Context-Specific Value Inference via Hybrid Intelligence

Enrico Liscio

TU Delft

Hybrid Intelligence

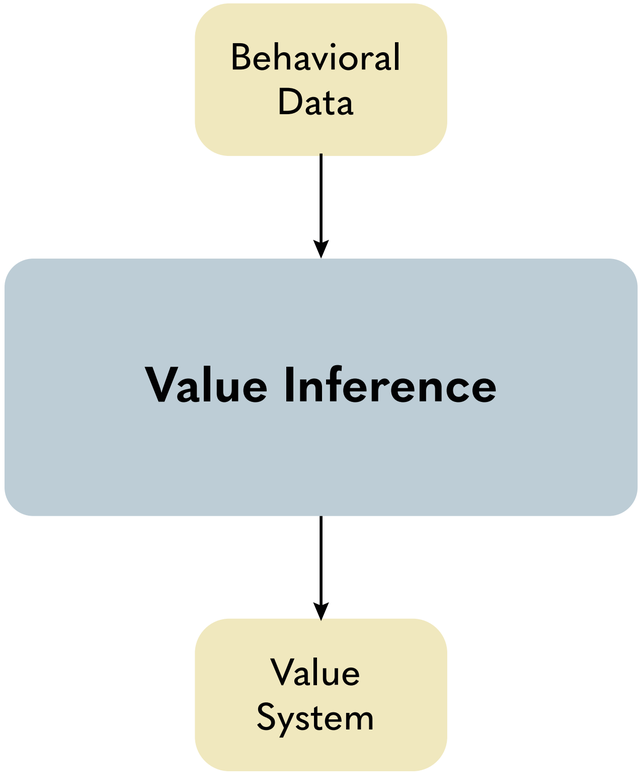
Value Alignment in Sociotechnical Systems

Value Inference

Which values are **relevant** to a decision-making **context**?

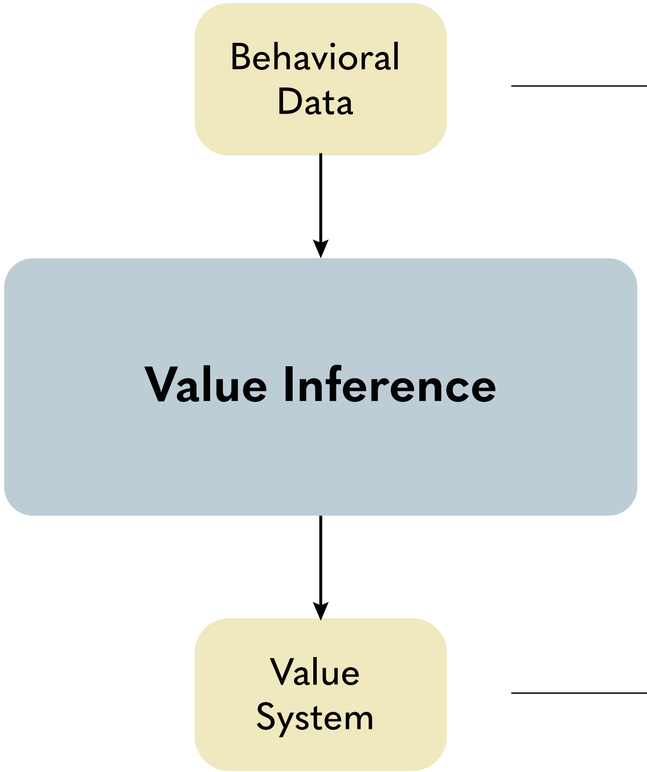
How do **different** stakeholders **prioritize** the relevant values?

Value Inference



Liscio et al. “Value Inference in Sociotechnical Systems.” *AAMAS*, 2023.

Value Inference

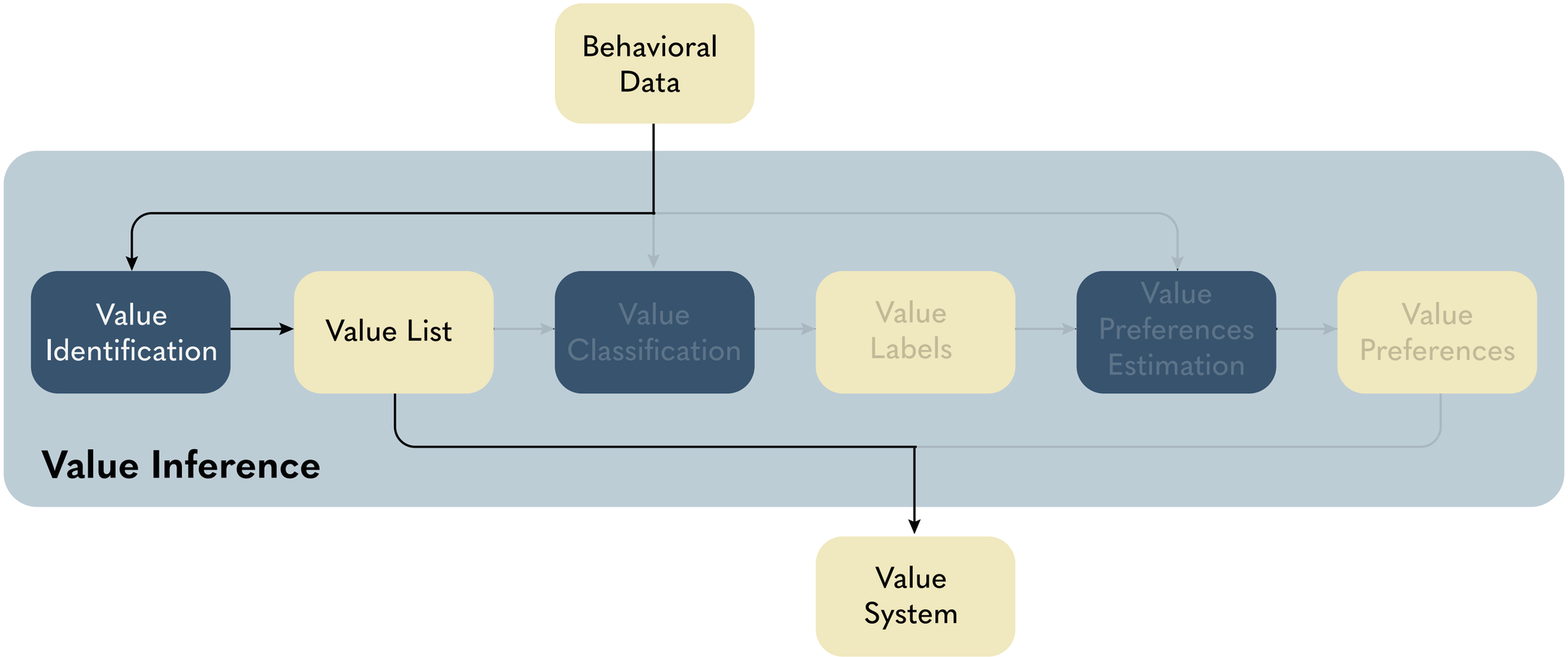


Stakeholders' **actions** (e.g., how they choose among alternative options) and **justifications** (how they motivate their choices).

The set of **relevant values** and the **preferences** that a stakeholder attibutes to them.

Liscio et al. "Value Inference in Sociotechnical Systems." *AAMAS*, 2023.

Value Identification



Value Identification

The challenge of identifying the values that are **relevant** to the decision-making context.

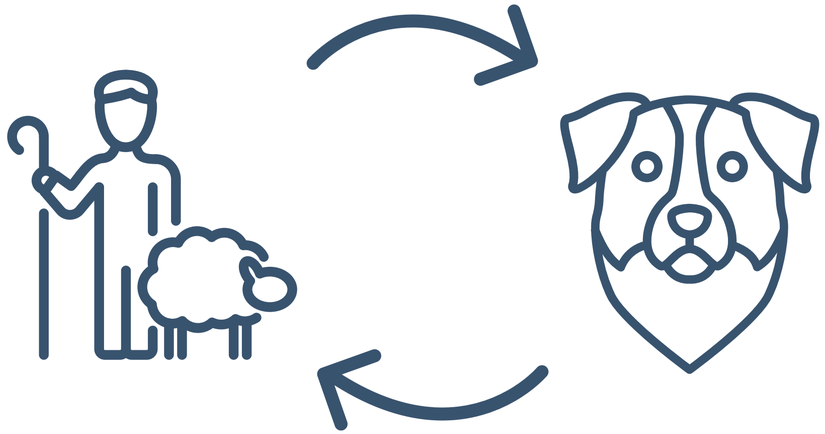
We strive to perform a **bottom-up** value identification that is based on the input of the relevant stakeholders.

AI helps us to **scale**, but we need (and want) **humans** to interpret the data.

Hybrid Intelligence

The **combination** of human and artificial intelligence can achieve more than the sum of the two.

Example: shepherd and shepherd's dog.

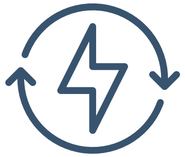


Example: Survey Data

Researchers in TU Delft performed two Participatory Value Evaluation **surveys** to gauge citizens' opinions on these topics:

COVID-19
(60k answers)

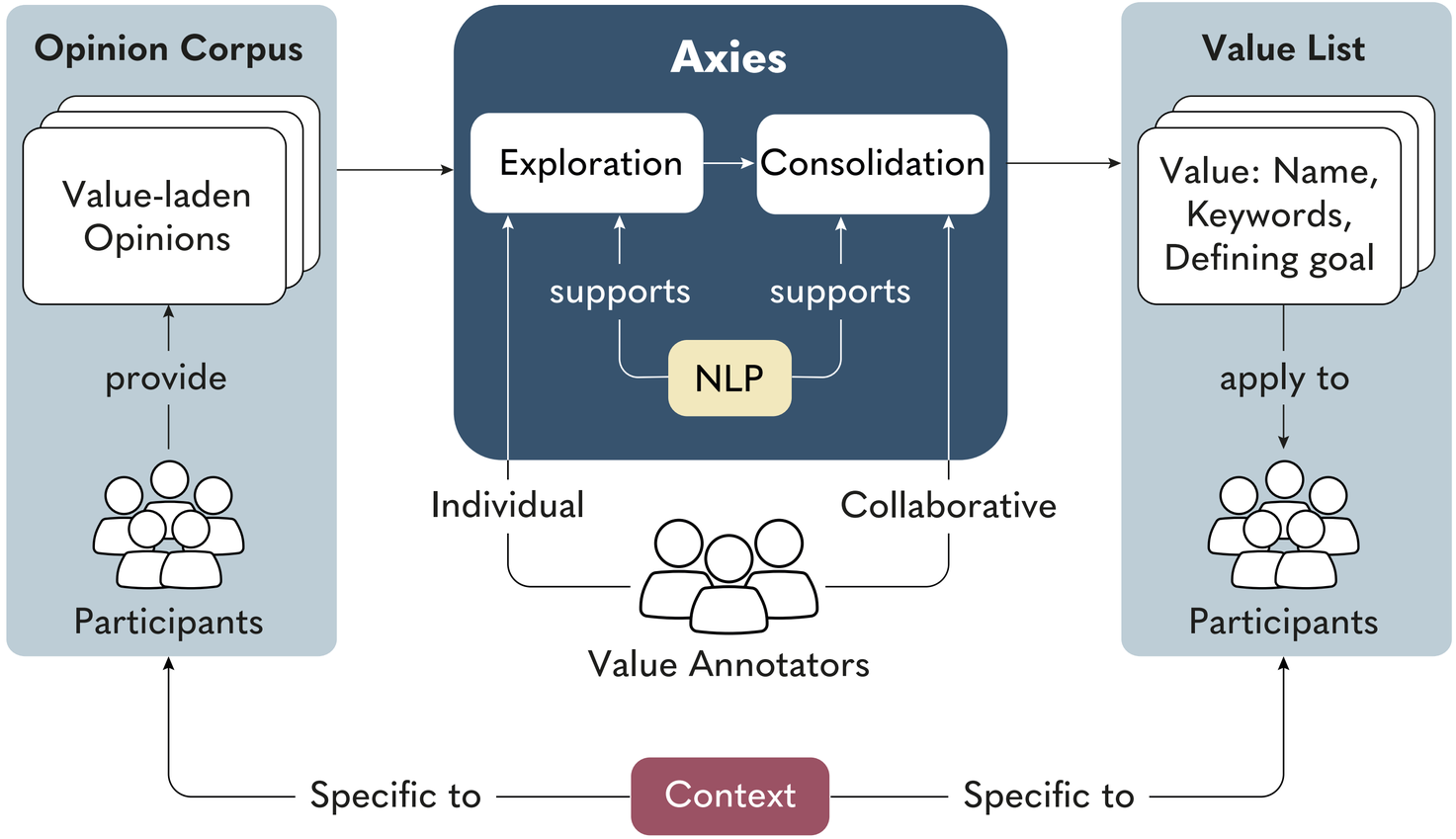
Mouter et al. "Public participation in crisis policymaking. How 30,000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures". *PLoSONE*, 2021.



Green Energy
(3k answers)

Itten et al. "When Digital Mass Participation Meets Citizen Deliberation: Combining Mini-and Maxi-Publics in Climate Policy-Making". *Sustainability*, 2022.

Axies Methodology



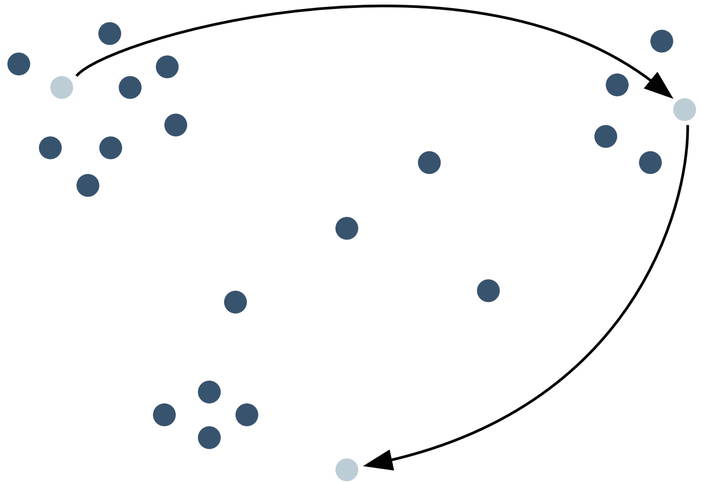
Liscio et al. "Axies: Identifying and Evaluating Context-Specific Values." *AAMAS*, 2021.

Liscio et al. "What Values Should an Agent Align With?" *JAAMAS*, 2022.

Axies Methodology

In the exploration phase, each annotator **independently** develops a value list.

The next survey answer to be analyzed is the **most different** from the already analyzed answers.

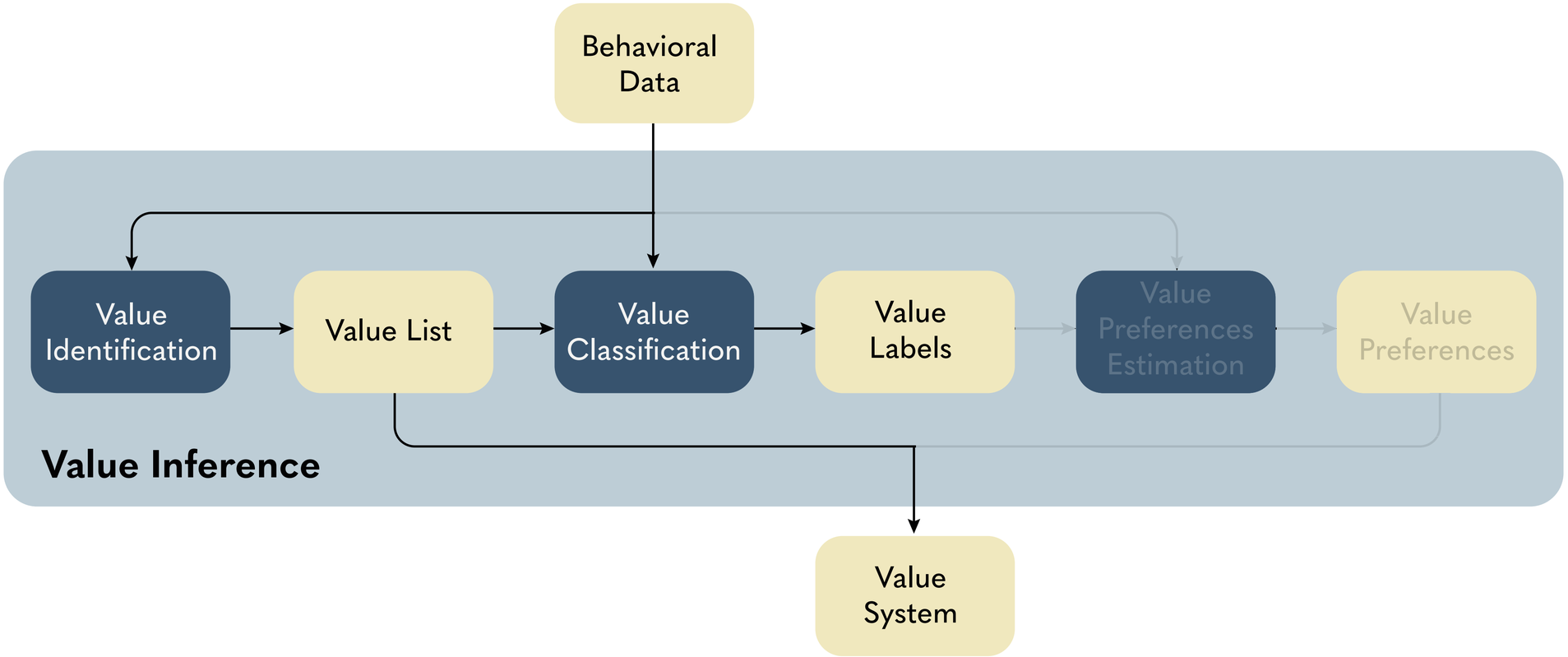


Axies Methodology

- Axies helps in identifying the values that are **relevant** to a decision-making context;

- Axies is a **HI methodology** where NLP and AL techniques guide **experts** in value identification.

Value Classification



Value Classification

The process of **detecting** value-laden content in natural language.

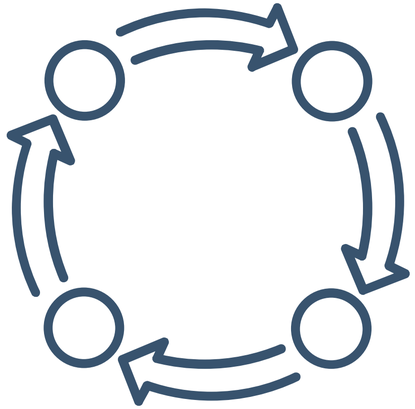
Value classification is **subjective** by nature, and highly domain-dependent. Here, we investigate the **domain dependency**.

Cross-Domain Value Classification

We perform cross-domain classification of moral values with the **Moral Foundation Twitter Corpus** (35k tweets).

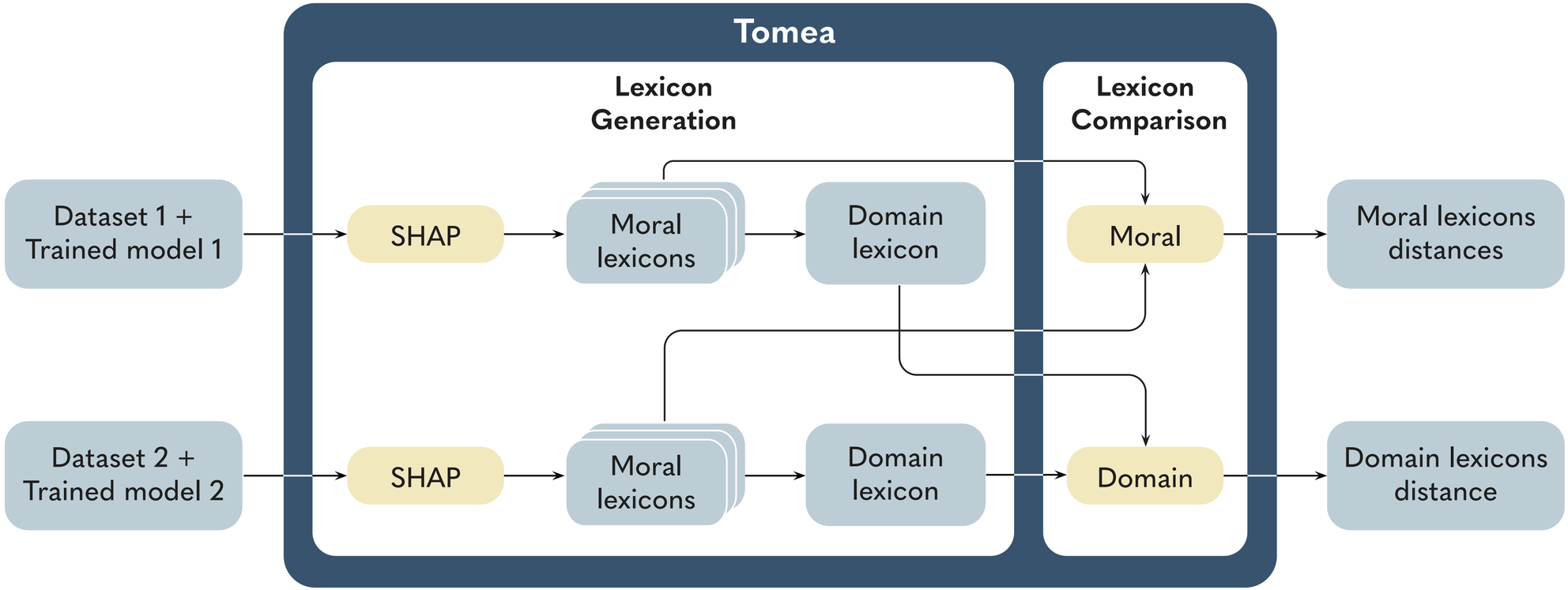
We evaluate across **seven domain** (e.g., #AllLives-Matter, #BlackLivesMatter, #hurricaneSandy) and four training modalities.

Our experiments show that models can (reasonably well) **generalize** to novel domains.



Liscio et al. "Cross-domain classification of moral values." *Findings of NAACL,* 2022.

What does a Classifier Learn about Morality?



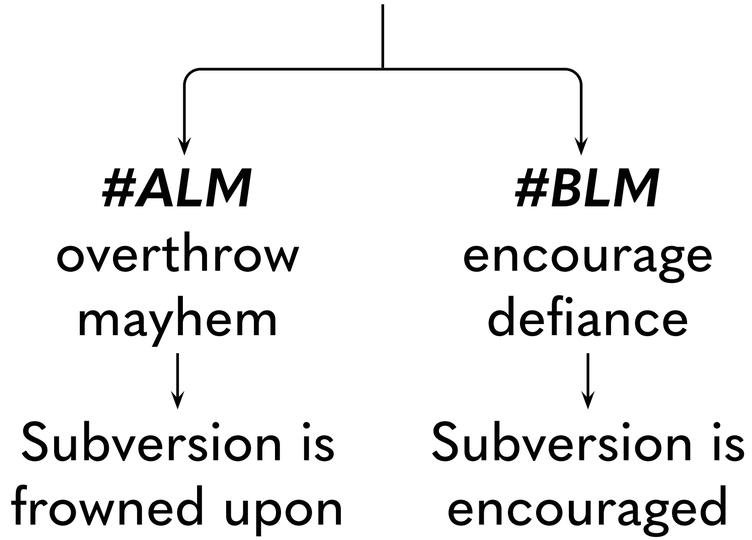
Liscio et al. "What does a text classifier learn about morality? An explainable method for cross-domain comparison of moral rhetoric." *Proceedings of ACL*, 2023.

What does a Classifier Learn about Morality?

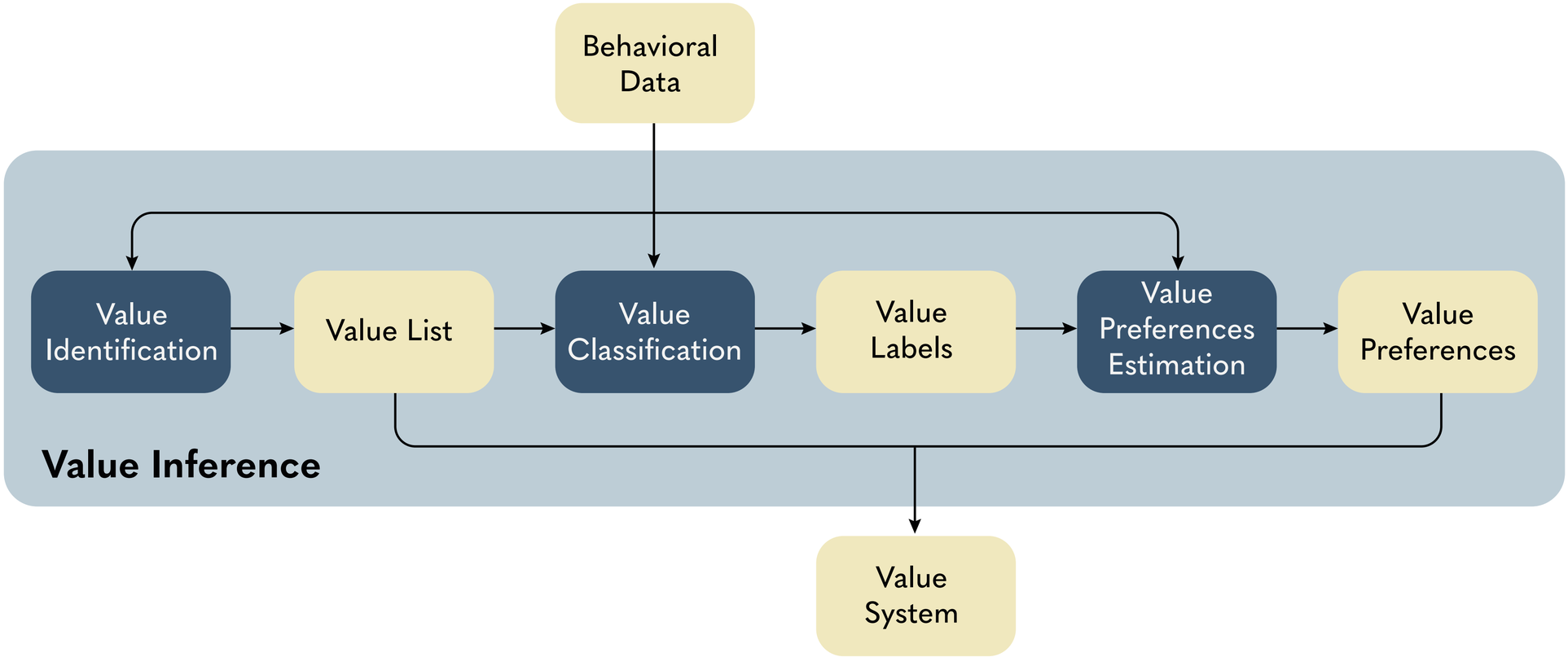
Language models recognize **small differences** in moral language across different domains.

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

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



Value Preferences Estimation



Value Preferences Estimation

The process of determining a **stakeholder's preferences** over a set of values based on their observed behavior.

Value preferences are estimated based on stakeholders' **actions** and the (processed) natural language **justifications** to their actions.

“Valuing is Deliberatively Consequential”

But what if actions and justifications are **inconsistent**?

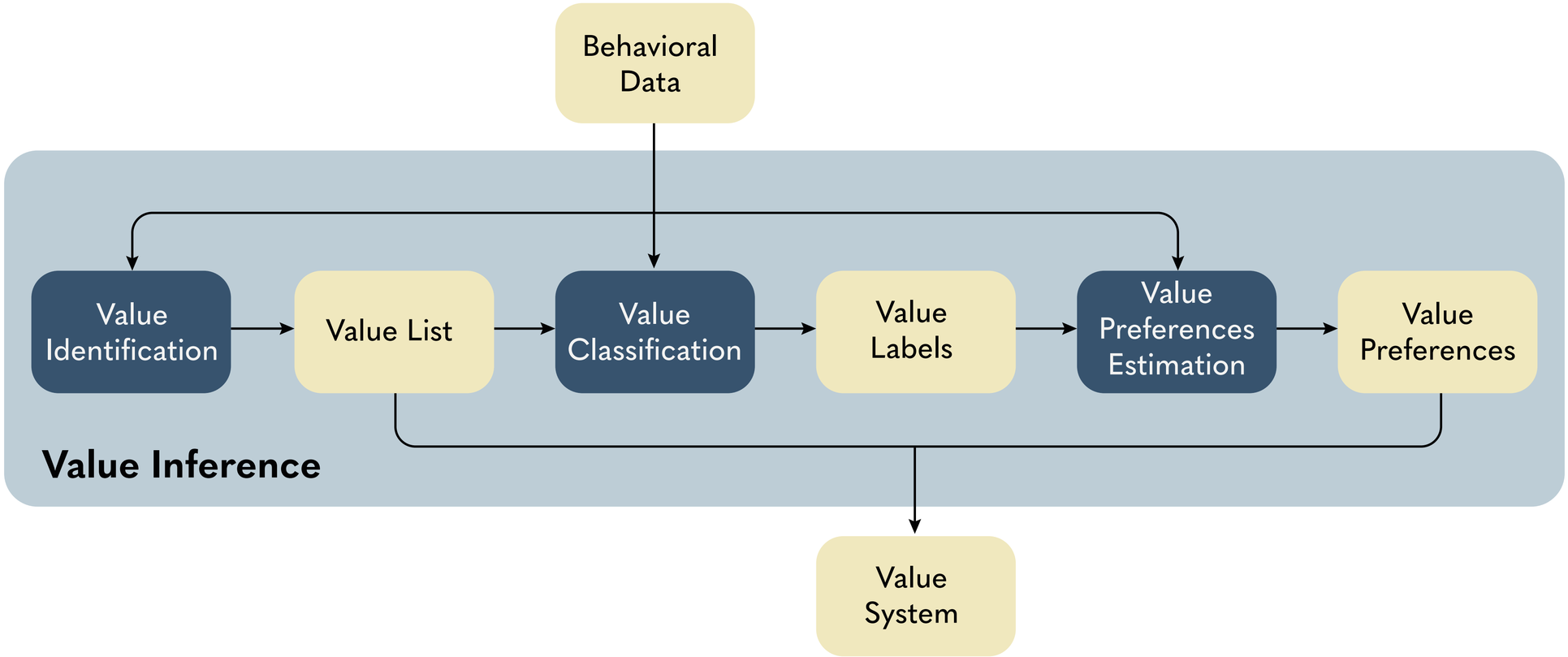
In case of conflicts between actions and justifications, we prioritize the preferences **estimated from justifications** over those estimated from actions.

This approach yields results that are **more aligned** with the value preferences estimated by humans.

Siebert et al., “Estimating Value Preferences in a Hybrid Participatory System.” *HHAI*, 2022.

Liscio et al., “Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems.” Under review at *JAIR*.

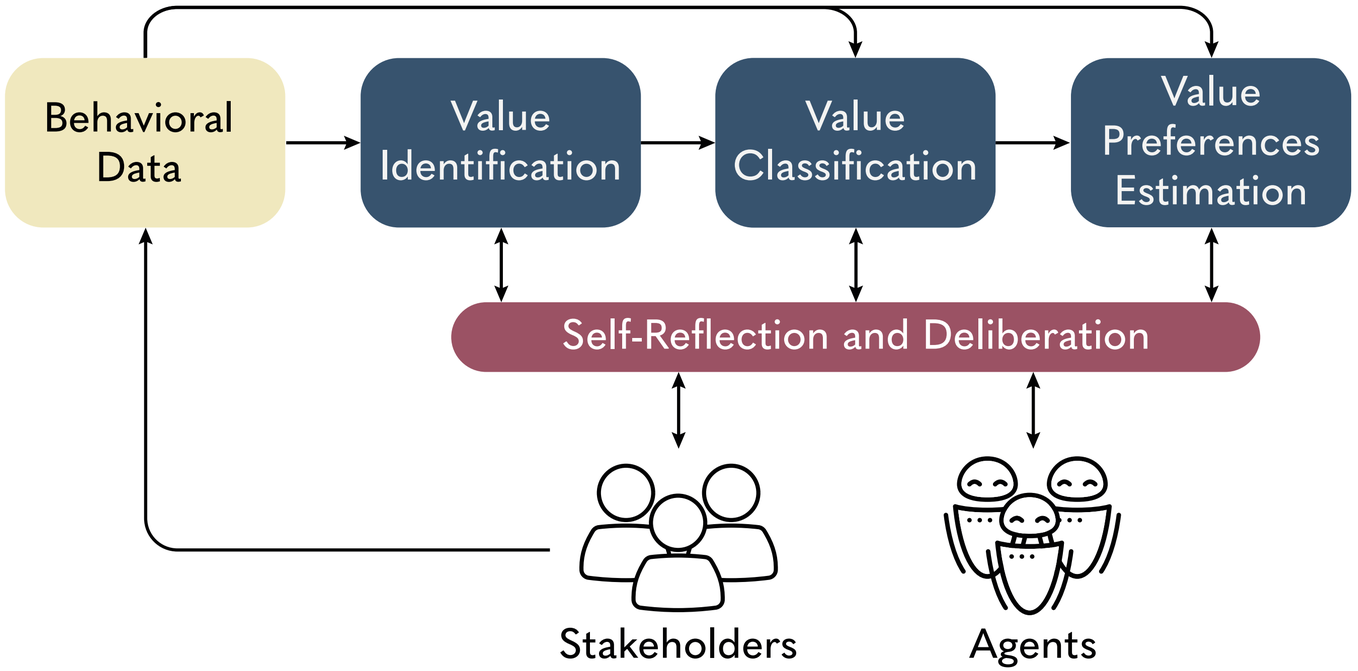
Value Inference



Observing is not enough

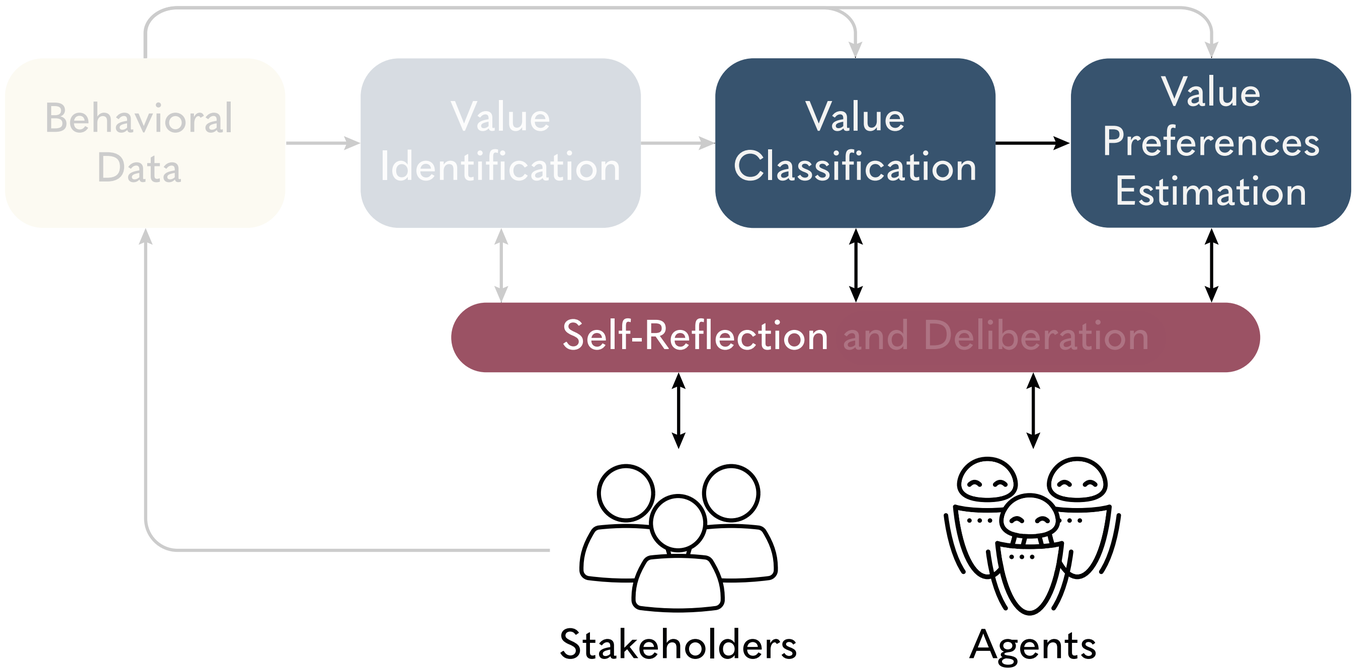
Value preferences are often **implicit** to ourselves, and thus not easily observable in behavioral data.

Hybrid Value Inference



Liscio et al. "Value Inference in Sociotechnical Systems." *AAMAS*, 2023.

Connecting Different Components



Liscio et al. “Value Inference in Sociotechnical Systems.” *AAMAS*, 2023.

Conclusion

- We introduced the **value inference** challenge and its components, focusing on its **context sensitivity**;
- We highlighted the importance of a **hybrid intelligence** approach;
- Remaining **challenges**: behavior observation/interpretation, context shifts identification, subjectivity, consent.

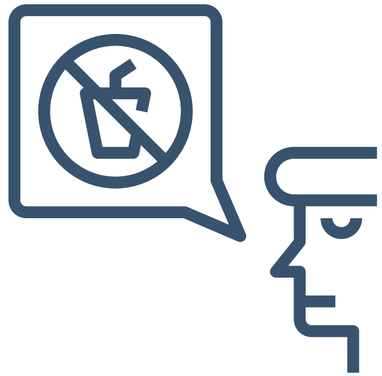
Practical Applications

- Support **policy-makers** in understanding the concerns of citizens;

  Lera-Leri, Roger X., et al. "Aggregating Value Systems for Decision Support." *Knowledge-Based Systems*, 2024.

- **Behavior change** support (e.g., learning to live with diabetes).

  de Boer, M.H., et al. "A contextual Hybrid Intelligent System Design for Diabetes Lifestyle Management", *ECAI - MRC workshop*, 2023.



Thank you!

Any Questions?

e.liscio@tudelft.nl

enricoliscio.github.io

Link to the value inference vision paper!