Context-Specific Values in Natural Language

Enrico Liscio
Pradeep K. Murukannaiah

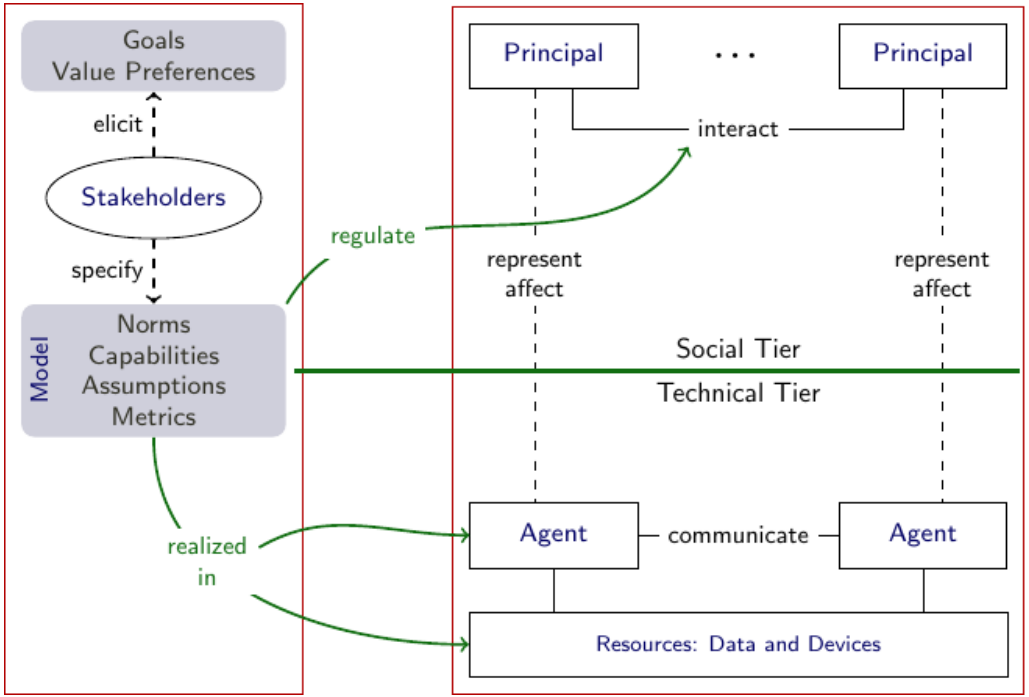
Table of contents

01

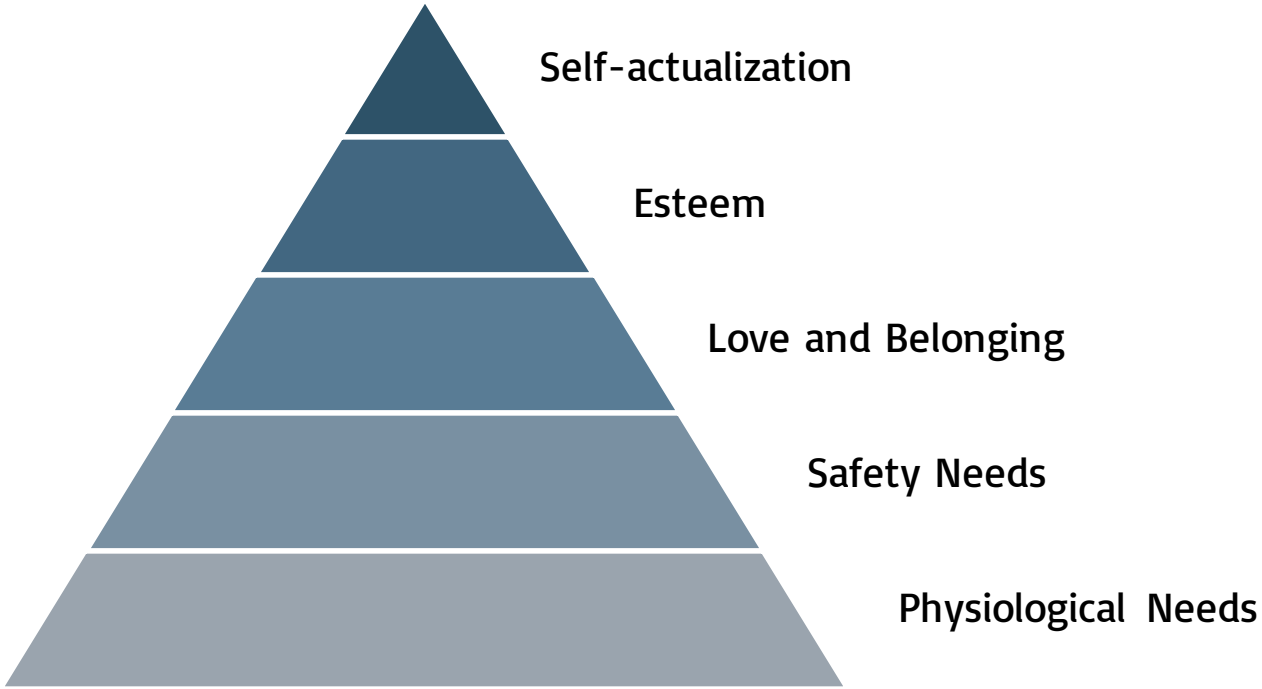
Values

Values define what we consider important in life.

AI in a Sociotechnical System

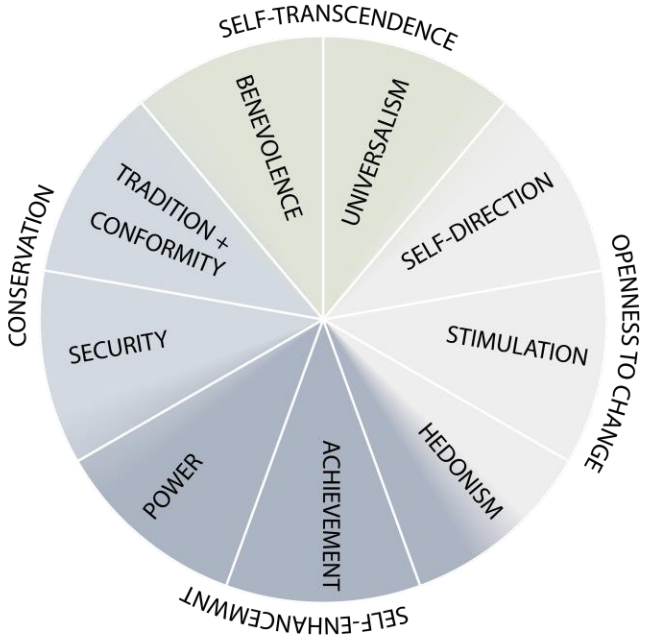


Basic Human Requirements



Basic Human Values

Schwartz Values



Moral Foundation Theory

Care/Harm

Fairness/Cheating

Loyalty/Betrayal

Authority/Subversion

Purity/Degradation

Features of Values

- Values refer to goals;
- Value beliefs are linked to affect;
- Value are standards of criteria;
- Values are ordered by importance;
- Value priorities guide actions;
- **Values transcend contexts.**

Context Dependency

Hypothesis:

Values transcend contexts.

Let's investigate that in practice!

02

Classify

Can we learn the value rhetoric behind a piece of text? How does the learned knowledge transfer across different contexts?

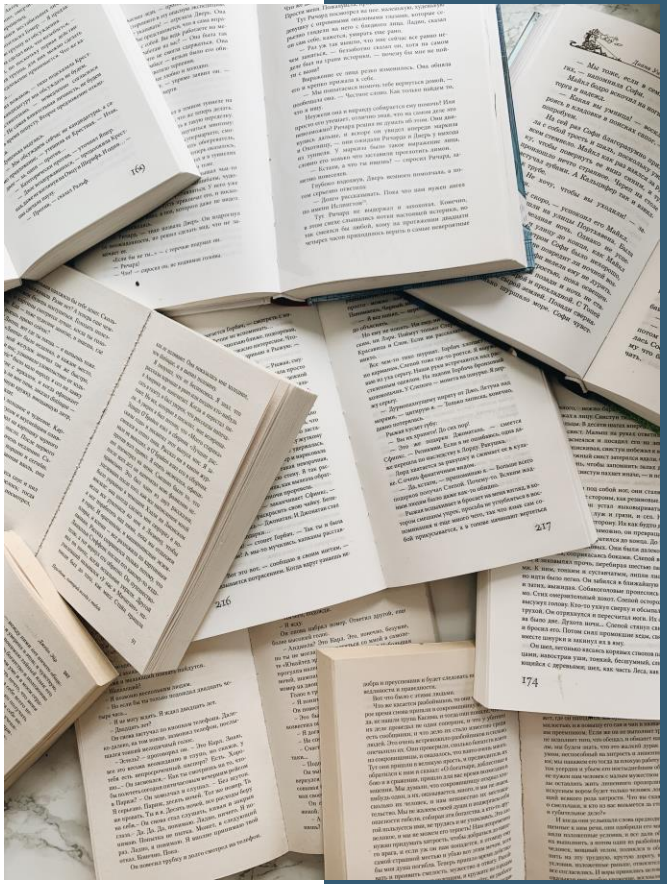


Photo by nadi borodina on Unsplash

Values in Natural Language

Value surveys are expensive and difficult to answer.

Estimating values from Natural Language allows:

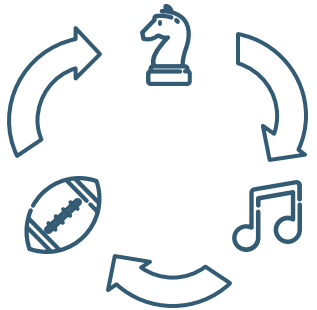
- → Humans to express values naturally;
- → Agents to have meaningful conversations with us.

Cross-Domain Classification

As we aim to classify values in different contexts, we investigate how knowledge is transferred across domains.

We inspect:

- → Generalizability
- → Transferability



E. Liscio, Alin E. Dondera, Andrei Geadau, Catholijn M. Jonker, Pradeep K. Murukannaiah. "Cross-Domain Classification of Moral Values". In *NAACL '22*, 2727-2745.

Dataset

We use the Moral Foundation Twitter Corpus (MFTC), composed of 35k tweets divided in seven datasets, annotated with the Moral Foundation Theory (MFT) values.

(ALM)	Police lives matter, all lives matter, peace and love people	→	Care
(BLM)	Which oppression is worse, sexism or racism?	→	Harm, Cheating
(Baltimore)	Baltimore Police will deliver an update on the #FreddieGray investigation. Listen live on WBAL	→	Nonmoral

Hoover, Joe, et al. "Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment." *Social Psychological and Personality Science* 11.8 (2020): 1057-1071.

MFTC Datasets

All Lives Matter

Baltimore Protests

Black Lives Matter

Hate Speech

2016 US Elections

MeToo Movement

Hurricane Sandy

MFT Values

Care/Harm

Fairness/Cheating

Loyalty/Betrayal

Authority/Subversion

Purity/Degradation

Generalizability

How well does a classifier perform on a novel domain?

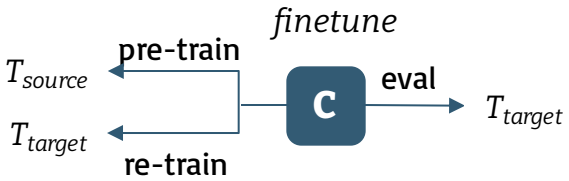
- → NLP models can decently generalize to novel domains;
- → Performances degrade on unbalanced target datasets.

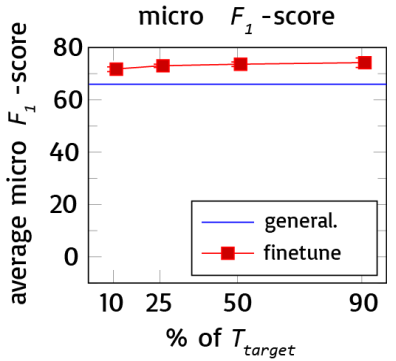
Finetuning

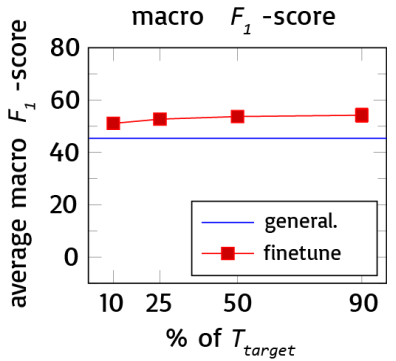
Does training on the target domain (finetune) help?

Finetuning leads to:

- → Better performance overall;
- → Better performances even with a small amount of data.







03

Explain

How does value rhetoric change across contexts?

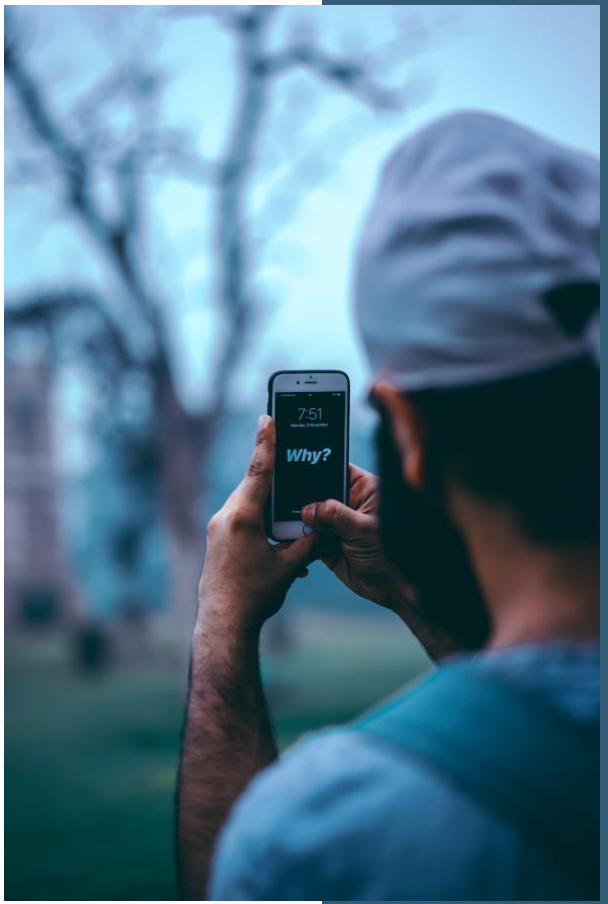


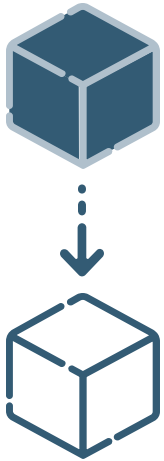
Photo by Dewang Gupta on Unsplash

Explainability

Explainable Artificial Intelligence (XAI) is aimed at providing explanations for decisions made by AI systems.

A *local* explanation provides justification for the system's prediction on a specific input.

A *global* explanation provides justification on the system's general predictive process.



Danilevsky, Marina, et al. "A Survey of the State of Explainable AI for Natural Language Processing." *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing.* 2020.

Value Rhetoric Explainability

Inspect whether the learnt value rhetoric is context specific in order to:

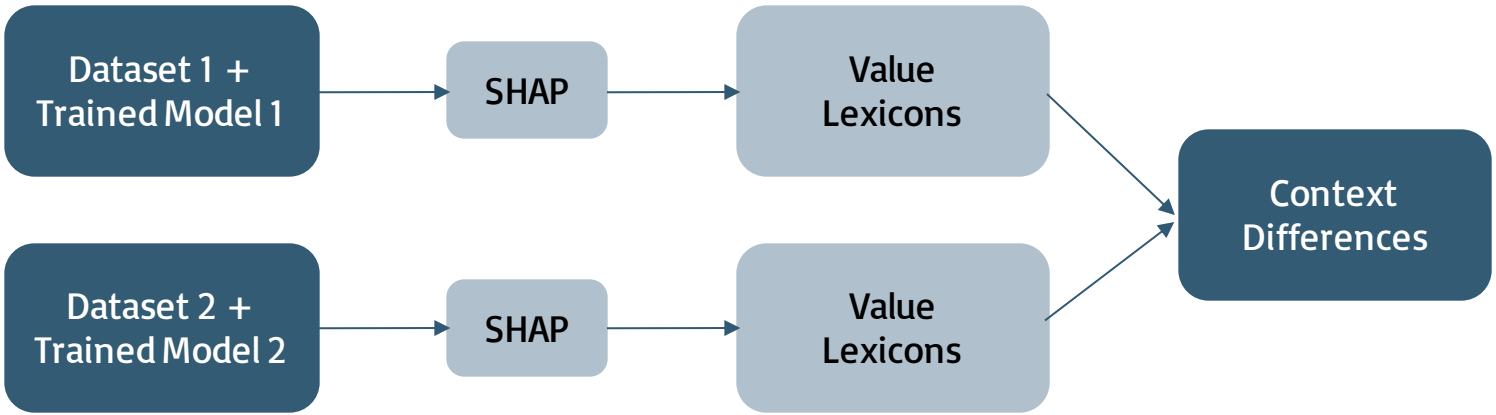
- (Dis)prove context specificity of values;
- Provide insight to social scientists and policy makers in moral reasoning across contexts.

Dataset - MFTC

MFTC Datasets
All Lives Matter
Baltimore Protests
Black Lives Matter
Hate Speech
2016 US Elections
MeToo Movement
Hurricane Sandy

MFT Values
Care/Harm
Fairness/Cheating
Loyalty/Betrayal
Authority/Subversion
Purity/Degradation

Value Rhetoric Comparison



Value Rhetoric Similarities

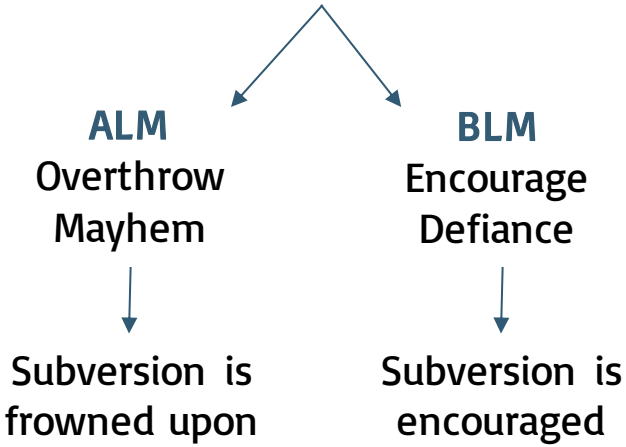
ALM and **BLM** generally have similar value rhetoric:

Fairness
Equality
Justice

Cheating
Fraud
Corruption

Value Rhetoric Differences

ALM and **BLM** generally have similar value rhetoric, but they differ for the value of *subversion*



Context Dependency

Hypothesis:

~~Values transcend contexts.~~*

Value expressions are context dependent!

04

Identify

Which values are relevant to a context?
How are they characterized in the context?

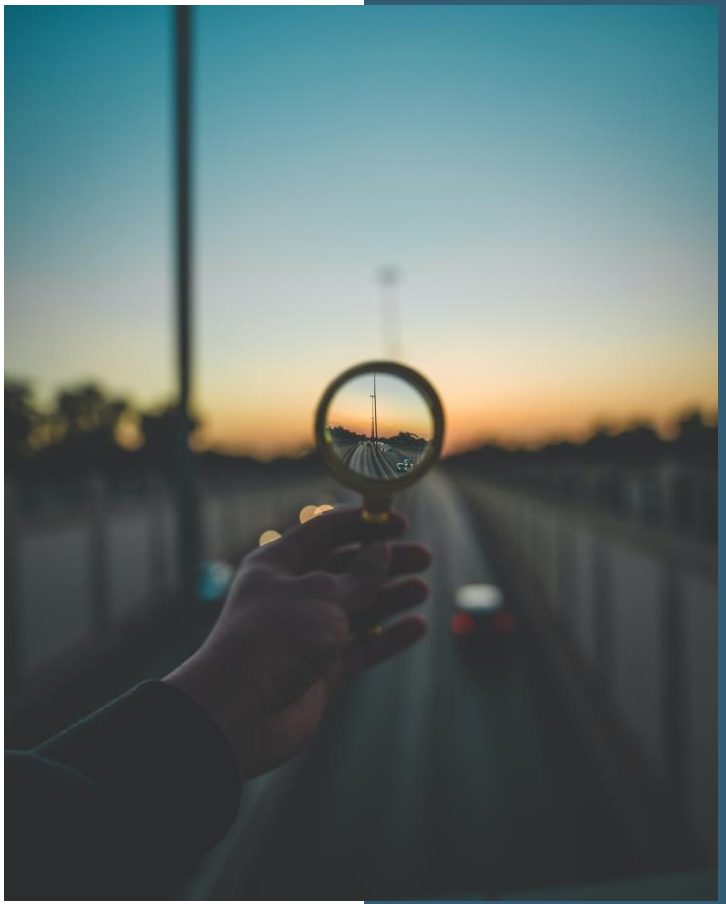


Photo by Yosef Futsum on Unsplash

Basic Human Values

- General and abstract
- Applicable across contexts
- Suitable for societal questions

Context-Specific Values

- Applicable to a context
- Defined within a context
- Suitable for concrete usage

Context-Specific Values

Context-specific values are applicable and defined within a context and are essential for concrete applications.

For example, think of the differences in these contexts:

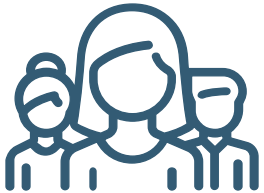
COVID-19

Green energy

Axies methodology

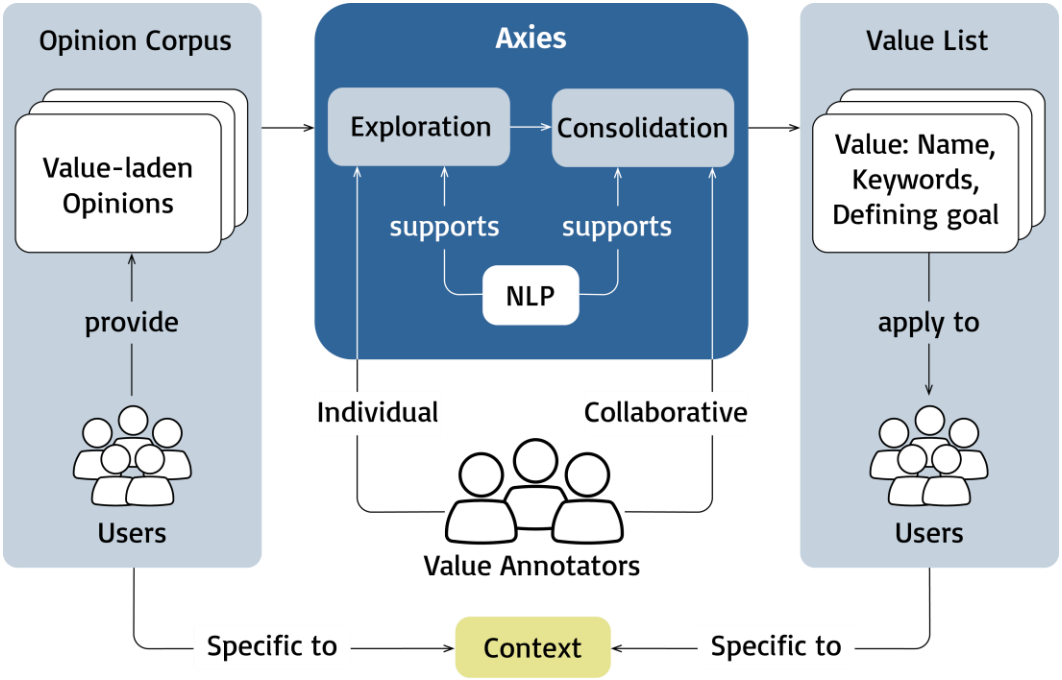
Axies is a hybrid (human+AI) methodology for identifying context-specific values, with the support of NLP techniques.

Axies simplifies and distributes the value identification process.



E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah. "What values should an agent align with?". In: *JAAMAS*, 36, 23, 2022.

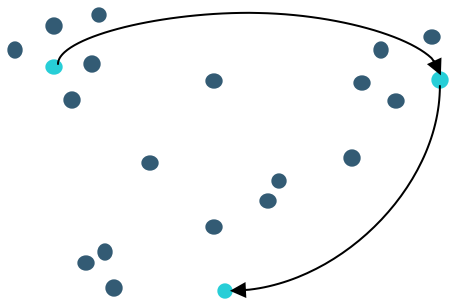
Axies methodology



Axies - Exploration

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

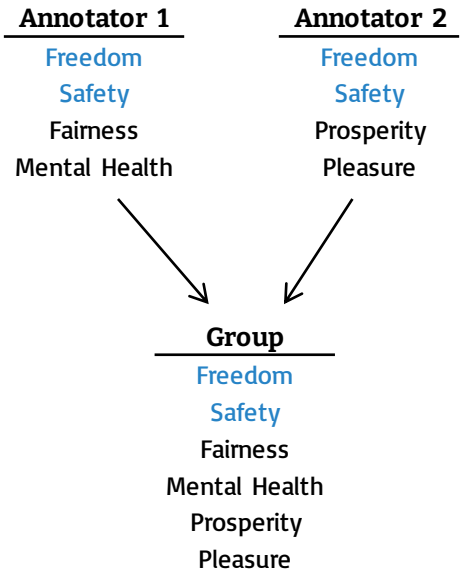
The next opinion to be analysed is the most different from the already analysed opinions.



Axies - Consolidation

The annotators in a group collaborate to merge their individual value lists.

Axies guides the annotators through the process via NLP moderation.



Evaluation

We perform Axies on two survey datasets:

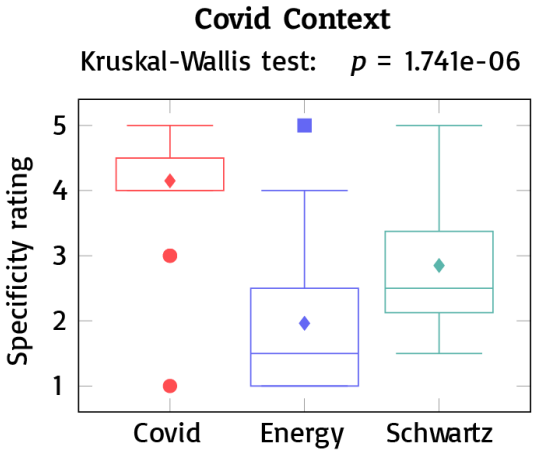
- COVID-19 (60,000 answers)
- Green Energy Transition (3,000 answers)

We ask ourselves:

- Does Axies yield context-specific values?
- What are the differences between Axies and basic values?

Results - Specificity

Axies yields values that are more context-specific than basic (Schwartz) values.

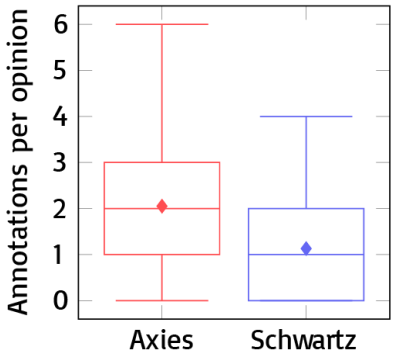


Results - Application

Laypeople annotate Axies values more often and with higher agreement.
This shows the suitability of context-specific values for practical applications.

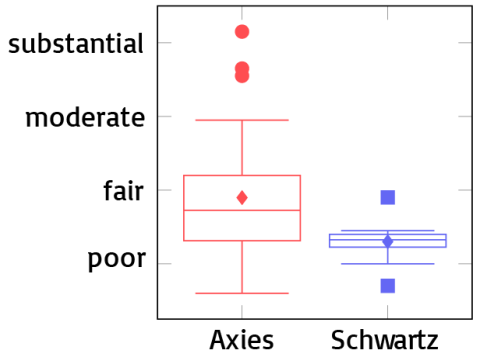
Covid Context

Wilcoxon's ranksum test: p = 2.384e-10
Cliff's delta: 0.43 (Medium)



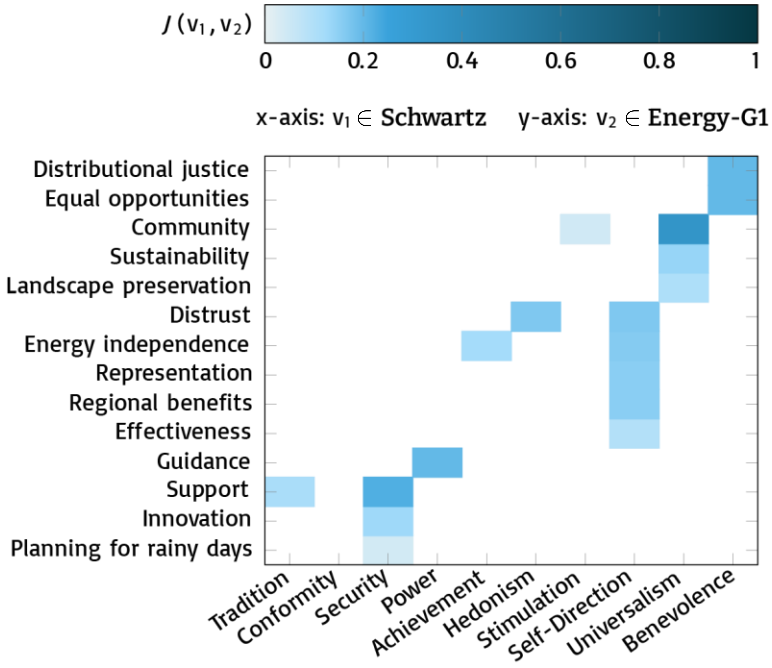
Covid Context

Welch's t-test: p = 0.02
Cliff's delta: 0.43 (Medium)



Results - Relationship

Only a few Schwartz values have a clear correspondence to Axies values, and values with a clear correspondence are often related to multiple Axies values.



Conclusions and Future Work

- Value expressions in language are context dependent;
- We propose Axies, a method for identifying context-specific values.

- Use Axies to find context-specific values, then annotate data with them;
- Address data scarcity and subjectivity in annotation collection.

Thanks!

Do you have any questions?

e.liscio@tudelft.nl
enricoliscio.github.io

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon and infographics & images by Freepik

Thanks to our collaborators!

Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, Alin E. Dondera, Andrei Geadau, Oscar Araque, Lorenzo Gatti, Kyriaki Kalimeri, Ionut Constantinescu

Presented Publications:

E. Liscio et al., **Axies: Identifying and Evaluating Context-Specific Values.** In *AAMAS '21*.

E. Liscio et al., **What values should an agent align with?** In: *JAAMAS* (2022).

E. Liscio et al., **Cross-Domain Classification of Moral Values.** In *NAACL '22*.