Tomea: an Explainable Method for Comparing Morality Classifiers across Domains

Enrico Liscio, Oscar Araque, Lorenzo Gatti,
Ionut Constantinescu, Catholijn M. Jonker,
Kyriaki Kalimeri, Pradeep K. Murukannaiah

TU Delft

Hybrid Intelligence

Universidad Politécnica de Madrid

ETH zürich

University of Twente.

ISI Foundation

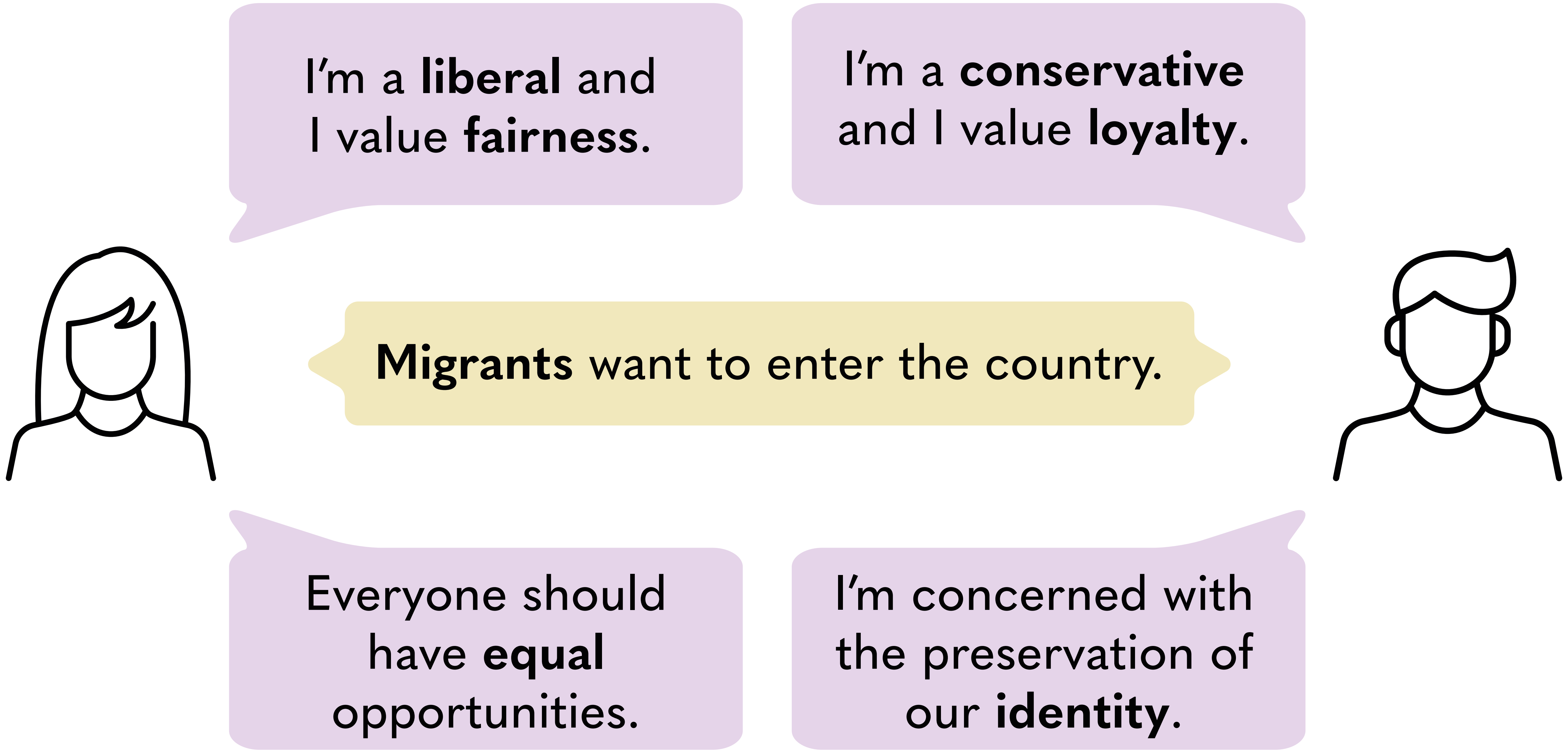
What is Morality?

Morality helps us **distinguish right from wrong**.

According to the **Moral Foundations Theory**, each situation can trigger one (or more) of these five moral elements:

care/harm
fairness/cheating
loyalty/betrayal
authority/subversion
purity/degradation

Moral Values Explain our Differences



I'm a **liberal** and I value **fairness**.

I'm a **conservative** and I value **loyalty**.

Migrants want to enter the country.

Everyone should have **equal** opportunities.

I'm concerned with the preservation of our **identity**.

Detecting Moral Values in Text

Detecting the moral rhetoric behind a statement allows artificial agents to recognize **individual differences** across humans.

Language models have been shown to be capable of recognizing moral rhetoric in text.

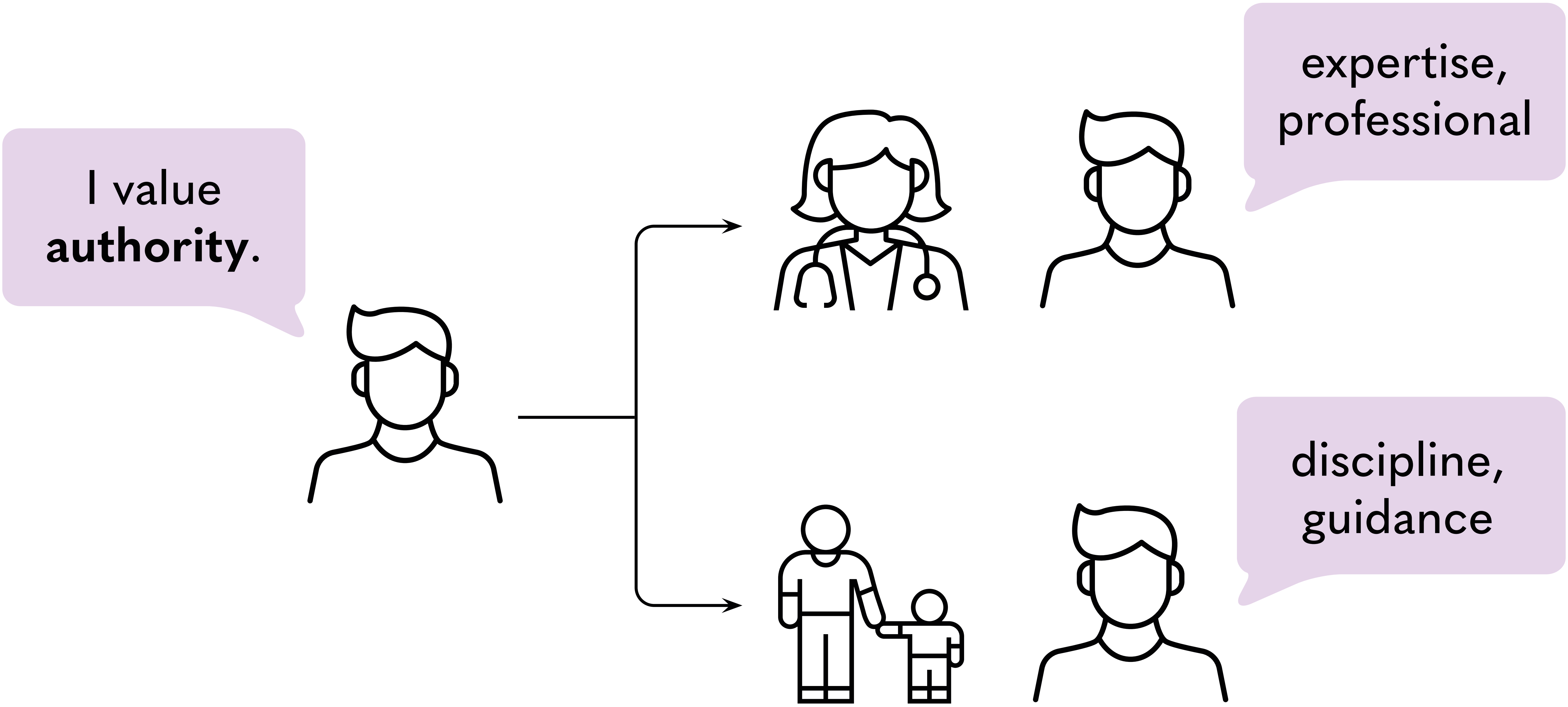
Moral Values Classification

“Police lives matter, all lives matter, peace and love people” → care

“Which oppression is worse, sexism or racism?” → harm, cheating

“Baltimore Police will deliver an update on the #FreddieGray investigation. Listen live on WBAL” → non-moral

Moral Rhetoric is Domain-Dependent

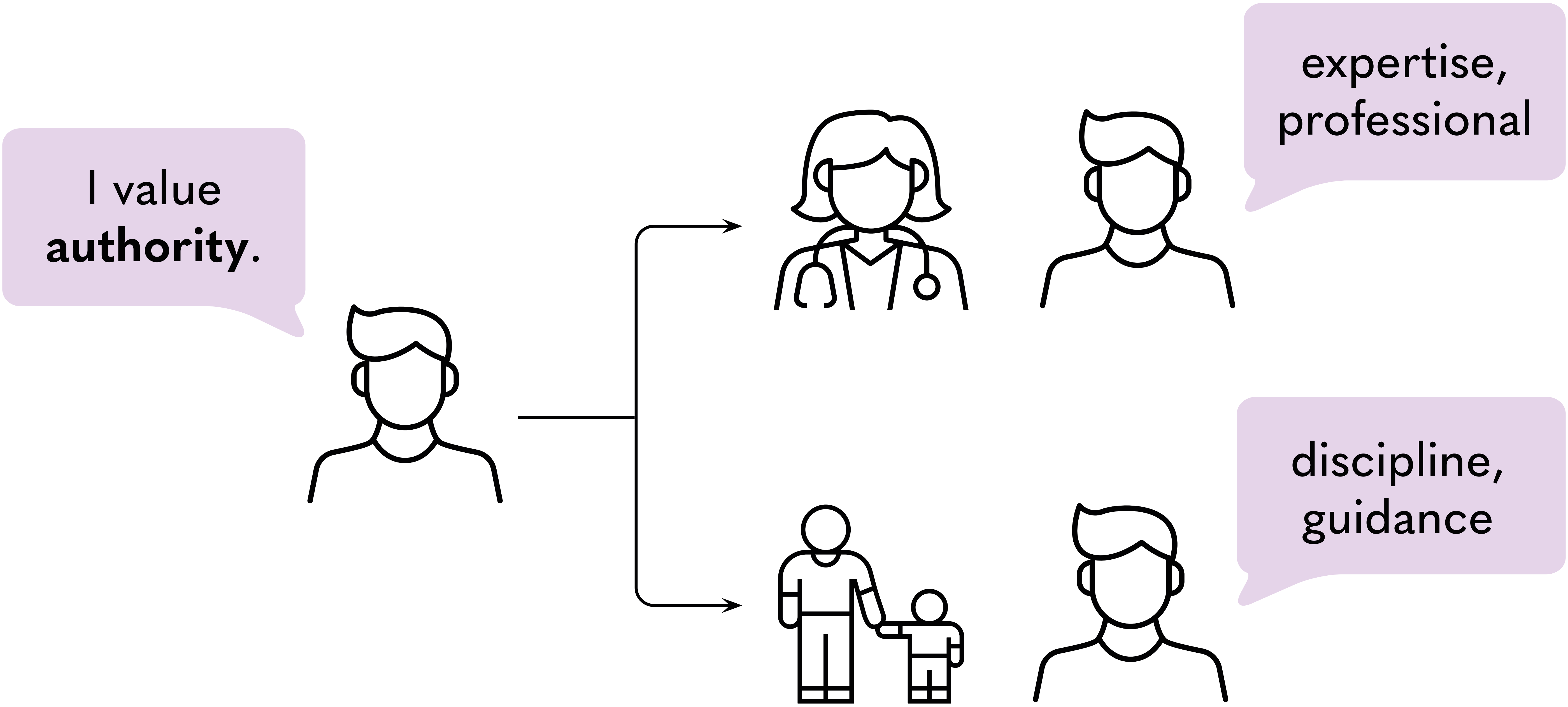


I value **authority**.

expertise, professional

discipline, guidance

Moral Rhetoric is Domain-Dependent

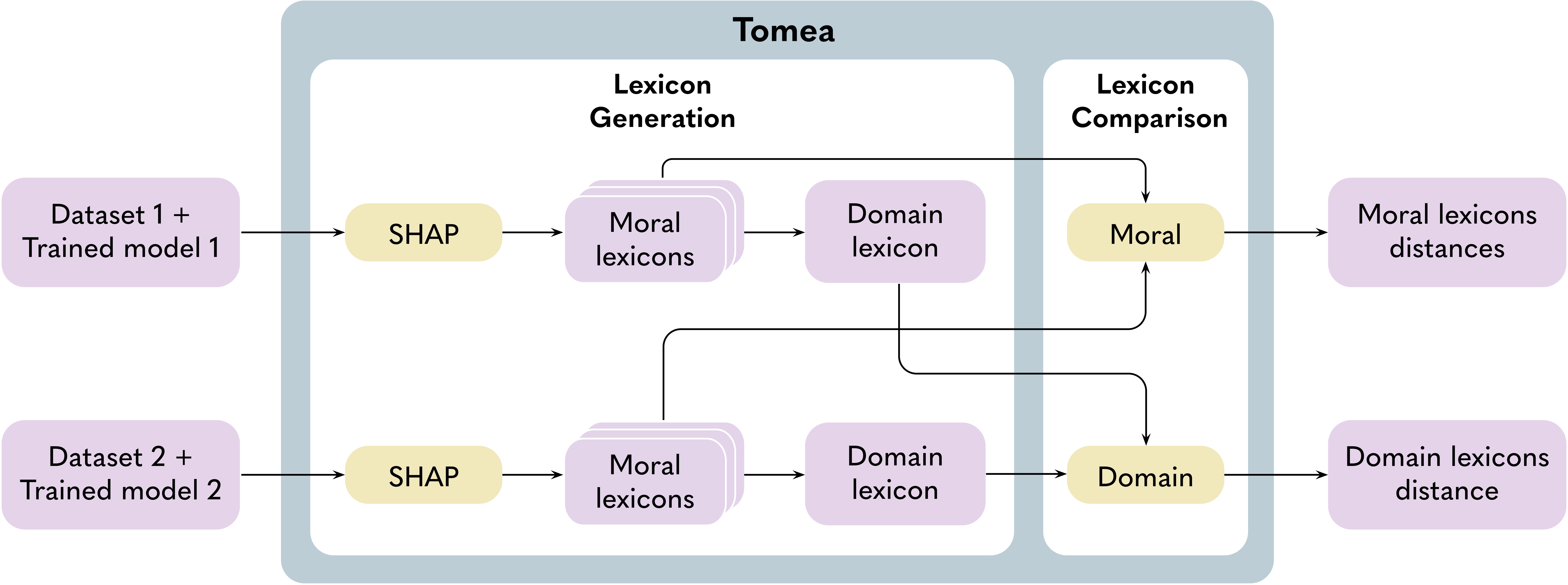

I value **authority.**

expertise, professional

discipline, guidance

Do language models detect domain-specific language?

Tomea: XAI Method



Cross-Domain Experiments

Cross-domain comparison of BERT trained in the **seven** domains of the Moral Foundation Twitter Corpus (35k tweets):

#hatespeech
#Baltimoreprotests
#ALM
#BLM
#MeToo
#hurricaneSandy
#elections2016

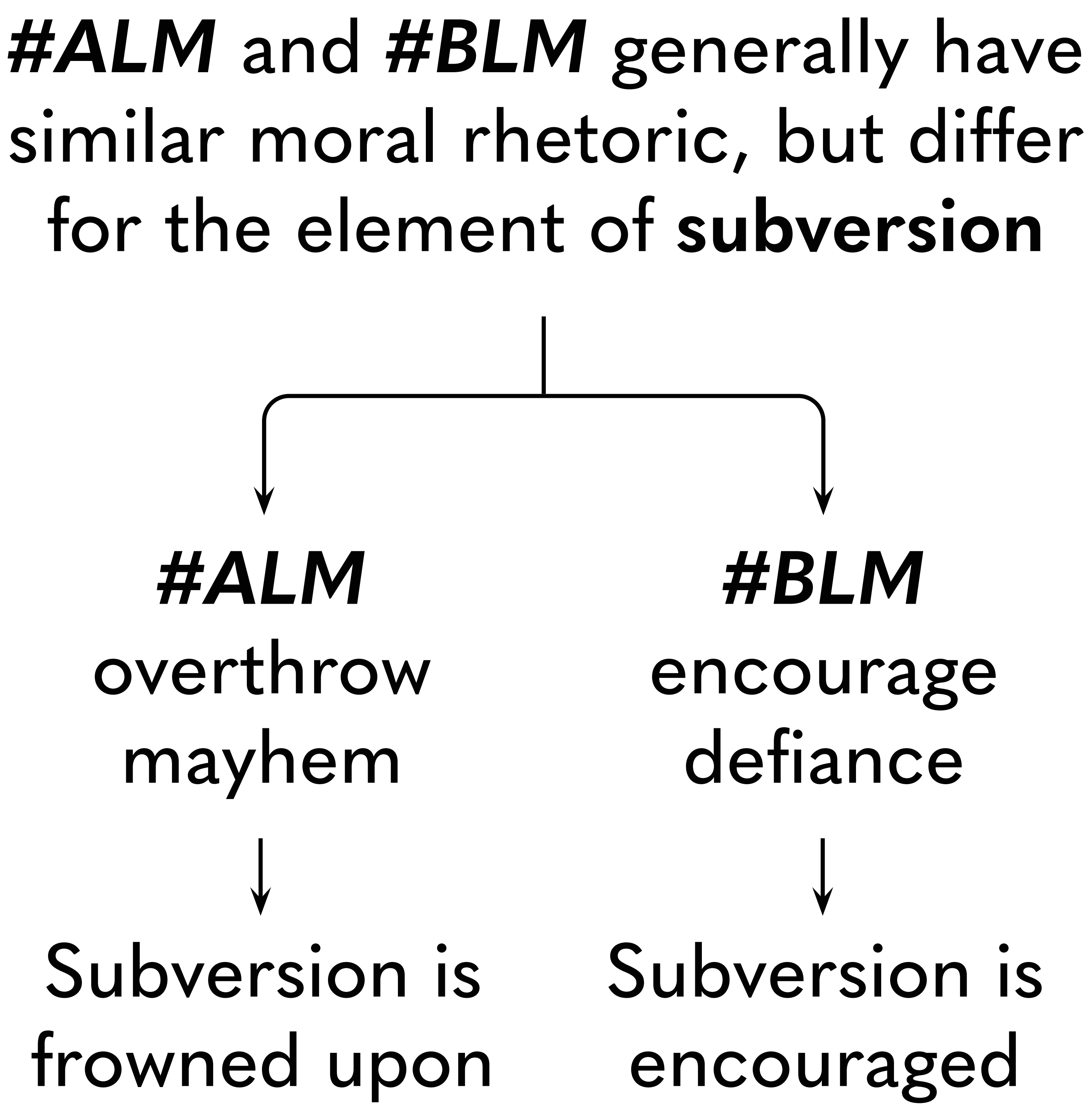
We perform **quantitative** and **qualitative** comparisons across domains.

Quantitative Comparisons

Crowd workers moderately agree with the fine-grained moral lexicon similarities between domains (correlation of 0.4).

High Tomea similarity between domains entails better **out-of-domain performance** of the models (correlation of 0.79).

Qualitative Comparisons



#ALM and ***#BLM*** generally have similar moral rhetoric, but differ for the element of **subversion**

#ALM
overthrow
mayhem

Subversion is
frowned upon

#BLM
encourage
defiance

Subversion is
encouraged

Takeaways

- Tomea is an XAI method that helps us compare how language models **represent morality across domains**.

- Our experiments with Tomea show that language models recognize **small differences** in moral language in different domains.

- Small but **critical differences** between domains may not affect quantitative results, but may **hinder usage** in a novel domain.

Thank you!

Any Questions?

Link to the paper!

✉ e.liscio@tudelft.nl

🌐 enricoliscio.github.io

Enrico Liscio, Oscar Araque, Lorenzo Gatti,
Ionut Constantinescu, Catholijn M. Jonker,
Kyriaki Kalimeri, Pradeep K. Murukannaiah

TUDelft

UNIVERSIDAD POLITÉCNICA DE MADRID
POLITÉCNICA

UNIVERSITY OF TWENTE.

Hybrid Intelligence

ETH zürich

ISI
ISI Foundation