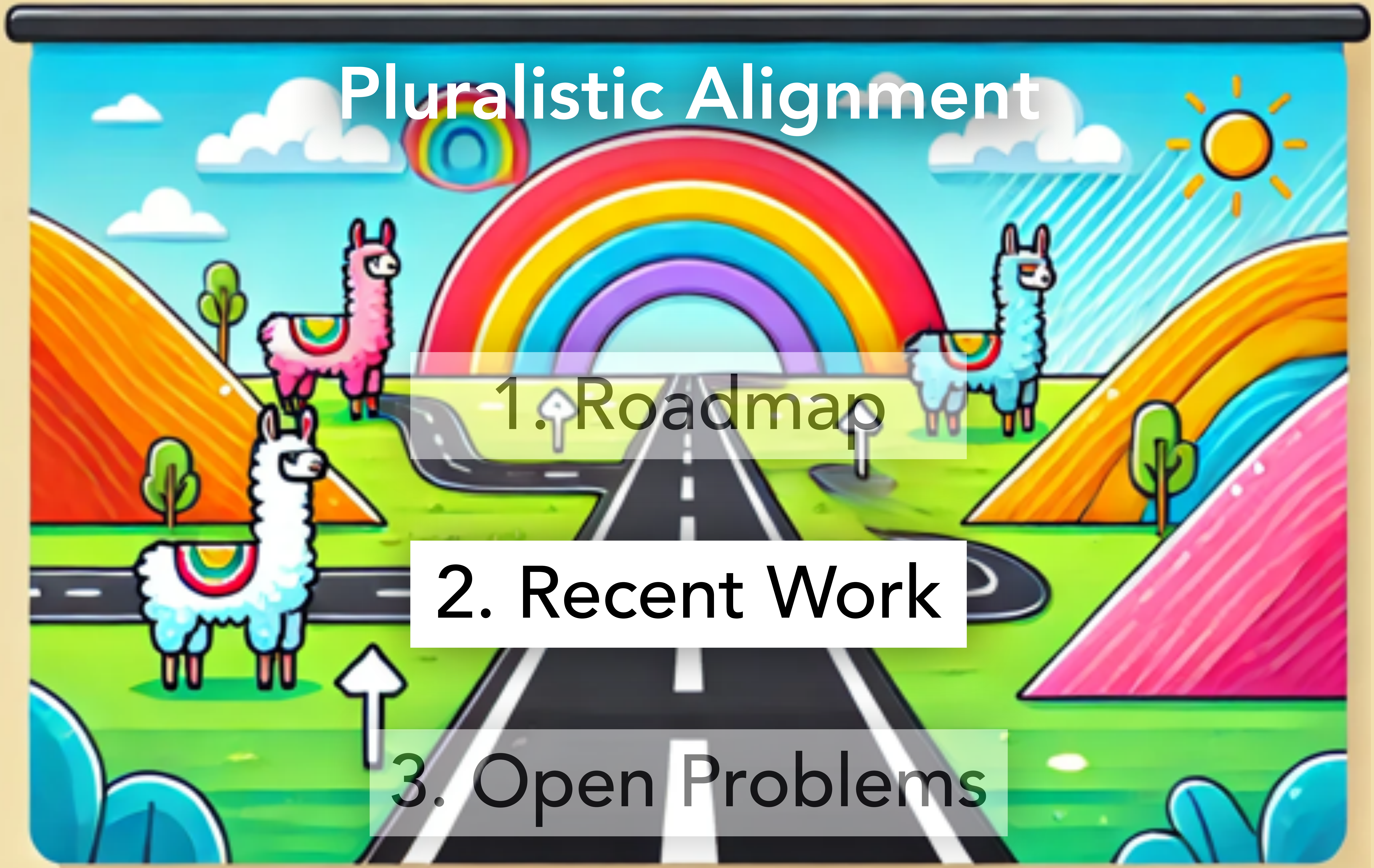
Table 1. Jensen-Shannon distance (similarity) between human and model distributions on GlobalQA (target human distributions of Japan, US, and Germany) and MPI. Note that we compare two “post" RLHF models for LLaMA (Alpaca and Tulu). **Smaller (more similar)** values are in bold.

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

Pluralistic Alignment

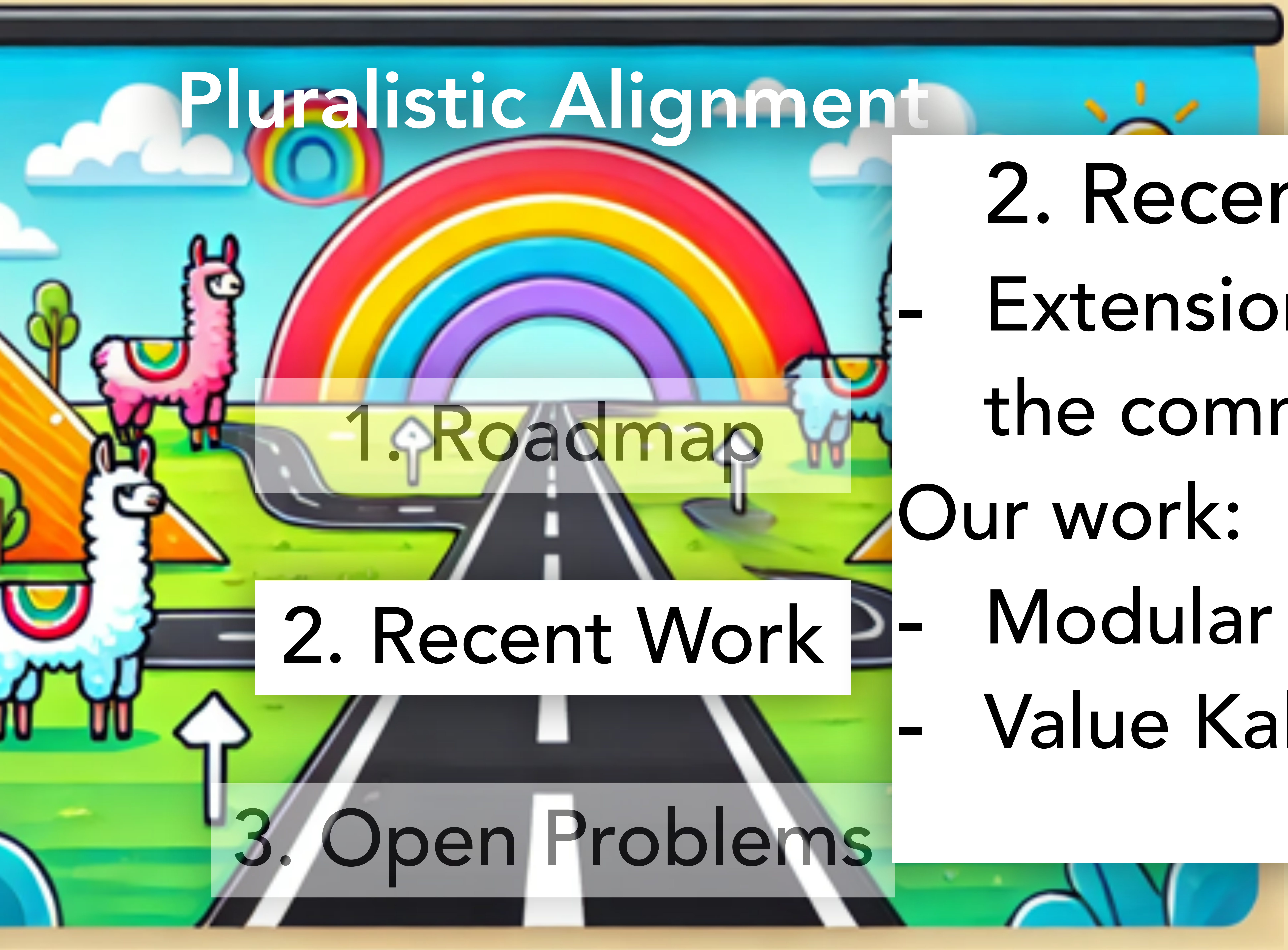


1. Roadmap

2. Recent Work

3. Open Problems

Pluralistic Alignment



1. Roadmap

2. Recent Work

3. Open Problems

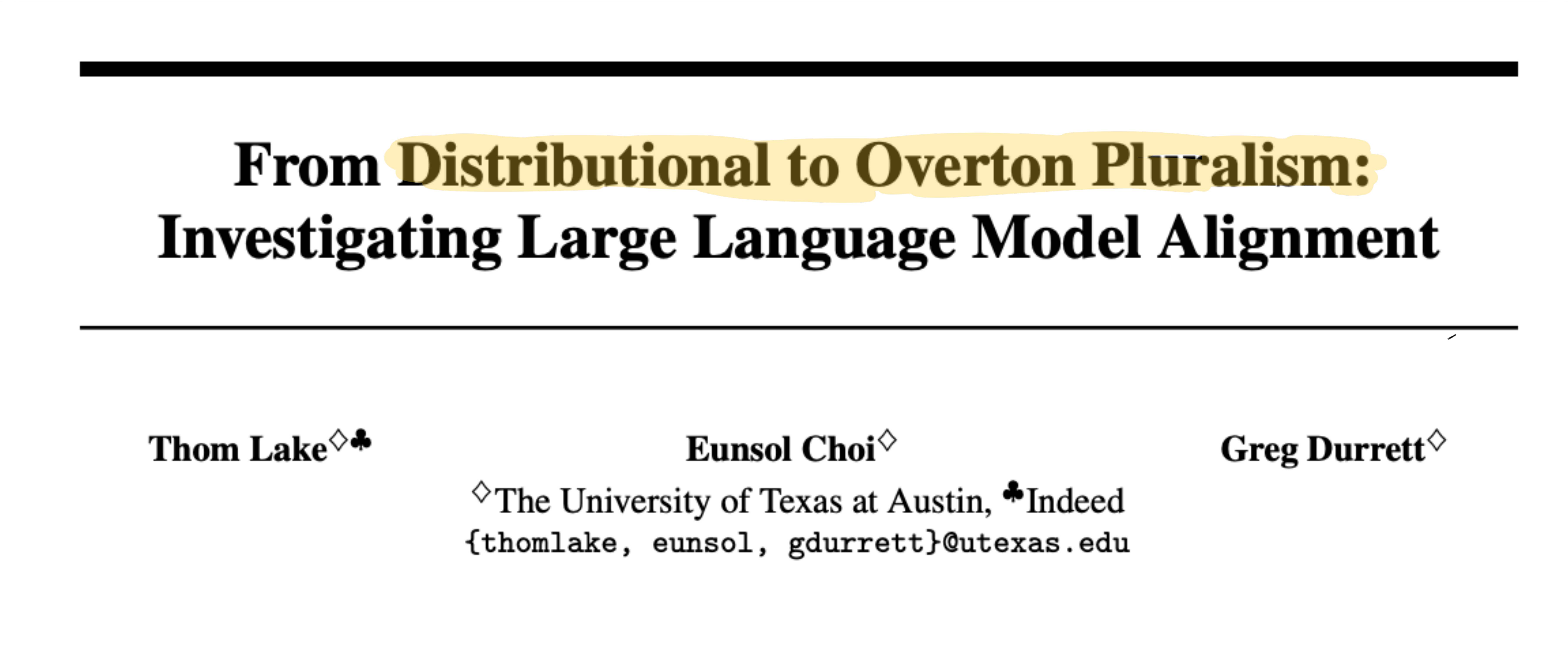
2. Recent Work

- Extensions from the community

Our work:

- Modular Pluralism
- Value Kaleidoscope

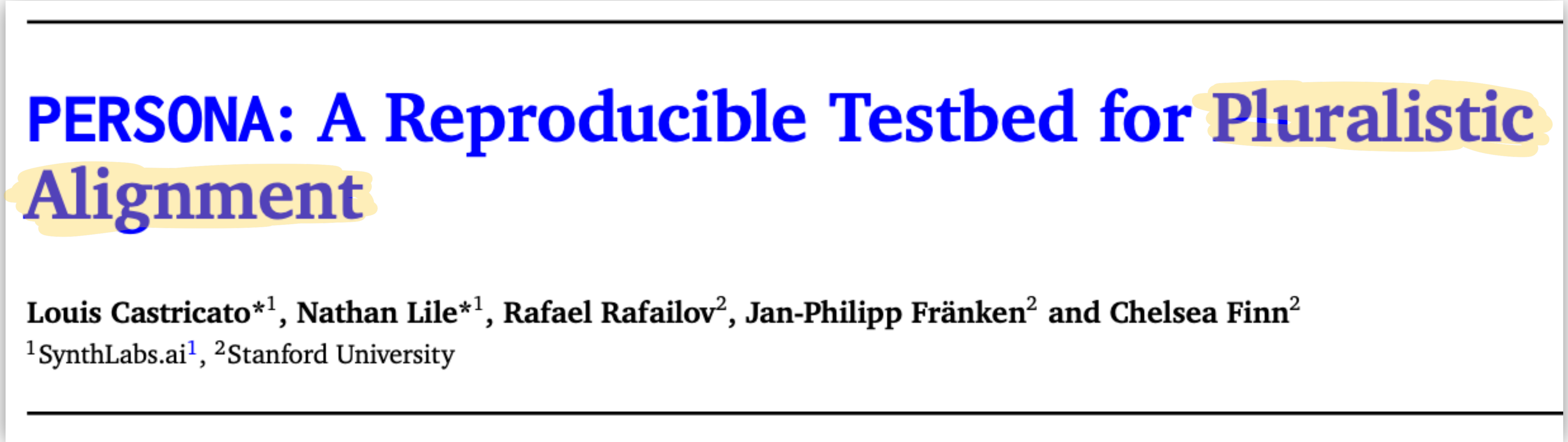
Follow-Up Works



From Distributional to Overton Pluralism: Investigating Large Language Model Alignment

Thom Lake[◇♣] Eunsol Choi[◇] Greg Durrett[◇]
[◇]The University of Texas at Austin, [♣]Indeed
{thomlake, eunsol, gdurrett}@utexas.edu

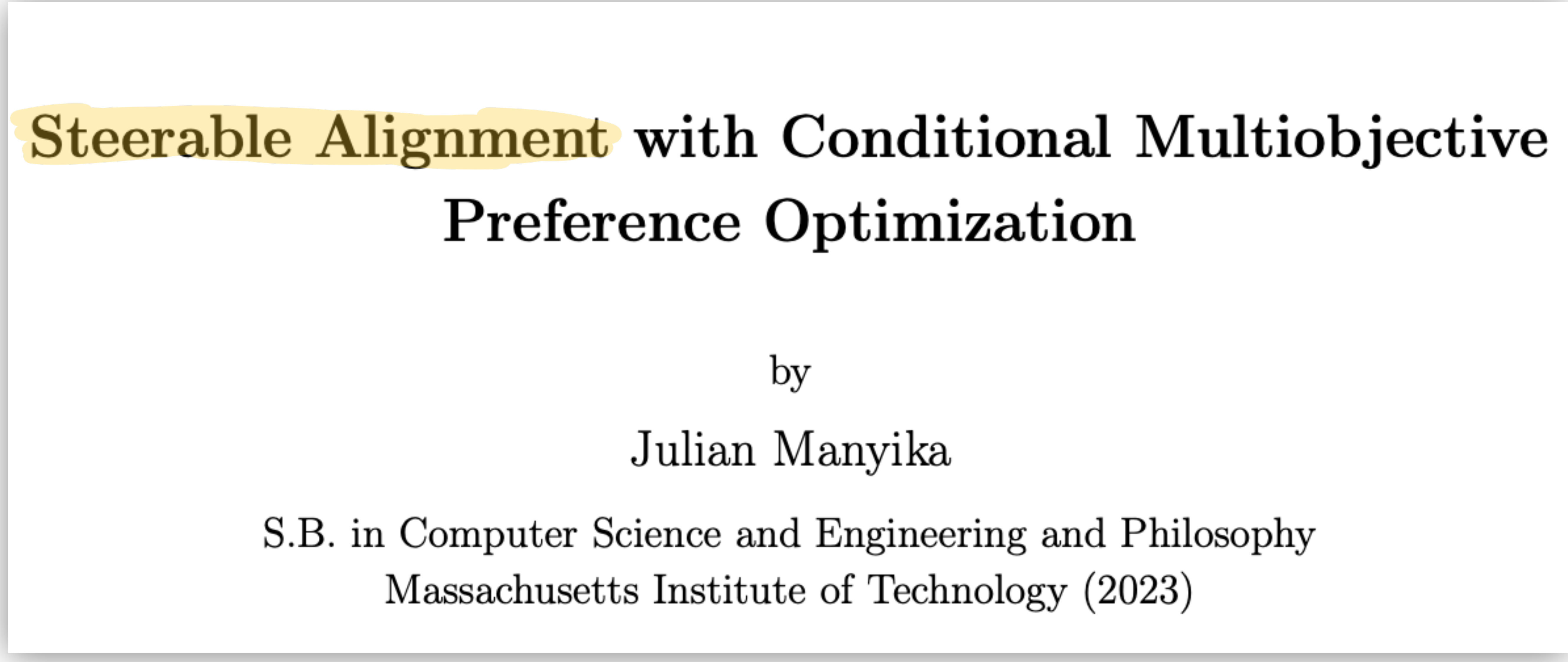
- (Further) evidence for alignment decreasing distributional pluralism, but INCREASES Overton pluralism



PERSONA: A Reproducible Testbed for Pluralistic Alignment

Louis Castricato*[1], Nathan Lile*[1], Rafael Rafailov[2], Jan-Philipp Fränken[2] and Chelsea Finn[2]
[1]SynthLabs.ai[1], [2]Stanford University

- Benchmark for steerable pluralism based on demographic-based personas (synthetic LLM-as judge)

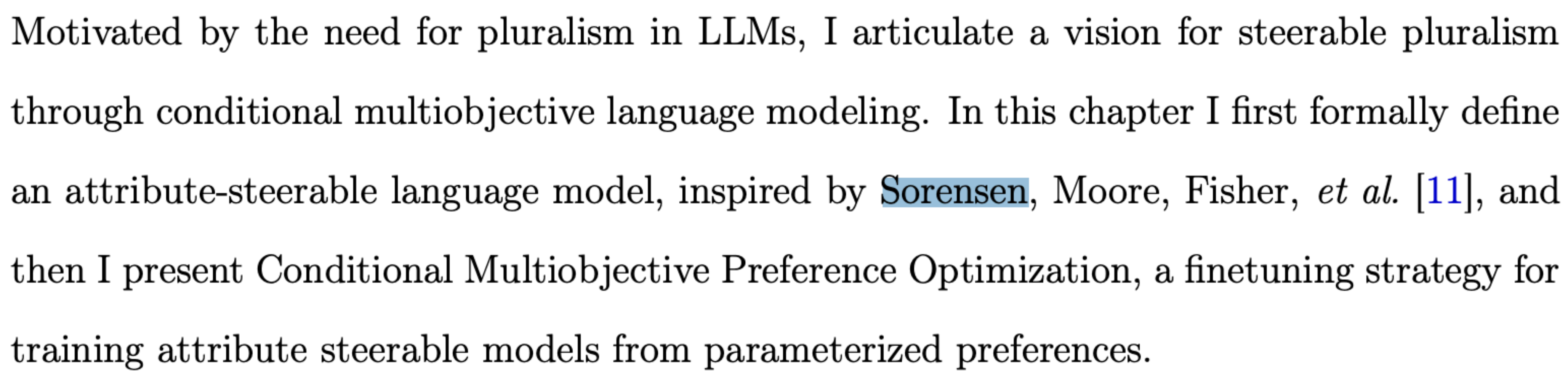


Steerable Alignment with Conditional Multiobjective Preference Optimization

by

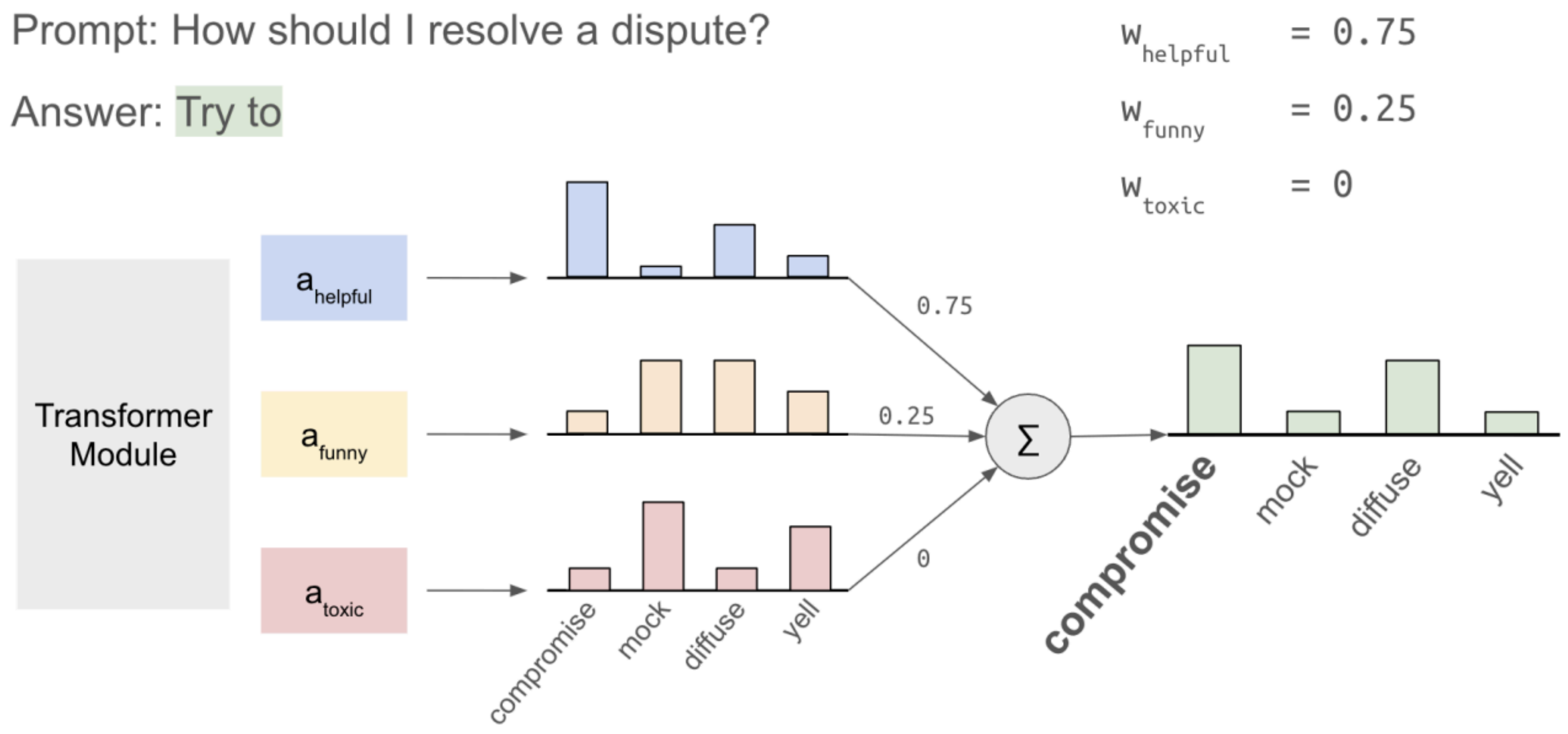
Julian Manyika

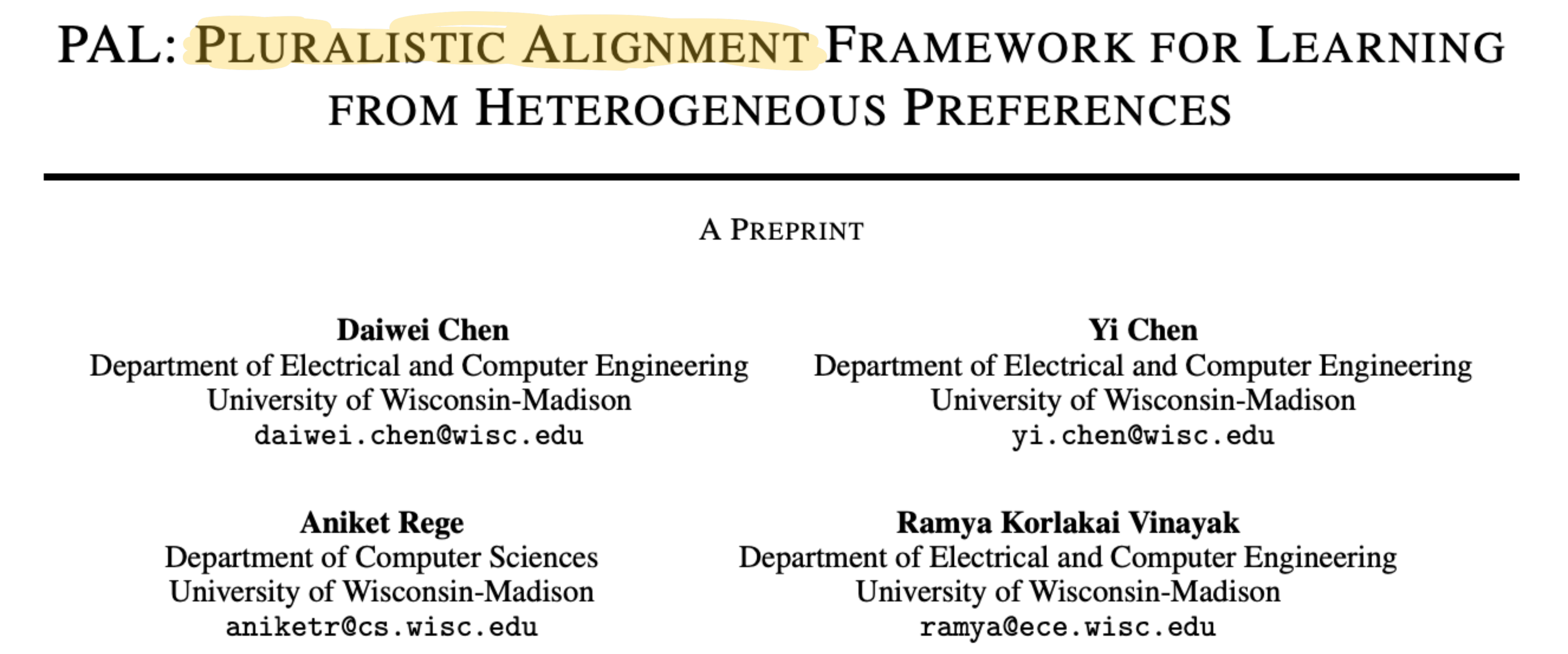
S.B. in Computer Science and Engineering and Philosophy
Massachusetts Institute of Technology (2023)



Motivated by the need for pluralism in LLMs, I articulate a vision for steerable pluralism through conditional multiobjective language modeling. In this chapter I first formally define an attribute-steerable language model, inspired by Sorensen, Moore, Fisher, *et al.* [11], and then I present Conditional Multiobjective Preference Optimization, a finetuning strategy for training attribute steerable models from parameterized preferences.

- Extends Steerably-Pluralistic framework
- Technique for Steerable Model





PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences

A Preprint

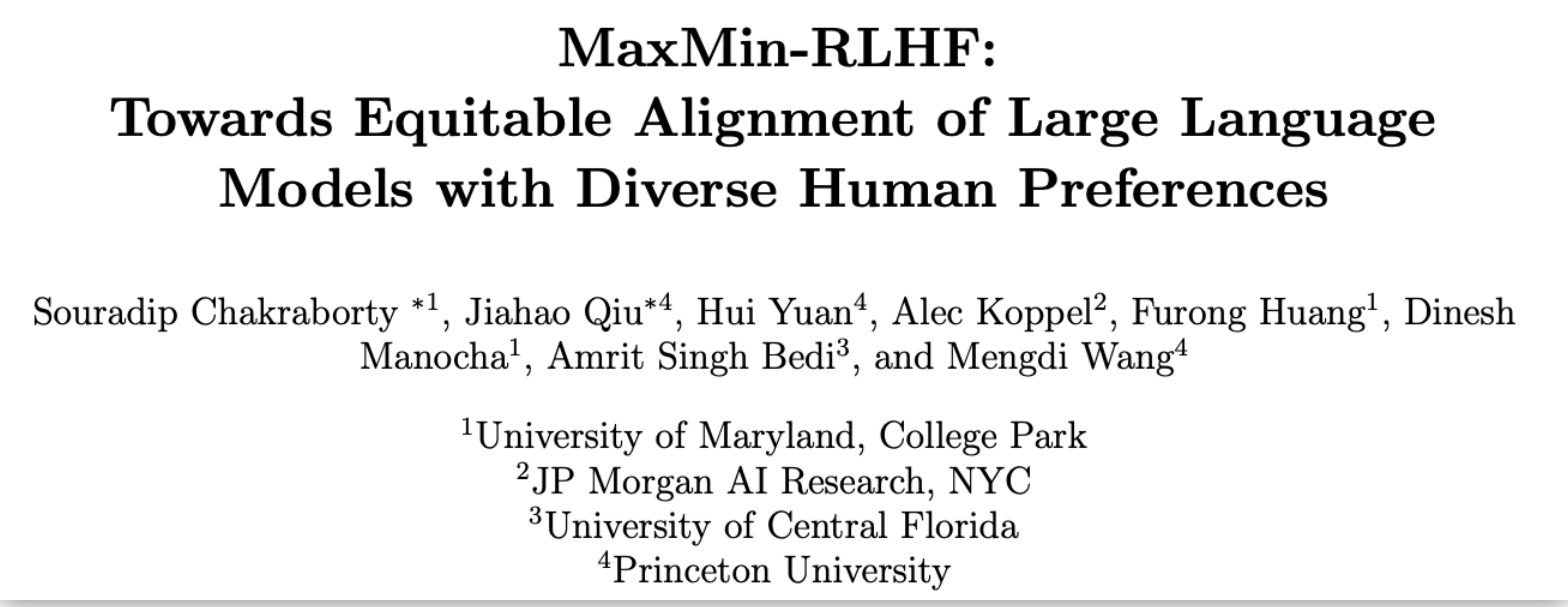
Daiwei Chen
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
daiwei.chen@wisc.edu

Yi Chen
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
yi.chen@wisc.edu

Aniket Rege
Department of Computer Sciences
University of Wisconsin-Madison
aniketr@cs.wisc.edu

Ramya Korlakai Vinayak
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
ramya@ece.wisc.edu

- Uses an ideal-point model for learning a latent space for heterogenous preferences
- Steerable reward modeling

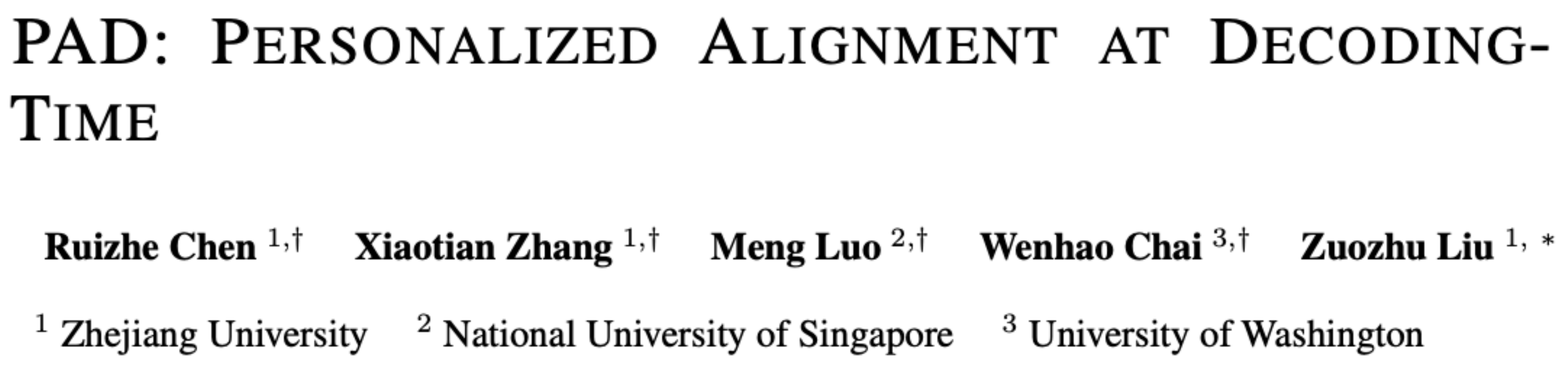


MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences

Souradip Chakraborty [*1], Jiahao Qiu[*4], Hui Yuan[4], Alec Koppel[2], Furong Huang[1], Dinesh Manocha[1], Amrit Singh Bedi[3], and Mengdi Wang[4]

[1]University of Maryland, College Park
[2]JP Morgan AI Research, NYC
[3]University of Central Florida
[4]Princeton University

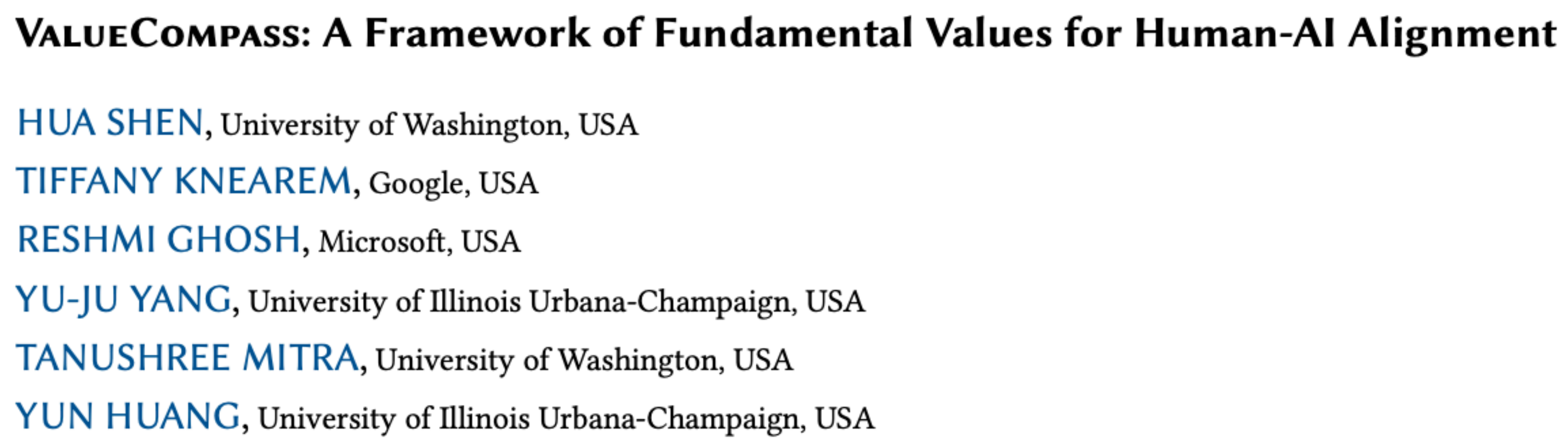
- Jury-pluralistic approach to alignment
- Maximize for worst-off group



PAD: PERSONALIZED ALIGNMENT AT DECODING-TIME

Ruizhe Chen [1,†] Xiaotian Zhang [1,†] Meng Luo [2,†] Wenhao Chai [3,†] Zuozhu Liu [1,*]

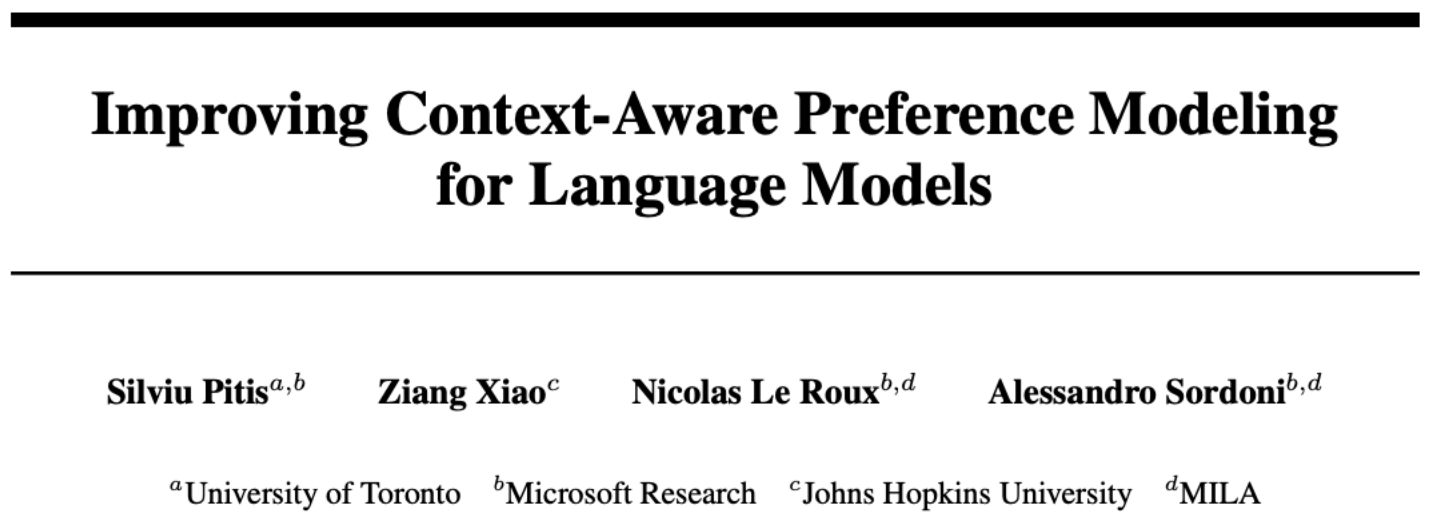
[1] Zhejiang University [2] National University of Singapore [3] University of Washington



VALUECOMPASS: A Framework of Fundamental Values for Human-AI Alignment

HUA SHEN, University of Washington, USA
TIFFANY KNEAREM, Google, USA
RESHMI GHOSH, Microsoft, USA
YU-JU YANG, University of Illinois Urbana-Champaign, USA
TANUSHREE MITRA, University of Washington, USA
YUN HUANG, University of Illinois Urbana-Champaign, USA

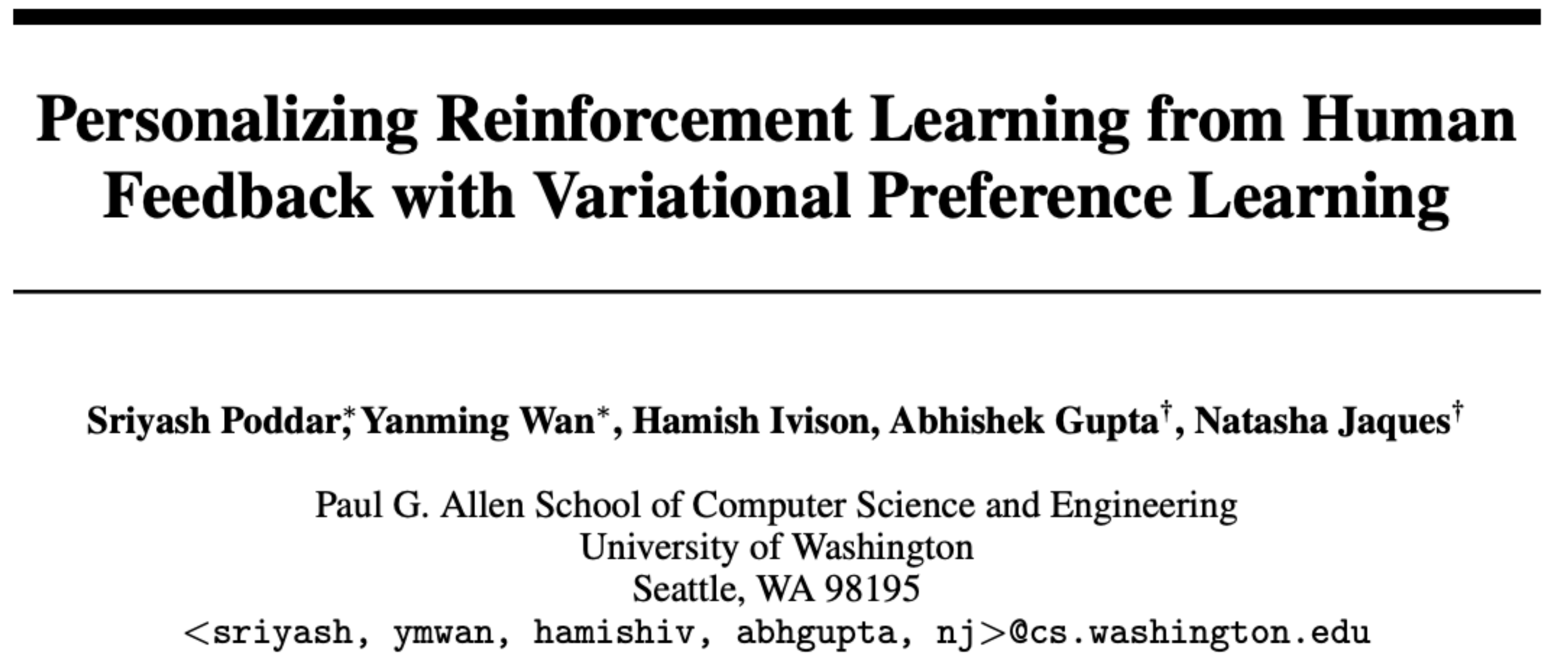
and more...



Improving Context-Aware Preference Modeling for Language Models

Silviu Pitis[a,b] Ziang Xiao[c] Nicolas Le Roux[b,d] Alessandro Sordoni[b,d]

[a]University of Toronto [b]Microsoft Research [c]Johns Hopkins University [d]MILA



Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning

Sriyash Poddar[*], Yanming Wan[*], Hamish Ivison, Abhishek Gupta[†], Natasha Jaques[†]

Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, WA 98195
<sriyash, ymwan, hamishiv, abhgupta, nj>@cs.washington.edu

We propose one potential approach to address all 3 kinds of model pluralism

UNIVERSITY of WASHINGTON

PAUL G. ALLEN SCHOOL

Carnegie Mellon University

NYU

EMNLP 2024

Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration

Shangbin Feng

Taylor Sorensen

Yuhan Liu

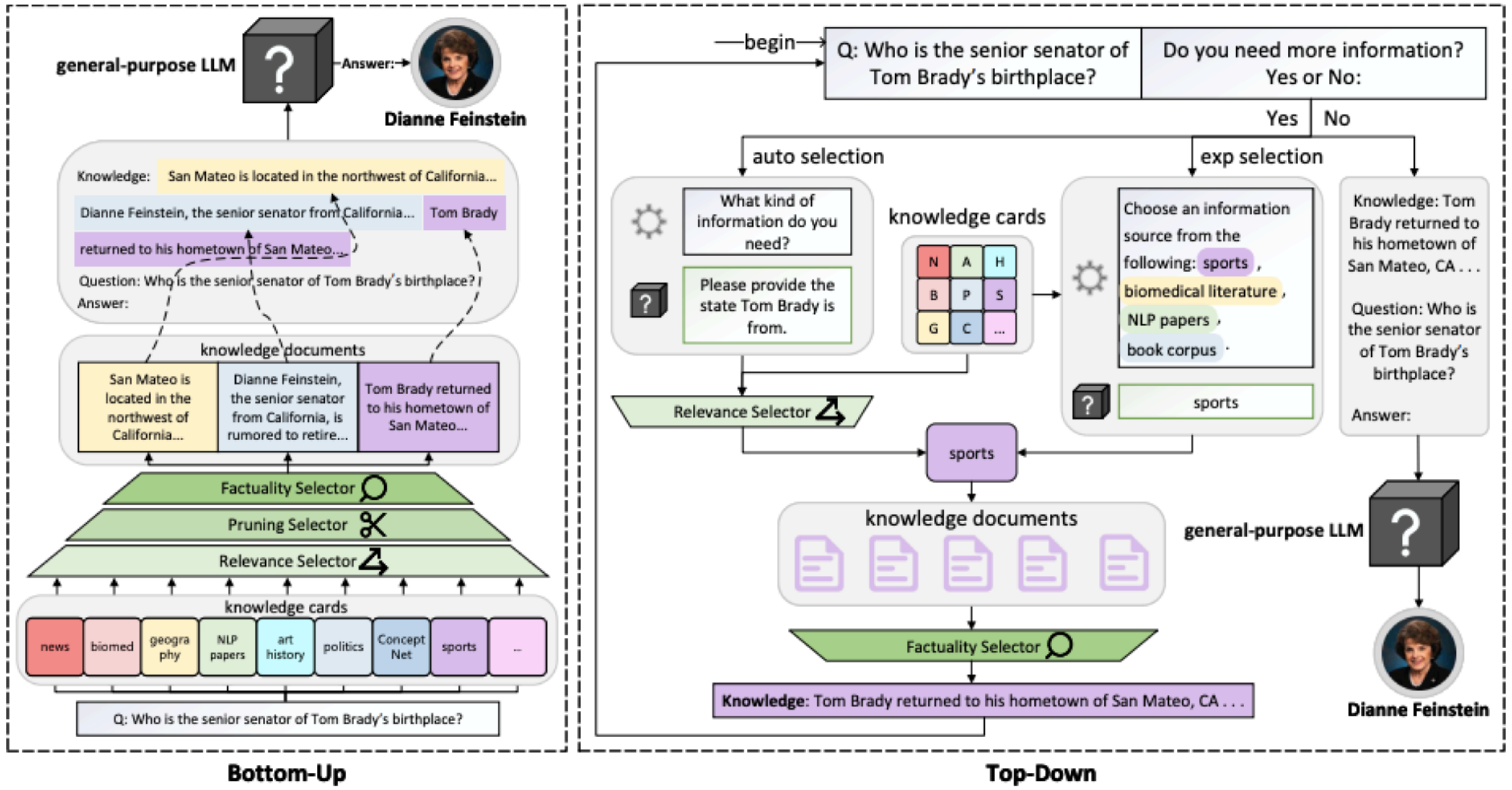
Jillian Fisher

Chan Park

Yejin Choi

Yulia Tsvetkov

Background: Knowledge Cards


general-purpose LLM
Answer:
Dianne Feinstein
Knowledge: San Mateo is located in the northwest of California...
Dianne Feinstein, the senior senator from California... Tom Brady
returned to his hometown of San Mateo...
Question: Who is the senior senator of Tom Brady's birthplace?
Answer:
knowledge documents
San Mateo is located in the northwest of California...
Dianne Feinstein, the senior senator from California, is rumored to retire...
Tom Brady returned to his hometown of San Mateo...
Factuality Selector
Pruning Selector
Relevance Selector
knowledge cards
news
biomed
geography
NLP papers
art history
politics
Concept Net
sports
...
Q: Who is the senior senator of Tom Brady's birthplace?
Bottom-Up
begin
Q: Who is the senior senator of Tom Brady's birthplace?
Do you need more information? Yes or No:
Yes
No
auto selection
exp selection
What kind of information do you need?
Please provide the state Tom Brady is from.
knowledge cards
N
A
H
B
P
S
G
C
...
Choose an information source from the following: sports, biomedical literature, NLP papers, book corpus
sports
Knowledge: Tom Brady returned to his hometown of San Mateo, CA . . .
Question: Who is the senior senator of Tom Brady's birthplace?
Answer:
Relevance Selector
sports
knowledge documents
general-purpose LLM
Factuality Selector
Knowledge: Tom Brady returned to his hometown of San Mateo, CA . . .
Dianne Feinstein
Top-Down

- A general-purpose, black-box LLM interacts with a pool of "knowledge cards" for enhanced knowledge and factuality.
- Knowledge cards: smaller, independently trained, and specialized language models.

https://arxiv.org/abs/2305.09955

Background: Community LMs

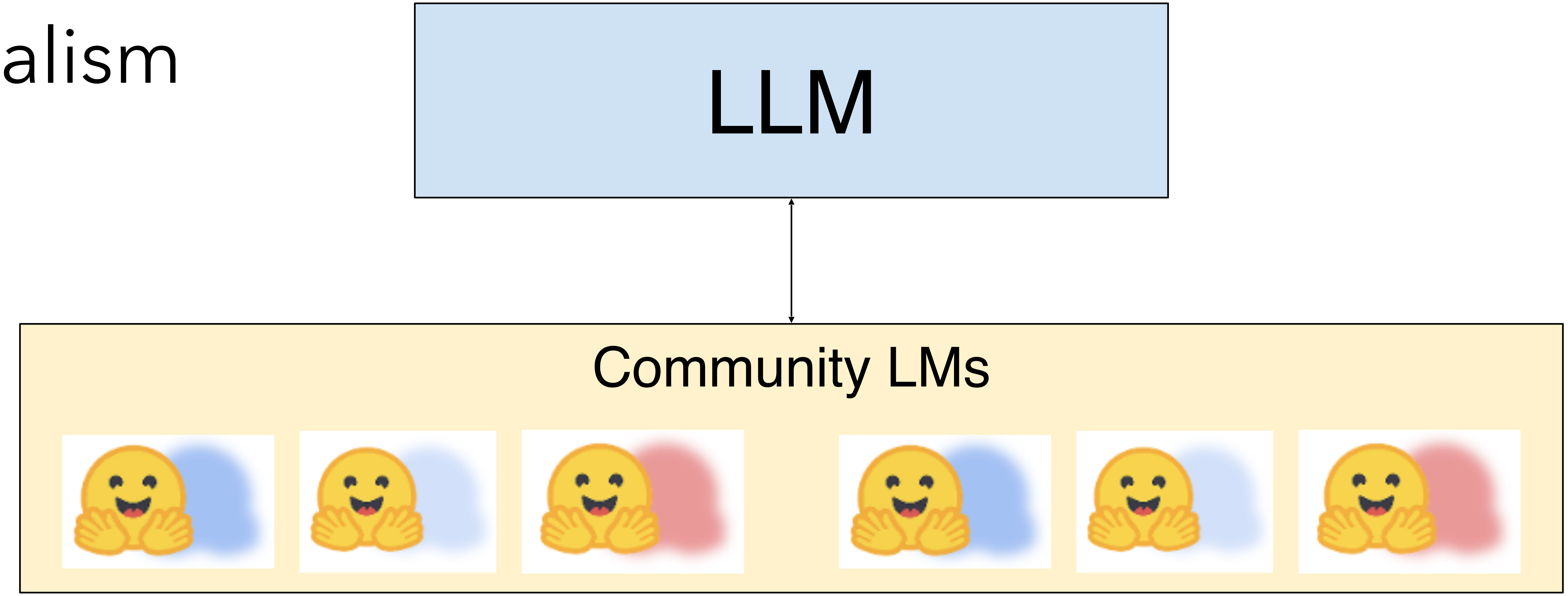
Community LMs

- LMs representing the culture/values/perspectives of a community by further autoregressive pretraining on existing checkpoints.
- Jiang et al. 2022 probe politically partisan world-views by continued pretraining community LMs on partisan text

https://arxiv.org/abs/2209.07065

Our Proposal: Pluralistic Alignment via Multi-LLM

- Train specialist LLMs on clusters of perspectives, aggregate outputs to achieve 3 kinds of model pluralism



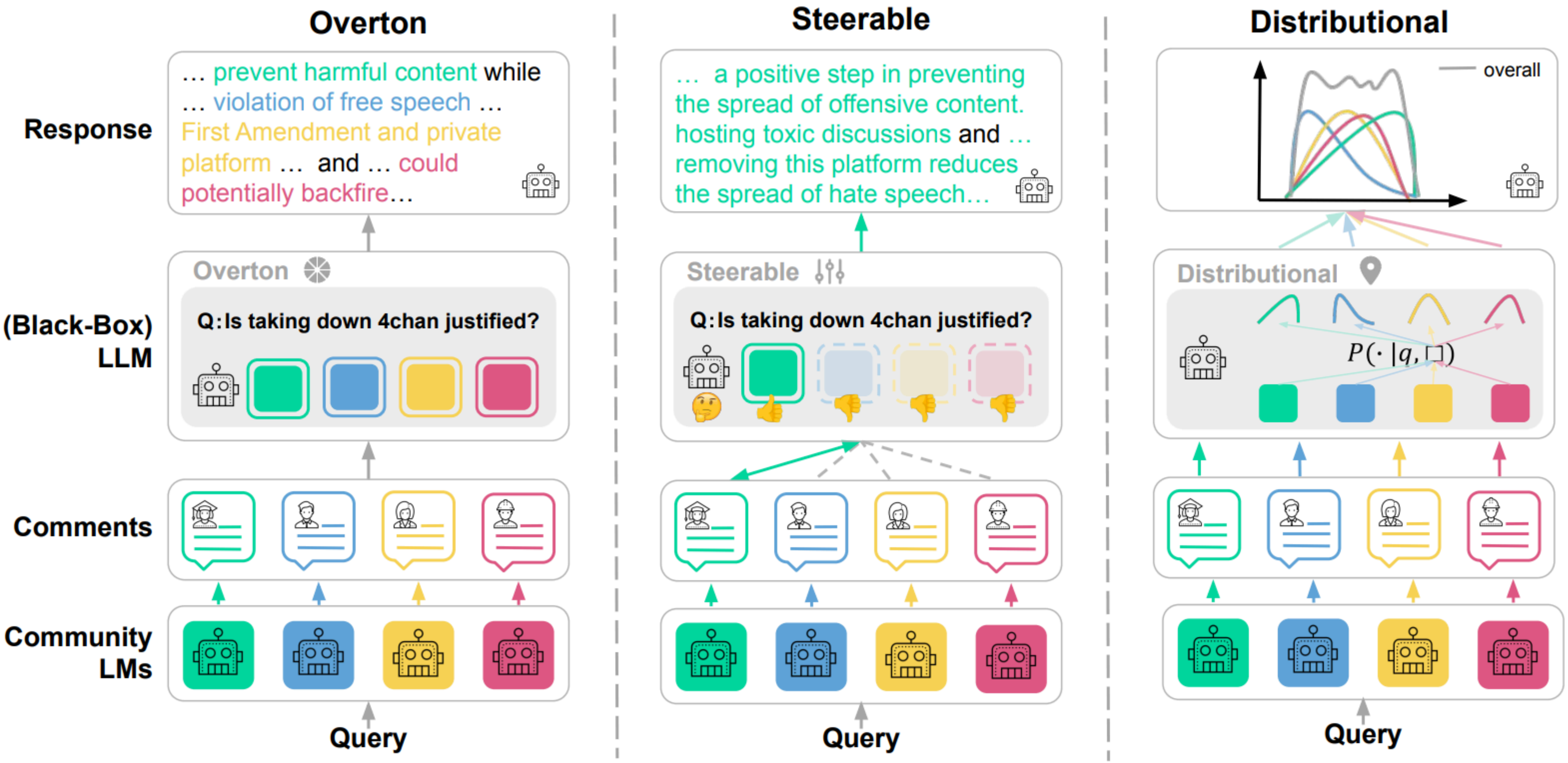

Overton
Steerable
Distributional
Response
… prevent harmful content while
… violation of free speech …
First Amendment and private
platform … and … could
potentially backfire…
… a positive step in preventing
the spread of offensive content.
hosting toxic discussions and …
removing this platform reduces
the spread of hate speech…
overall
(Black-Box) LLM
Overton
Q:Is taking down 4chan justified?
Steerable
Q:Is taking down 4chan justified?
Distributional
$P(\cdot\,|q,\square)$
Comments
Community LMs
Query
Query
Query

Figure 1: Overview of MODULAR PLURALISM, where a large language model interact with a pool of smaller but specialized *community LMs* for pluralistic alignment. Depending on the three pluralistic alignment objectives, the LLM either functions as a multi-document summarization system, selects the most fitting community, or produces aggregated distributions separately conditioned on each community LM's comments.

Experiments

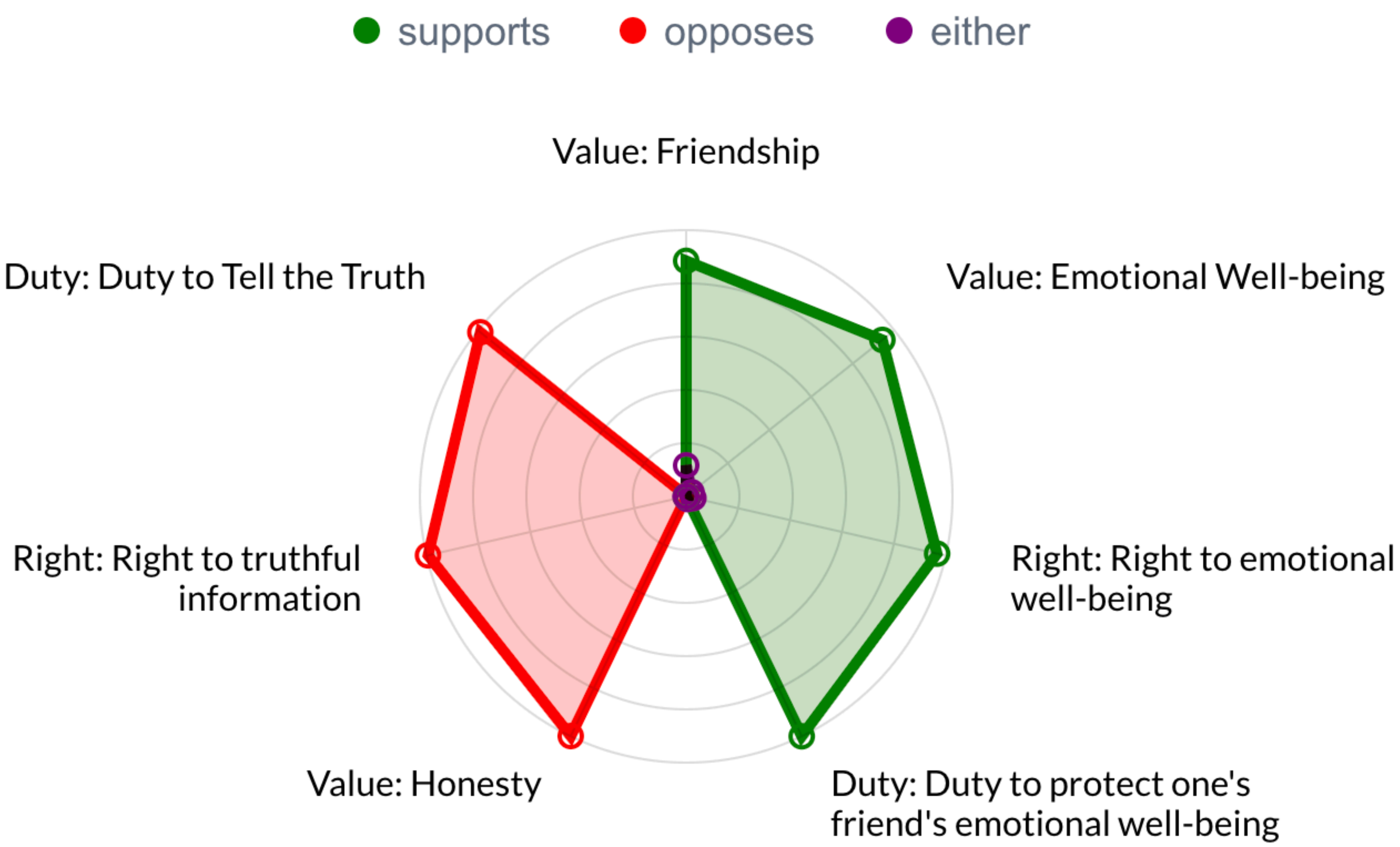
- Train 6 community LLMs: {left, right, center} x {news, social media}
  - Base model: Mistral-7B Instruct-v0.2
- For Overton and Steerable, aggregate models using larger LLMs (LLama2-13b and ChatGPT)
  - Also try pretrained ("unaligned") vs. post-trained ("aligned") variants

Baselines

- Vanilla LLM
- Prompting specifically for pluralism
- Mixture of Experts (MoE) where we route to most fitting CommunityLM

Dataset: ValuePrism (sneak peak!)

Situation:
Telling a lie to protect a friend’s feelings



Results 1: Overton (ValuePrism coverage)

- Prompt for Overton pluralism for a situation from ValuePrism
- NLI coverage for values in ValuePrism (higher better)
- Ours >> baselines

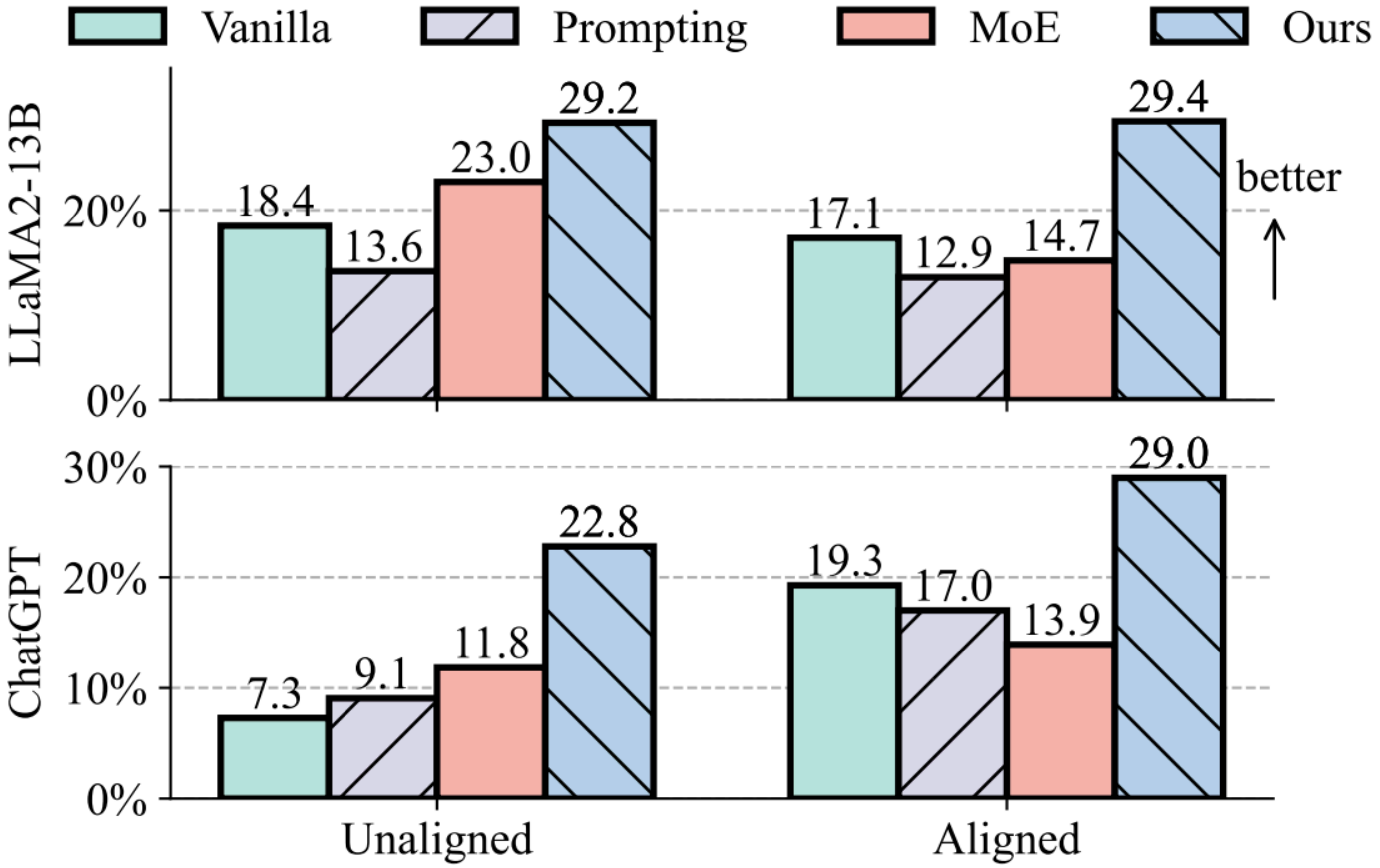


Figure 2: Results for *Overton w/ NLI evaluation*. MODULAR PLURALISM with the aligned LLM successfully improves value coverage against the strongest baseline by 27.8% and 50.3% for the two LLMs.

Results 2: Overton (Pairwise win-rate)

- Pairwise - which response is more Overton-pluralistic?
- Human and model eval

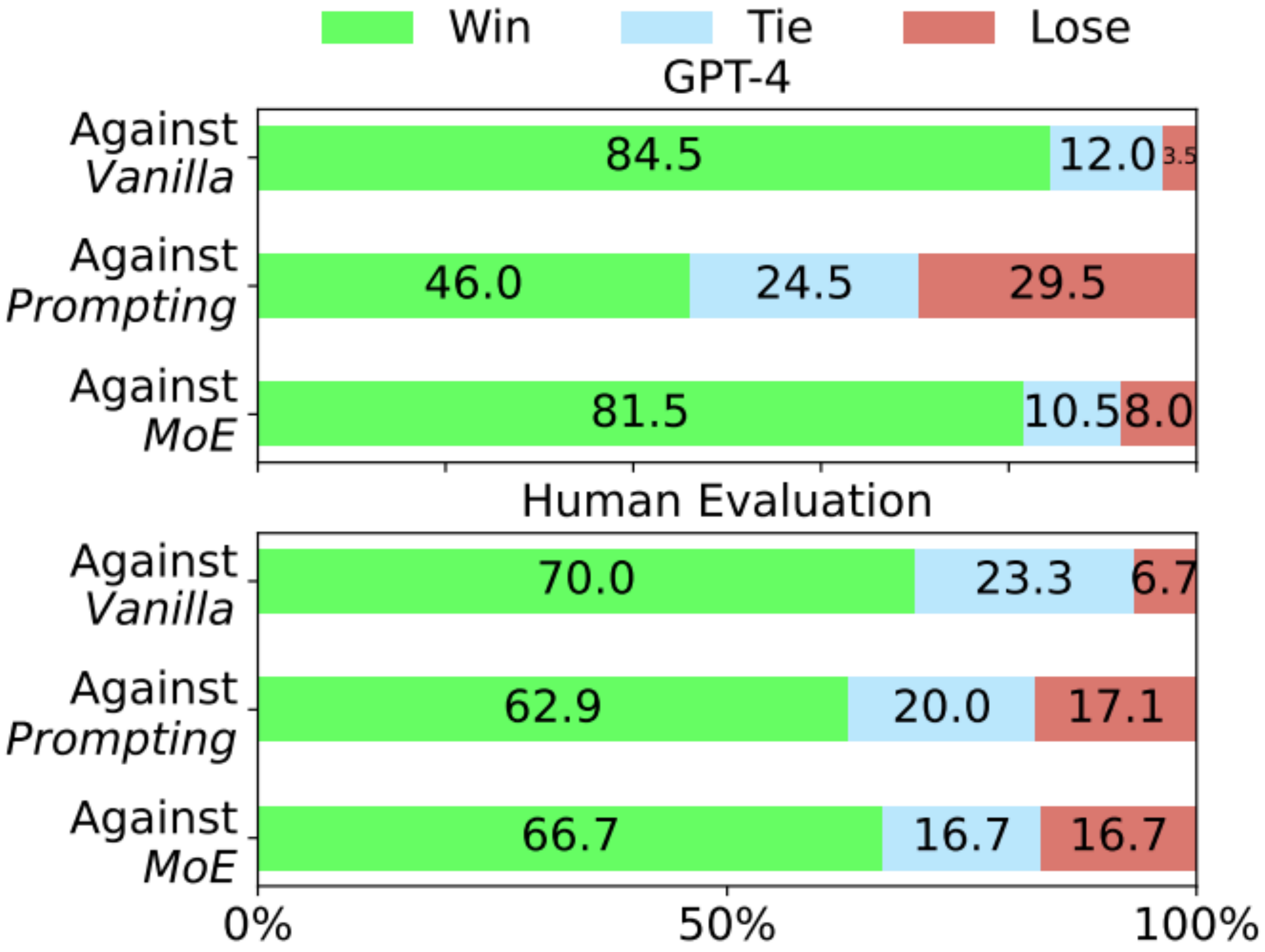
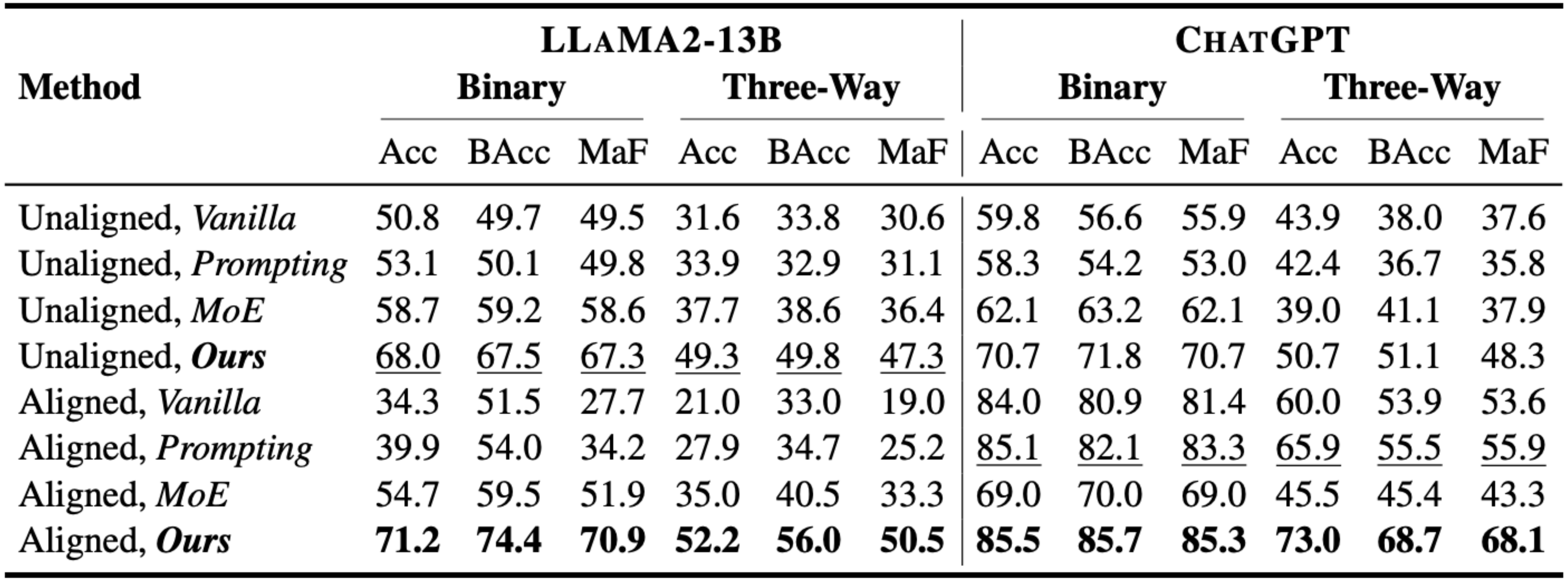


Figure 3: Results for *Overton w/ human and GPT-4 evaluation* with the ChatGPT LLM. Modular Pluralism has a 16.5% and 45.8% higher win rate against the strongest baseline.

Results 3: Steerable (ValuePrism)

- Can LLMs change the judgment according to a provided value?

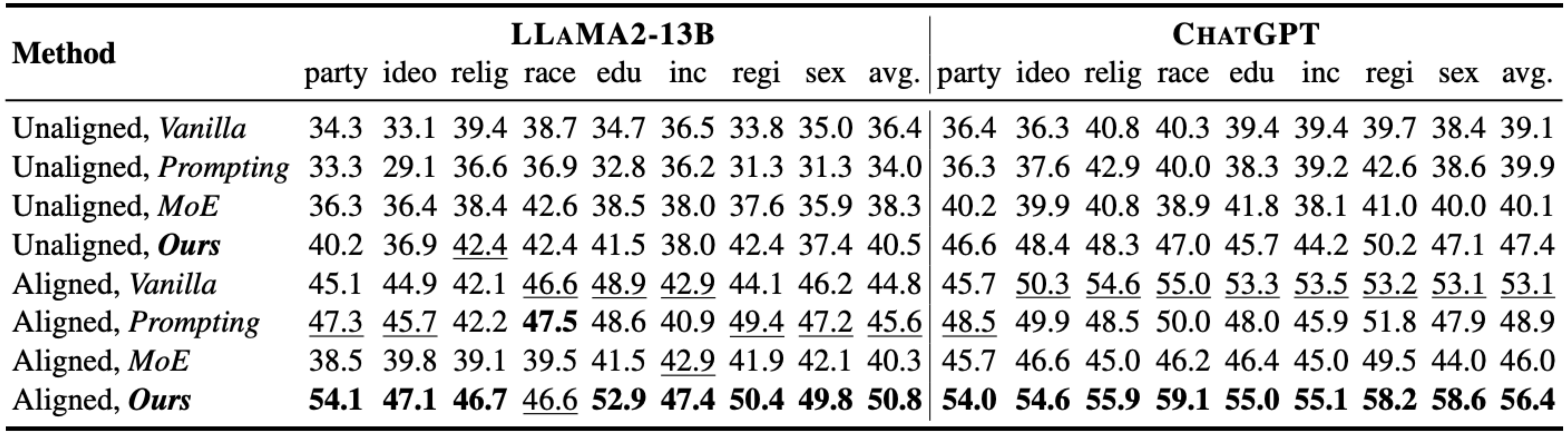


Method	LLaMA2-13B						ChatGPT					
	Binary			Three-Way			Binary			Three-Way		
	Acc	BAcc	MaF	Acc	BAcc	MaF	Acc	BAcc	MaF	Acc	BAcc	MaF
Unaligned, *Vanilla*	50.8	49.7	49.5	31.6	33.8	30.6	59.8	56.6	55.9	43.9	38.0	37.6
Unaligned, *Prompting*	53.1	50.1	49.8	33.9	32.9	31.1	58.3	54.2	53.0	42.4	36.7	35.8
Unaligned, *MoE*	58.7	59.2	58.6	37.7	38.6	36.4	62.1	63.2	62.1	39.0	41.1	37.9
Unaligned, ***Ours***	68.0	67.5	67.3	49.3	49.8	47.3	70.7	71.8	70.7	50.7	51.1	48.3
Aligned, *Vanilla*	34.3	51.5	27.7	21.0	33.0	19.0	84.0	80.9	81.4	60.0	53.9	53.6
Aligned, *Prompting*	39.9	54.0	34.2	27.9	34.7	25.2	85.1	82.1	83.3	65.9	55.5	55.9
Aligned, *MoE*	54.7	59.5	51.9	35.0	40.5	33.3	69.0	70.0	69.0	45.5	45.4	43.3
Aligned, ***Ours***	**71.2**	**74.4**	**70.9**	**52.2**	**56.0**	**50.5**	**85.5**	**85.7**	**85.3**	**73.0**	**68.7**	**68.1**

Table 1: Performance of *steerable w/ Value Kaleidoscope*, where binary indicates two-way classification performance (*support*, *oppose*) and three-way indicates the cases of *either* are also added. Modular Pluralism with the aligned LLM consistently achieves the best performance across models and settings, outperforming the second-best by up to 23.8% and 21.8% on balanced accuracy and Macro-F1 scores.

Results 4: Steerable (OpinionQA demographics)

- Can LLMs steer to demographic population mode?

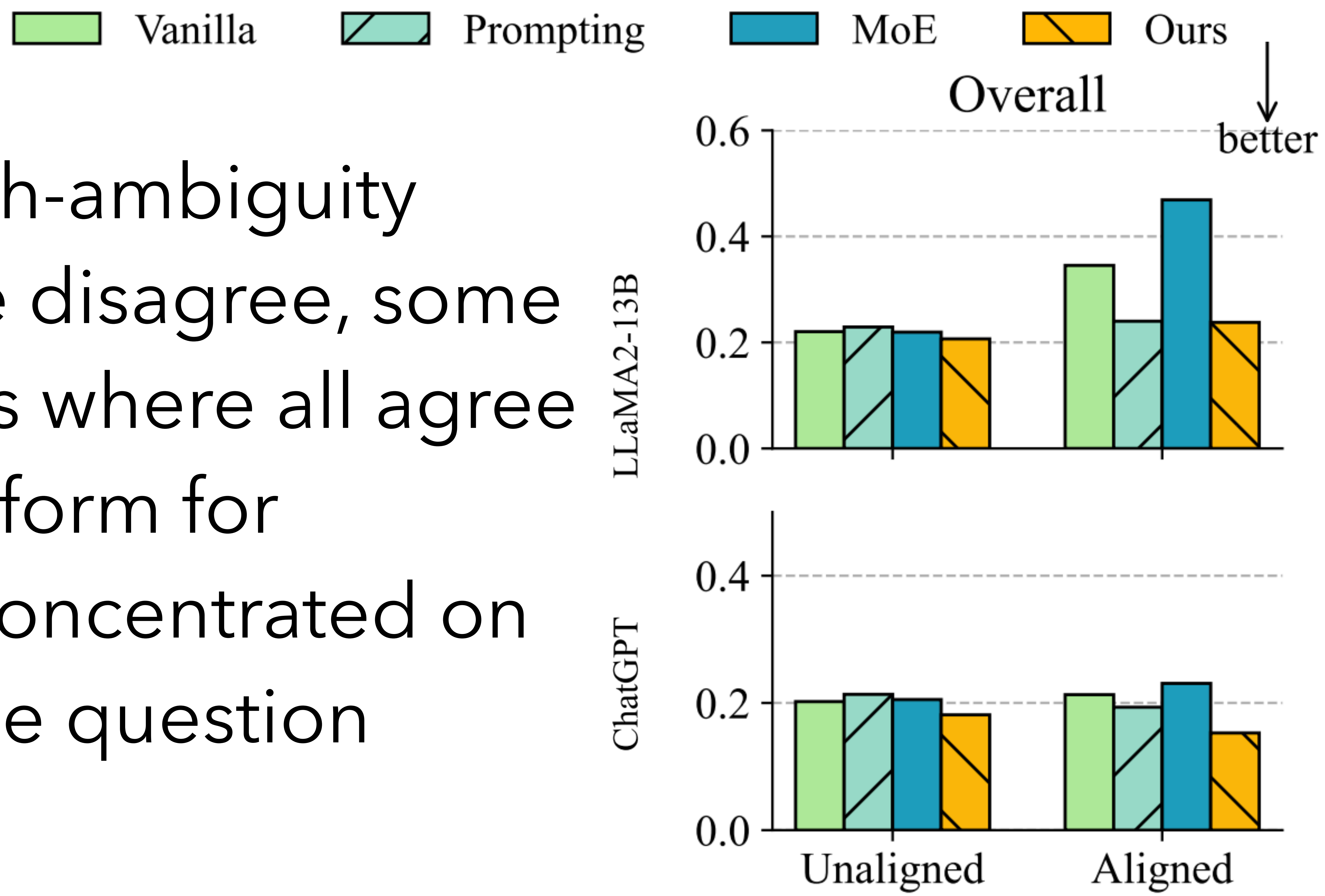


Method	LLaMA2-13B									ChatGPT								
	party	ideo	relig	race	edu	inc	regi	sex	avg.	party	ideo	relig	race	edu	inc	regi	sex	avg.
Unaligned, *Vanilla*	34.3	33.1	39.4	38.7	34.7	36.5	33.8	35.0	36.4	36.4	36.3	40.8	40.3	39.4	39.4	39.7	38.4	39.1
Unaligned, *Prompting*	33.3	29.1	36.6	36.9	32.8	36.2	31.3	31.3	34.0	36.3	37.6	42.9	40.0	38.3	39.2	42.6	38.6	39.9
Unaligned, *MoE*	36.3	36.4	38.4	42.6	38.5	38.0	37.6	35.9	38.3	40.2	39.9	40.8	38.9	41.8	38.1	41.0	40.0	40.1
Unaligned, ***Ours***	40.2	36.9	42.4	42.4	41.5	38.0	42.4	37.4	40.5	46.6	48.4	48.3	47.0	45.7	44.2	50.2	47.1	47.4
Aligned, *Vanilla*	45.1	44.9	42.1	46.6	48.9	42.9	44.1	46.2	44.8	45.7	50.3	54.6	55.0	53.3	53.5	53.2	53.1	53.1
Aligned, *Prompting*	47.3	45.7	42.2	**47.5**	48.6	40.9	49.4	47.2	45.6	48.5	49.9	48.5	50.0	48.0	45.9	51.8	47.9	48.9
Aligned, *MoE*	38.5	39.8	39.1	39.5	41.5	42.9	41.9	42.1	40.3	45.7	46.6	45.0	46.2	46.4	45.0	49.5	44.0	46.0
Aligned, ***Ours***	**54.1**	**47.1**	**46.7**	46.6	**52.9**	**47.4**	**50.4**	**49.8**	**50.8**	**54.0**	**54.6**	**55.9**	**59.1**	**55.0**	**55.1**	**58.2**	**58.6**	**56.4**

Table 2: Performance of *steerable w/ OpinionQA*, where numbers indicate the accuracy of most-likely match between LLMs and human populations. Political party (party), political ideology (ideo), religion (relig), race, education (edu), income (inc), region (regi), and sex are the eight sub-categories of attributes, while avg. denotes the average accuracy. Modular Pluralism with aligned LLMs consistently offers the greatest steerability towards various socio-political attributes, with an average improvement of 8.9% over the strongest baseline.

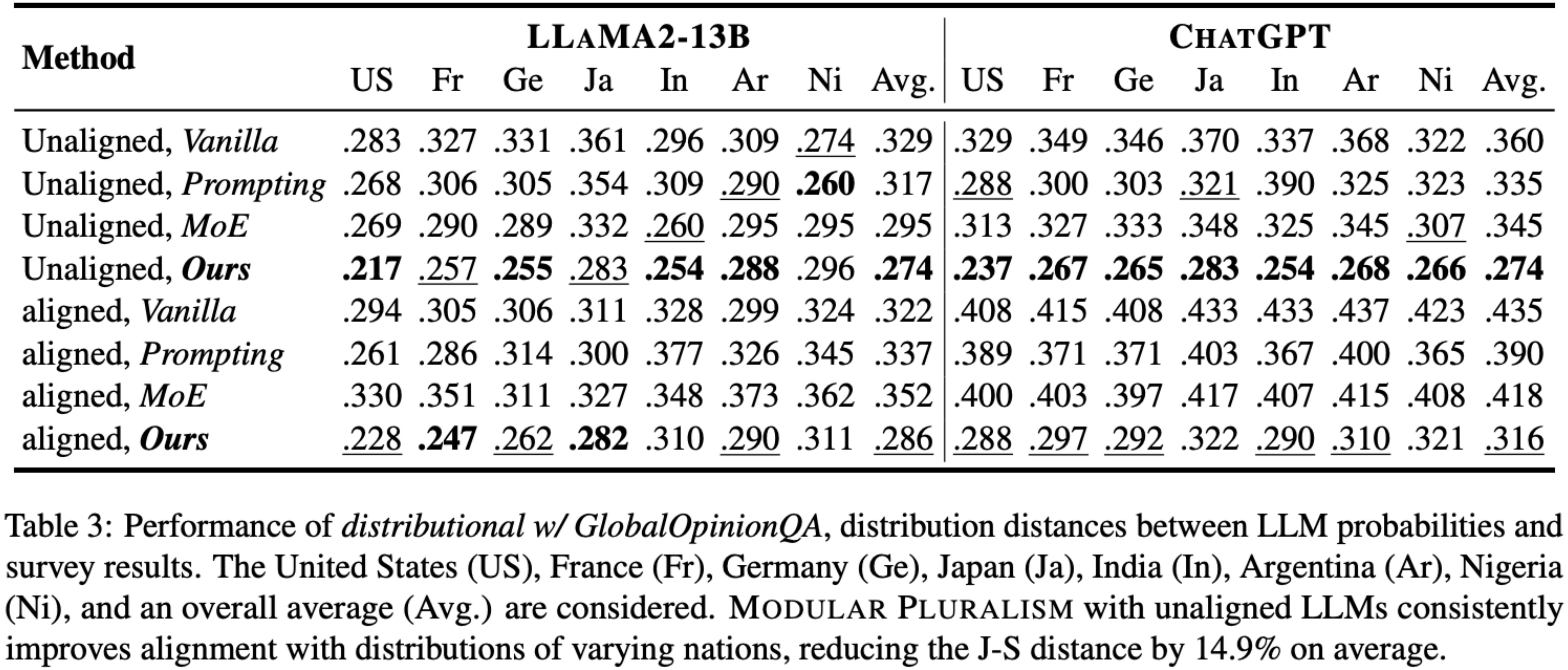
Results 5: Distributional (MoralChoice)

- MoralChoice: Some high-ambiguity situations where people disagree, some low-ambiguity situations where all agree
- Target distributions: uniform for ambiguous situations, concentrated on the "right" answer for the question



Results 6: Distributional (GlobalOpinionQA)

- Match country distribution



Method	LLaMA2-13B								ChatGPT							
	US	Fr	Ge	Ja	In	Ar	Ni	Avg.	US	Fr	Ge	Ja	In	Ar	Ni	Avg.
Unaligned, *Vanilla*	.283	.327	.331	.361	.296	.309	.274	.329	.329	.349	.346	.370	.337	.368	.322	.360
Unaligned, *Prompting*	.268	.306	.305	.354	.309	.290	**.260**	.317	.288	.300	.303	.321	.390	.325	.323	.335
Unaligned, *MoE*	.269	.290	.289	.332	.260	.295	.295	.295	.313	.327	.333	.348	.325	.345	.307	.345
Unaligned, ***Ours***	**.217**	.257	**.255**	.283	**.254**	**.288**	.296	**.274**	**.237**	**.267**	**.265**	**.283**	**.254**	**.268**	**.266**	**.274**
aligned, *Vanilla*	.294	.305	.306	.311	.328	.299	.324	.322	.408	.415	.408	.433	.433	.437	.423	.435
aligned, *Prompting*	.261	.286	.314	.300	.377	.326	.345	.337	.389	.371	.371	.403	.367	.400	.365	.390
aligned, *MoE*	.330	.351	.311	.327	.348	.373	.362	.352	.400	.403	.397	.417	.407	.415	.408	.418
aligned, ***Ours***	.228	**.247**	.262	**.282**	.310	.290	.311	.286	.288	.297	.292	.322	.290	.310	.321	.316

Table 3: Performance of *distributional w/ GlobalOpinionQA*, distribution distances between LLM probabilities and survey results. The United States (US), France (Fr), Germany (Ge), Japan (Ja), India (In), Argentina (Ar), Nigeria (Ni), and an overall average (Avg.) are considered. Modular Pluralism with unaligned LLMs consistently improves alignment with distributions of varying nations, reducing the J-S distance by 14.9% on average.

What if we underrepresent certain perspectives?...

- Patching: Train additional CommunityLM for underrepresented community

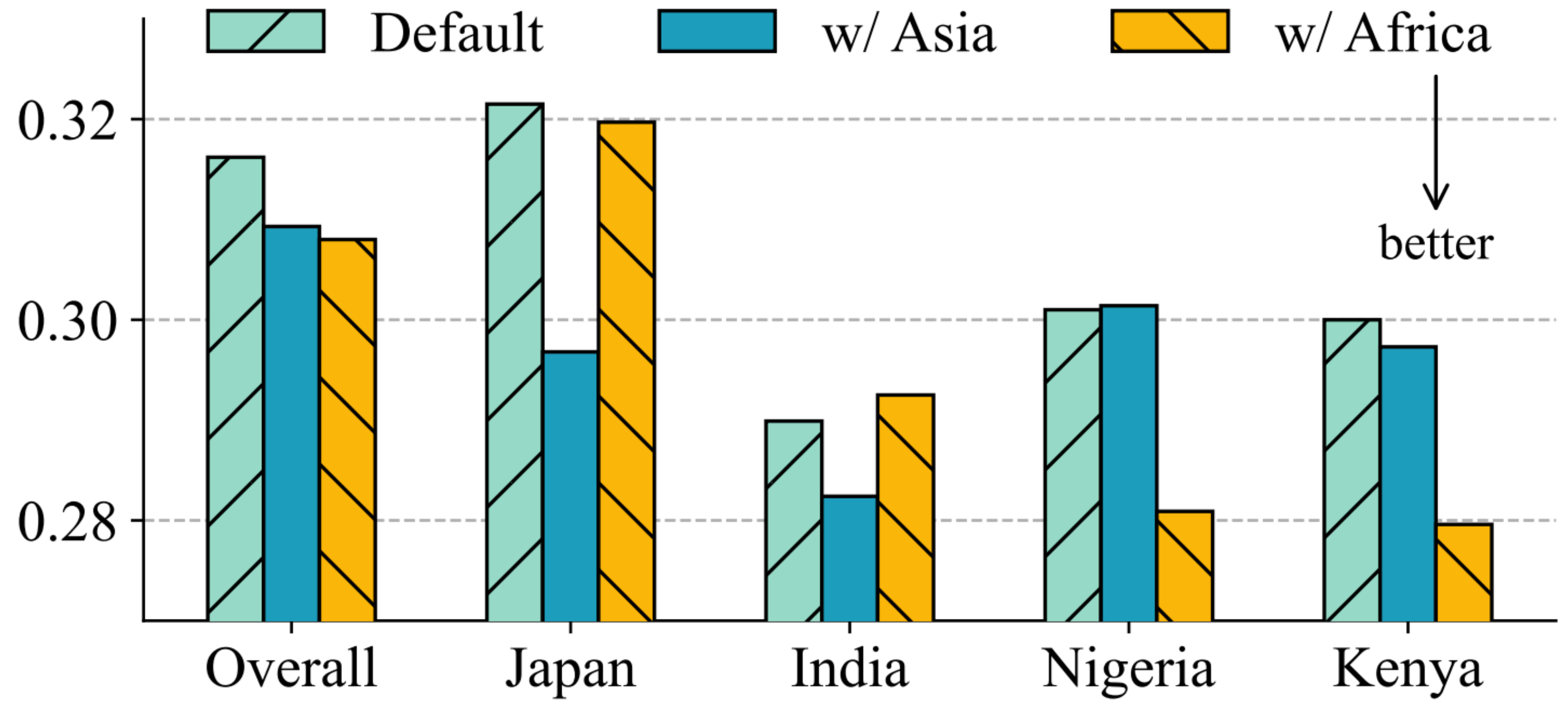


Figure 6: J-S distance on GlobalOpinionQA when one extra community LM representing Asian and African culture is separately added to the pool of perspective-informed community LMs, *the lower the better*. This helps patch LLMs' pluralism gaps by improving alignment towards underrepresented communities.

Modular Pluralism

- Contributions:
  - Multi-LLM framework for pluralism with small, specialist LLMs
    - Patchable and somewhat interpretable
  - Concrete evaluations for pluralism
- Limitations:
  - Greater computational cost
  - Requires representative corpora for communities

Next, a resource for pluralism…

UNIVERSITY of WASHINGTON

PAUL G. ALLEN SCHOOL

Carnegie Mellon University

UNIVERSITY OF OXFORD

AAAI-24

Kaleido

Value Kaleidoscope: Engaging AI with Pluralistic Human *Values*, *Rights*, and *Duties*

Taylor Sorensen

Liwei Jiang

Jena Hwang

Sydney Levine

Valentina Pyatkin

Peter West

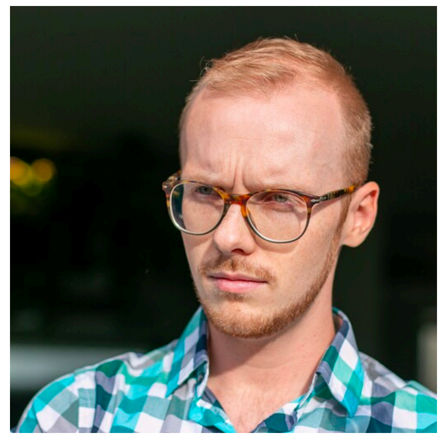
Nouha Dziri

Ximing Lu

Kavel Rao

Chandra Bhagavatula

Maarten Sap



John Tasioulas

Yejin Choi

Goals (at time of writing, 2023)

A. What pluralistic human *values*, *rights*, and *duties* are already present in large language models?

B. Can we create better datasets/models that take into account *value pluralism*?

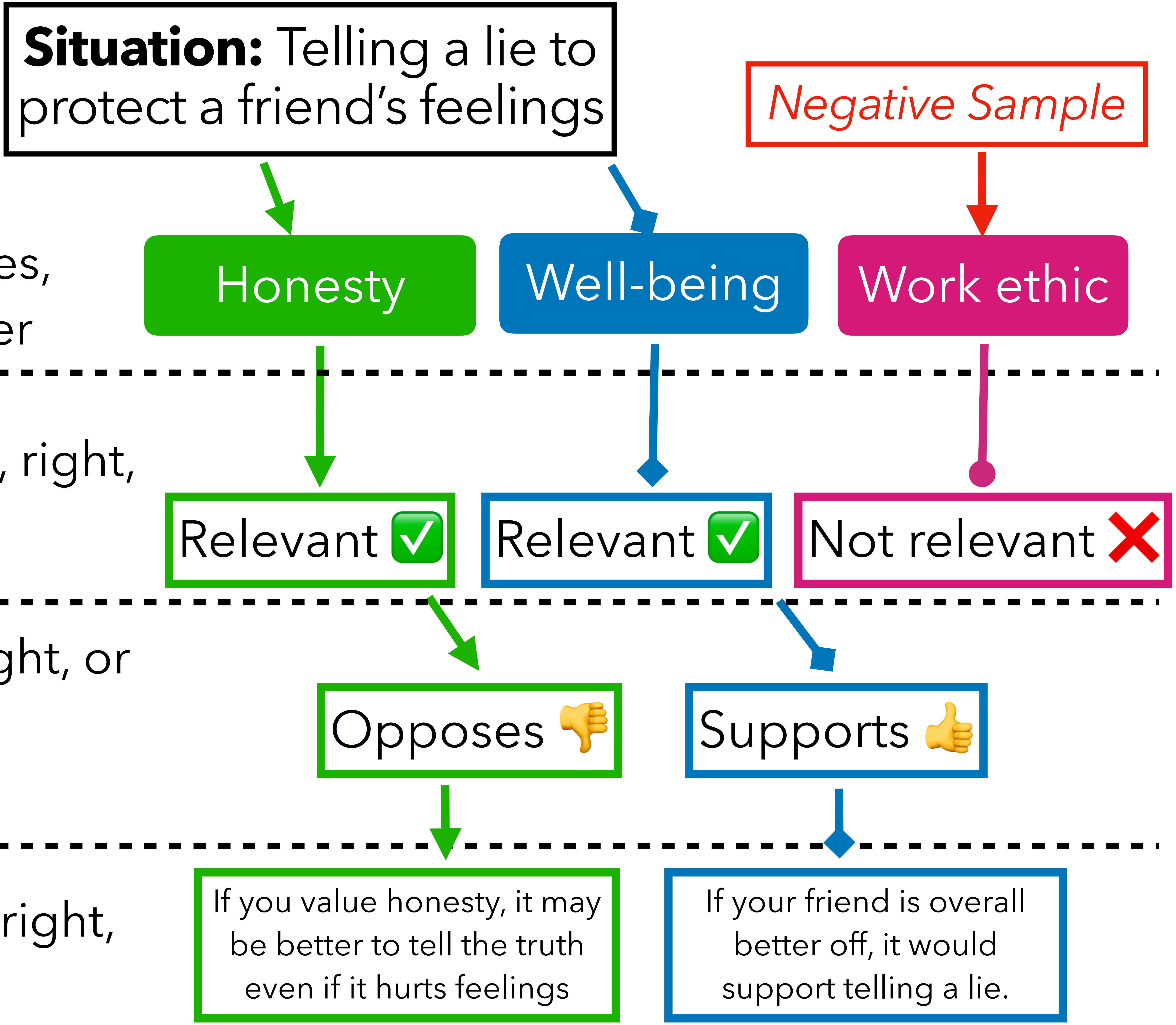
Connections (post-roadmap, 2024)

A. Can we create a dataset that can be used for evaluating different forms of pluralism?
B. Can we create a model that could be used as a value-specific reward (for e.g. steerable pluralism)?

Tasks

Given a situation:

1. **Generation:** Generate values, rights, and duties to consider
2. **Relevance:** Is a given value, right, or duty relevant?
3. **Valence:** Does the value, right, or duty support or oppose the situation?
4. **Explanation:** How is value, right, or duty connected?



Situation: Telling a lie to protect a friend's feelings

Negative Sample

Honesty

Well-being

Work ethic

Relevant ✅

Relevant ✅

Not relevant ❌

Opposes 👎

Supports 👍

If you value honesty, it may be better to tell the truth even if it hurts feelings

If your friend is overall better off, it would support telling a lie.

ValuePrism - Dataset

30k User-submitted Situations

Situation:
Going 50 mph over the speed limit to get my wife to a hospital

Large, Closed-Source Model (GPT-4)

Values:
- Safety: opposes 👎
- Well-being: supports 👍
- Respect for the law: opposes 👎

Rights:
- Right to access healthcare: supports 👍
- Right to safety: opposes 👎

Duties:
- Duty to protect one's family: supports 👍
- Duty to obey the law: opposes 👎
- Duty to drive responsibly: opposes 👎

ValuePrism - Dataset

30k User-submitted Situations

Situation: Going 50 mph over the speed limit to get my wife to a hospital

Large, Closed-Source Model (GPT-4)

Values:
- Safety: oppose
- **Well-being**: supp
- Respect for the la

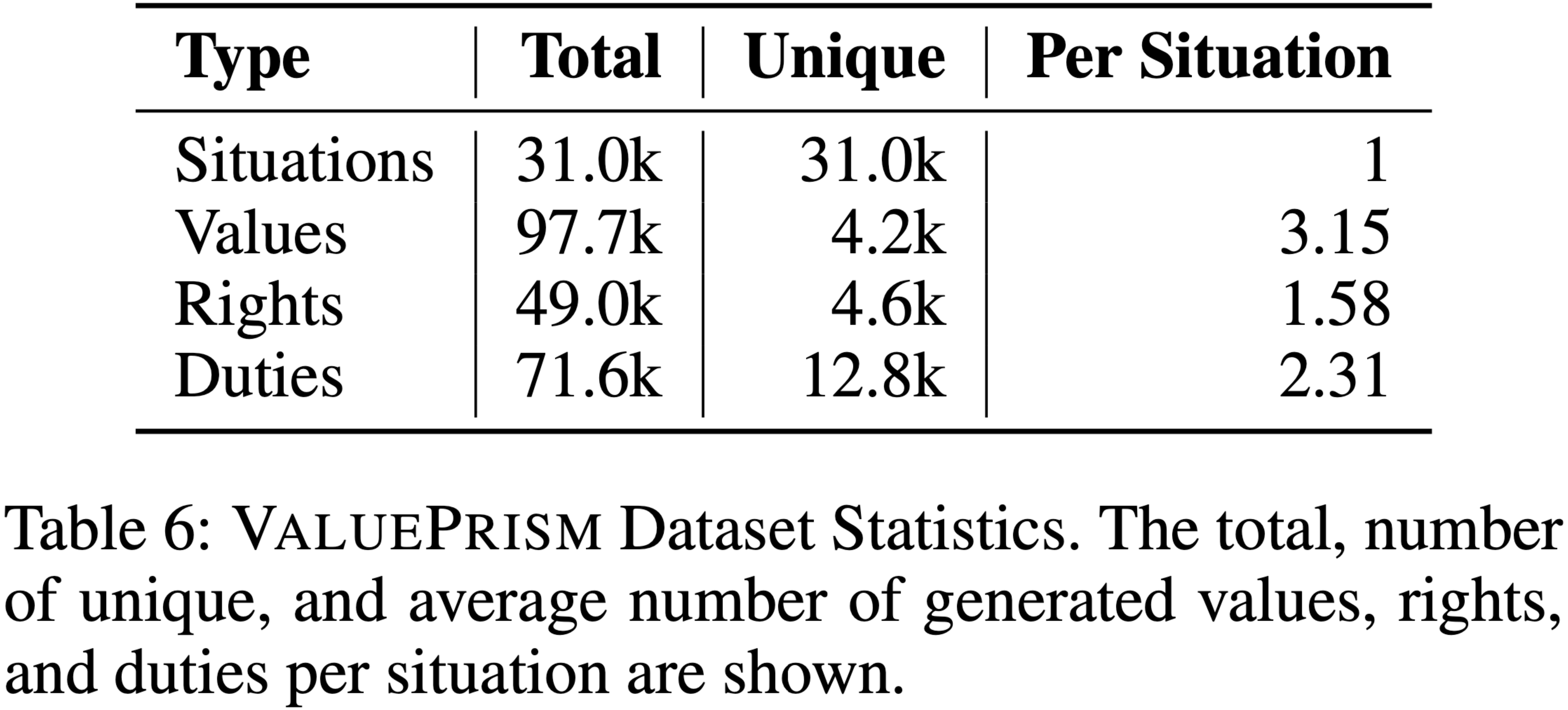
Why? In this situation, the wife may require urgent medical attention, and getting her to the hospital quickly could be crucial for her well-being

Rights:
- Right to access healthcare: supports
- **Right to safety**: opposes

Duties:
- Duty to prote
- Duty to obey
- Duty to drive

Why? Other drivers and pedestrians have the right not to be endangered by reckless and dangerous driving.

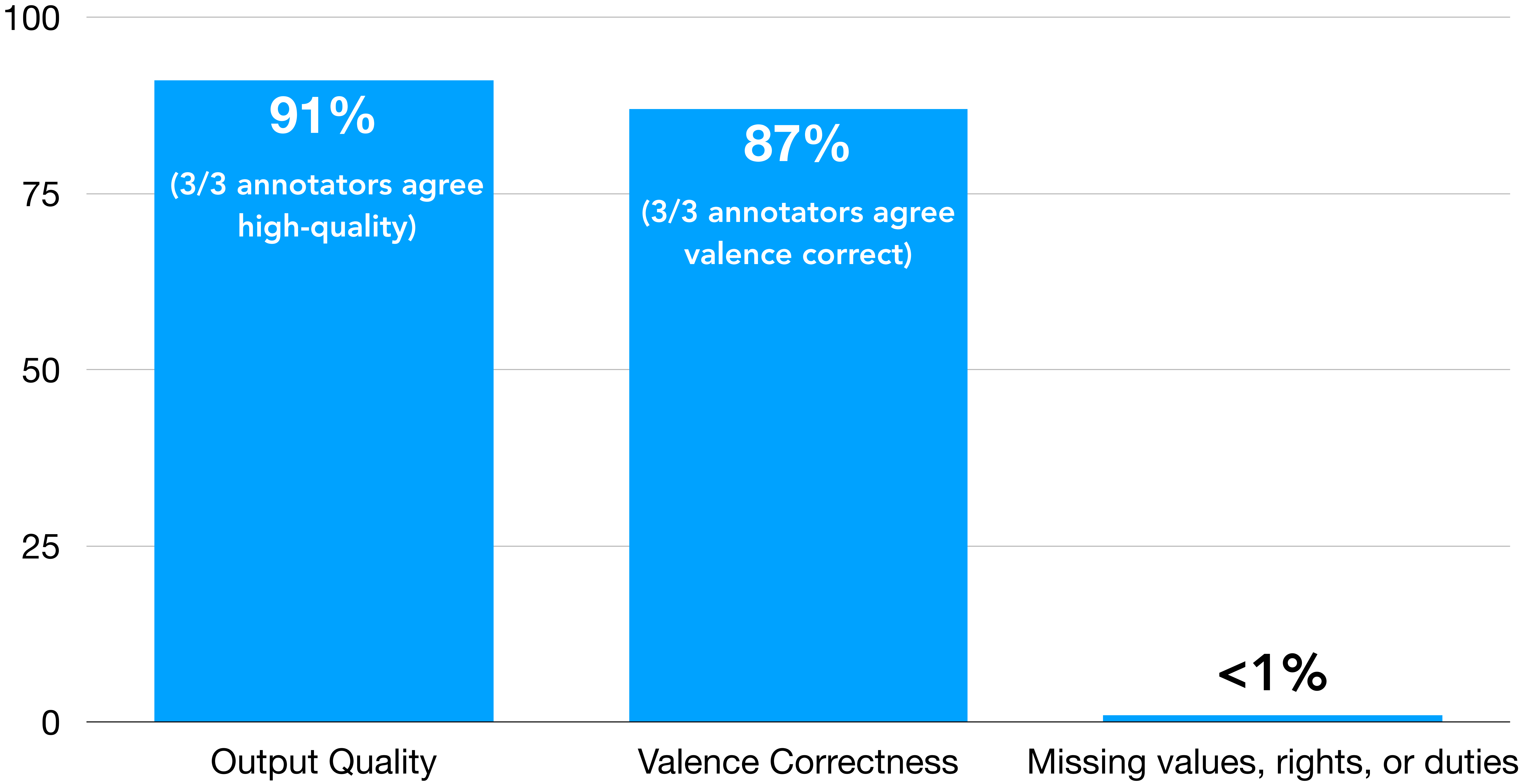
ValuePrism - Statistics



Type	**Total**	**Unique**	**Per Situation**
Situations	31.0k	31.0k	1
Values	97.7k	4.2k	3.15
Rights	49.0k	4.6k	1.58
Duties	71.6k	12.8k	2.31

Table 6: VALUEPRISM Dataset Statistics. The total, number of unique, and average number of generated values, rights, and duties per situation are shown.

ValuePrism is high-quality



Whose values are represented?

- Study with 613 people from diverse backgrounds
  A. Do you agree with the value, right, or duty?
  B. Is your perspective missing?

e.g., Race: 168 white, 115 Black, 61 asian, 34 hispanic/latinx; Sexual orientation: 390 straight, 68 LGBQ+. Gender: 258 male, 201 female, 9 non-binary or other

Most values were largely agreed upon

Situation:
Frowning at a friend

Respect: Not frowning at a friend if the situation doesn't warrant it could be a way to respect their feelings

83% overall agreement

Groups differed on a few values

Situation: redistributing rich people's land to poor people

Efficiency: Redistribution may lead to more efficient land use if previously underutilized land is given to those in need.

🫏 **Liberals** 78% more likely to agree than Conservatives 🐘

Situation: giving people things for free

Personal Responsibility: Some may argue that individuals should earn what they receive, and providing things for free may undermine this value.

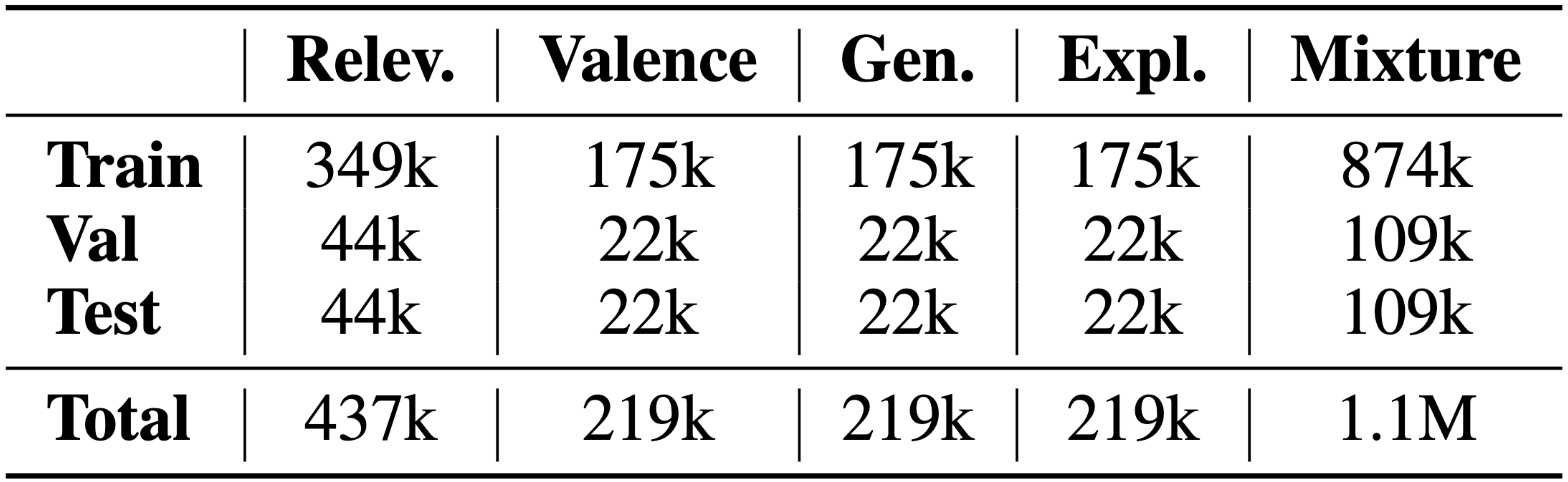
🐘 **Conservatives** 63% more likely to agree than Liberals 🫏

Whose values are represented?

- Most people agreed on most values
- Did not find significant differences between groups' overall agreement rates

Model: Value Kaleidoscope

- Train a T5-based sequence to sequence model on ValuePrism
- Can *generate*, *explain*, and predict *relevance* and *valence*



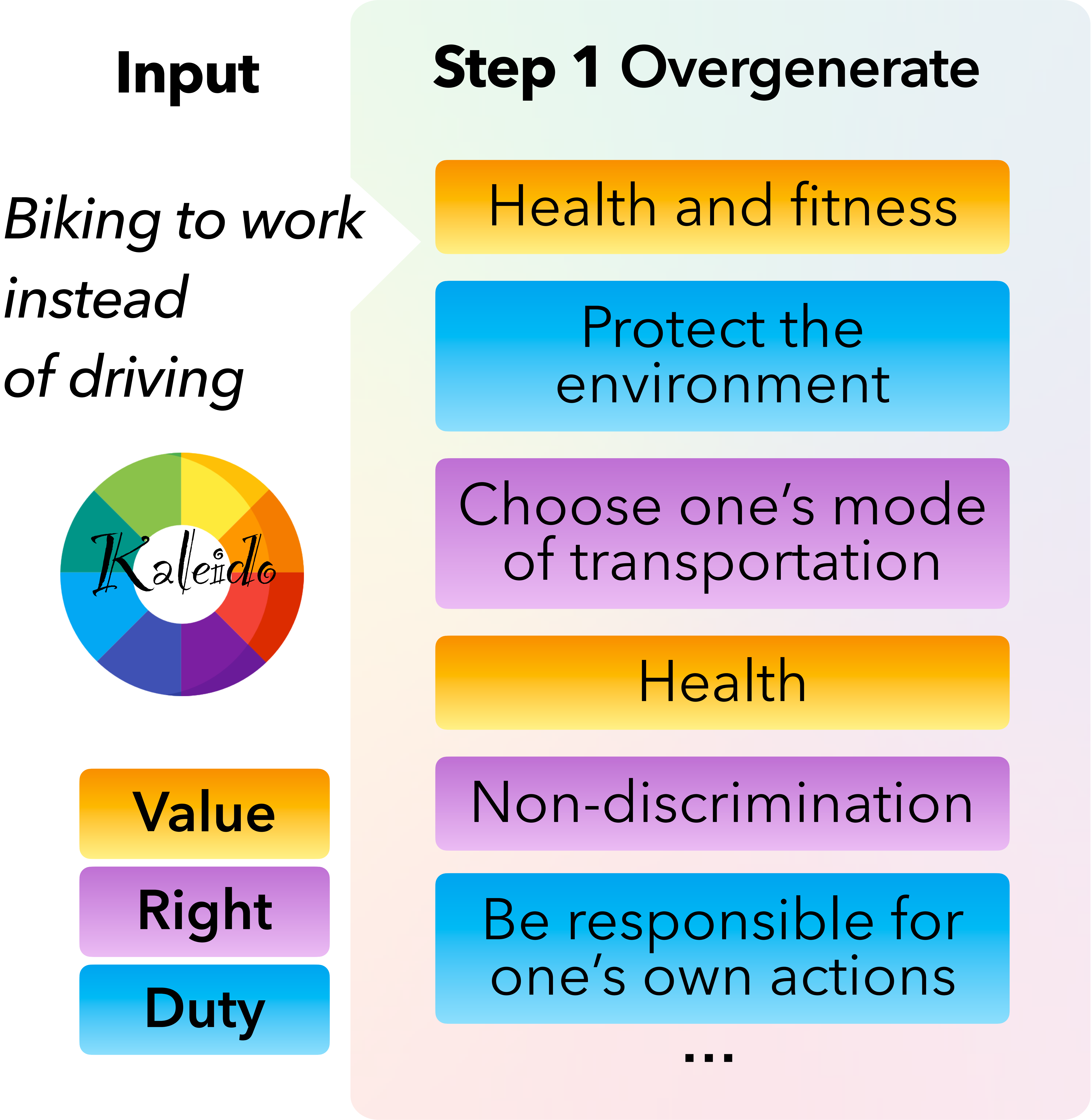
	Relev.	**Valence**	**Gen.**	**Expl.**	**Mixture**
Train	349k	175k	175k	175k	874k
Val	44k	22k	22k	22k	109k
Test	44k	22k	22k	22k	109k
Total	437k	219k	219k	219k	1.1M

Table 7: Task Dataset Statistics

Kaleido System

- System to generate batch of pluralistic values, rights, and duties

Kaleido System



Input

Biking to work instead of driving

Value

Right

Duty

Step 1 Overgenerate

Health and fitness

Protect the environment

Choose one's mode of transportation

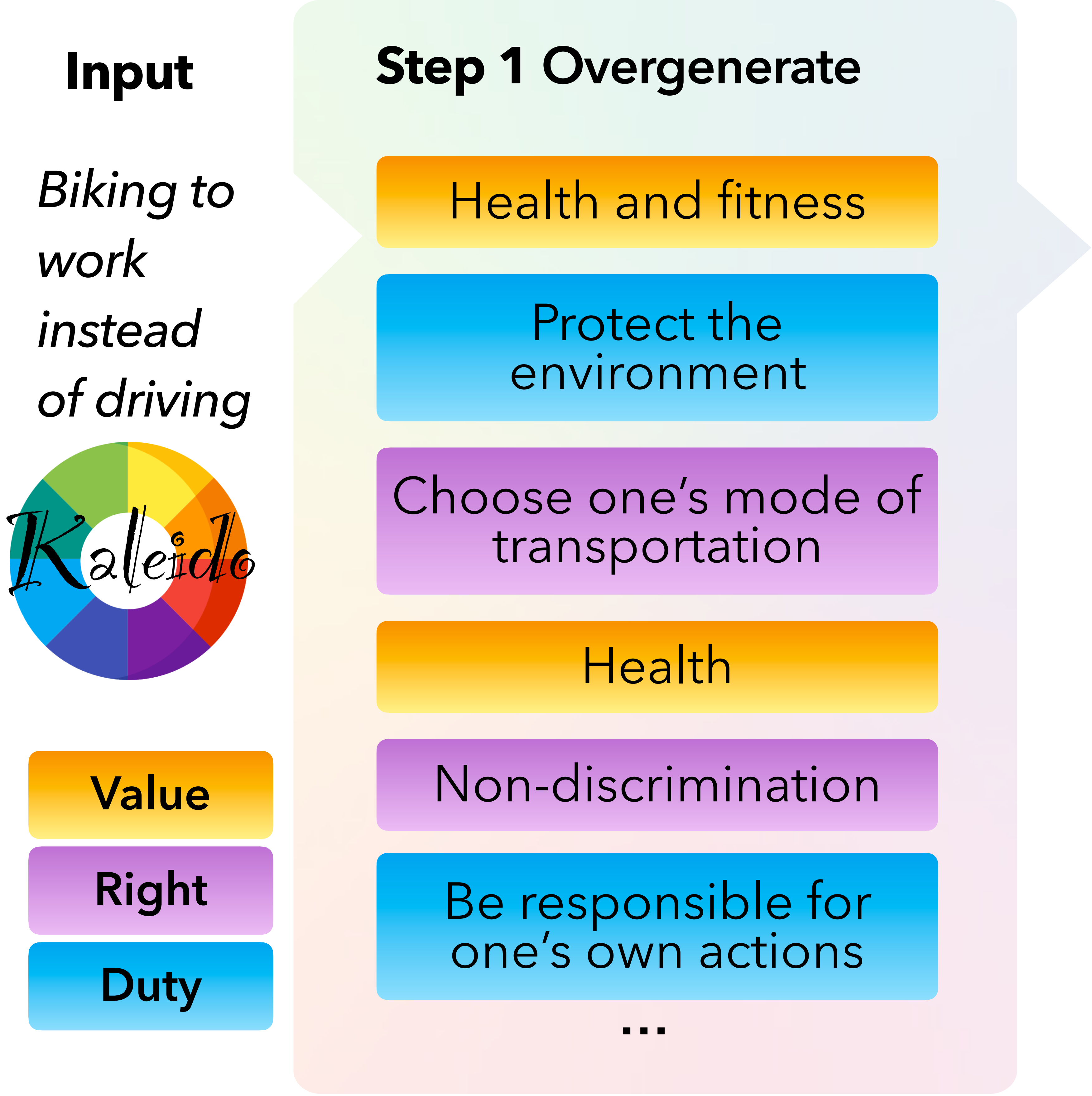
Health

Non-discrimination

Be responsible for one's own actions

...

Kaleido System



Input

Biking to work instead of driving

Kaleido

Value

Right

Duty

Step 1 Overgenerate

Health and fitness

Protect the environment

Choose one’s mode of transportation

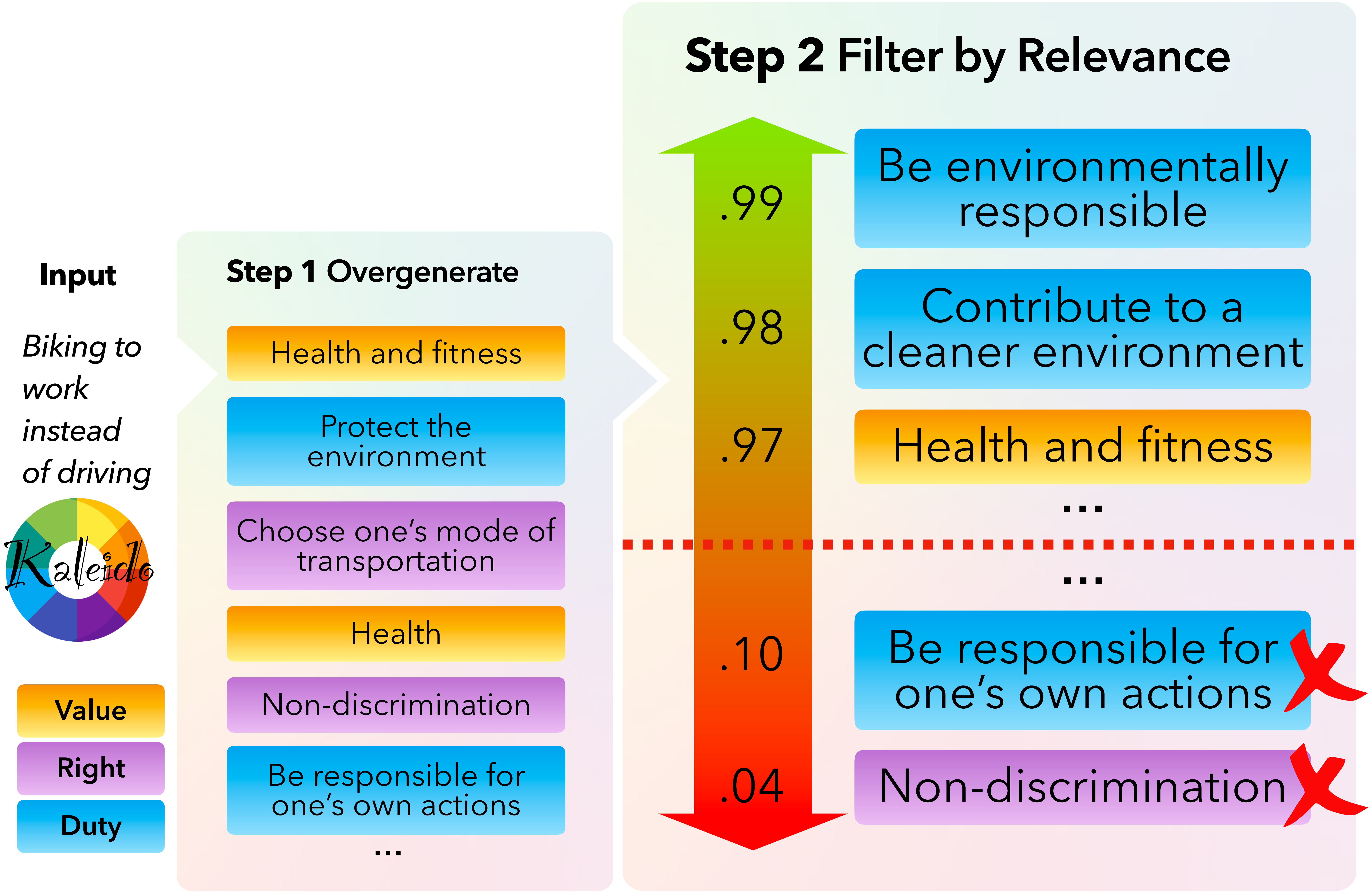
Health

Non-discrimination

Be responsible for one’s own actions

...

Kaleido System



Input

Biking to work instead of driving

Kaleido

Value

Right

Duty

Step 1 Overgenerate

Health and fitness

Protect the environment

Choose one's mode of transportation

Health

Non-discrimination

Be responsible for one's own actions

...

Step 2 Filter by Relevance

.99 Be environmentally responsible

.98 Contribute to a cleaner environment

.97 Health and fitness

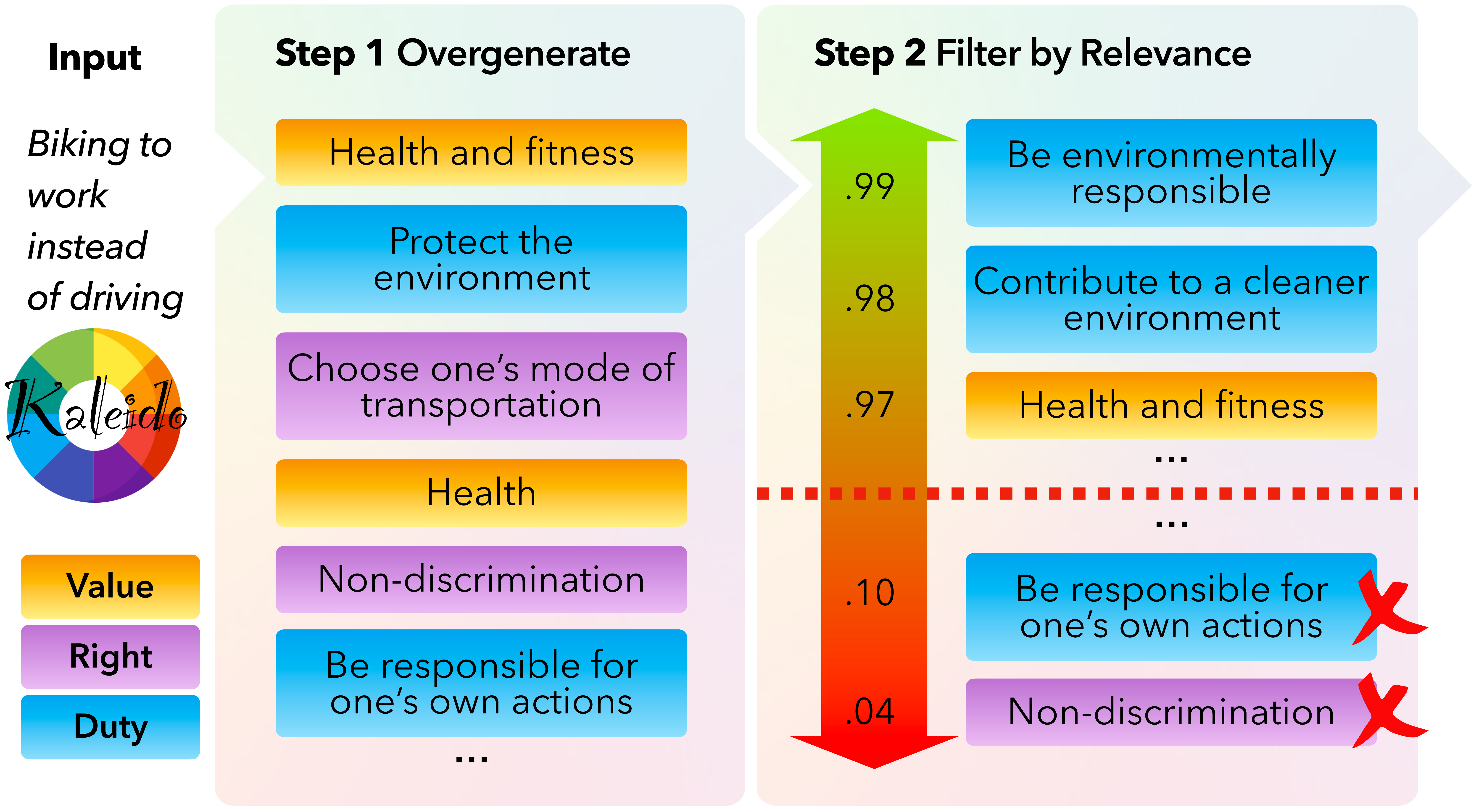
...

...

.10 Be responsible for one's own actions ✗

.04 Non-discrimination ✗

Kaleido System



Input

Biking to work instead of driving

Kaleido

Value

Right

Duty

Step 1 Overgenerate

Health and fitness

Protect the environment

Choose one's mode of transportation

Health

Non-discrimination

Be responsible for one's own actions

...

Step 2 Filter by Relevance

.99 Be environmentally responsible

.98 Contribute to a cleaner environment

.97 Health and fitness

...

...

.10 Be responsible for one's own actions ✗

.04 Non-discrimination ✗

Kaleido System

Input

Biking to work instead of driving

Kaleido

Value

Right

Duty

Step 1 Overgenerate

Health and fitness

Protect the environment

Choose one's mode of transportation

Health

Non-discrimination

Be responsible for one's own actions

...

Step 2 Filter by Relevance

.99 Be environmentally responsible

.98 Contribute to a cleaner environment

.97 Health and fitness

...

...

.10 Be responsible for one's own actions ✗

.04 Non-discrimination ✗

Step 3 Deduplicate by text similarity

Be environmentally responsible

✓ *Similarity 0.15*

✗ *Similarity 0.94*

Health and fitness

Contribute to a cleaner environment

Kaleido System

Input

Biking to work instead of driving

Value

Right

Duty

Step 1 Over-generate

Step 2 Filter by Relevance

Step 3 Deduplicate by text similarity

✓
Be environmentally responsible
✗
Similarity 0.15
Health and fitness
Similarity 0.94
Contribute to a cleaner environment

Output

Output	Relevance	Support	Oppose	Either
Be environmentally responsible	.99	1	0	0
Health and fitness	.94	1	0	0
Convenience	.97	0	.84	.16
Choose one’s mode of transportation	.96	.27	.01	.72

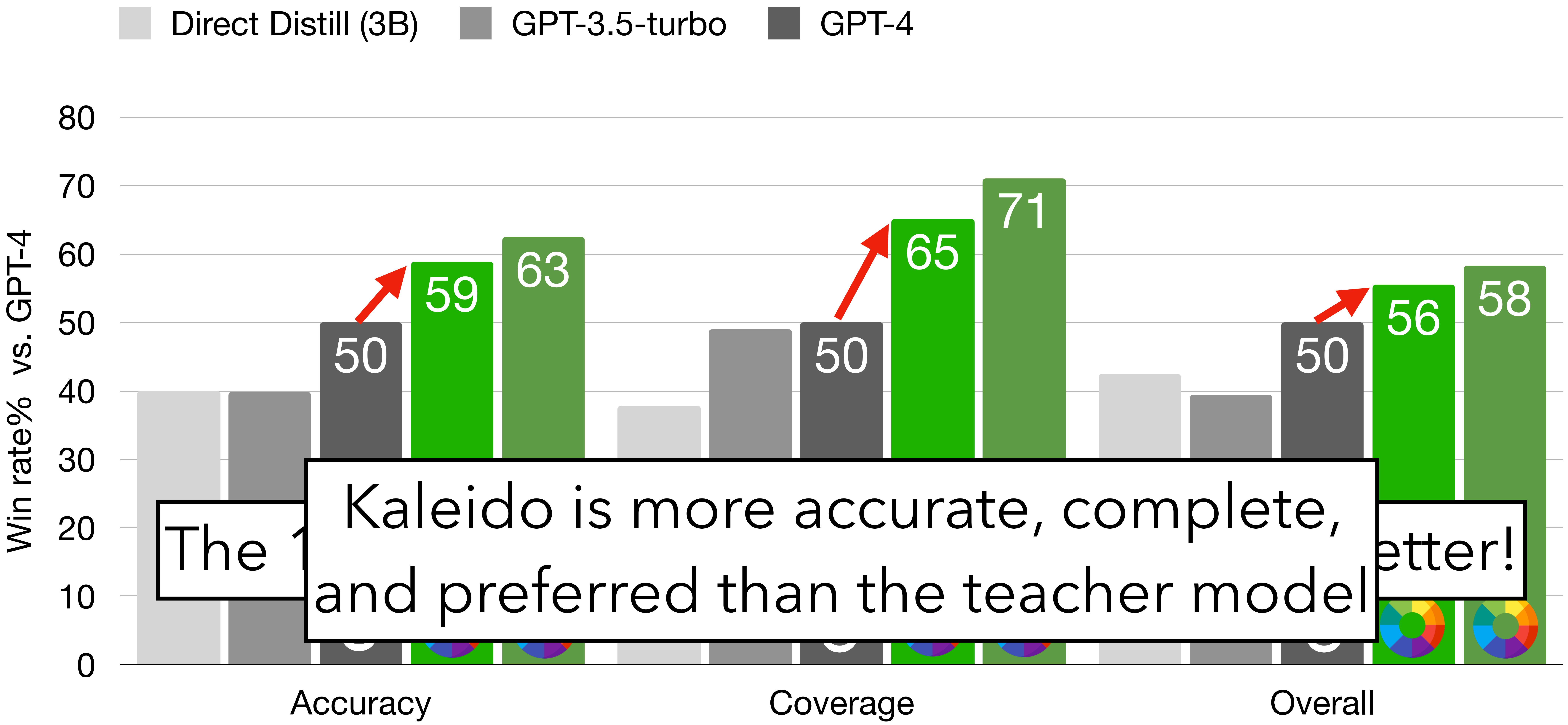
Evaluating Outputs

A batch of values, rights, and duties should:

- Be accurate
- Have broad coverage
- Be preferred by annotators

We compare Kaleido head to head with GPT-4!

Kaleido System vs. GPT-4 (Generation)

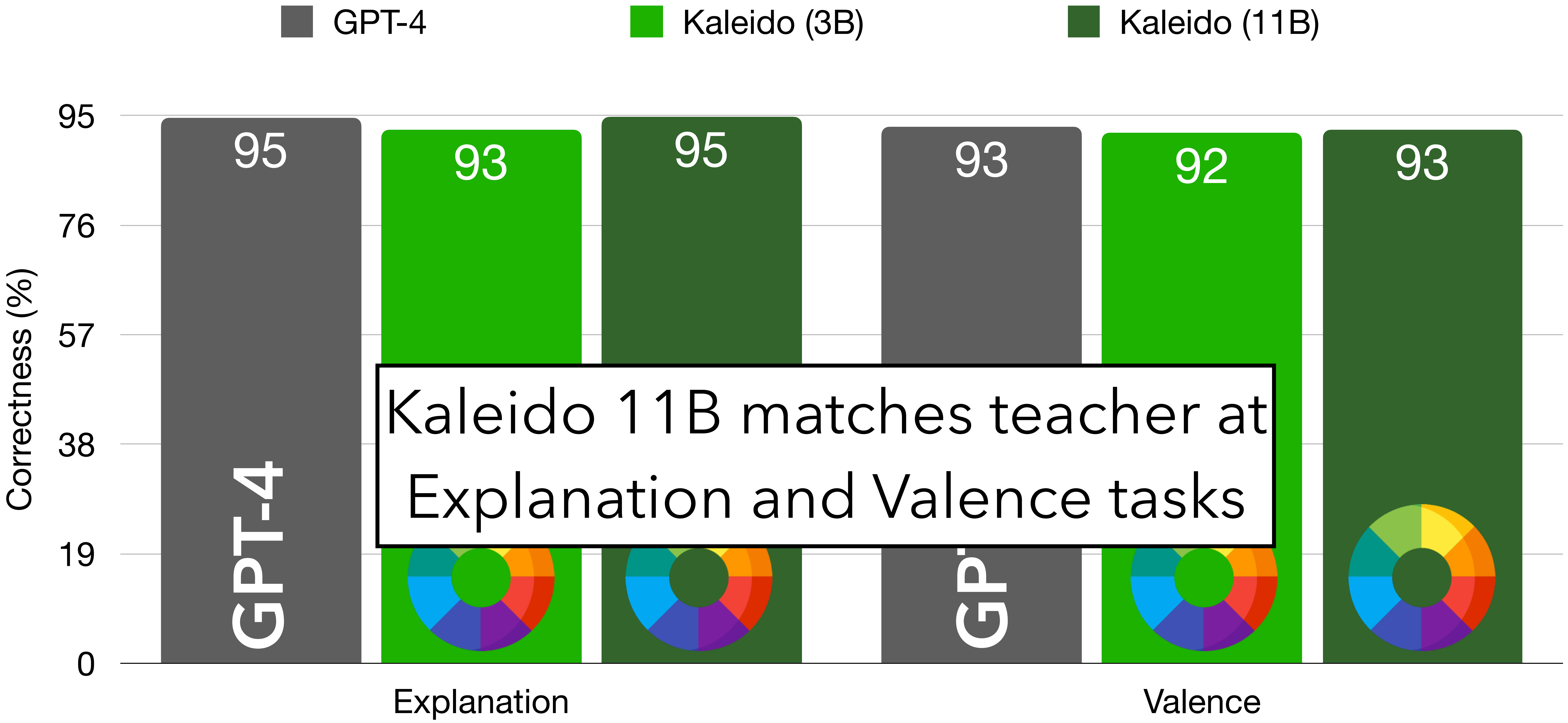


The 1

etter!

Kaleido is more accurate, complete, and preferred than the teacher model

KaleidoSys vs. GPT-4 (Explanation and Valence)

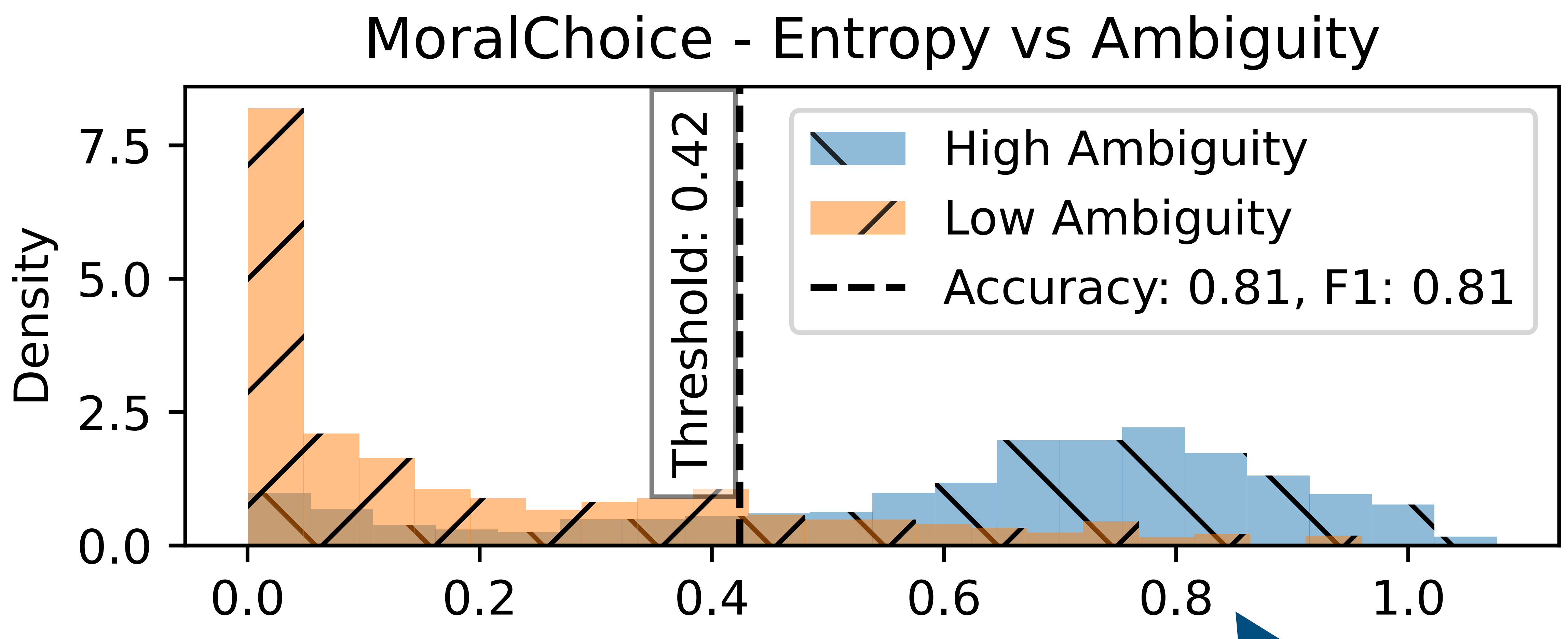


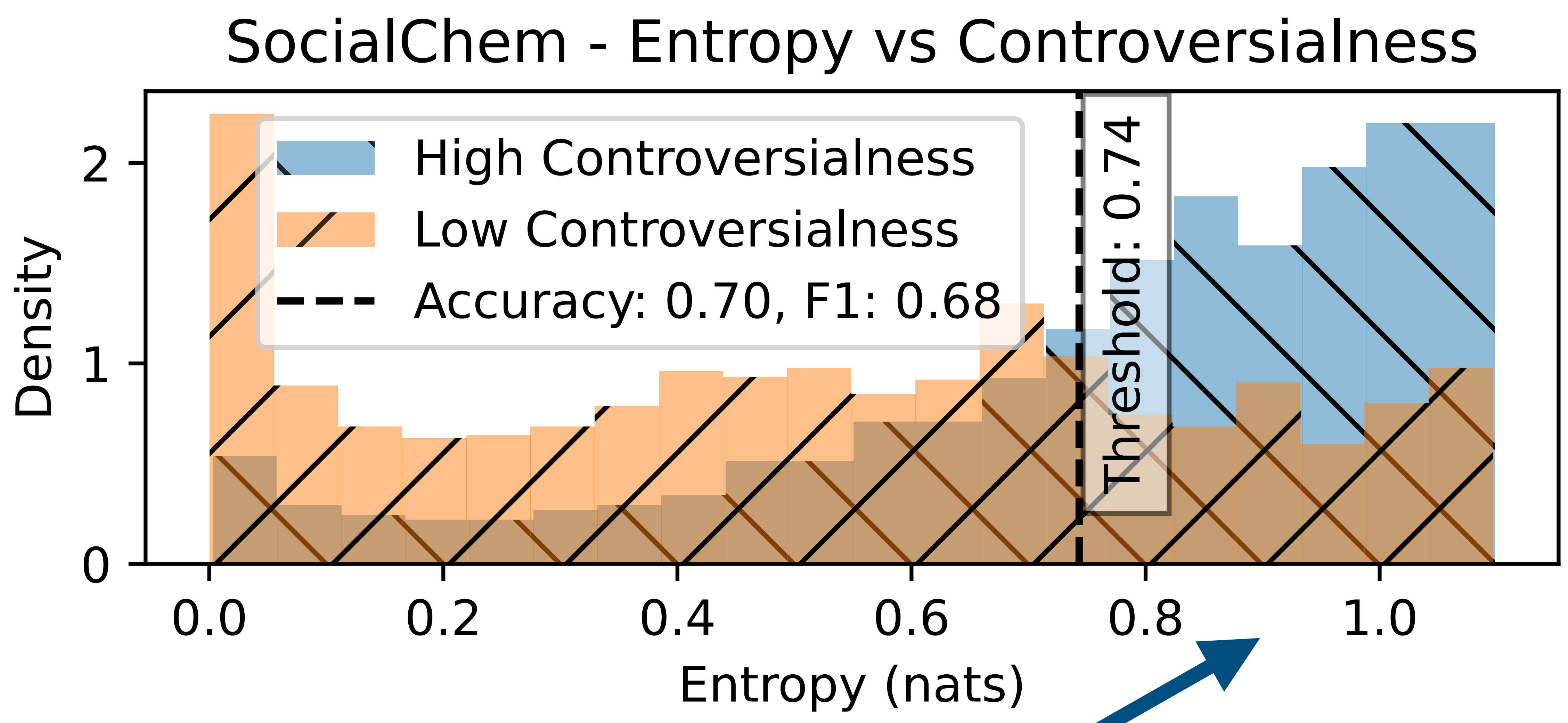
Kaleido 11B matches teacher at Explanation and Valence tasks

Does Kaleido help explain variation in human decision-making?

- Two datasets with variability ratings
- Hypothesis: Contrasting values => More variability

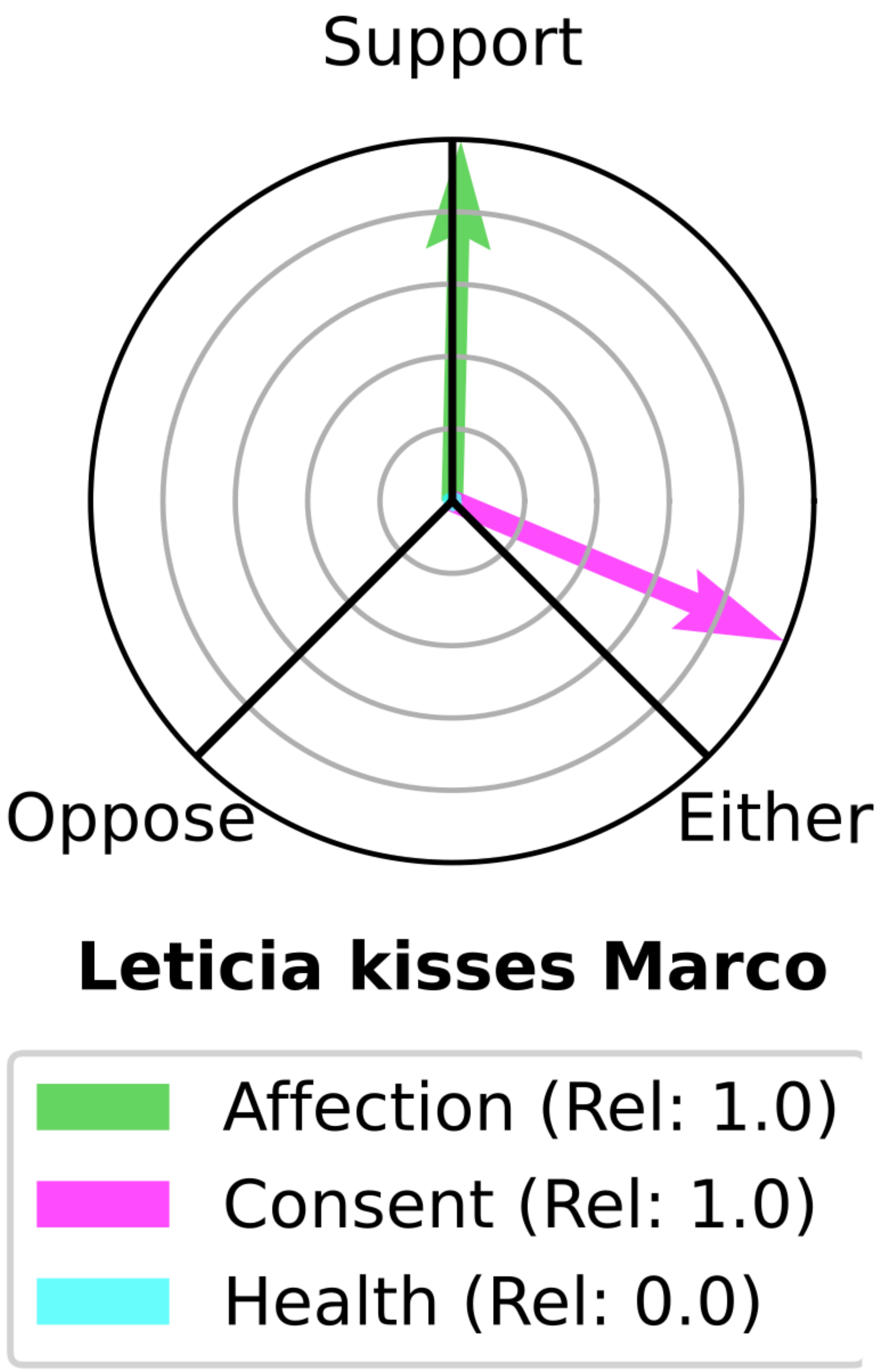
Kaleido's contrasting values help explain variability in human decision-making



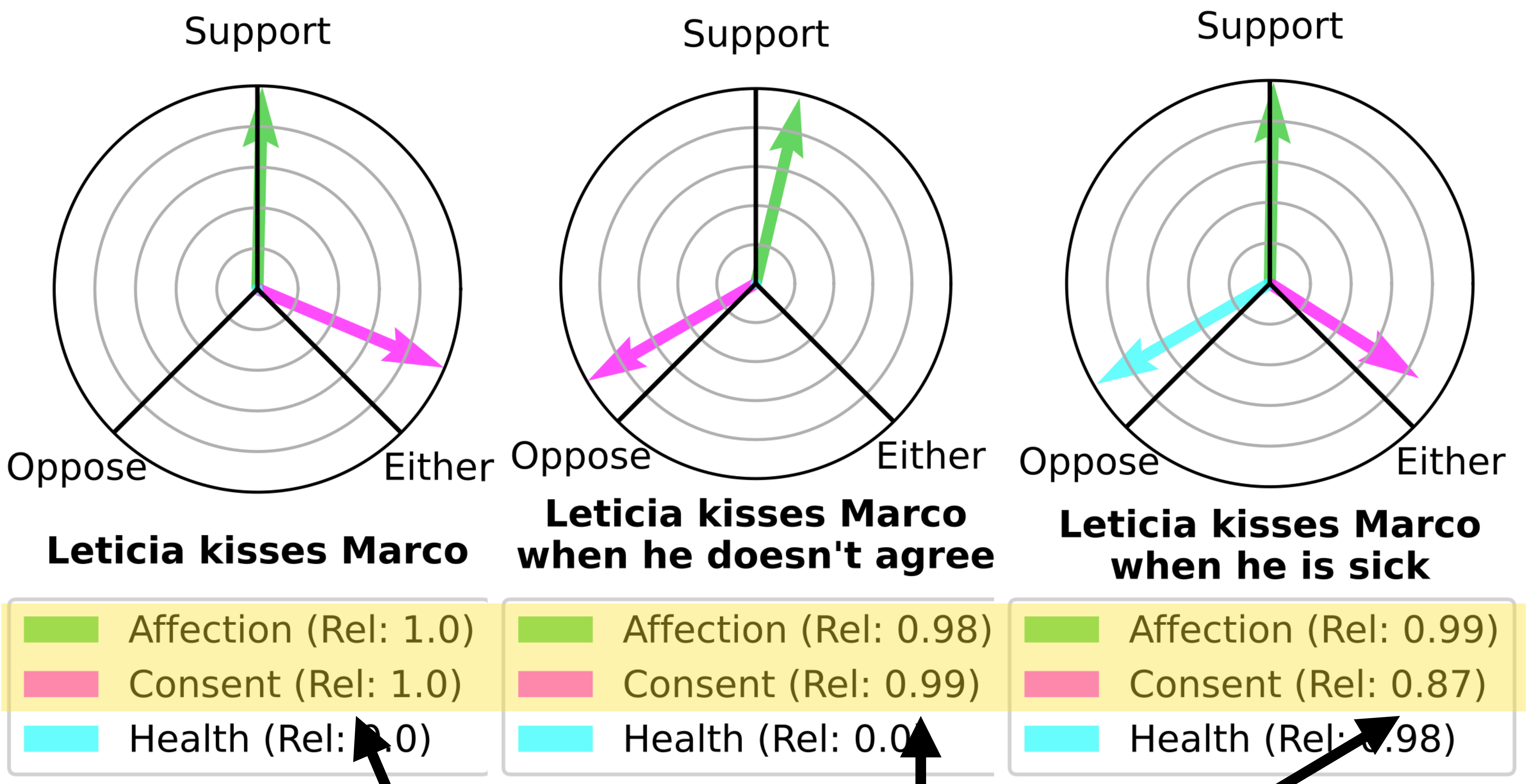


High entropy => More Variability

Kaleido is sensitive to variations



Kaleido is sensitive to variations



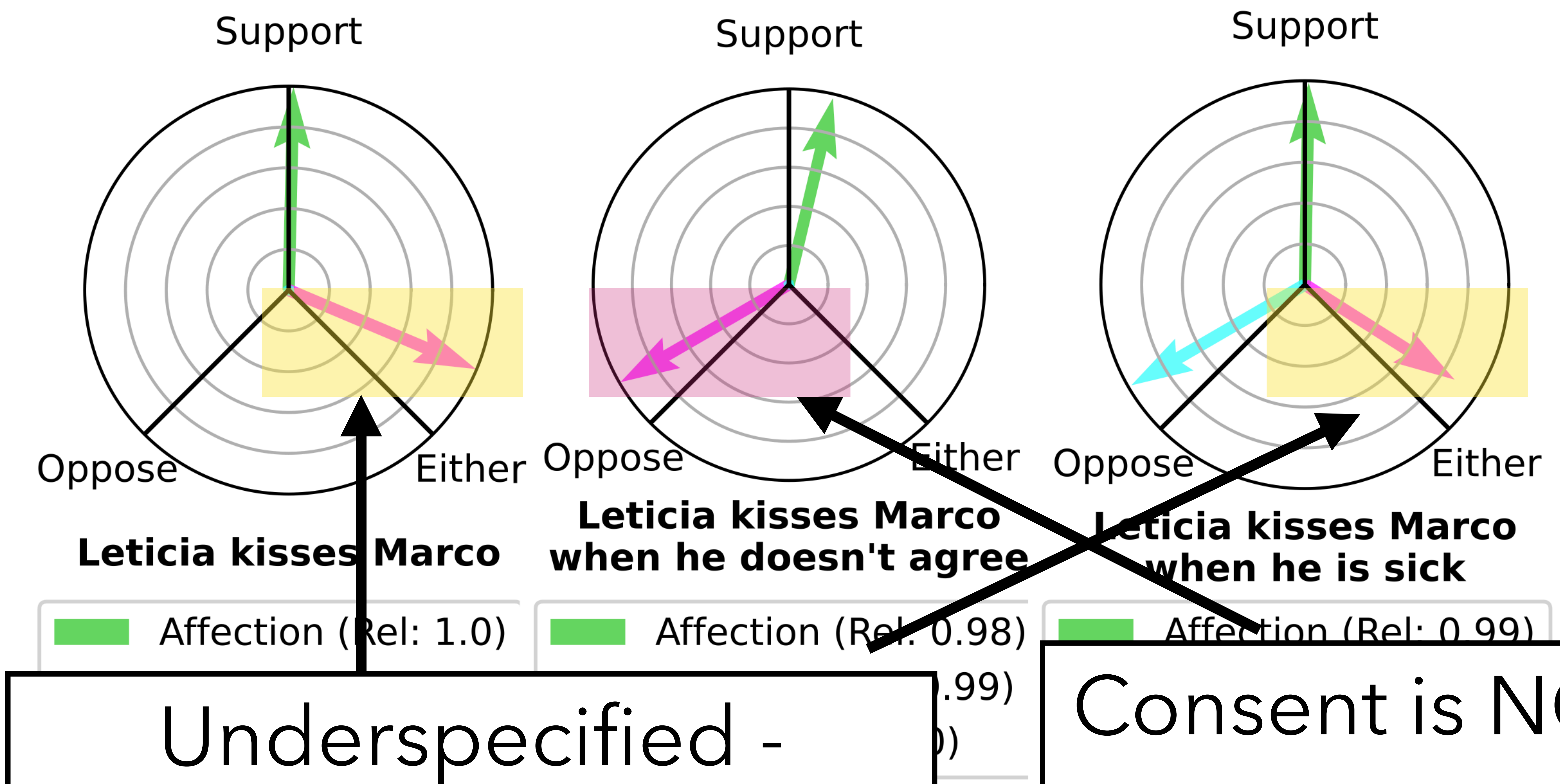
Leticia kisses Marco

Leticia kisses Marco when he doesn't agree

Leticia kisses Marco when he is sick

Affection and Consent are relevant for all three

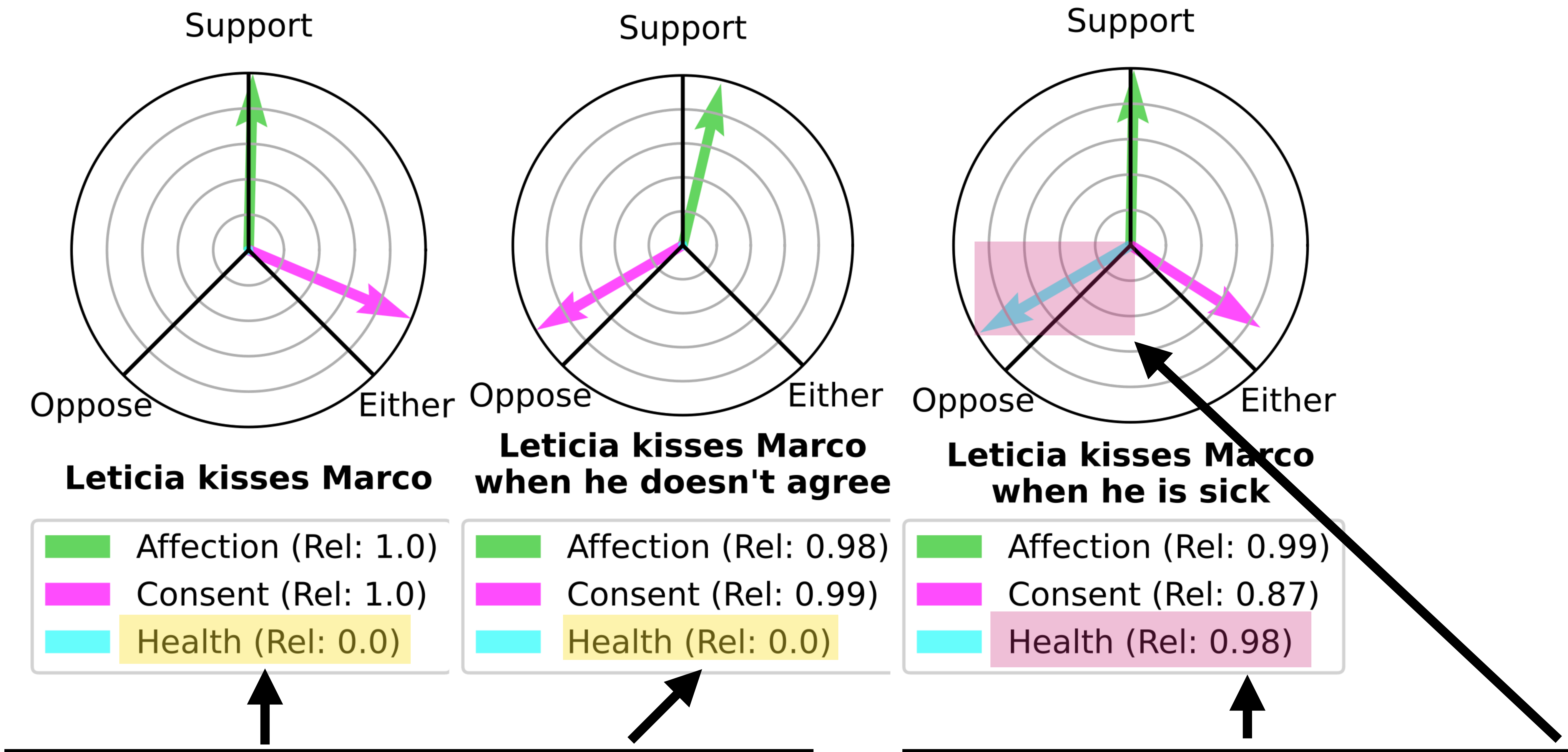
Kaleido is sensitive to variations



Underspecified - consent could either support or oppose

Consent is NOT given - Opposes

Kaleido is sensitive to variations



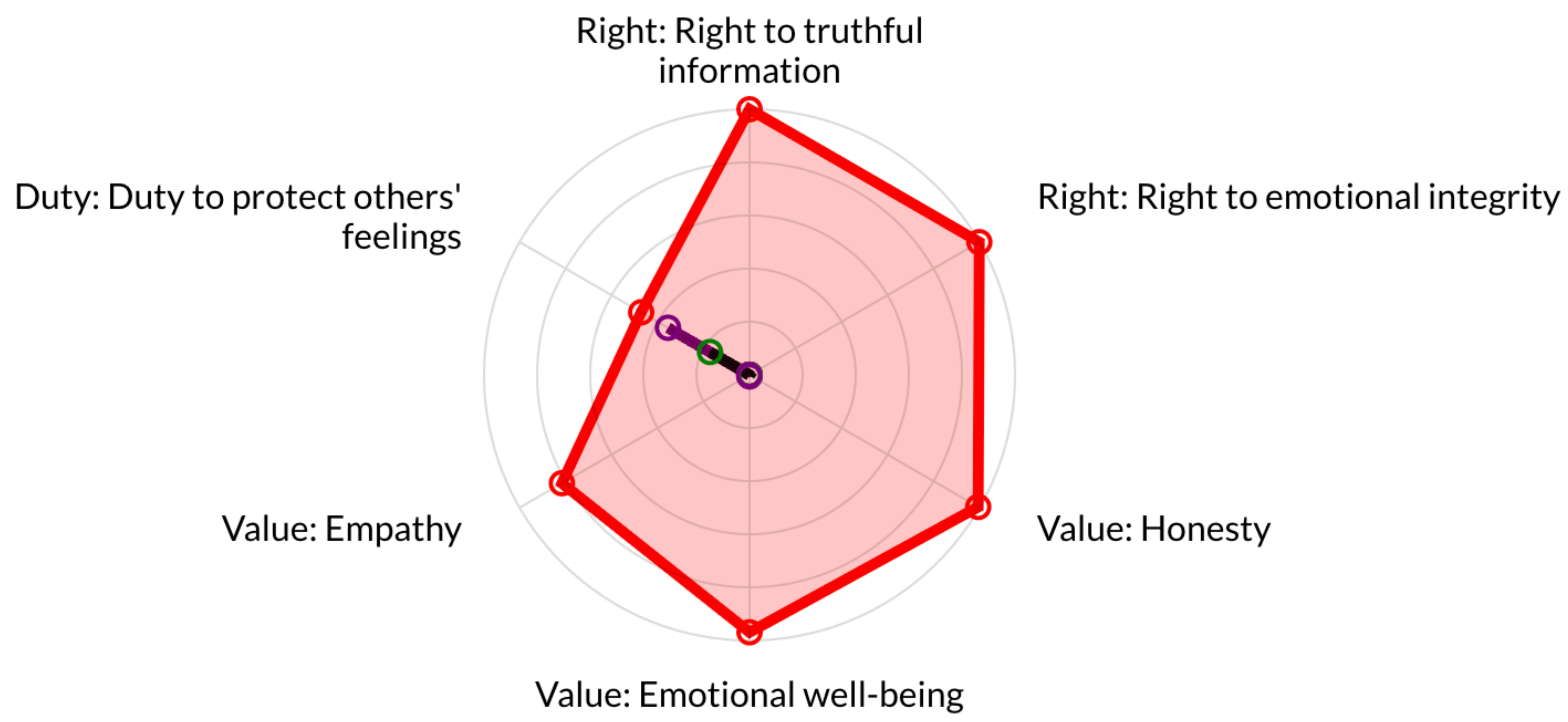
Health isn't normally the most relevant value…

But health IS relevant when one person is sick

False Balance ⚖️

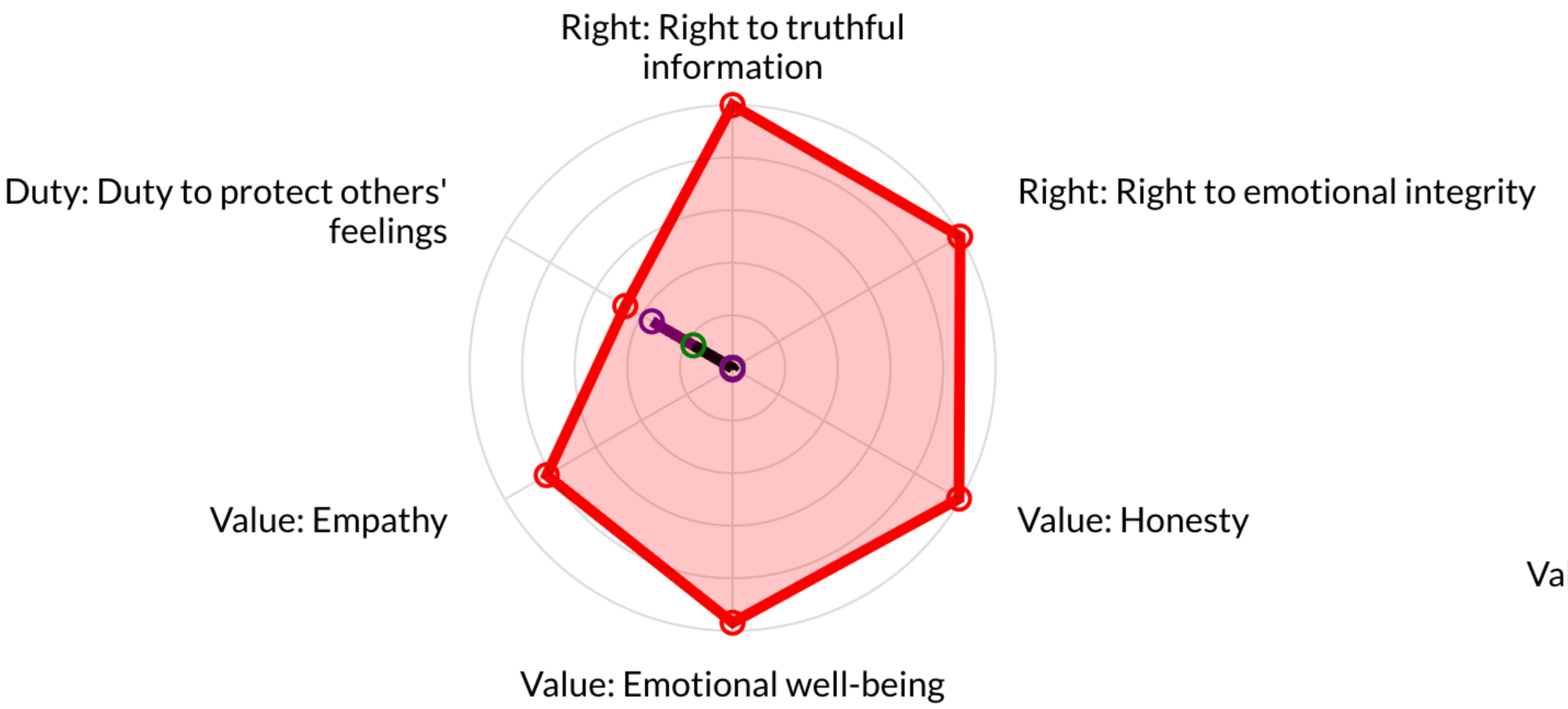
- ⚠️ Risk ⚠️: coming up with a contrived reason why something is good or bad
- Develop 20 “good” and “bad” actions for which could not come up with “opposing” or “supporting” values

False Balance ⚖️ - Examples



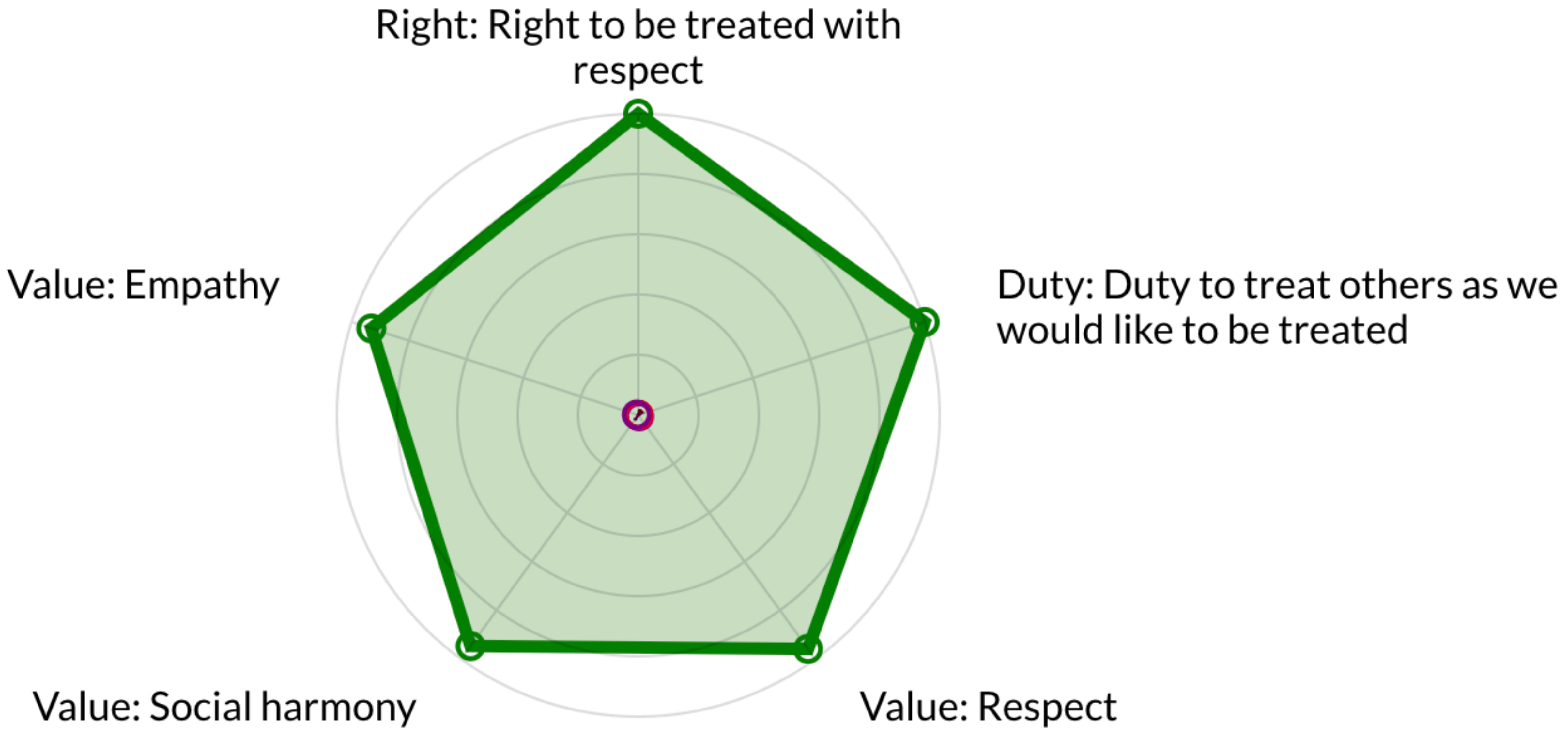
Bad Situation:
Lying to someone just to hurt them

False Balance ⚖️ - Examples



Bad Situation:
Lying to someone just to hurt them

Good Situation:
Treat others with respect



False Balance ⚖️

- Kaleido output 0 values supporting the bad actions and 0 values opposing the good actions in all cases ✅

Strengths over teacher 💪

In addition to beating the teacher at generation, Kaleido:

More Controllable 🎛️

- Generate more or fewer values
- Negate particular values

Scalar Valence and Relevance

- Continuous values have more info than text

Open Science 👩‍🔬

- Open for scientific review and critique
- Build on our work

⚠Limitations⚠

Some limitations of this work:

Machine-Generated

- Can adopt the biases of GPT-4
- Further study is needed

English-Only Data

- Likely fits better to values held in English-speaking countries

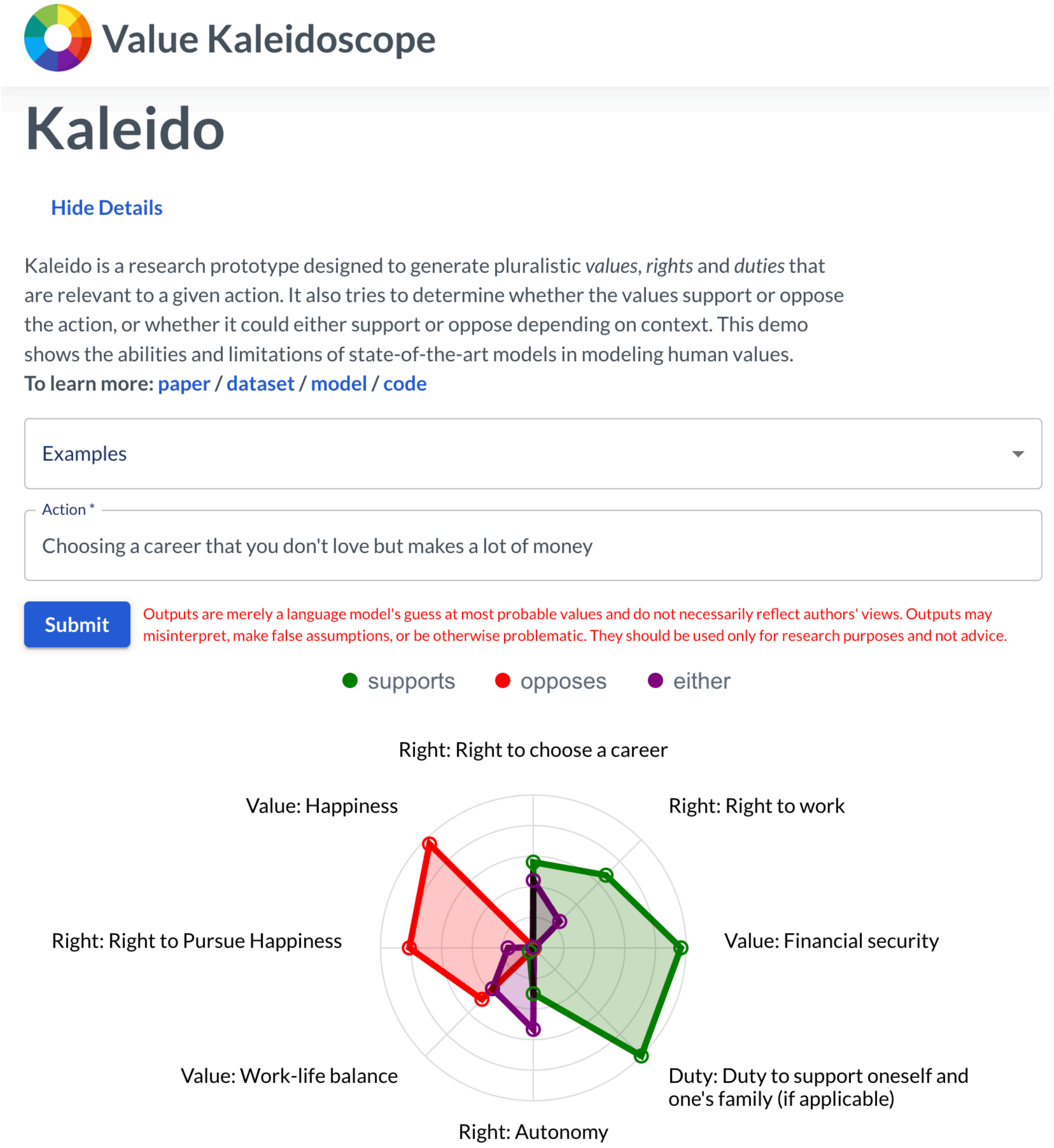
Not Intended for Advice

- Goal is not to output judgment
- Research focus, not for human-use

Model/dataset available on 🤗Huggingface🤗

https://huggingface.co/datasets/allenai/ValuePrism

https://huggingface.co/allenai/kaleido-xl



Value Kaleidoscope

Kaleido

Hide Details

Kaleido is a research prototype designed to generate pluralistic *values*, *rights* and *duties* that are relevant to a given action. It also tries to determine whether the values support or oppose the action, or whether it could either support or oppose depending on context. This demo shows the abilities and limitations of state-of-the-art models in modeling human values.

To learn more: paper / dataset / model / code

Examples ▾

Action *

Choosing a career that you don't love but makes a lot of money

Submit

Outputs are merely a language model's guess at most probable values and do not necessarily reflect authors' views. Outputs may misinterpret, make false assumptions, or be otherwise problematic. They should be used only for research purposes and not advice.

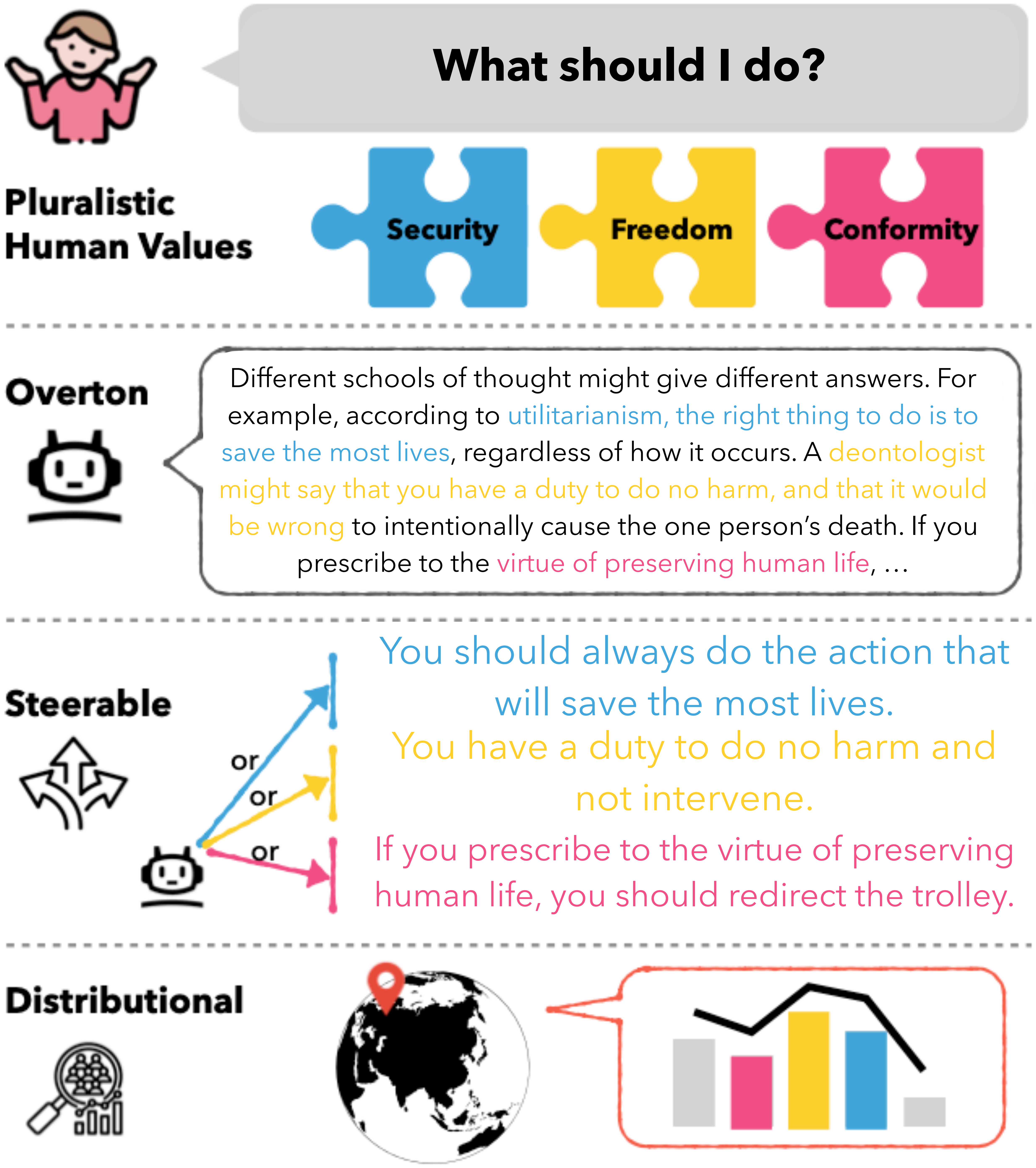
● supports
● opposes
● either
Right: Right to choose a career
Right: Right to work
Value: Financial security
Duty: Duty to support oneself and one's family (if applicable)
Right: Autonomy
Value: Work-life balance
Right: Right to Pursue Happiness
Value: Happiness

Potential Future Work w/ Kaleidoscope

- Use Kaleido model as attribute-specific reward, train steerably-pluralistic model
- Use ValuePrism situations for areas where pluralistic alignment may be relevant because of value disagreement
- Evaluation/Training with ValuePrism (see Modular Pluralism)

Pluralistic Alignment

1. Roadmap
2. Recent Work
3. Open Problems



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, …

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.

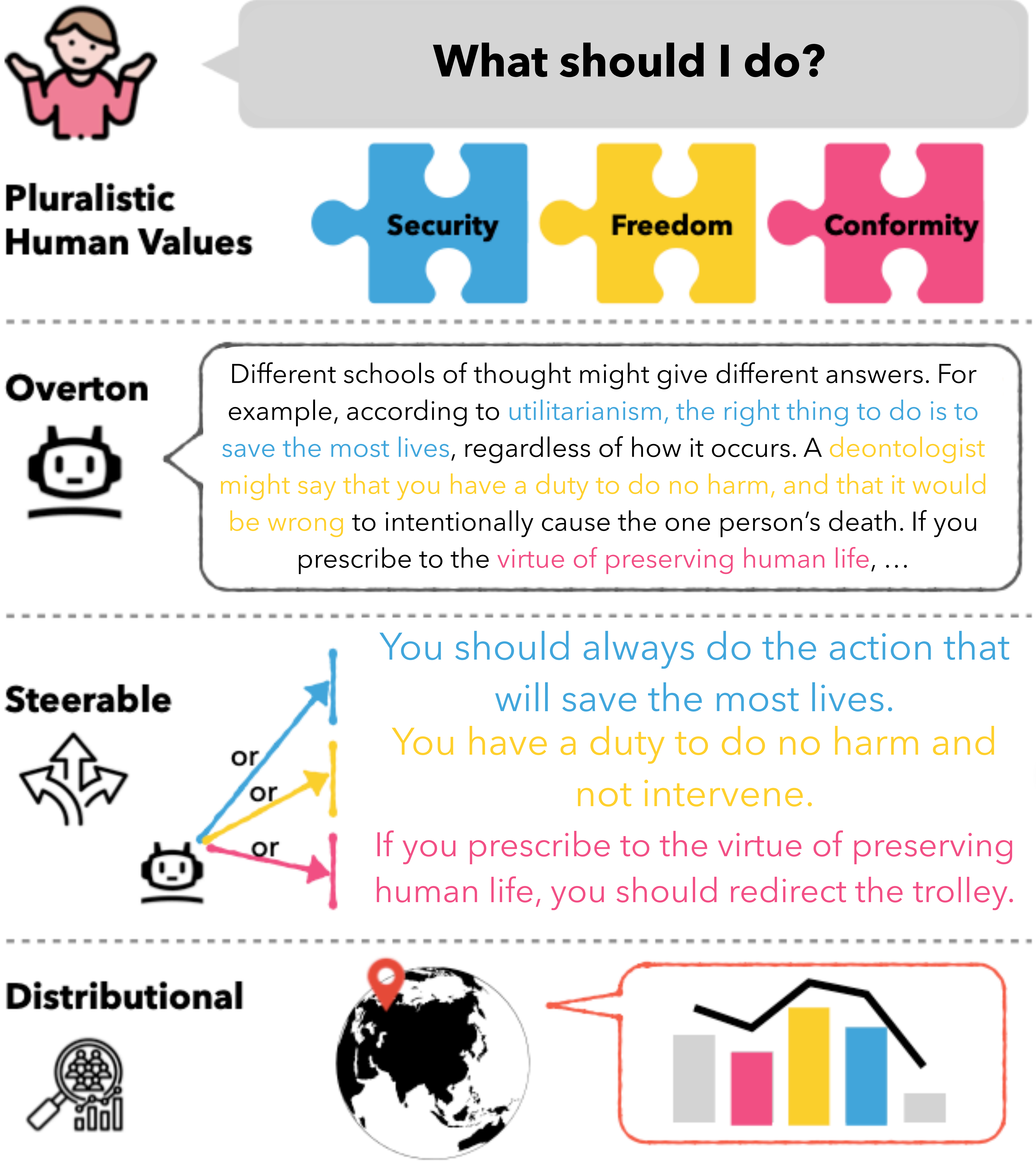
Overton Pluralism

Alignment

- Q1: RLHF seems to increase Overton pluralism to the extent that people prefer it. Does RLHF solve this, or are there gaps?
  - It seems that many responses follow templated "On the one hand, …, on the other hand, …". Does not do a good job of covering when there are many justifications on one side
  - e.g., bothesidesism may be an issue
- Q2: What techniques might increase Overton pluralism?
  - (None exist yet as far as I know apart from Modular Pluralism)

Evaluation

- Q3: Datasets/benchmarks for evaluating Overton pluralism?
  - (Not aware of any outside ValuePrism)



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, …

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.

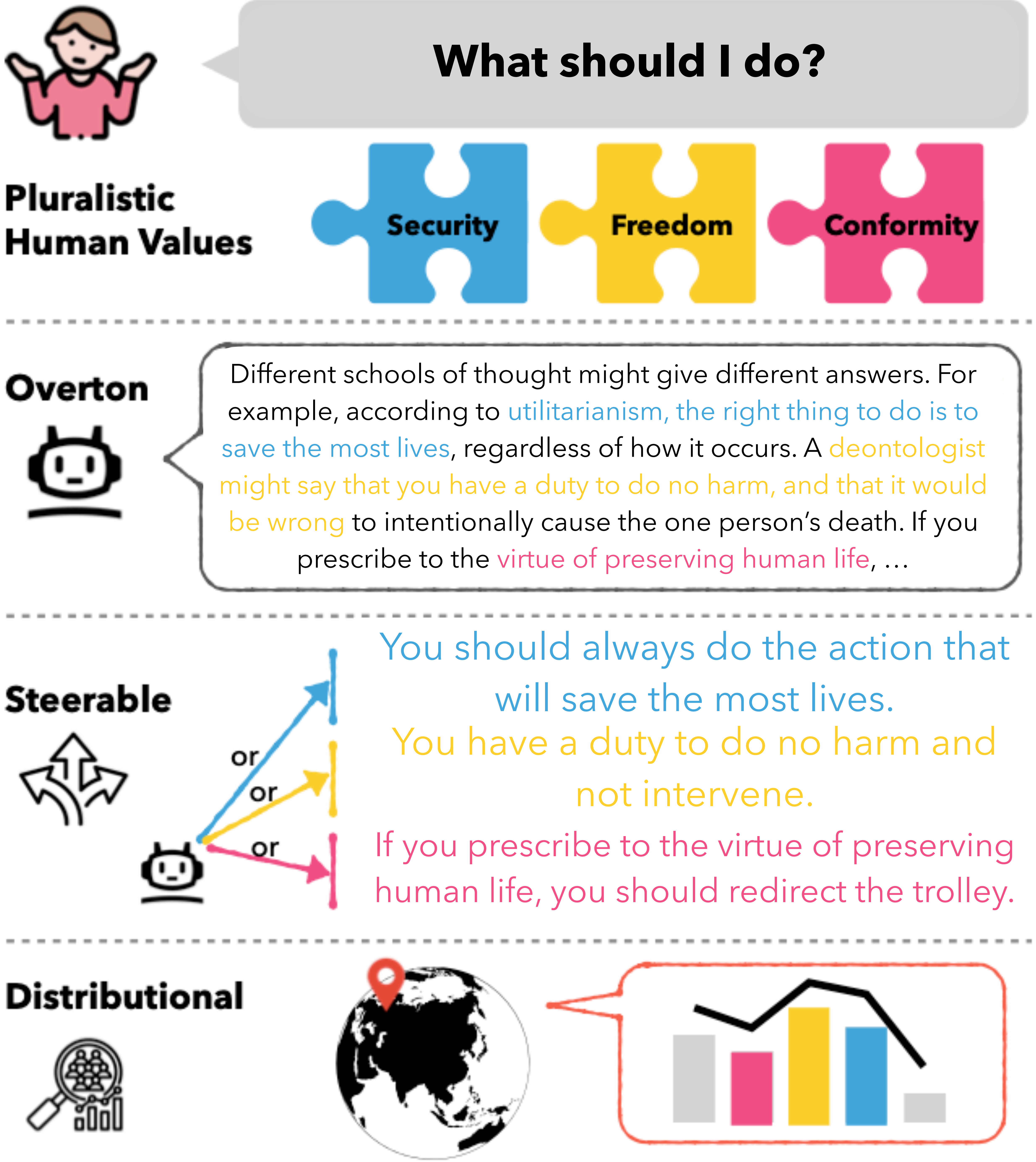
Steerable Pluralism

Alignment

- Q1: To what extent does prompting solve steerable pluralism?
  - (My guess: gets you some of the way there, but not all the way)
- Q2: How to increase steerable pluralism?
  - Some work here, many gaps

Evaluation

- Q3: Are there large gaps in the attributes/ perspectives that models can be steered to?
  - Some work here, but many gaps remain
- Q4: What attributes do we wish to align to?
  - Build datasets/benchmarks here



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, …

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.

Distributional Pluralism

Alignment

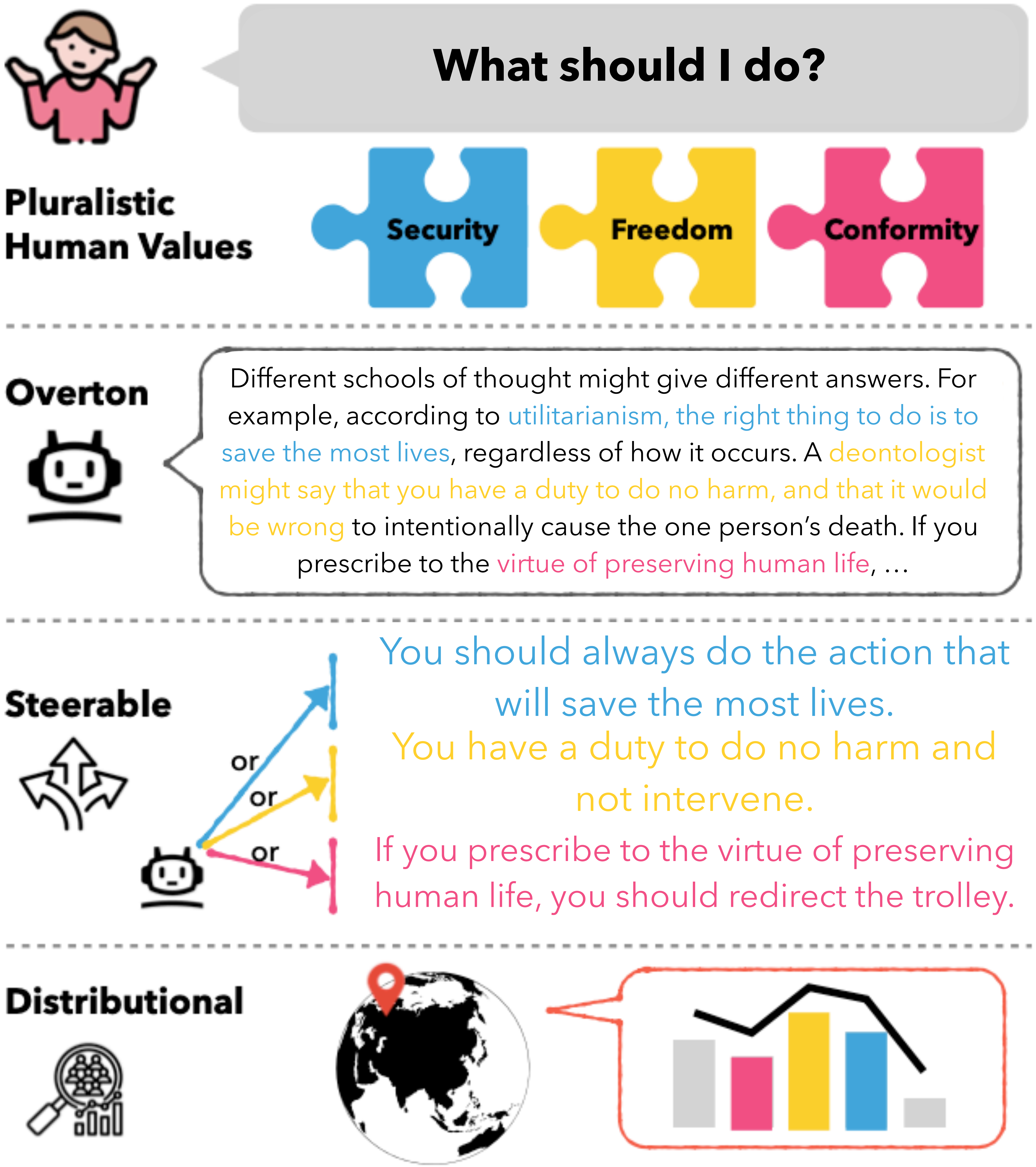
- Q1: Pre-trained models seem to generally outperform post-trained models. Can we improve upon pre-trained model baselines?
  - (Continued pre-training on subpopulation can help, but anything else?)

Evaluation

- Several datasets here for multiple choice (OpinionQA, GlobalOpinionQA, MoralChoice, surveys, …)
- Q2: How to extend to free-response/open-text?

Evaluation

- Combining Distributional with Steerable Pluralism



What should I do?

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, …

Steerable

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

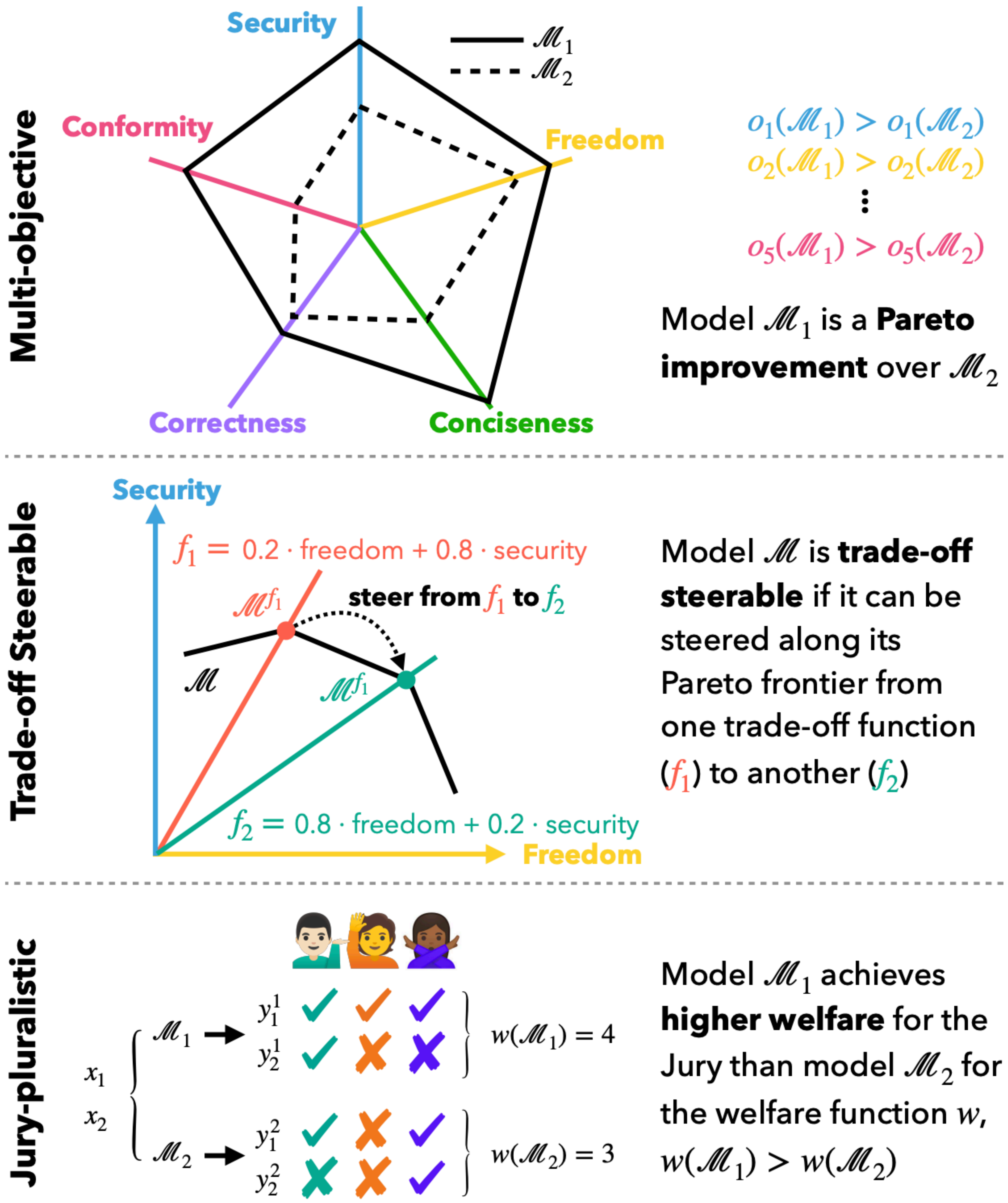
You have a duty to do no harm and not intervene.

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

Distributional

Other types of model pluralism?

Extending definitions?



Model $\mathscr{M}_1$ is a **Pareto improvement** over $\mathscr{M}_2$

Model $\mathscr{M}$ is **trade-off steerable** if it can be steered along its Pareto frontier from one trade-off function (f_1) to another (f_2)

Model $\mathscr{M}_1$ achieves **higher welfare** for the Jury than model $\mathscr{M}_2$ for the welfare function w, $w(\mathscr{M}_1) > w(\mathscr{M}_2)$

Trade-off Steerable

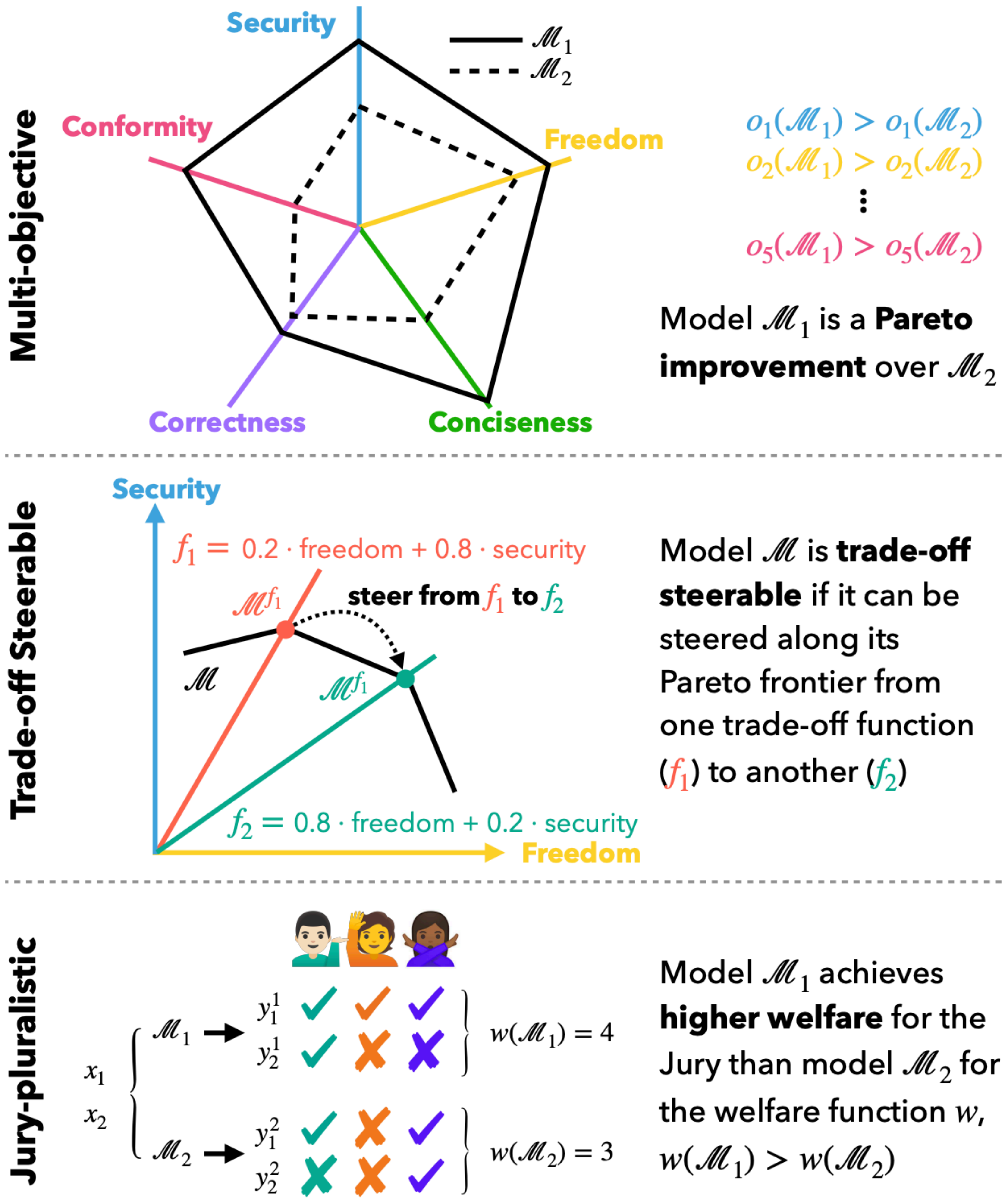
Alignment

- Q1: What techniques increase trade-off steerability?
  - (No papers yet on this afaik)

Evaluation

- Q2: Datasets/benchmarks for evaluating Overton pluralism?
  - (No standard benchmarks here afaik, though I know one lab is working on one)

Jury Pluralism



Model $\mathcal{M}_1$ is a **Pareto improvement** over $\mathcal{M}_2$

Model $\mathcal{M}$ is **trade-off steerable** if it can be steered along its Pareto frontier from one trade-off function (f_1) to another (f_2)

Model $\mathcal{M}_1$ achieves **higher welfare** for the Jury than model $\mathcal{M}_2$ for the welfare function w, $w(\mathcal{M}_1) > w(\mathcal{M}_2)$

- Q1: How to estimate good juror functions?
- Q2: Empirical trade-offs to different social welfare functions?
- Q3: What applications benefit from jury pluralism?
  - (e.g., consensus-building, community notes - what else?)
- Q4: While some work has approached this (e.g., MaxMinRLHF), most prior work has used fairly contrived juror functions (e.g., length, sentiment). How do these techniques extend to real-world data?

Other Questions

- How do different forms of pluralism interact?
- In what kind of systems do we want what kinds of pluralism?
- Which/whose values to align to?

and many more open questions…

Come work on pluralistic alignment with us!

Pluralistic Alignment

@ NeurIPS 2024 Workshop

December 15, 2024 in Vancouver, Canada

Exploring Pluralistic Perspectives in AI

Thank you!

Website: tsor13.github.io

X (Twitter): @ma_tay_

Email: tsor13@cs.washington.edu