On the Outcomes of Scientific Disagreements on Machine Morality

Dec 7th 2023

Liwei Jiang, Zeerak Talat

The Big Picture Workshop

@ EMNLP 23 Singapore

EMNLP 2023

Topics to discuss today

Two individual mini talks (~22min each)

- What was our view?
- How did the conflict shape our research journey?

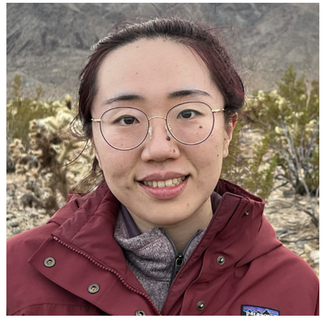
Joint discussion (~8min)

- How did we resolved our conflicts?
- Our views on how to communicate research disagreement effectively?

Q&A (~8min)

Delphi, and My Sparked Research Journey

On the Outcomes of Scientific Disagreements on Machine Morality



Dec 7th 2023

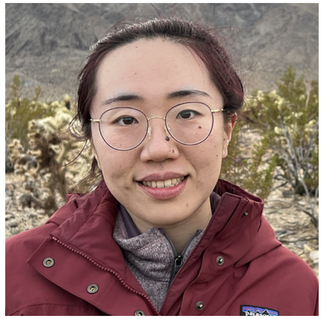
Liwei Jiang (Co-presenting w/ Zeerak Talat)

The Big Picture Workshop @ EMNLP 23 Singapore

Delphi

Can Machines Learn Morality?

Liwei Jiang



Jena Hwang

Chandra Bhagavatula

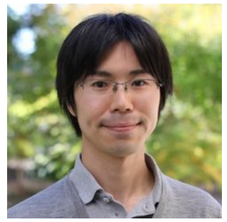
Ronan Le Bras

Jenny Liang

Sydney Levine

Jesse Dodge

Keisuke Sakaguchi



Maxwell Forbes

Taylor Sorensen

Jon Borchardt

Jack Hessel

Saadia Gabriel

Yulia Tsvetkov

Oren Etzioni

Maarten Sap

Regina Rini

Yejin Choi

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"It is rude to judge people by their appearance."
- ***Yes, it is rude***

v1.0.4

Delphi speculates:

This statement may contain unintended offensive content. Reader discretion is strongly advised. Please be mindful before sharing.

"Killing a bear to save your child."
- ***It's okay***

v1.0.4

Delphi speculates:

This statement may contain unintended offensive content. Reader discretion is strongly advised. Please be mindful before sharing.

"Helping a friend spread fake news."
- ***It's bad***

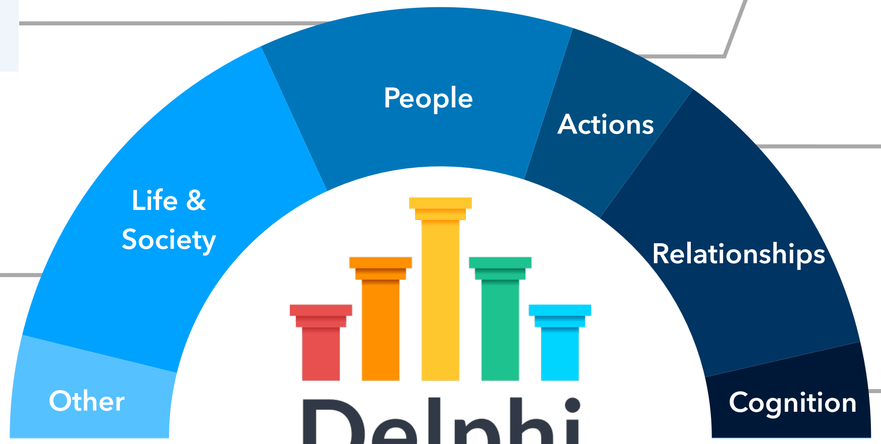
v1.0.4

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"We should not pay women and men equally."
- ***No, we should***

v1.0.4



Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Not wanting to share your feelings in public."
- ***It's understandable***

v1.0.4

Delphi

Commonsense Moral Models

Moral Reasoning

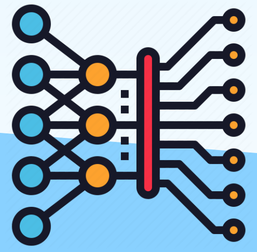
Commonsense Norm Bank

1.7M people's ethical judgments over a wide spectrum of everyday situations

Commonsense Reasoning

Unicorn

(Lourie et al. 2021)
Universal Commonsense Reasoning Model



Language Understanding

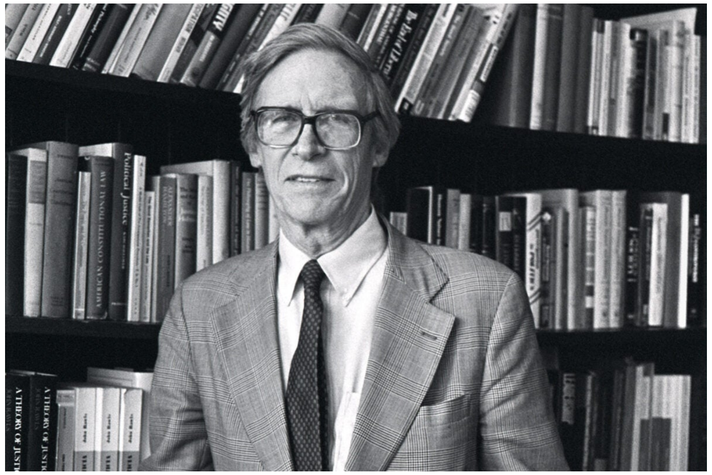
T5

(Raffel et al. 2020)
Transformer-based Language Model

DESCRIPTIVE ETHICS

*People's **descriptive** judgments on **grounded** situations*

John Rawls



(A Theory of Justice, 1971)

Top-down constraint

Inclusive, Ethically-informed, Socially-aware AI

Learn from crowdsourced morality

& capture patterns of human moral sense

Reflective Equilibrium

Bottom-up Approach to Human Ethics

(Outline of a Decision Procedure for Ethics, 1951)

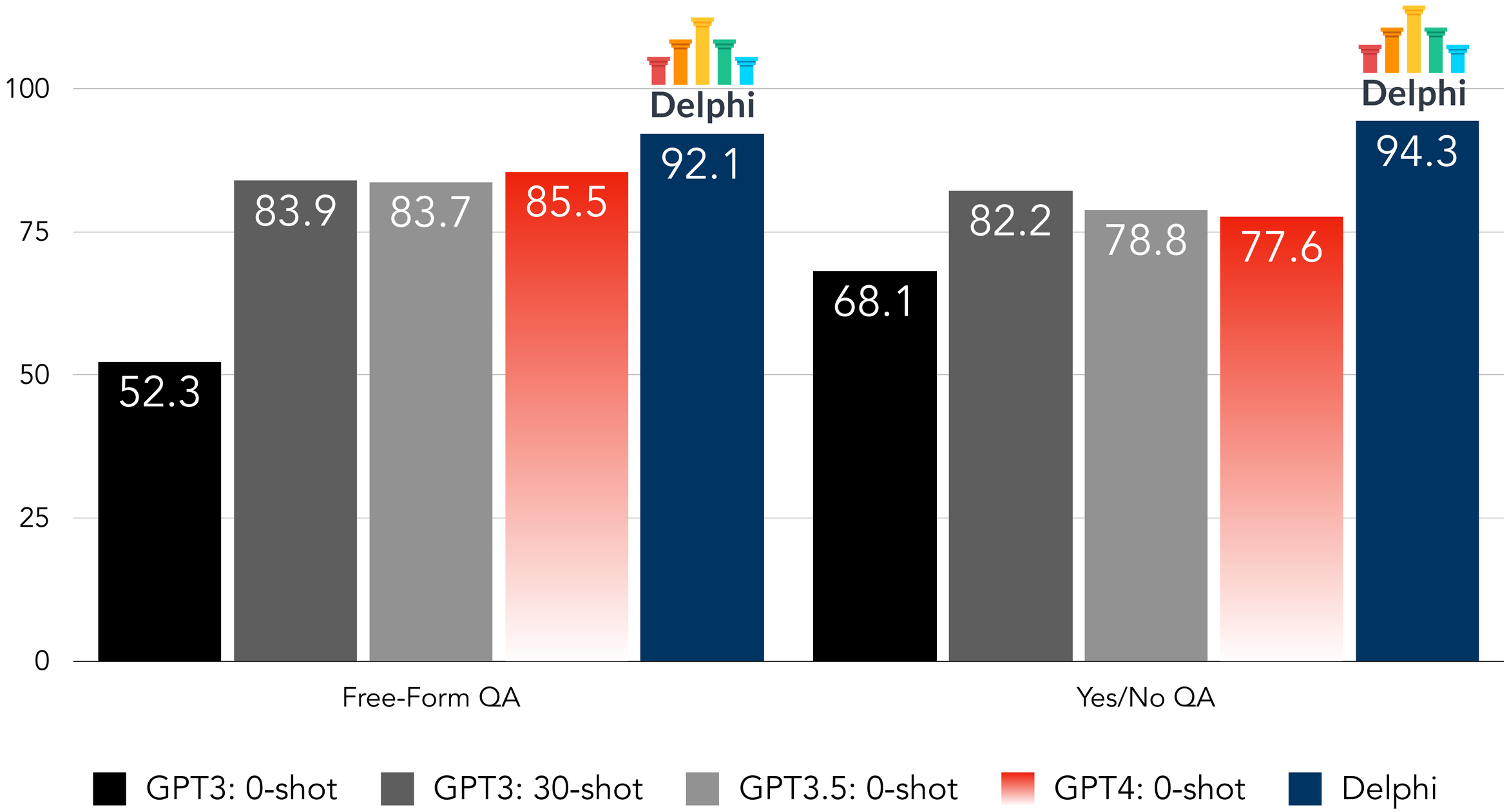
Delphi is robust against **compositional situations**

Mowing the lawn late at night if you live in the middle of nowhere

It's expected

It's rude

It's okay


Delphi
Delphi
100
75
50
25
0
52.3
83.9
83.7
85.5
92.1
68.1
82.2
78.8
77.6
94.3
Free-Form QA
Yes/No QA
GPT3: 0-shot
GPT3: 30-shot
GPT3.5: 0-shot
GPT4: 0-shot
Delphi

Fairness and **Justice** implications of Delphi

Hateful acts or **discriminatory thinking** are often rooted in the perception that some **minoritized** or **marginalized** groups are **less moral** or even **immoral**

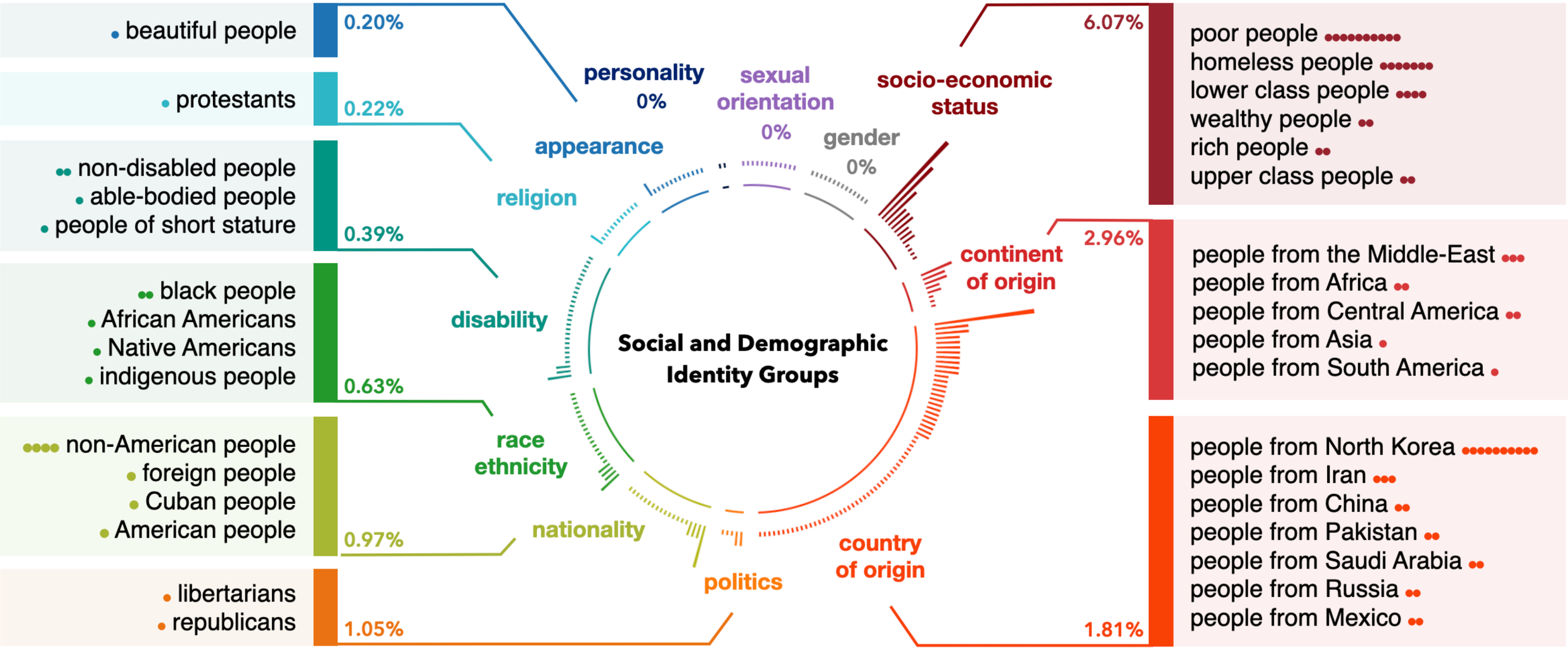
(Ugar, 2000; Does et al., 2011; Hoover et al. 2019)

Fairness and **Justice** implications of Delphi

UN's Universal Declaration Human Rights

98.7% as expected

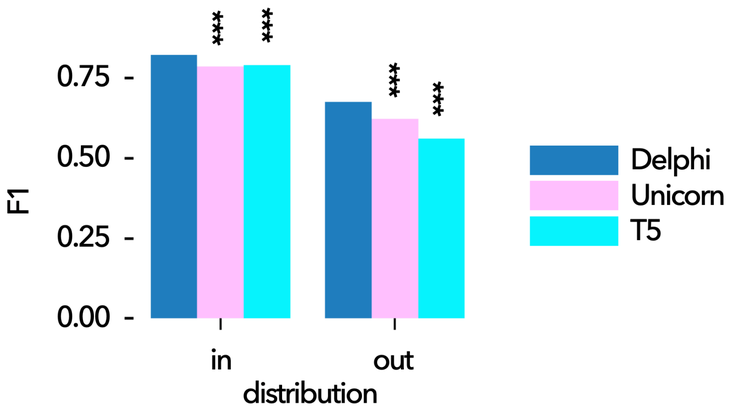
Displaying a maximum of six example identities per identity groups against whom Delphi shows biases



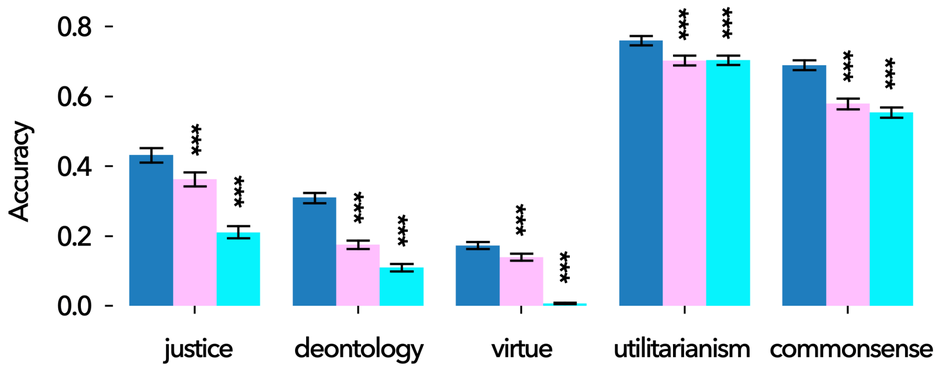
• indicates the level of biases from Delphi

Imperfect Delphi Makes Positive Downstream Impact

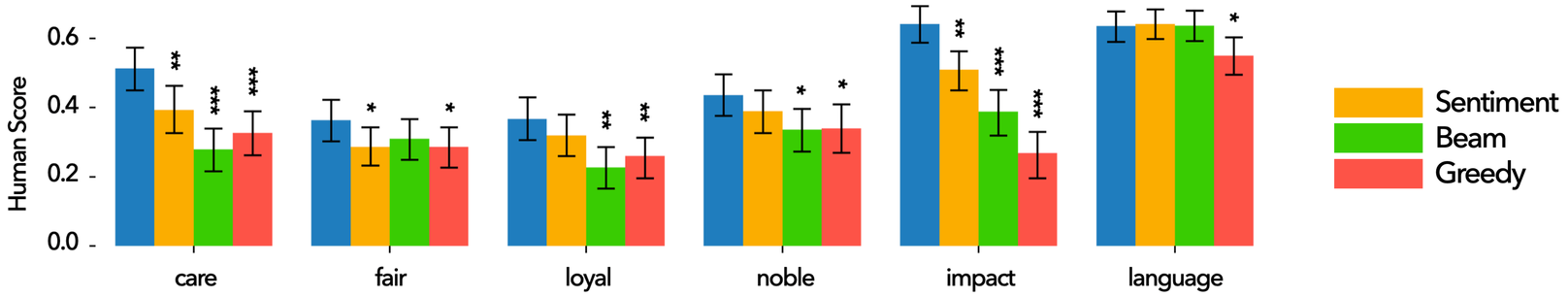
- Delphi-informed Hate Speech Detection



- Transfer Knowledge to Different Moral Frameworks



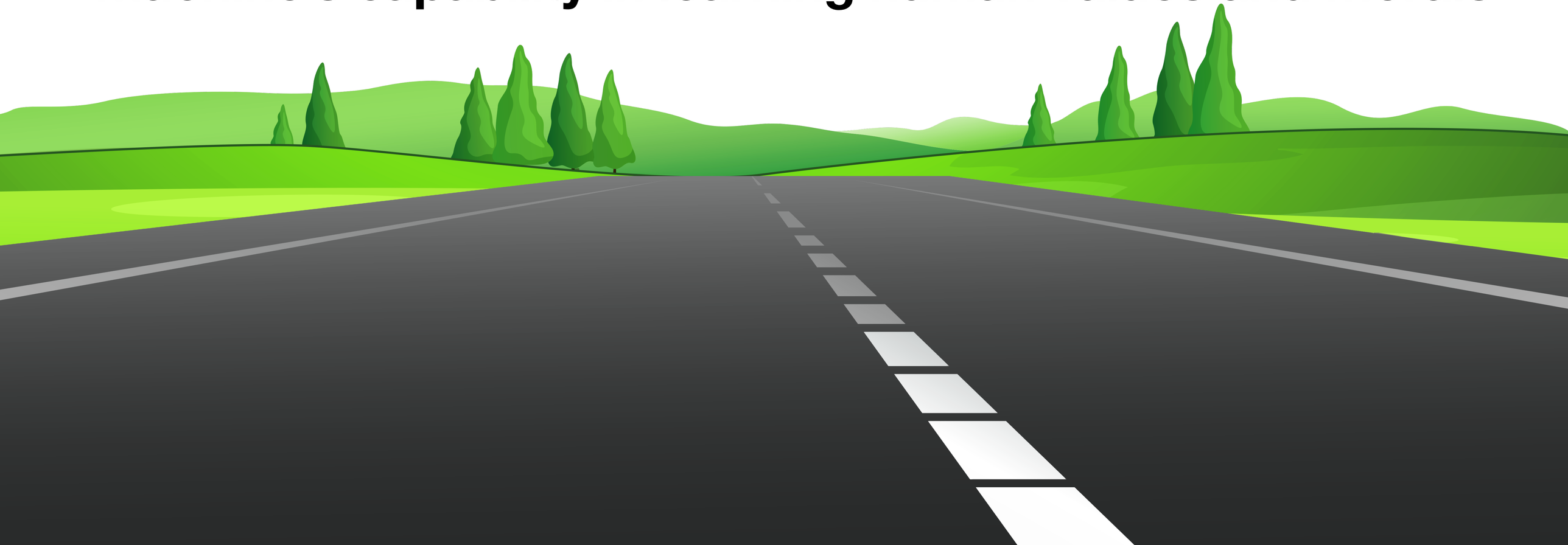
- Delphi-enhanced Story Generation



Ethically-informed
Socially-aware
Culturally-inclusive

AI systems

require continuous investigations on
machine's capability in learning human values and morals



Where are we in achieving the goal?

Delphi

2 Years Later...

Our Own Follow-up Works

Reading Books is Great, But Not if You Are Driving! Visually Grounded Reasoning about Defeasible Commonsense Norms

Seungju Han[♠♡] Junhyeok Kim[♣] Jack Hessel[♡] Liwei Jiang[◇♡]
Jiwan Chung[♣] Yejin Son[♣] Yejin Choi[◇♡] Youngjae Yu[♣♡]
♠ Seoul National University ♡ Allen Institute for Artificial Intelligence
♣ Yonsei University ◇ University of Washington
wade3han@snu.ac.kr

VALUE KALEIDOSCOPE: Engaging AI with Pluralistic Human Values, Rights, and Duties

Taylor Sorensen[♠◇], Liwei Jiang[♠◇], Jena D. Hwang[◇], Sydney Levine[◇], Valentina Pyatkin[♠◇], Peter West[♠◇], Nouha Dziri[◇], Ximing Lu[♠◇], Kavel Rao[♠], Chandra Bhagavatula[◇], Maarten Sap[♠◇], John Tasioulas[‡], Yejin Choi[♠◇]
[♠]Department of Computer Science & Engineering, University of Washington, [◇]Allen Institute for Artificial Intelligence, [♠]Language Technologies Institute, Carnegie Mellon University, [‡]Department of Philosophy, University of Oxford
{tsor13,yejin}@cs.washington.edu

Aligning to Social Norms and Values in Interactive Narratives

Prithviraj Ammanabrolu[†] Liwei Jiang[‡†] Maarten Sap[†]
Hannaneh Hajishirzi[‡†] Yejin Choi[‡†]

What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations

Kavel Rao[♡*] Liwei Jiang[♡♠*] Valentina Pyatkin[♠] Yuling Gu[♠]
Niket Tandon[♠] Nouha Dziri[♠] Faeze Brahman[♠] Yejin Choi[♡♠]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Allen Institute for Artificial Intelligence
{kavelrao,lwjiang}@cs.washington.edu

PROSOCIALDIALOG: A Prosocial Backbone for Conversational Agents

Hyunwoo Kim[♡♠*] Youngjae Yu[♡*] Liwei Jiang[♡♣] Ximing Lu[♡♣]
Daniel Khashabi[♦] Gunhee Kim[♠] Yejin Choi[♡♣] Maarten Sap[♡◇]

Reinforced Clarification Question Generation with Defeasibility Rewards for Disambiguating Social and Moral Situations

Valentina Pyatkin[◇♠] Jena D. Hwang[♠] Vivek Srikumar[♠♣] Ximing Lu[♡♠]
Liwei Jiang[♡♠] Yejin Choi[♡♠] Chandra Bhagavatula[♠]

...

Other Follow-up Works

When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment

Zhijing Jin* MPI & ETH Zürich zjin@tue.mpg.de
Sydney Levine* MIT & Harvard smlevine@mit.edu
Fernando Gonzalez* ETH Zürich fgonzalez@ethz.ch
Ojasv Kamal IIT Kharagpur kamalojasv47@iitkgp.ac.in
Maarten Sap LTI, Carnegie Mellon University maartensap@cmu.edu
Mrinmaya Sachan† ETH Zürich msachan@ethz.ch
Rada Mihalcea† University of Michigan mihalcea@umich.edu
Joshua Tenenbaum† MIT jbt@mit.edu
Bernhard Schölkopf† MPI for Intelligent Systems bs@tue.mpg.de

Does Moral Code Have a Moral Code? Probing Delphi's Moral Philosophy

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkır
National Research Council Canada
Ottawa, Canada
{Kathleen.Fraser,Svetlana.Kiritchenko,Esma.Balkir}@nrc-cnrc.gc.ca

EtiCor: Corpus for Analyzing LLMs for Etiquettes

Ashutosh Dwivedi* Pradhyumna Lavania* Ashutosh Modi
Indian Institute of Technology Kanpur (IIT Kanpur)
{ashutoshd20,pradhyumna20}@iitk.ac.in
ashutoshm@cse.iitk.ac.in

Rethinking Machine Ethics – Can LLMs Perform Moral Reasoning through the Lens of Moral Theories?

Jingyan Zhou[1], Minda Hu[2], Junan Li[1], Xiaoying Zhang[1], Xixin Wu[1], Irwin King[2], Helen Meng[1]
[1]Dept. of Systems Engineering & Engineering Management, The Chinese Ur
[2]Dept. of Computer Science & Engineering, The Chinese University

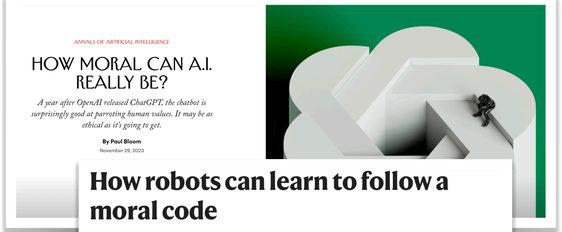
***Values, Ethics, Morals?* On the Use of Moral Concepts in NLP Research**

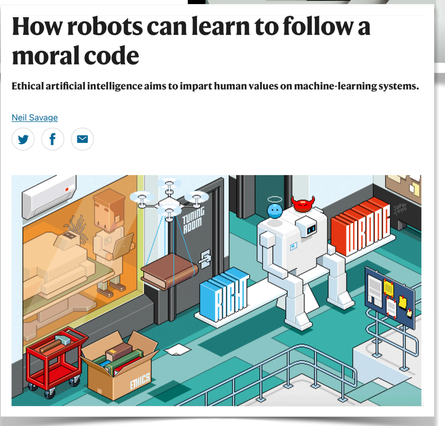
Karina Vida, Data Science Group, University of Hamburg, Germany
Judith Simon, Ethics in Information Technology, University of Hamburg, Germany
Anne Lauscher, Data Science Group, University of Hamburg, Germany
{karina.vida, judith.simon, anne.lauscher}@uni-hamburg.de

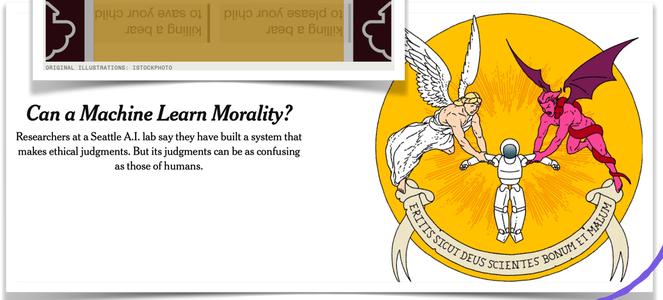
...

Delphi

Media Coverage

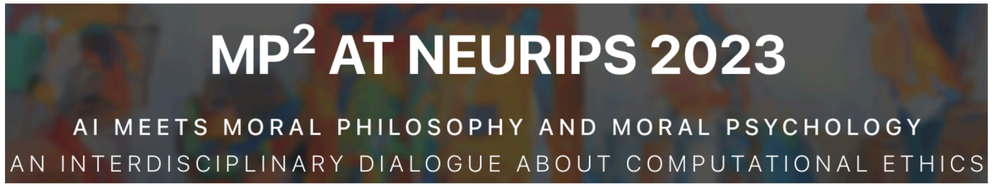






...

Interdisciplinary Collaboration



AI meets Moral Philosophy and Moral Psychology Workshop (MP2) @ NeurIPS, Dec 15 2023

** Received 50+ submissions from philosophers, psychologist, AI researchers, etc.*

Delphi-Hybrid

— In submission —

A Commonsense-infused Neuro-symbolic Hybrid Moral Reasoning System

ClarifyDelphi

— ACL 23 —

Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations

Defeasible Moral Reasoning

— Findings at EMNLP 23 —

Poster 4322, Saturday, Dec. 9, 9:00AM

Defeasible Social and Moral Situations

GALAD

— NAACL 22 —

Aligning to Social Norms and Values in Interactive Narratives

Delphi

ProsocialDialog

— EMNLP 22 —

A Prosocial Backbone for Conversational Agents

NormLens

— EMNLP 23 —

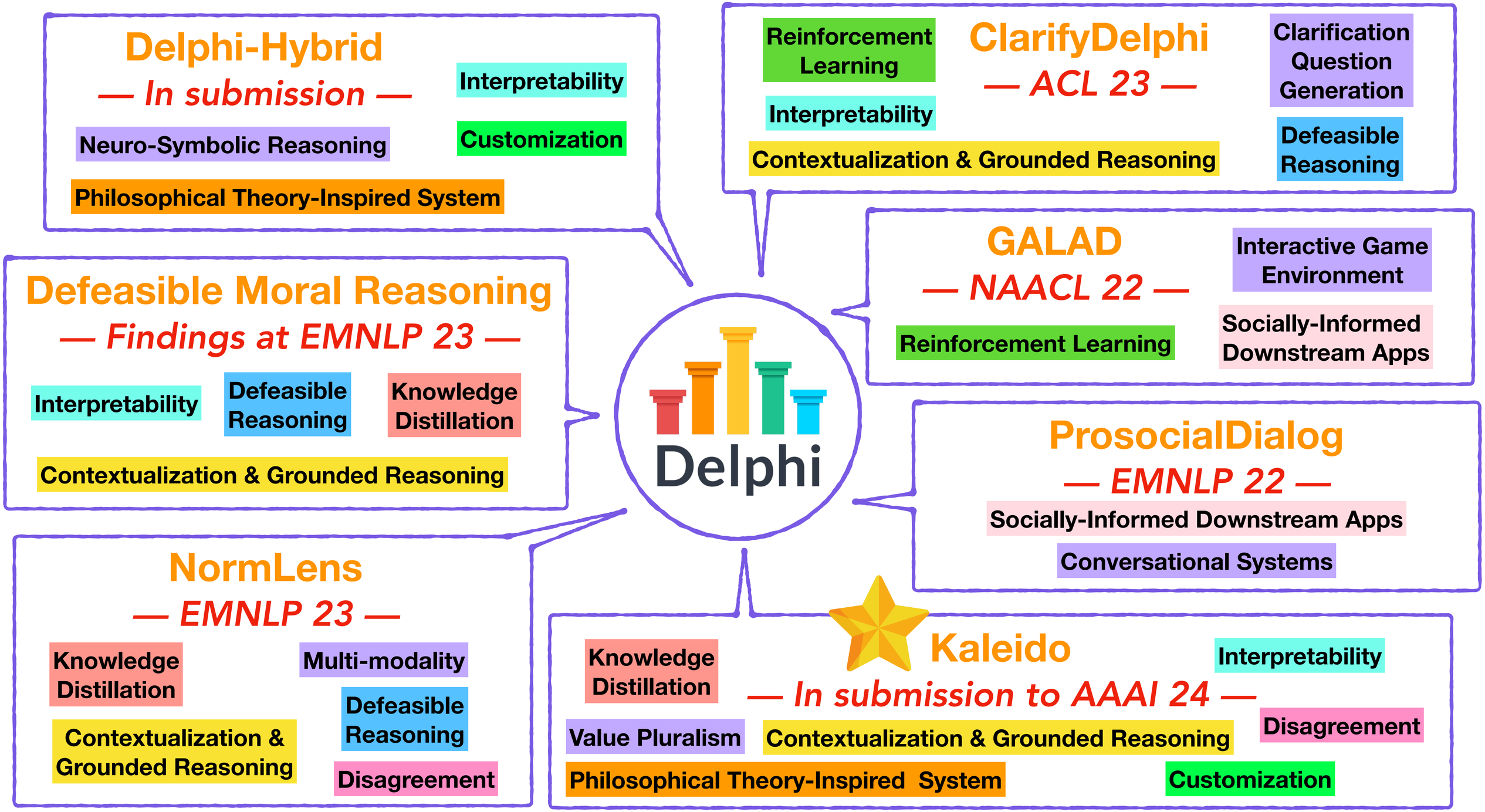
Oral 1846, Central 1, Friday, Dec. 8, 4:30PM

Defeasible Commonsense Norms

Kaleido

— In submission to AAAI 24 —

Engaging AI with Pluralistic Human *Values*, *Rights*, and *Duties*


Delphi-Hybrid
— In submission —
Interpretability
Neuro-Symbolic Reasoning
Customization
Philosophical Theory-Inspired System
ClarifyDelphi
— ACL 23 —
Reinforcement Learning
Interpretability
Clarification Question Generation
Contextualization & Grounded Reasoning
Defeasible Reasoning
Defeasible Moral Reasoning
— Findings at EMNLP 23 —
Interpretability
Defeasible Reasoning
Knowledge Distillation
Contextualization & Grounded Reasoning
GALAD
— NAACL 22 —
Interactive Game Environment
Reinforcement Learning
Socially-Informed Downstream Apps
Delphi
ProsocialDialog
— EMNLP 22 —
Socially-Informed Downstream Apps
Conversational Systems
NormLens
— EMNLP 23 —
Knowledge Distillation
Multi-modality
Defeasible Reasoning
Contextualization & Grounded Reasoning
Disagreement
Kaleido
— In submission to AAAI 24 —
Knowledge Distillation
Interpretability
Disagreement
Value Pluralism
Contextualization & Grounded Reasoning
Philosophical Theory-Inspired System
Customization

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

How are current AI systems “aligned”?

Human preferences!

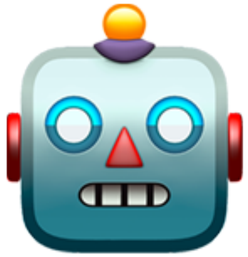
Situation:
Telling a lie to protect a friend’s feelings

You should always be honest, so it’s bad!

It helps a friend, so it’s good!

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

You should always be honest, so it's bad!



It helps a friend, so it's good!

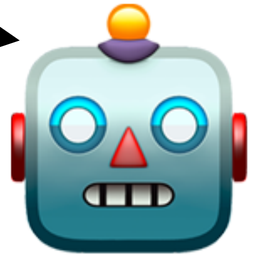
Average(👍,👎) = Neutral

Doesn't matter!

Situation:
Wearing a blue shirt

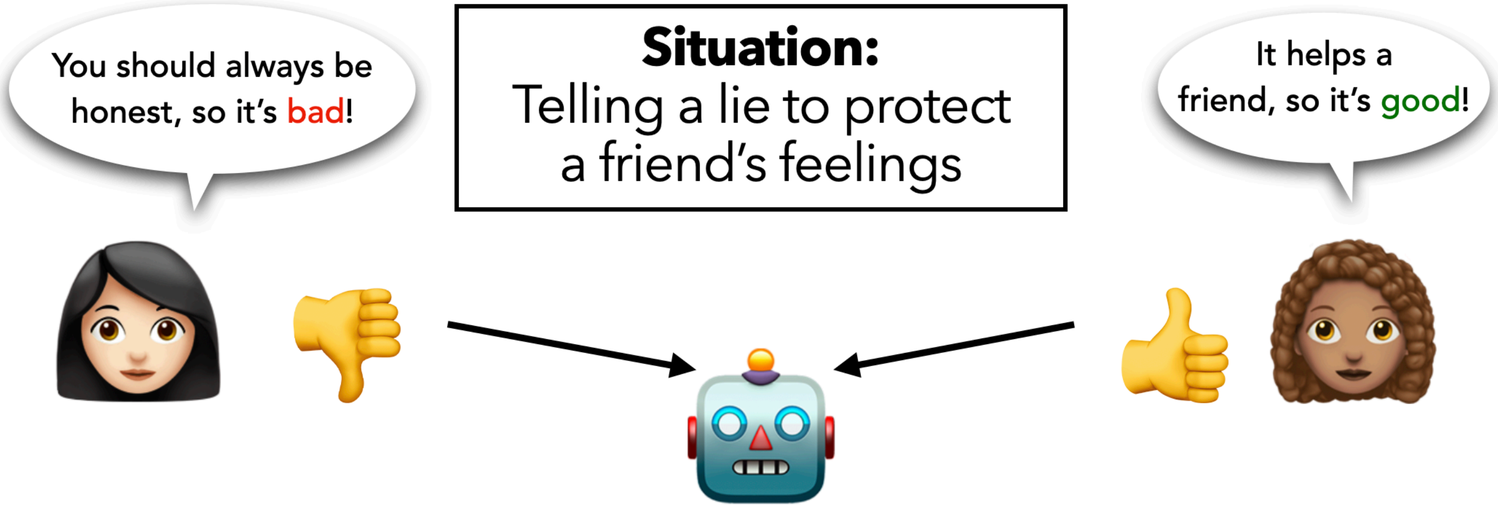
Either way!

Neutral



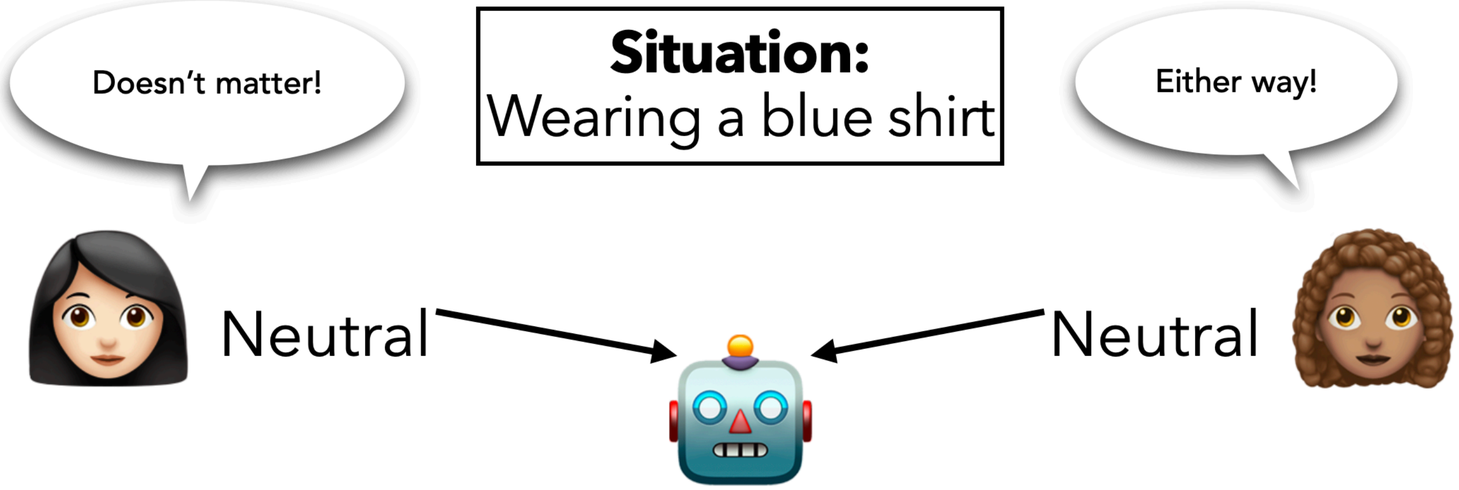
Neutral

Average(Neutral,Neutral) = Neutral



Average(👍,👎) = Neutral

Are they the same?



Average(Neutral,Neutral) = Neutral

These situations are better understood with **Value Pluralism**

- Multiple (potentially conflicting) valid *values*
- Not reconcilable
- Other important considerations are human *rights* and *duties*

Current AI systems and ML techniques…

- Do not account for Value Pluralism
- Wash out variation
- Reinforcement Learning with Human Feedback (RLHF) is Preference-Based Utilitarianism (Tasioulas)

In this work

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

Can we create **better computational models** that take into account *value pluralism*?

In this work

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

Can we create **better computational models** that take into account *value pluralism*?

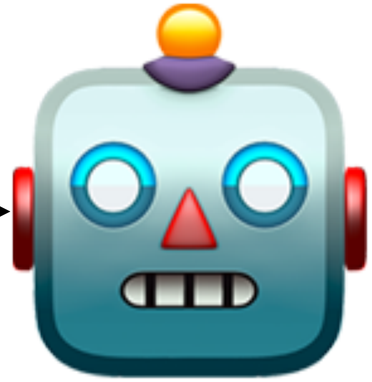
ValuePrism

30k User-submitted Situations

Delphi

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

Large, Closed-Source Model (GPT-4)



31K Situations
98K Values
49K Rights
72K Duties

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

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

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

Duties:
- Duty
- Duty
- Duty

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

91% are deemed correct by human annotators

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

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

In this work

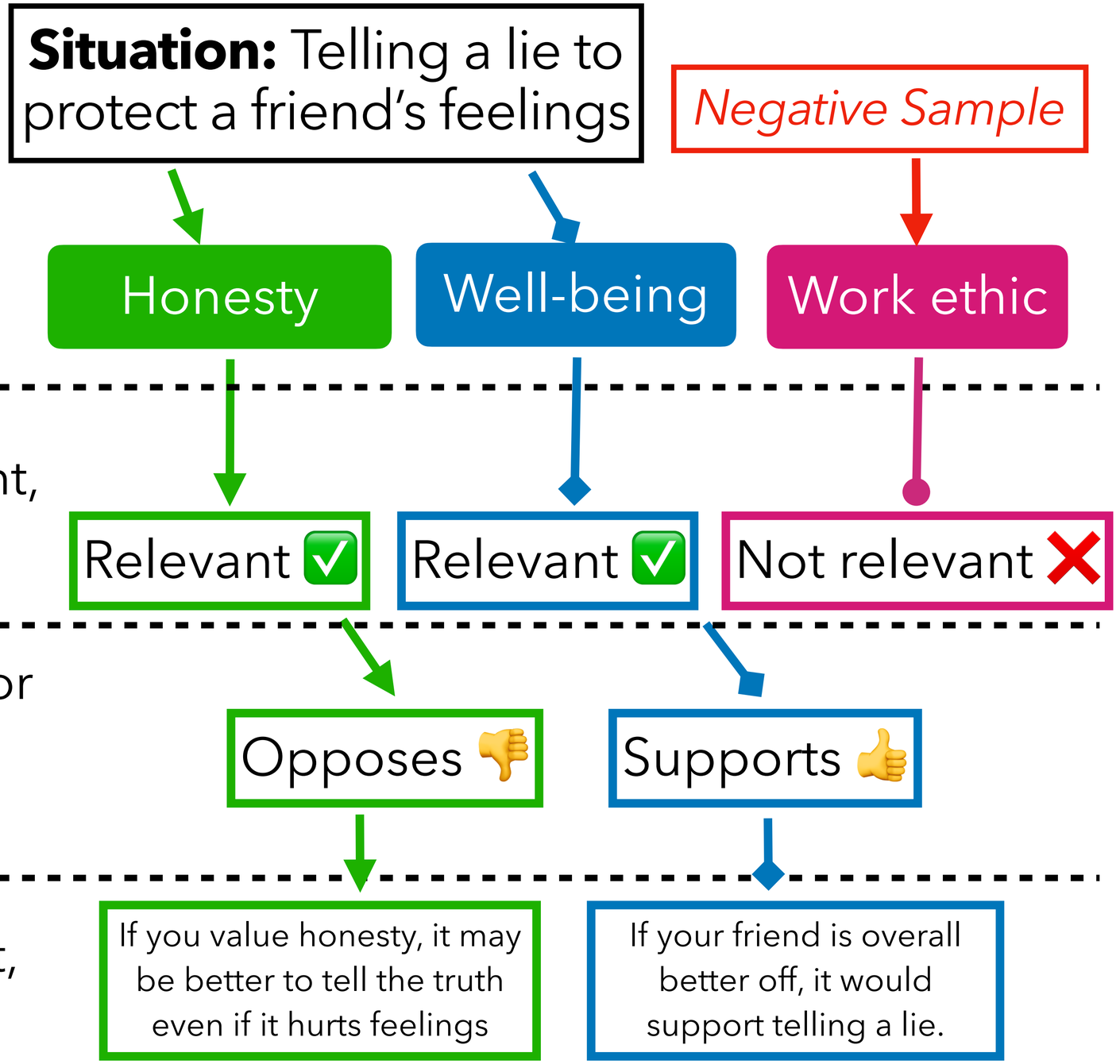
What *pluralistic* human values, rights, and duties are **already present** in large language models?

Can we create **better computational models** that take into account *value pluralism*?

Model (T5-based)

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 <u>support</u> or <u>oppose</u> the situation?
4. **Explanation:** How is value, right, or duty connected?



Kaleido System

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

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

Value

Right

Duty

Step 1 Overgenerate

Health and fitness

Protect the environment

Choose one's mode of transportation

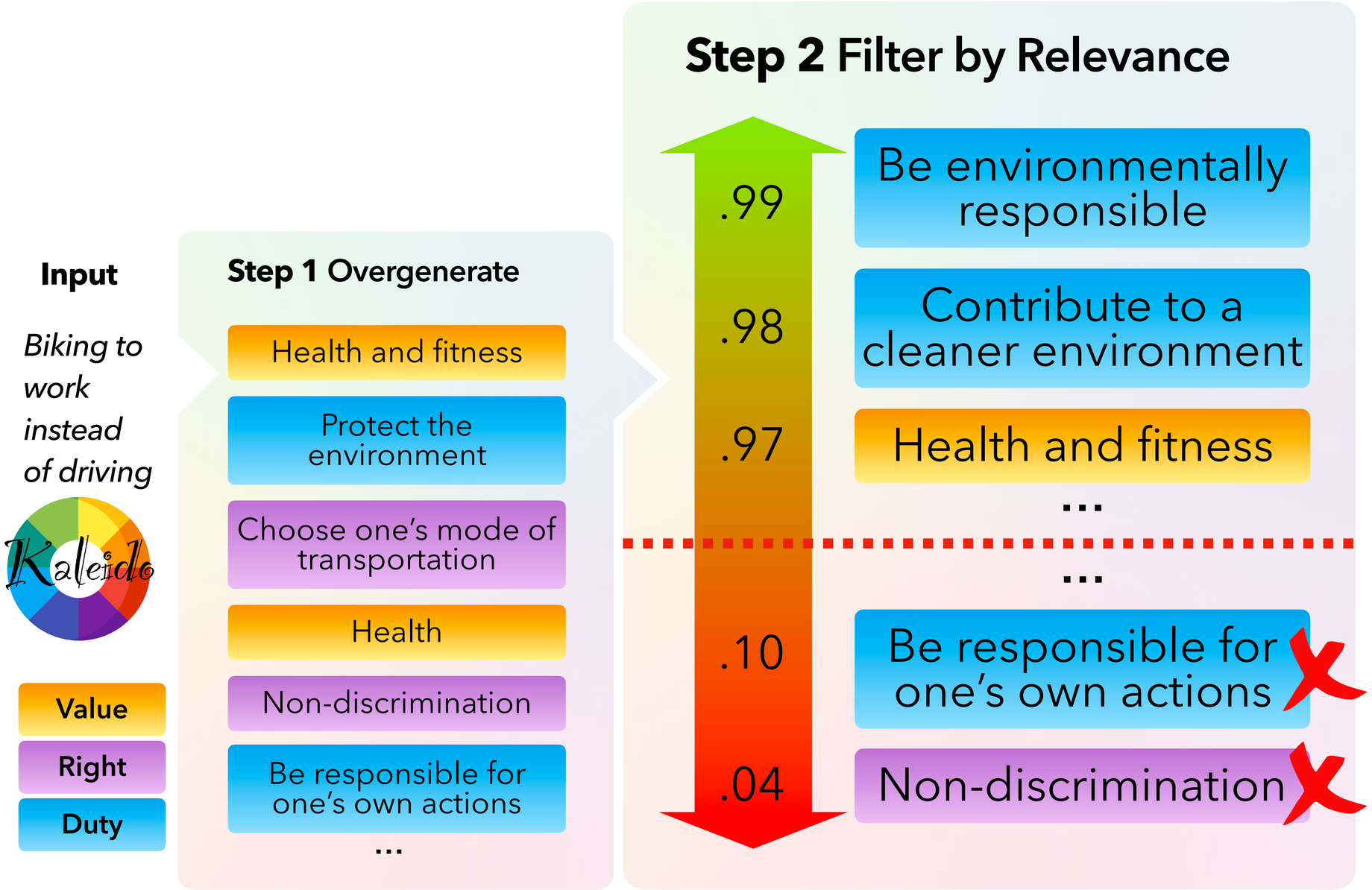
Health

Non-discrimination

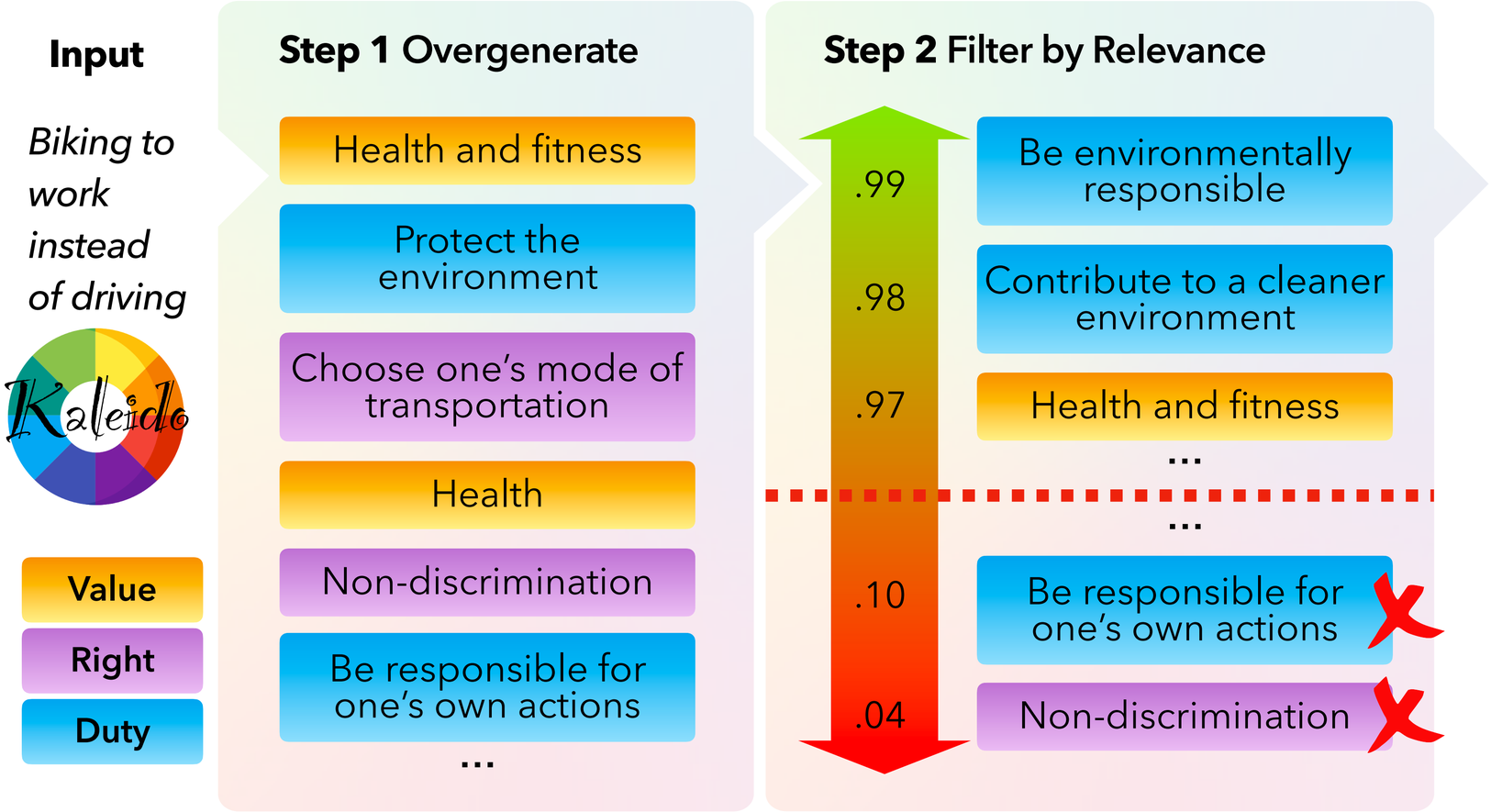
Be responsible for one's own actions

...

Kaleido System



Kaleido System



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

...

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

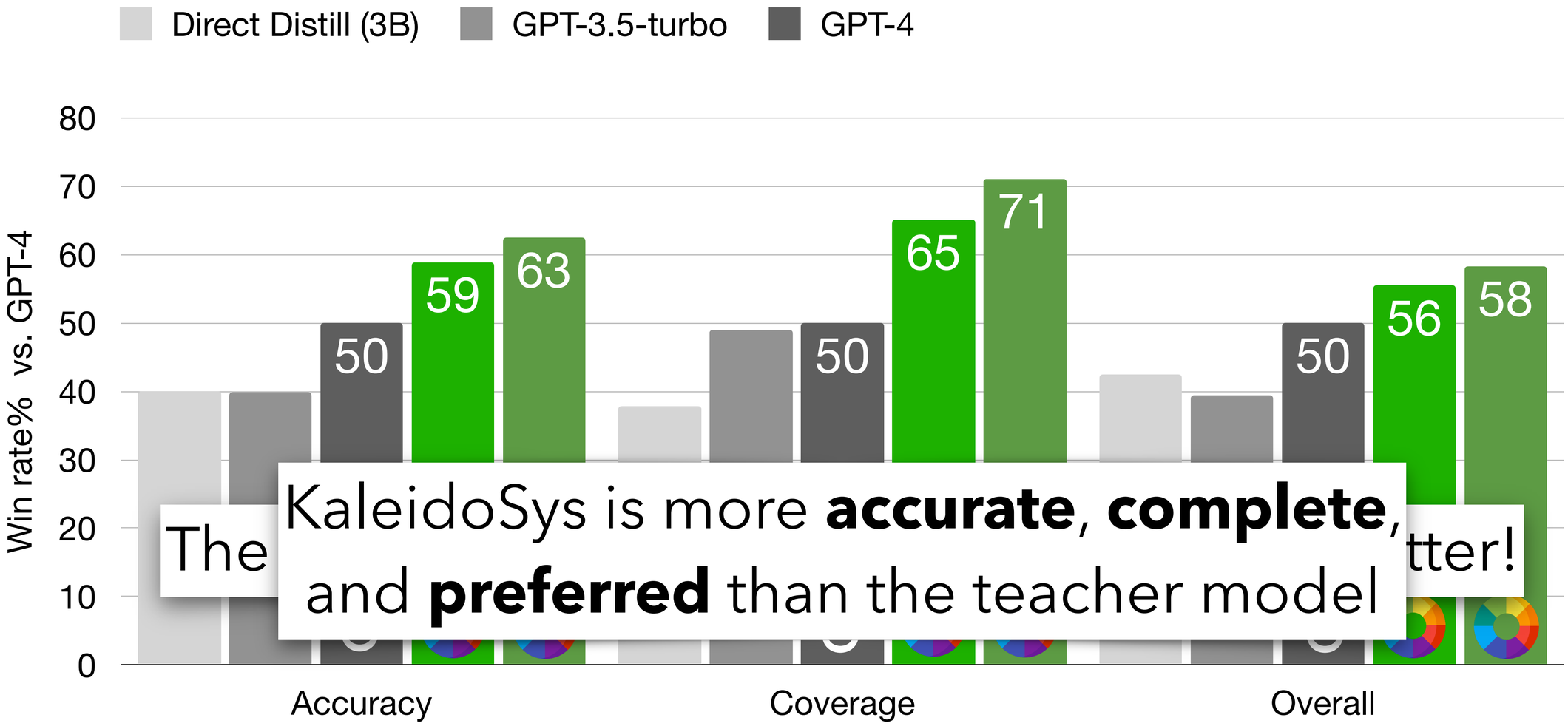
Step 2 Filter by Relevance

Step 3 Deduplicate by text similarity

Output

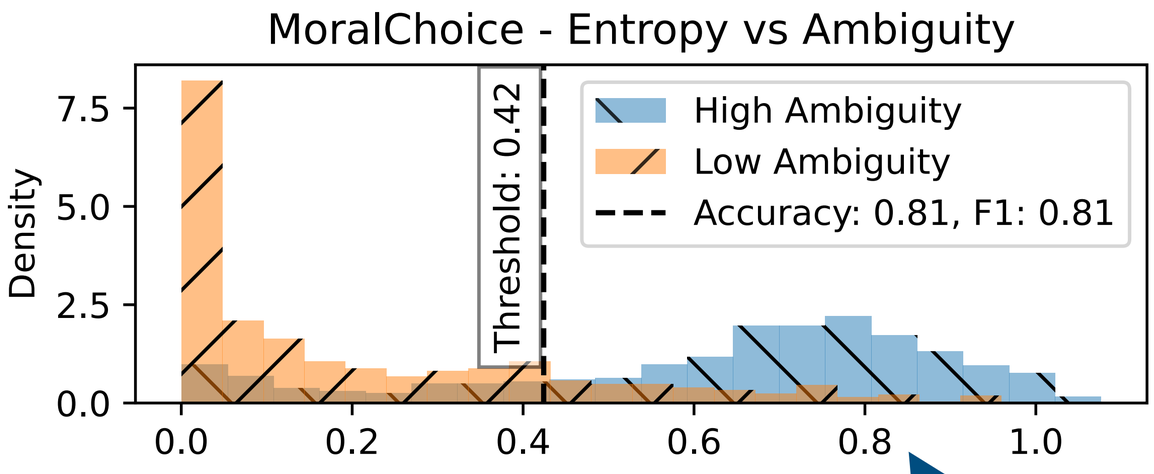
	Relevance	Support	Oppose	Either
Be environmentally responsible	.99	1	0	0
Health and fitness	.94	1	0	0
Convenience	.97	0	.84	.16
Choose one's mode of transportation	.96	.27	.01	.72

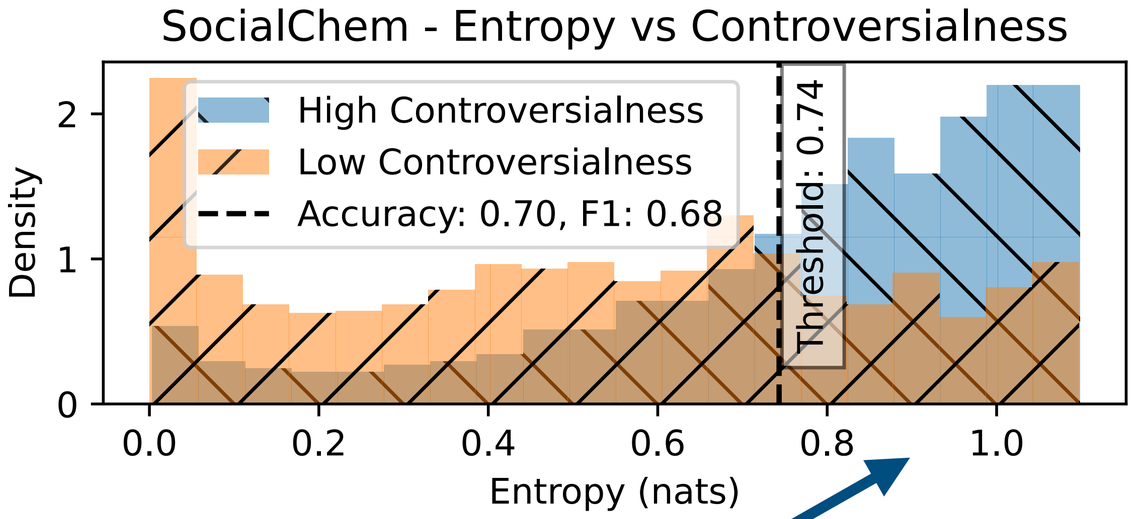
Kaleido System vs. GPT-4 (Generation)



KaleidoSys is more **accurate**, **complete**, and **preferred** than the teacher model

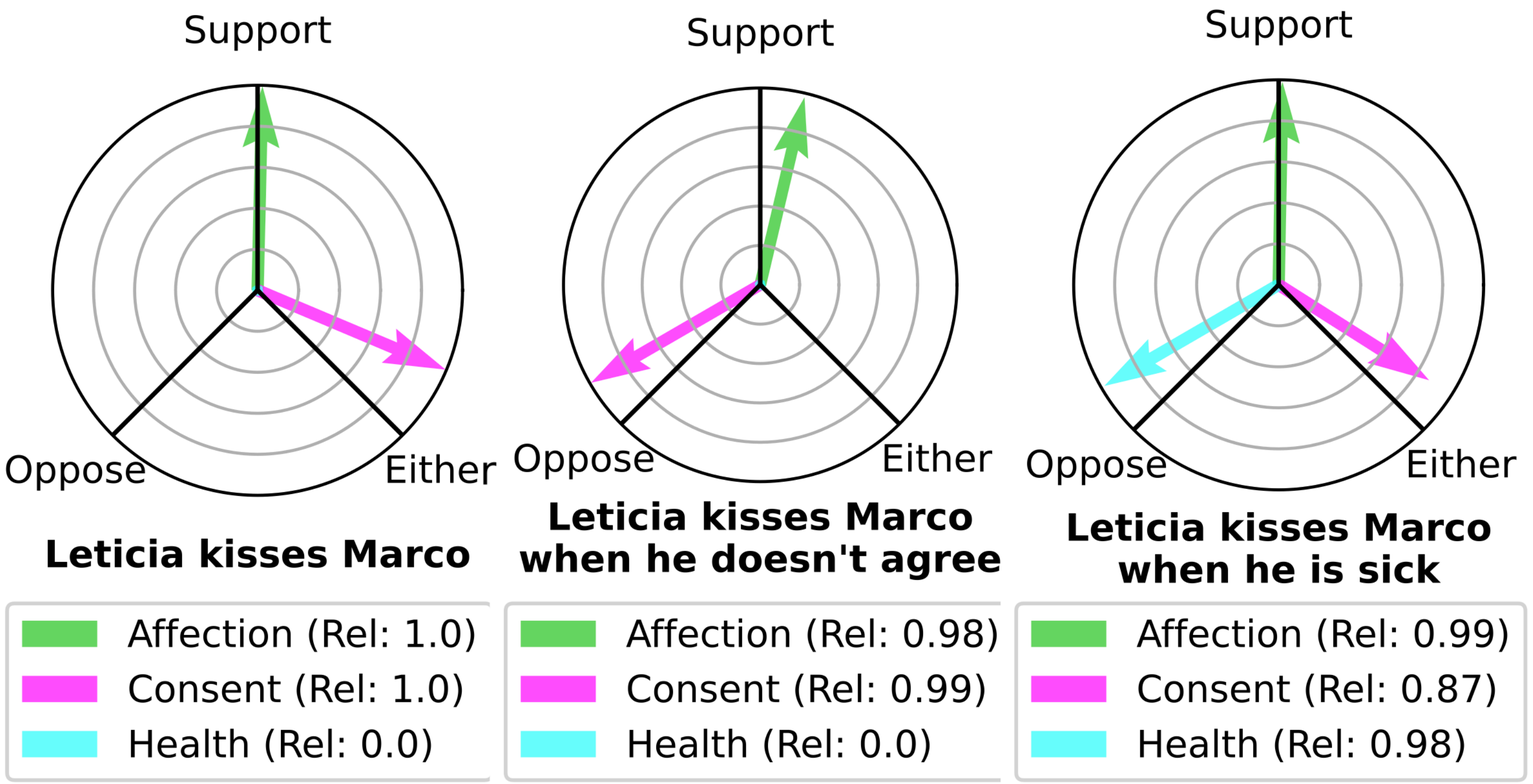
Kaleido's contrasting values help explain variability in human decision-making





High entropy => More Variability

Kaleido is sensitive to variations



Declaration of Human Rights 

Matches for 97.5% of the UDHR's articles

UDHR	ValuePrism
Everyone has the right to a nationality	Right to nationality
Everyone, without any discrimination , has the right to equal pay for equal work.	Right to equal pay
Everyone has the right of equal access to public service in his country.	Right to access services
Everyone has the right to rest and leisure, including reasonable limitation of working hours and periodic holidays with pay.	Right to engage in leisure activities

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

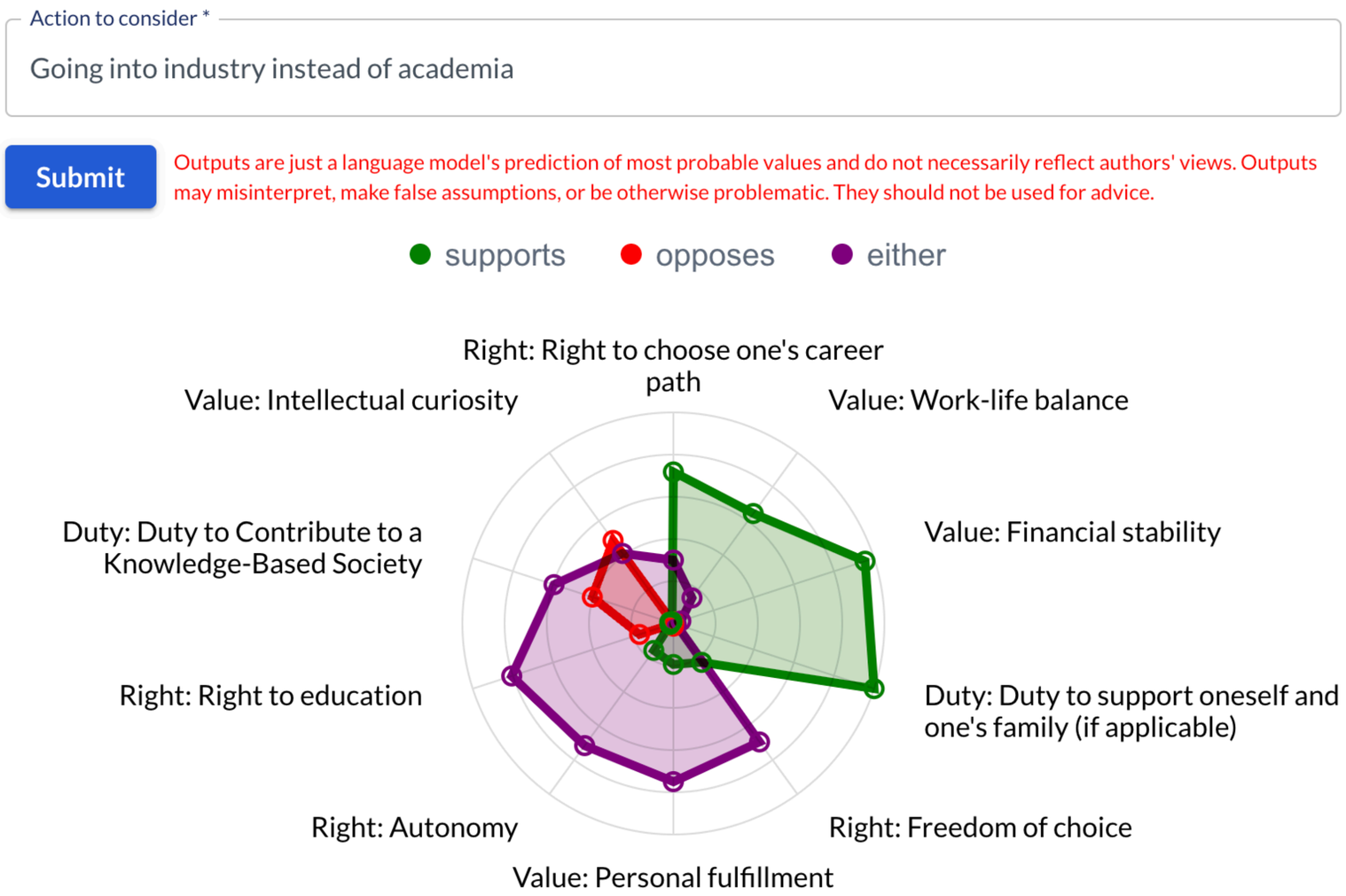
Not Intended for Advice

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

We hope Kaleido serves as a first step to better model **pluralistic human values, rights, and duties**

Demo:

kaleido.allen.ai



Where are we heading towards in the future?

Many unsolved mysteries in AI

… and Humanity

Open Research Challenges

… when we try to find **morally salient factors** that impact human moral decision-making

… when we try to define what moral **understanding & reasoning** means for humans

… when we try to identify how **multi-cultural norms** are manifested in human society

… when we wonder how to advance AI alignment to accommodated **pluralistic human values or conflicted views in society**

… when we try to quantify **the disparate impact of biases or toxicity** on different people

…

Open Research Challenges

… when we try to find **morally salient factors** that impact **human** moral decision-making

… when we try to define what moral **understanding & reasoning** means for **humans**

… when we try to identify how **multi-cultural norms** are manifested in **human society**

… when we wonder how to advance AI alignment to accommodated **pluralistic human values or conflicted views in society**

… when we try to quantify **the disparate impact of biases or toxicity** on different **people**

…

Do we (as not only <u>AI researchers</u> but in general as <u>humans</u>) understand humans well enough to advance AI to the next level?

Current Paradigm in Human → AI

Taking **inspirations** from existing knowledge about humans to **model "intelligence" in machines**

e.g., chain-of-thought prompting, dual-process reasoning with system-1/2), evaluate models on human capabilities

Current Paradigm in AI → Human

AI

Computational linguistics, computational psychology, computational social science…

Applied Disciplines (in Humanity)

In turn, AI benefits sciences by developing useful **models, tools, and methods** that can be used to **simplify** and **bolster** the existing approaches in many **applied disciplines**

e.g., vaccine development, educational evaluation tools, assist psychotherapy, analyzing big data for social phenomenon

Current Paradigm in AI → Human

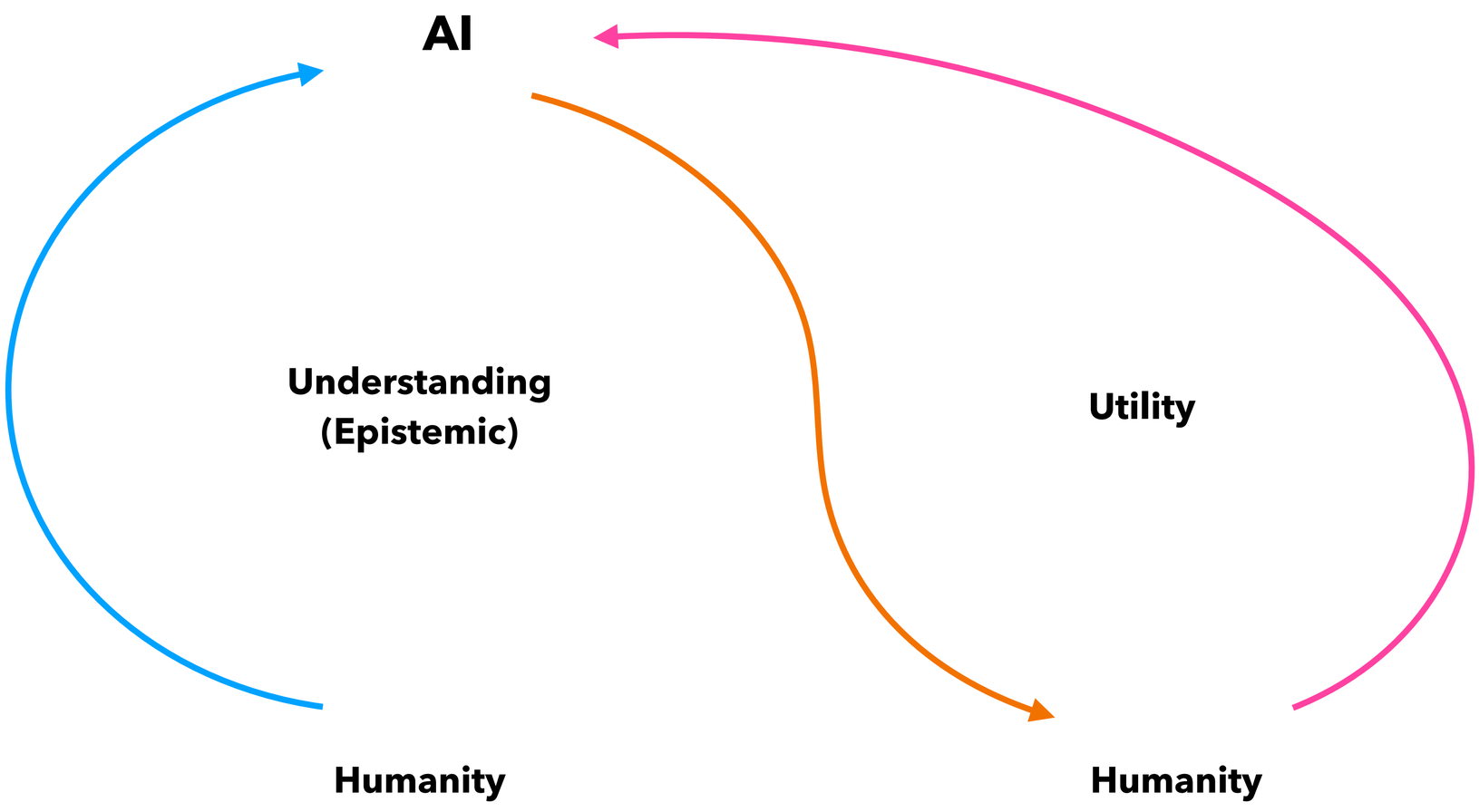
AI

Feedback to AI to improve its **utility**.
Develop better AI to improve human **experiences** (e.g., education, finance, scientific paper reading). There are disciplines like **HCI** that specializes in this feedback loop

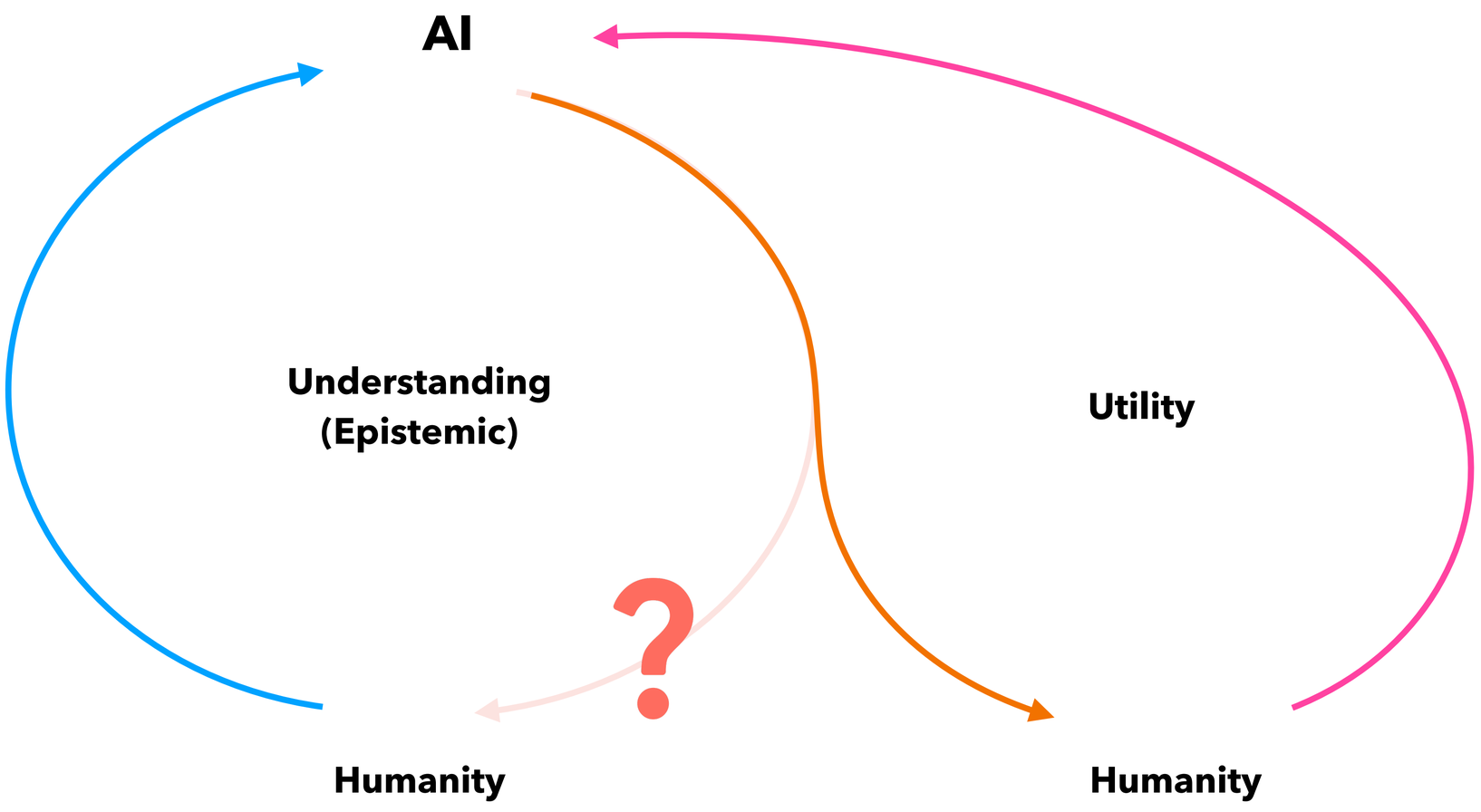
Applied Disciplines (in Humanity)

Sometimes, the insights and results taken from the application can also feedback into the **further development of AI tools**.

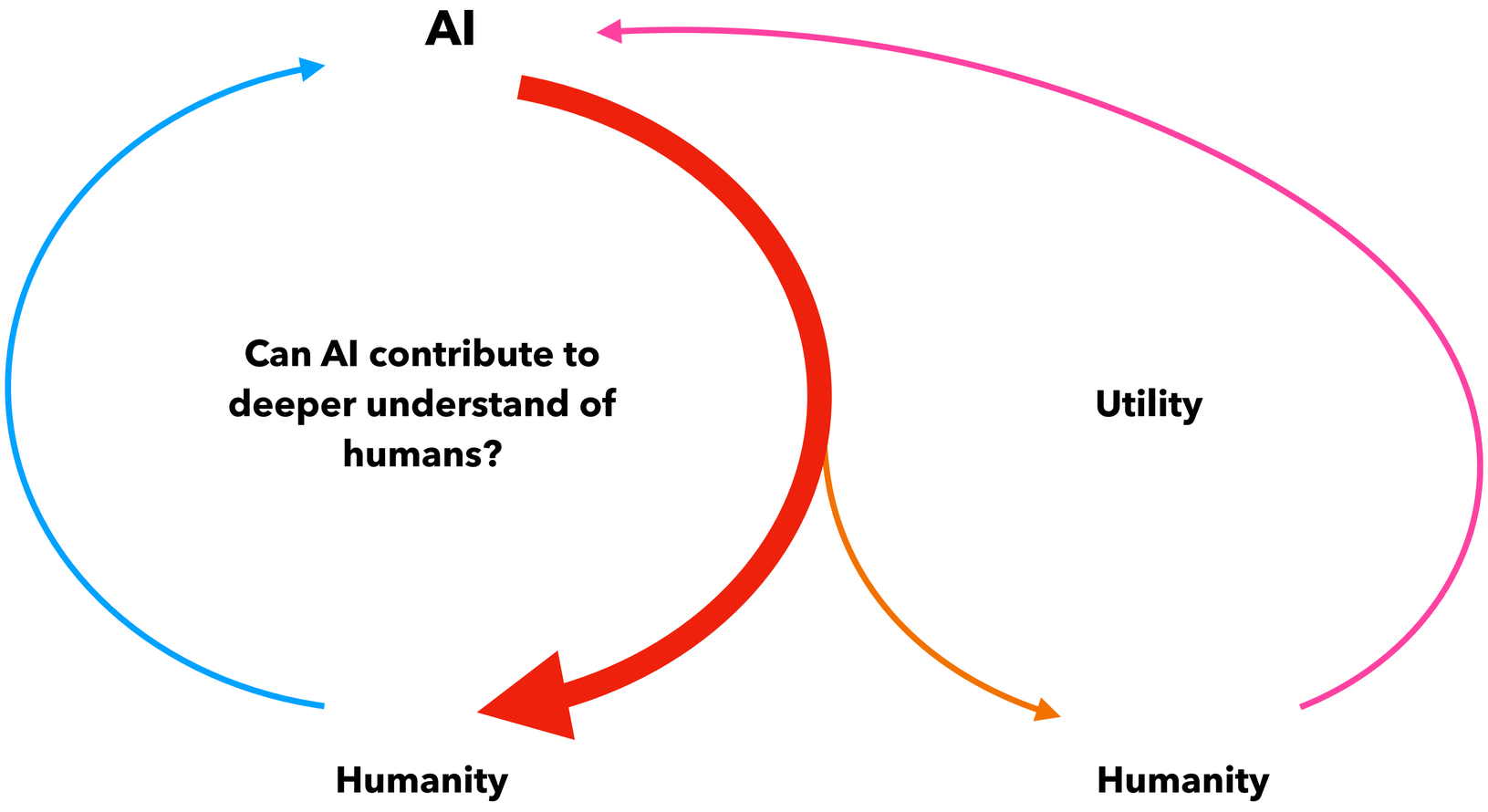
Current Paradigm



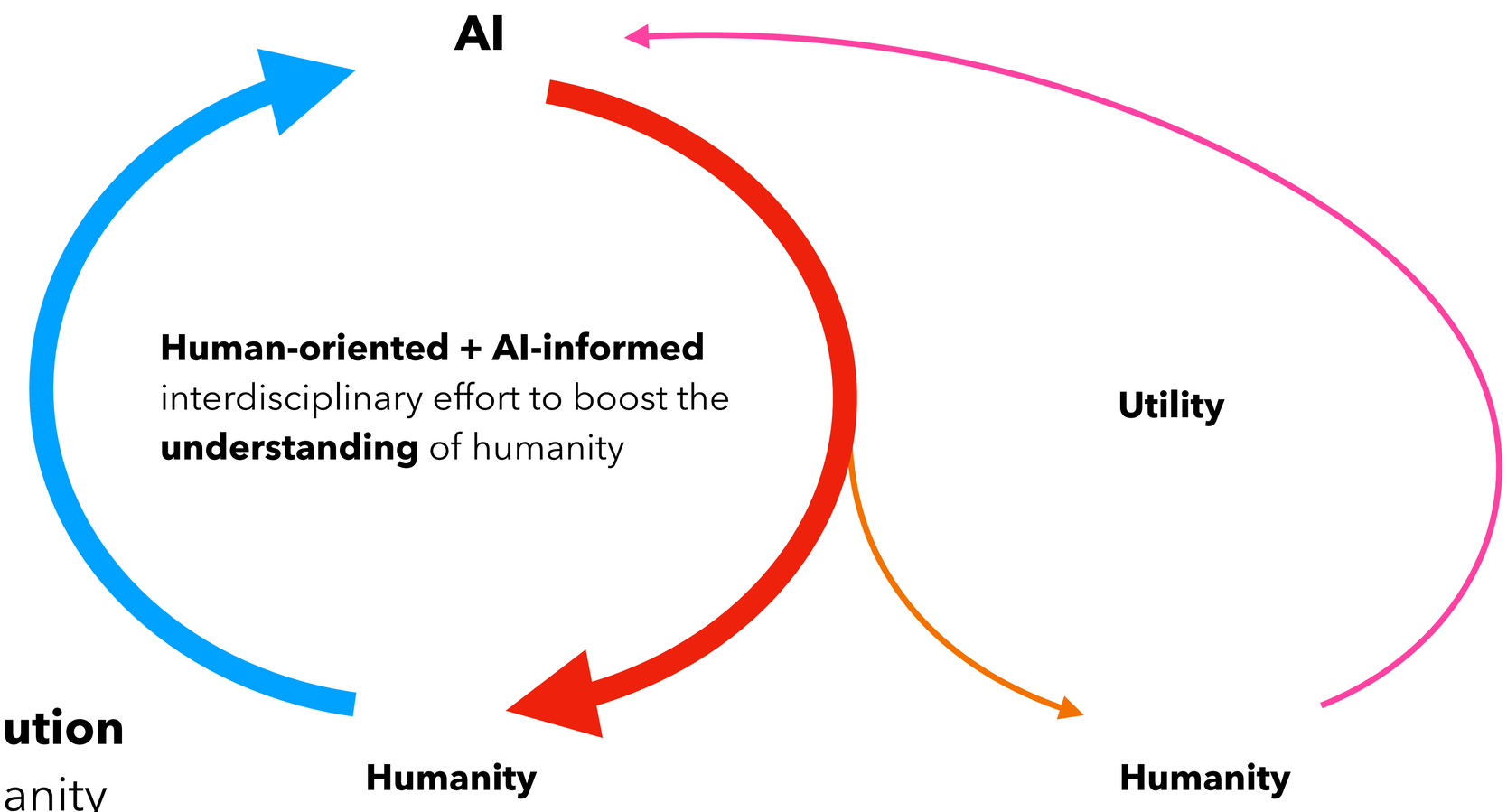
Missing piece?



Can the <u>process of AI development</u> contribute to deeper understand of humans?



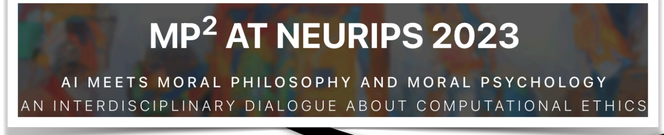
AI ⟷ Humanity

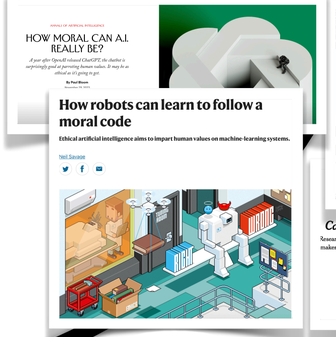


A Co-evolution
AI & Humanity

We (as AI/ML/NLP researchers) need better ways to **approach human-facing challenges** and **engage interdisciplinary knowledge** in building better AI!

Conversely, we can **contribute to the understanding of humans** too via building AI.






Thank You!

Thank You!

Liwei Jiang

lwjiang@cs.washington.edu

University of Washington

Allen Institute for AI

Happy to chat anytime!

What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations

Kavel Rao[♡*] Liwei Jiang[♡♠*] Valentina Pyatkin[♠] Yuling Gu[♠]
Niket Tandon[♠] Nouha Dziri[♠] Faeze Brahman[♠] Yejin Choi[♡♠]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Allen Institute for Artificial Intelligence
{kavelrao,lwjiang}@cs.washington.edu

— Findings at EMNLP 23 —

Poster 4322, Saturday, Dec. 9, 9:00AM

Reading Books is Great, But Not if You Are Driving! Visually Grounded Reasoning about Defeasible Commonsense Norms

Seungju Han[♠♡] Junhyeok Kim[♣] Jack Hessel[♡] Liwei Jiang[♢♡]
Jiwan Chung[♣] Yejin Son[♣] Yejin Choi[♢♡] Youngjae Yu[♣♡]
♠ Seoul National University ♡ Allen Institute for Artificial Intelligence
♣ Yonsei University ♢ University of Washington
wade3han@snu.ac.kr

— EMNLP 23 —

Oral 1846, Central 1, Friday, Dec. 8, 4:30PM

MP^2 AT NEURIPS 2023

AI MEETS MORAL PHILOSOPHY AND MORAL PSYCHOLOGY

AN INTERDISCIPLINARY DIALOGUE ABOUT COMPUTATIONAL ETHICS

AI meets Moral Philosophy and Moral Psychology Workshop (MP2) @ NeurIPS, Dec 15 2023

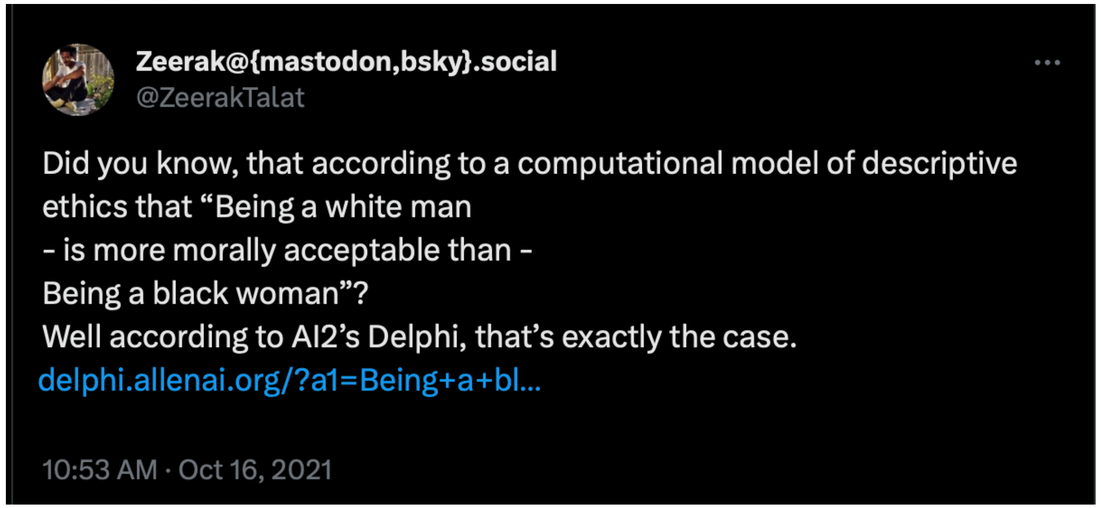
Zee's Part of the Talk

- Overview of timeline
- Our considerations around the response
- Work arising since

A Timeline of Objections

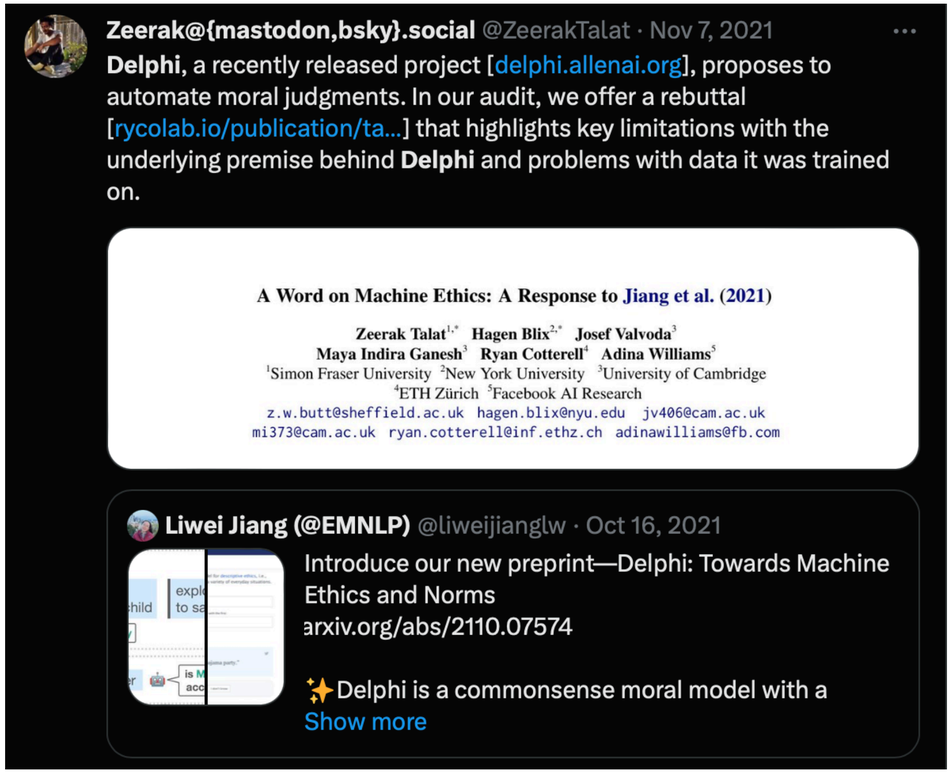
~3 AM Oct -16, 2021: A friend asks me if I've seen Delphi

~3:50 AM (local time): I tweet



A Timeline of Objections

~*3 AM Oct -16, 2021:* A friend asks me if I've seen Delphi
~3:50 AM: I tweet
~Oct 20: Initial call w/ co-authors to discuss response
~Oct 20 - Nov 6: Drafting Response
    Notable moments
        Consider dropping it
Nov 7: Release of our initial draft
Second & last time I tweet about it



October 20 - Nov 7

Wanted to drop it each time there was a lull in discourse
Goal: Write a critique and response
    But also try to avoid being overly harsh

Outcomes

A lot of hurt feelings
A lot (too much?) discourse online
Strategically good/bad decisions

Fundamental Disagreements

Science and conflict go hand in hand
Care work is essential to this
    Care work != accepting bad faith behavior
Honesty is key
Demands cannot exist

Work since

Making clear the assumptions held in NLP
Truly believe: We can be better as a field

Event	**Driving at the maximum speed of the road**	
Value	Efficiency	support
	Safety	oppose
	Rule-following	oppose
	Law-abiding behavior	oppose
Right	Right to freedom of movement	either
	Right to safety	oppose
	Right to life and physical integrity	oppose
Duty	Duty to follow traffic laws	oppose
Event	**Gambling with my boss**	
Value	Professionalism	oppose
	Loyalty	oppose
	Personal enjoyment	support
	Financial well-being	oppose
Right	Right to work in a safe and professional environment	oppose
	Autonomy	either
Duty	Duty to maintain a professional relationship with your boss	oppose
	Duty to uphold company policies	oppose
Event	**someone cuts me off. i honk ten times**	
Value	Courtesy	oppose
	Respect for others	oppose
	Safety	support
	Harmony	oppose
Right	Right to peaceful enjoyment of the road	oppose
	Right to not be subjected to harassment	oppose
	Right to safety	oppose
Duty	Duty to express displeasure	support
	Duty to be a considerate driver	oppose

Table 1: Example outputs from KALEIDO^{SYS}.

Sorensen et al. "Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties" 2023.

A person whose reflection is being distorted by mirrors.
Source: funplanners.com

Boris Karloff as Frankenstein's Monster.
Source: Frankenstein (1931)

Discussion

How did we resolve our conflicts?

Discussion

Our views on how to communicate research disagreement effectively?

Questions?