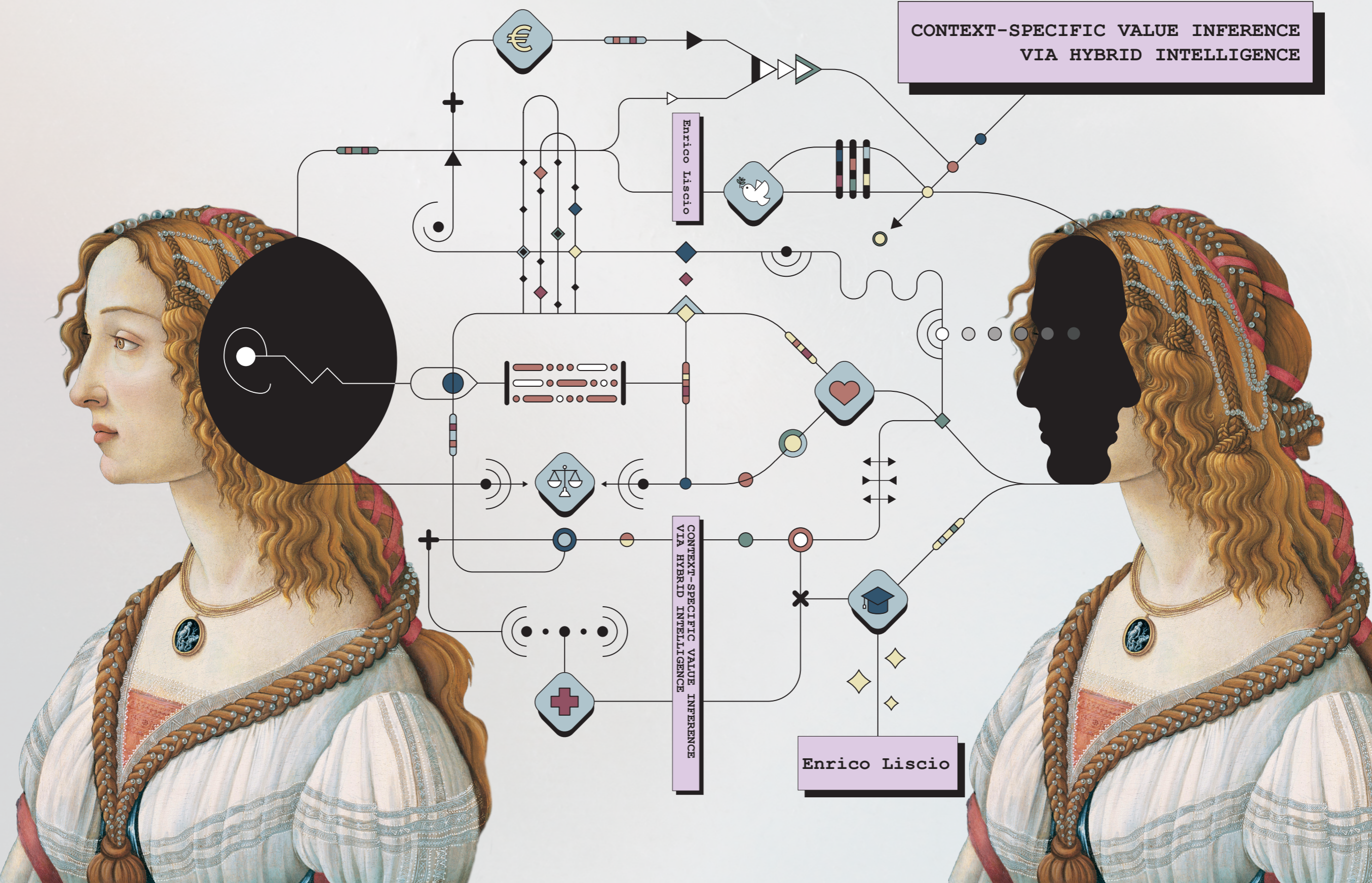


CONTEXT-SPECIFIC VALUE INFERENCE
VIA HYBRID INTELLIGENCE



Enrico Liscio

CONTEXT-SPECIFIC VALUE INFERENCE
VIA HYBRID INTELLIGENCE

Enrico Liscio

**CONTEXT-SPECIFIC VALUE INFERENCE
VIA HYBRID INTELLIGENCE**

Enrico LISCIO

CONTEXT-SPECIFIC VALUE INFERENCE VIA HYBRID INTELLIGENCE

Dissertation

for the purpose of attaining the degree of doctor
at Delft University of Technology,
by the authority of Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Monday, 22nd April 2024, at 12:30 o'clock

by

Enrico LISCIO

Master of Science in Systems & Control,
Delft University of Technology, the Netherlands,
born in Forlì, Italy

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. C.M. Jonker	Delft University of Technology, <i>promotor</i>
Dr. P.K. Murukannaiah	Delft University of Technology, <i>copromotor</i>

Independent members:

Prof. dr. ir. I.R. van de Poel	Delft University of Technology
Prof. dr. P.T.J.M. Vossen	Vrije Universiteit Amsterdam
Prof. dr. P. Yolum Birbil	Utrecht University
Prof. dr. M.P. Singh	North Carolina State University, United States of America
Prof. dr. M.A. Neerincx	Delft University of Technology, <i>reserve member</i>

SIKS Dissertation Series No. 2024-13.

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.



Keywords: Values, Natural Language Processing, Morality, Ethics, Explainable AI, Active Learning, Hybrid Intelligence

Printed by: Proefschriftspecialist | <https://www.proefschriftspecialist.nl/>

Cover by: Kamila Waszkowiak

Style: TU Delft House Style, with modifications by Moritz Beller
<https://github.com/Inventitech/phd-thesis-template>

Copyright © 2024 by Enrico Liscio

ISBN 978-94-6384-559-5

An electronic version of this dissertation is available at <http://repository.tudelft.nl/>.

CONTENTS

Summary	ix
Samenvatting	xi
1 Introduction	1
1.1 A Novel Value Inference Approach.	3
1.1.1 Behavioral Data	4
1.1.2 Value Identification	4
1.1.3 Value Classification	5
1.1.4 Value Preferences Estimation.	6
1.2 Hybrid Value Inference.	7
1.3 A Cross-Cutting Contribution	9
2 Background and Related Works	11
2.1 Human Values	12
2.2 Values in AI	14
2.2.1 Values in Natural Language Processing.	14
2.2.2 Values in Engineering Agents and Multiagent Systems	15
2.3 Other Relevant AI Concepts	16
2.4 Datasets	17
2.4.1 Participatory Value Evaluation (PVE).	17
2.4.2 Moral Foundation Twitter Corpus (MFTC)	18
I Value Identification	21
3 Identifying and Evaluating Context-Specific Values	23
3.1 Introduction	24
3.2 Axes Methodology	26
3.2.1 Value Exploration	28
3.2.2 Value Consolidation	30
3.3 Experiments	31
3.3.1 Experiment 1: Value Lists	32
3.3.2 Experiment 2: Context-Specificity	33
3.3.3 Experiment 3: Comprehensibility, Consistency, Relationship, and Application.	33
3.3.4 Statistical Analyses.	35

3.4	Results and Discussion	36
3.4.1	Value Lists	36
3.4.2	Context-Specificity	36
3.4.3	Comprehensibility	38
3.4.4	Consistency	40
3.4.5	Relationship	41
3.4.6	Application	44
3.4.7	Threats to Validity	45
3.5	Conclusions and Future Directions	46
II	Value Classification	49
4	Cross-Domain Classification of Moral Values	51
4.1	Introduction	52
4.2	Experimental Setup	53
4.2.1	Cross-Domain Evaluation	54
4.2.2	Comparisons	54
4.2.3	Metrics	55
4.3	Results and Discussion	55
4.3.1	General Trends	55
4.3.2	Generalizability	56
4.3.3	Transferability	57
4.3.4	Catastrophic Forgetting	58
4.3.5	Misclassification Errors	58
4.3.6	Annotators Agreement	59
4.4	Conclusions and Directions	61
5	An Explainable Method for Cross-Domain Comparison of Moral Values	63
5.1	Introduction	64
5.2	The Tomea Method	65
5.2.1	Moral and Domain Lexicons	66
5.2.2	Lexicon Generation	66
5.2.3	Lexicon Comparison	66
5.3	Experiment Design	67
5.3.1	Model Training	67
5.3.2	Pairwise Comparisons	67
5.3.3	Evaluation	68
5.4	Results and Discussion	69
5.4.1	Cross-Domain Comparisons	69
5.4.2	Crowd Evaluation	70
5.4.3	Out-of-Domain Performance	71
5.4.4	Qualitative Analysis	72
5.5	Conclusions and Directions	73

III	Value Estimation	75
6	Value Preferences Estimation in Hybrid Participatory Systems	77
6.1	Introduction	78
6.2	Background	80
6.2.1	Data	80
6.2.2	Formalization	81
6.3	Method.	83
6.4	Experimental Setting.	85
6.4.1	Evaluation procedure.	86
6.5	Results and Discussion.	86
6.6	Conclusion and Future Directions	88
IV	Conclusions	91
7	Closing the Loop with a Hybrid Intelligence Approach	93
7.1	Introduction	94
7.2	Method.	96
7.3	Experimental Setting.	97
7.3.1	Evaluation Procedure	98
7.4	Results and Discussion.	98
7.5	Conclusion and Future Directions	100
8	Contributions and Future Work	103
8.1	The Value Inference Framework	104
8.1.1	Value Identification	104
8.1.2	Value Classification	105
8.1.3	Value Estimation.	106
8.2	Hybrid Value Inference.	106
8.3	Challenges and Opportunities	107
8.4	Limitations.	109
8.5	Societal Implications	110
9	Beyond Value Inference	113
9.1	Aggregating Value Systems for Decision Support.	114
9.1.1	Method.	114
9.1.2	Results	116
9.1.3	Takeaways.	117
9.2	Real-World Applications (being) Developed	118
V	Appendices	119
A	Identifying and Evaluating Context-Specific Values	121
A.1	Experiments Protocol	121
A.1.1	Experiment 1: Value Lists	121
A.1.2	Experiment 2: Specificity.	121
A.1.3	Experiment 3: Comprehension and Consistency	122

A.2	Web Platform	123
A.3	Extended results	125
A.3.1	Value lists	125
A.3.2	Comprehensibility	125
B	Cross-Domain Classification of Moral Values	137
B.1	Experimental Details	137
B.1.1	Data Preprocessing.	137
B.1.2	Hyperparameters.	137
B.1.3	Computing Infrastructure	137
B.1.4	Random Seeds	138
B.1.5	Artifacts Usage.	138
B.2	Extended Results.	139
B.2.1	Model Comparison.	139
B.2.2	Composition of the Source Dataset.	140
C	An Explainable Method for Cross-Domain Comparison of Moral Values	145
C.1	Experimental Details	145
C.1.1	Data Preprocessing.	145
C.1.2	Hyperparameters.	145
C.1.3	Model Training.	145
C.1.4	Computing Infrastructure	146
C.1.5	Random Seeds	146
C.1.6	Artifacts Usage.	146
C.2	Crowd Evaluation	147
C.2.1	Annotation Job Layout	147
C.2.2	Quality Control	147
C.2.3	User demographics.	148
C.3	Extended Results.	149
C.3.1	m -distances	149
C.3.2	Correlation by Domain and Element	149
C.3.3	Qualitative Analysis	151
D	Closing the Loop with a Hybrid Intelligence Approach	153
D.1	NLP Model Training	153
	Bibliography	155
	SIKS Dissertations	181
	Acknowledgments	191
	Curriculum Vitæ	193
	List of Publications	195

SUMMARY

Human values are the abstract motivations that drive our opinions and actions. AI agents ought to align their behavior with our value preferences (the relative importance we ascribe to different values) to co-exist with us in our society. However, value preferences differ across individuals and are dependent on *context*. To reflect diversity in society and to align with contextual value preferences, AI agents must be able to discern the value preferences of the relevant individuals by interacting with them. We refer to this as the **value inference** challenge, which is the focus of this thesis. Value inference entails several challenges and the related work on value inference is scattered across different AI subfields. We present a comprehensive overview of the value inference challenge by breaking it down into three distinct steps and showing the interconnections among these steps.

We start by addressing **value identification**, the challenge of identifying the set of values relevant to a decision-making process. We recognize that the set of relevant values is dependent on the decision-making context, and propose a method that combines human and artificial intelligence to identify context-specific values. Our method employs Natural Language Processing techniques to assist human annotators in systematically identifying context-specific values in a corpus composed of value-laden opinions.

Next, we tackle **value classification**, the challenge of detecting value-laden content in natural language. We evaluate how language models can classify values in the text, and investigate how the dependency on context impacts the classification performance. First, we perform a cross-context evaluation of the performance of value classifiers. Then, we propose an Explainable AI method for investigating the extent to which language models learn context-specific expressions of values.

Third, we focus on **value preferences estimation**, the challenge of estimating how humans prioritize the values that are relevant to the decision-making context. We propose and compare methods to estimate value preferences based on an individual's choices and the justifications they provide for their choices. We follow the rationale that, when conflicts are detected between the values that support one's choices and one's justifications, the values that support the justifications should be prioritized.

Relying solely on AI methods to infer values may not yield accurate estimates, due to the implicit nature of human value preferences. Humans must be actively engaged for a successful value inference. To this end, we propose a **Hybrid Intelligence** (HI) vision where human and artificial intelligence complement each other during the value inference process. We then introduce an HI approach that fosters self-reflection on values by connecting value classification and value preferences estimation. With our innovative perspective and experiments, we advocate for a HI approach to guide AI agents in inferring individuals' context-specific value preferences. This thesis lays a foundation for fostering harmonious co-existence between artificial and human agents in human-AI societies. Building on our work, several applications are being developed, ranging from support for deliberative policy-making at a municipal level to behavior change for diabetes patients.

SAMENVATTING

Menselijke waarden zijn de abstracte motivaties die onze meningen en handelingen sturen. AI-agenten zouden hun gedrag moeten afstemmen op onze waardevoorkeuren (het relatieve belang dat we hechten aan verschillende waarden) om samen met ons in onze samenleving te kunnen bestaan. Waardevoorkeuren verschillen echter tussen individuen en zijn afhankelijk van de *context*. Om de diversiteit in de samenleving te weerspiegelen en om aan te sluiten bij de contextuele waardevoorkeuren, moeten AI-agenten de waardevoorkeuren van de relevante individuen kunnen onderscheiden door met hen te interageren. We noemen dit de uitdaging van **waarde-inferentie**, de focus van dit proefschrift. Waarde-inferentie brengt verschillende uitdagingen met zich mee en het gerelateerde werk aan waarde-inferentie is verspreid over meerdere AI deelgebieden. We geven een uitgebreid overzicht van de waarde-inferentie-uitdaging door deze op te splitsen in drie afzonderlijke stappen en de onderlinge verbanden tussen deze stappen te laten zien.

We beginnen met het aanpakken van **waarde-identificatie**, de uitdaging van het identificeren van de set van waarden die relevant zijn voor een besluitvormingsproces. We erkennen dat de verzameling relevante waarden afhankelijk is van de context van de besluitvorming en stellen een methode voor die menselijke en kunstmatige intelligentie combineert om contextspecifieke waarden te identificeren. Onze methode maakt gebruik van Natural Language Processing-technieken om menselijke annoteerders te helpen bij het systematisch identificeren van contextspecifieke waarden in een corpus dat bestaat uit meningen die waarden bevatten.

Vervolgens pakken we **waarde-classificatie** aan, de uitdaging om inhoud met waarden te detecteren in natuurlijke taal. We evalueren hoe goed taalmodellen waarden in de tekst kunnen classificeren en onderzoeken hoe de contextafhankelijkheid de classificatieprestaties beïnvloedt. Eerst voeren we een contextoverschrijdende analyse uit van de prestaties van waardeclassificatiemodellen. Vervolgens stellen we een uitlegbare AI-methode voor om te onderzoeken in hoeverre taalmodellen de contextspecifieke uitdrukking van waarden leren.

Ten derde richten we ons op de **waardevoorkeureninschatting**, de uitdaging om in te schatten hoe mensen prioriteit geven aan de waarden die relevant zijn voor de beslissingscontext. We ontwikkelen en vergelijken methodes om waardevoorkeuren te schatten op basis van de keuzes van een individu en de rechtvaardigingen die ze daarbij geven. We volgen de redenering dat, wanneer conflicten worden gedetecteerd tussen de waarden onderliggend aan iemands keuzes en de rechtvaardigingen, de waarden die de rechtvaardigingen ondersteunen voorrang moeten krijgen.

Alleen vertrouwen op AI-methodes om waarden af te leiden levert mogelijk geen nauwkeurige schattingen op, vanwege de impliciete aard van menselijke waardevoorkeuren. Mensen moeten actief betrokken worden voor een succesvolle waarde-inferentie. Daarom stellen we een **Hybride Intelligentie** (HI) visie voor waarbij menselijke en kunstmatige intelligentie elkaar aanvullen tijdens het proces van waarde-inferentie. Vervolgens intro-

duceren we een HI-benadering die zelfreflectie over waarden bevordert door het verbinden van waardeclassificatie met het inschatten van waardevoorkeuren. Met ons innovatieve perspectief en onze experimenten pleiten we voor een HI-benadering om AI-agenten te begeleiden bij het afleiden van de contextspecifieke waardevoorkeuren van individuen. Dit proefschrift legt een basis voor het bevorderen van een harmonieuze co-existentie tussen kunstmatige en menselijke agenten in mens-AI samenlevingen. Voortbouwend op ons werk worden er verschillende toepassingen ontwikkeld, variërend van ondersteuning voor het maken van beleid op gemeentelijk niveau tot gedragsverandering voor diabetespatiënten.

1

INTRODUCTION

Human mythology teems with tragic stories of eager characters being granted wishes that lead to disaster. Examples range from Draupadi, the Pandavas' wife, in the Hindu Mahabharata [24] to the tragic story of King Midas [226]. These tales warn us to be careful what we wish for, as *unintended consequences* can hide behind every desire. A similar lesson applies to the engineering of intelligent machines. Ranging from Erehwon, an 1872 dystopic novel that discusses the ban of intelligent machines [45], to a modern classic such as Terminator [48], storytellers warn us that superintelligent machines may purposely seek to end humanity. However, philosophers and experts in Artificial Intelligence (AI) instead warn us that the greatest threat is posed by a superintelligent AI agent that could *unintentionally* wreak havoc in our society. Bostrom [40] famously describes the paperclip apocalypse scenario, centered around a superintelligent AI agent tasked with the goal of producing paperclips. A determined superintelligent agent may end up turning everything it can into paperclips and deplete the planet's resources, even to achieve such an apparently mundane goal. How do we prevent this doom scenario from happening?

Researchers have proposed to address this challenge by engineering AI agents that do not just align with our instructions but with our deepest interests [40, 260]. For example, AI agents ought to understand that our desire for paperclips is subject to the condition that it is achieved without harming other humans. This challenge is referred to as *value alignment* [100, 261, 283], which postulates that we should engineer AI agents that align with our human *values*. Human values are the abstract motivations that drive our opinions and actions [268], spanning concepts such as fairness and self-determination. It is our *value system* (which can be described as our *value preferences* over different relevant values [270, 276, 331]) that guides our actions. For instance, both the values of freedom and safety are relevant to the discussion of mouth mask mandates to limit the spread of a pandemic—some of us prioritize freedom over safety and some others safety over freedom, leading to different stances. That is, different individuals, influenced by their socio-cultural environment [78], may hold different value systems. AI agents cannot be simply pre-loaded with a fixed value system to which their behavior should align, as this system would likely reflect only a small subset of (powerful) individuals. Further, how each of us prioritizes values is dependent on *context*—that is, our value preferences can change based on e.g., situation, location, and interlocutors [126]. For example, one might generally prioritize freedom over safety, but change their preferences in the context of a pandemic. To reflect the diversity of the stakeholders in society and to align with contextual value systems, AI agents will need to be equipped with the ability to discern the value systems of the relevant stakeholders by interacting with them. We refer to this as the *value inference* challenge, which constitutes the core of this thesis:

How can an AI agent infer an individual's value system in a decision-making context?

This introductory chapter is structured as follows. Section 1.1 introduces the value inference challenge and our novel approach for performing value inference in a human-AI society. Section 1.2 extends our approach by proposing a Hybrid Intelligence vision. Throughout these two sections, we describe how each chapter of this thesis contributes toward the challenges outlined in our approach. Finally, Section 1.3 concludes the chapter by motivating how our novel approach helps in bridging different AI communities.

1.1 A NOVEL VALUE INFERENCE APPROACH

In a sociotechnical system (STS), i.e., a society where humans and AI agents co-exist, AI agents ought to align their behavior with the value systems of the relevant human stakeholders. Value inference refers to the challenge of inferring the value systems of the relevant stakeholders in a decision-making context.

Value inference is a complex task that is composed of multiple elements. There is an increasing body of AI literature that touches upon different aspects of value inference, ranging from the semi-automatic identification of the values that are relevant to a decision-making context [42, 318] to the detection of value-laden natural language [15, 150]. However, real-world applications often require a combination of these functionalities. In this section, we introduce a holistic view of how the pieces of value inference fit together.

We propose a value inference approach that starts with the observation of stakeholders' *behavior*, i.e., the choices they make and the natural language justifications they provide for their choices. Then, we identify three fundamental steps of value inference as (1) *identification* (which values are relevant to a decision-making context?), (2) *classification* (what are the values underlying a natural language justification?), and (3) *estimation* (what are an individual's value preferences?). The output of value inference is a stakeholder's *value system*, i.e., their preferences over the set of values that are relevant to the decision-making context. Figure 1.1 outlines value inference as a modular framework consisting of the fundamental steps we identified to go from the behavioral data to the value system. The dark blocks in Figure 1.1 represent the *processes* and the light blocks represent the *information* the processes consume or produce. Our framework's modularization has two advantages. First, the separation of concerns into processes delineates research challenges. Second, the interdependencies between processes expose research challenges that can otherwise fall through the gaps. For example, although value classification influences value preferences estimation, these connections are largely unexplored.

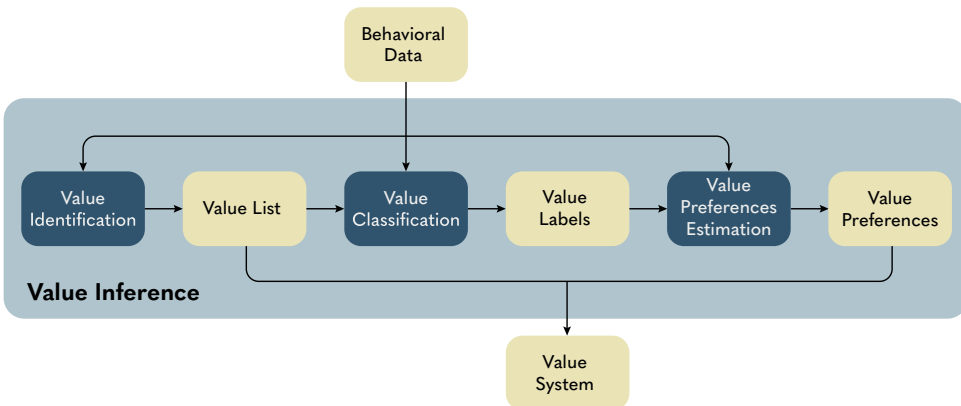


Figure 1.1: Value inference processes (dark-colored blocks) and information (light-colored blocks). Value inference starts by observing behavioral data (i.e., stakeholders' choices and natural language justifications provided for their choices). Then, the values relevant to the decision-making context are identified and classified in the justifications. Finally, value preferences are estimated based on the choices and the values classified in the justifications. Value inference results in a stakeholder's value system, i.e., their value preferences over the set of relevant values.

In the rest of this section, we detail the framework introduced in Figure 1.1. In Subsection 1.1.1, we discuss what constitutes behavioral data and how it can be observed. Then, Subsections 1.1.2, 1.1.3, and 1.1.4 describe the value inference processes and introduce the current state-of-the-art (which we expand in **Chapter 2**). For each process, we outline the research gaps and indicate how this thesis addresses them.

We recognize that value inference, as a purely AI task where a sequence of computational methods is applied to behavioral data, is not likely to yield good estimates of a value system. This is because value systems are often implicit in humans [125, 172, 286] and are, thus, not easily observable in the behavioral data. Hence, we must actively engage humans for successful value inference. We address these challenges by proposing a Hybrid Intelligence (HI) [7] approach, where human and artificial intelligence augment each other. We outline our HI vision for value inference in Section 1.2.

1.1.1 BEHAVIORAL DATA

In our framework, values are inferred from behavioral data. We consider stakeholders' *actions*, e.g., how they choose over competing alternatives [35, 302, 323] or solve a problem [114, 220, 260], as behavioral data. However, value preferences are often implicit in actions, and inferring values solely based on actions is difficult. Since language is an important means of value expression, the value preferences underlying our actions can be observed in the natural language *justifications* we provide for those actions [107, 266]. Thus, we can exploit the observation of both actions and justifications as the behavioral data that is input to the value inference framework.

Observing and processing human behavior constitutes a field of study on its own, as exemplified by Raman [246]. However, this thesis does not explore this field. Instead, we employ datasets that contain value-laden choices and/or justifications of relevant stakeholders (Section 2.4). We treat this data as the input to the value inference processes and focus our contributions on the value inference challenge. Future research endeavors could expand upon our work by integrating value inference with the observation of value-laden behavior, as we further elaborate in Section 8.3.

1.1.2 VALUE IDENTIFICATION

Value identification refers to the process of identifying the set of values relevant to a decision-making context.

State-of-the-Art Lists of basic human values, applicable across cultures and contexts, have been proposed by ethicists [254, 268] and psychologists [107]. However, such lists have been shown to be too generic for practical applications [167, 179, 237] and have been identified by experts without active stakeholder participation. Value Sensitive Design (VSD) [96] proposes participatory methods for identifying stakeholders' values, e.g., Tuomela et al. [299] employ sensory ethnography to identify the values of users of a smart home energy management system. Data-driven methods for identifying values have also been proposed. Boyd et al. [42] demonstrate that values identified from free-response language (e.g., Facebook status messages) yield better predictive coverage of real-world behavior than values extracted from self-report questionnaires such as the Schwartz Value Survey.

Building on [42], Wilson et al. [318] describe a crowd-powered algorithm to generate a hierarchy of general values.

Research Gap Research suggests that not all basic values are relevant to all contexts [167, 179, 237, 268]. Further, an individual’s value system may not be consistent across contexts [69, 312]. That is, how an individual interprets and prioritizes values depends on *context*. For instance, one might generally value freedom over safety but prioritize safety over freedom during a global pandemic. Thus, we advocate for the identification of context-specific values, i.e., values applicable and defined within a context. Context-specific values are deemed essential for the concrete use and analysis of values (e.g., designing policies) as argued by an increasing body of literature [6, 215, 303]. A data-driven approach to the identification of context-specific values would allow AI agents to dynamically identify the relevant values across different contexts.

Our Contribution In **Part I**, we tackle the challenge of identifying the values that are relevant to a decision-making context. We recognize that value identification necessitates human oversight, as the judgment of what is relevant ought to lie in humans. Further, it is paramount that such a decision is taken with the involvement of a large group of stakeholders, so as to identify a representative set of values. With this in mind, we envision value inference to be performed by humans with the support of AI systems, so that AI systems can simplify the process and let humans make only a few high-level decisions. To this end, in **Chapter 3**, we propose *Axies*, a hybrid method for identifying context-specific values. *Axies* employs NLP techniques to guide humans through a crowdsourced dataset of value-laden natural language justification with the goal of identifying the values that are relevant to the stakeholders who provided the justifications and the context under examination. We then compare the resulting context-specific values to a set of general values (Schwartz’s basic values [268]) and show that context-specific values are more suitable for concrete applications.

1.1.3 VALUE CLASSIFICATION

Value classification refers to the process of detecting value-laden content in natural language.

State-of-the-Art Value classification has been addressed from both lexicon-based and supervised approaches. Lexicon-based methods exploit value lexicons to detect values in natural language. Value lexicons are generated manually [105], via semi-automated methods [18, 131, 251, 318], or expanded from an initial seed via NLP techniques [19, 239]. Value lexicons are used to detect values in natural language through word count software [229] or similarity in embedding space [26, 101, 273]. However, adapting a lexicon to a novel domain is a significant additional effort as it requires identifying words that are relevant and removing words that are not relevant in the novel domain. Other methods have approached value classification by employing the supervised classification paradigm [15, 130, 150, 173, 206]. A textual dataset is annotated with values belonging to a value taxonomy, and the labels are used to train a supervised model. This approach leverages the abilities of large language models and shows better performance and better generalization

to novel domains than lexicon-based approaches [133], but may necessitate large amounts of contextual training data.

Research Gap NLP methods have been shown to be capable of recognizing value-laden content in natural language. However, the way in which we express value is context-dependent. Additional research is required to investigate the extent to which language models can transfer across contexts—that is, evaluate the performance of a language model trained in a context and tested in a different context. This research is necessary to unveil whether all-purpose language models can be used across contexts, or whether context-specific models are needed to grasp the different value representations across contexts.

Our Contribution In **Part II**, we examine the context sensitivity of value classifiers. We start by evaluating the extent to which a language model can perform out-of-context value classification. In **Chapter 4**, we perform a cross-context evaluation across seven contexts—that is, we train a state-of-the-art language model with data collected in context A and evaluate it on data collected in context B, with all possible combinations of the seven contexts. We compare four training settings (e.g., by training it only with data collected in context A, or by training it with data collected in context A and further training it with a small portion of data collected in context B). Our results show that language models can generalize to novel contexts, however introducing some classification errors. In **Chapter 5**, we investigate the errors that are introduced in cross-context classification. To this end, we propose Tomea, a method for comparing how language models represent value concepts across contexts. Tomea provides a comparison of value representations across contexts that is based on an explainable AI method. In this way, Tomea returns both a qualitative and quantitative comparison of value rhetoric, allowing for a deeper investigation of differences across contexts. Our experiments show that language models commit infrequent mistakes—however, when inspected qualitatively, the mistakes can be critical, possibly hindering the usage of a language model in a novel context. These results stress the importance of finetuning language models with contextual data.

1.1.4 VALUE PREFERENCES ESTIMATION

Value estimation is the process of determining a stakeholder's preferences over a set of values based on their observed behavior.

State-of-the-Art Value preferences are typically represented as rankings over a fixed set of values [268, 270, 331]. Existing approaches estimate individuals' value preferences with survey instruments such as the Portrait Value Questionnaire [268], Schwartz Value Survey [268], Value Living Questionnaire [317], and Moral Foundations Questionnaire [106]. However, surveys are criticized for not grounding value preferences to a context [167, 237]. Directly asking humans about their value preferences through questionnaires often leads to incomplete and hypothetical answers that do not reflect real-life behavior [41]. A complicating factor is that value preferences have been recognized to be context-specific [44, 126, 159], where context refers to factors such as actors, actions, and judges [267]. Other approaches follow the principles of VSD by combining self-reported surveys with participatory design [171, 237]. VSD methods situate value estimation in a design

context by, e.g., showing relevant photos [167, 237] or videos [299]. Yet, the need for human moderation limits the scale in which VSD methods can be applied. In contrast, Inverse Reinforcement Learning (IRL) [220] learns humans' reward functions based on the observed actions, and Cooperative IRL (CIRL) [114] augments IRL with human feedback. However, IRL assumes that humans are aware of their reward functions and is criticized for the infeasibility of estimating an individual's rationality and value preferences simultaneously [201].

Research Gap Existing methods estimate value preferences based on stakeholders' choices (e.g., answers to a questionnaire). However, as language is our preferred way to express values [107, 266], we envision value preferences estimation to be based on both the choices and the natural language justifications that individuals provide for those choices. However, value preferences revealed from one's choices and one's justifications may be inconsistent [41, 222]. Thus, the value preferences estimation method would need to include methods to address these (potential) inconsistencies.

Our Contribution **Part III** focuses on the estimation of individual's value preferences. In **Chapter 6**, we propose and compare methods to estimate value preferences from the choices and the textual justifications provided in a survey. The compared methods estimate a survey's participant value preferences based on (1) their choices alone, (2) their textual justifications alone, or (3) a combination of their choices and textual justifications. We operationalize the philosophical stance that "valuing is deliberatively consequential" [266]. That is, if a participant's choice is based on a deliberation of value preferences, the value preferences can be observed in the justification the participant provides for the choice. Thus, in case of conflicts between choices and justifications, our methods prioritize the preferences estimated from justifications over those estimated from choices. We show that this approach produces results that are more similar to the value preferences that humans estimate when compared to the results obtained by employing choices or justifications alone.

1.2 HYBRID VALUE INFERENCE

Value inference cannot be performed solely via computational methods (e.g., machine learning from human behavioral data). Since value reasoning is cognitively challenging [167, 238] and implicit in human thinking [125, 172, 286], values may not be explicitly evident in behavioral data. Further, often humans can express their values only in concrete situations, and values can be emergent [138]. Thus, humans should be systematically guided through the processes of *self-reflection* [172, 237] and *deliberation* [76, 115] to become aware of their value systems and how they change based on context. This makes value inference a Hybrid Intelligence (HI) endeavor [7], requiring human and artificial intelligence to augment each other. Figure 1.2 shows an overview of the HI framework we envision.

Self-Reflection Humans must be made aware of values and guided through value reasoning via a process of self-reflection [172, 237]. Self-reflection can be achieved by creating *feedback loops* among the components in our framework. That is, based on the observed

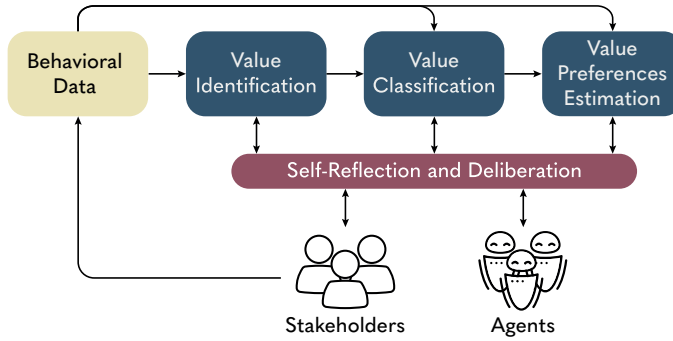


Figure 1.2: A HI value inference framework where agents guide humans in self-reflection and deliberation by situating value reasoning in the stakeholders' interpreted behavior. This interaction in turn affects stakeholders' behavior, creating a feedback loop in the value inference framework.

behavior and the inferred values, AI agents can interact with humans and help them reflect on their value systems. Agents can facilitate self-reflection by *situating* value reasoning in specific contexts and behaviors, e.g., by asking concrete questions such as what motivated a human to choose a specific action in a context, as opposed to asking generic and hypothetical questions over value preferences.

Deliberation In addition to self-reflection, deliberating with others [76, 115] and confronting individuals with different value systems [265] help us discover our own value systems. To this end, an increasing number of digital deliberation platforms have been proposed [156, 275]. However, the deliberation quality in unmoderated platforms is often poor, due to polarization and lack of inclusivity [43, 152]. AI-supported human moderation improves deliberation quality [155] but requires large numbers of human moderators. Recently, artificial moderating agents [112, 113] have been proposed to facilitate large-scale deliberation, e.g., a moderating agent can automatically add targeted comments to foster back-and-forth discussions and increase the depth of deliberation.

A Motivating Example We introduce a hypothetical example to demonstrate how self-reflection and deliberation could be fostered in a hybrid value inference framework. Consider a participatory decision-making situation in which policy makers consult the relevant stakeholders to create COVID-19 regulations. In this case, there is a large variety of stakeholders, including ordinary citizens, healthcare providers, transit authorities, small businesses, and so on. The policy makers seek regulations that satisfy technical constraints (e.g., beds available in the intensive care units) but also align with the stakeholders' value systems. To infer the stakeholders' values about potential COVID-19 regulations, policy makers set up a digital deliberation on the issue [117], where participants discuss the impact of proposed regulations on the healthcare system and the society, and they may vote on different proposals. Artificial agents moderate the discussion by fostering idea-sharing and confrontation to increase the deliberation quality.

Value inference can be initially performed based on the participants' behavior on the platform, and subsequently refined through self-reflection and deliberation. For each

stakeholder (Amber), Amber's agent investigates whether the inferred value system is correct. The system can be incorrectly inferred because (1) the set of identified values does not fully represent Amber's value sentiment (which requires revisiting value identification), (2) Amber's justifications have been misinterpreted (which requires revisiting value classification), or (3) Amber disagrees with some parts of the estimated value preferences (which requires revisiting value preferences estimation). Next, Amber's agent can guide her in reflecting on the inferred value system. For example, if the inferred value system is inaccurate because not enough input has been provided in the deliberation, the agent may ask Amber for additional value-laden input through targeted questions (e.g., asking a justification for a specific upvote). The agent can additionally provide explanations about the value inference processes or show the values that were classified from the arguments proposed by Amber. Through this systematically guided reflection, Amber becomes aware of her value system and the agent has obtained a validated preference profile of Amber. Finally, Amber and her agent may initiate discussions with other stakeholders and their agents to adjust the value inference processes. For instance, the value list may have to be updated, the language model for value classification may need to account for a minority language, or the relative importance that the AI agent attributes to actions and justifications needs to be adjusted. Importantly, the adjustment of the value inference processes should not be fully automatic. The involvement of relevant stakeholders is essential for meaningful human control [277] on the value inference framework. After the agents have validated the value systems of all participants, clustering techniques may be applied to provide an anonymized summarization of the opinions of the participants to inform the decision-making process at the political level.

Our Contribution Chapter 7 describes a HI method that fosters self-reflection in stakeholders by connecting two value inference processes, value classification and value preferences estimation. Similar to the participatory deliberation setting described above, we introduce our method in the context of a survey where participants make choices and provide natural language justifications for their choices. We employ the methods introduced in Part II to classify the values underlying the justifications, and the methods introduced in Part III to estimate participants' value preferences based on their choices and the values classified in their justifications. In Part III, we address inconsistencies between choices and justifications by prioritizing the values that support the justifications over those that support the choices. Here, instead, we intend to investigate these inconsistencies. We propose a strategy that guides the interaction between AI agents and stakeholders with the intent of disambiguating these inconsistencies, i.e., by asking the stakeholders to validate the correctness of the value labels that were detected in their justifications by the value classifier, in an active learning fashion. We compare our method to state-of-the-art active learning strategies but find no significant differences. We conclude the chapter by reflecting on the lessons learned by introducing our proposed disambiguation strategy.

1.3 A CROSS-CUTTING CONTRIBUTION

Our contribution lies at the intersection of the research areas of Autonomous Agents and MultiAgent Systems (AAMAS) and Natural Language Processing (NLP), as reflected by the

different publication venues of our works. On the one hand, we propose a vision that is centered around AI agents that co-exist with humans in a sociotechnical system, designing mechanisms that guide AI agents in inferring stakeholders' value systems by interacting with them. On the other hand, language sits at the core of our approach, as that is the preferred medium that humans use to articulate their value preferences.

We are already observing an increase in cross-pollination across these two fields. The 2023 edition of the International Conference on Autonomous Agents and Multiagent Systems (the most acclaimed conference in the AAMAS community) saw an increase in NLP applications. Yejin Choi gave a keynote talk on teaching commonsense knowledge to language models [56] and NLP appeared through works on virtual agents [325], vision-language navigation [116], language grounding [314], and explainable AI [253]. Similarly, the 2023 Annual Meeting of the Association for Computational Linguistics (the most acclaimed conference in the NLP community) has seen an increase in works centered around autonomous agents [160, 287] and personal assistants [37, 258, 274]. Furthermore, the concept of human values is gaining traction in both communities, ranging from the engineering of value-aligned normative systems [204, 270] to the generation of morally framed arguments [15]. Chapter 2 provides an in-depth overview of the presence of human values in the AAMAS and NLP communities.

Value inference and value alignment require a combination of the expertise of these two communities. Both communities will benefit from realizing that the personalized communication between humans and AI agents is key to the design of value-aligned agents and sociotechnical systems.

2

2

BACKGROUND AND RELATED WORKS

In this Chapter, we introduce the related works and background necessary to back the remainder of the Thesis. Section 2.1 details the notion of human values and introduces examples of its application outside the field of AI. Section 2.2 describes applications of the concept of human values in the field of AI. Section 2.3 introduces other concepts in the field of AI that are fundamental for the work we describe in this thesis. Finally, Section 2.4 describes the two main datasets that we use in our experiments.

2.1 HUMAN VALUES

The notion of human values has been at the center of psychological and sociological studies in the last century [13, 87, 107, 128, 254, 268]. Schwartz [268] offers a complete overview of the concept of human values. Basic human values are defined as what we consider important in life, beliefs linked inextricably to affect which refer to desirable goals that motivate action. Schwartz proposes the value taxonomy displayed in Figure 2.1, where tensions between competing values are displayed at opposite ends of the circumplex—for instance, the values of universalism and achievement are conflicting in that universalism seeks the well-being of others, whereas achievement seeks self-realization. Graham et al. [107] propose the Moral Foundation Theory (MFT), composed of a set of moral values (Table 2.1) that are intended to deconstruct our morality (i.e., the internal compass that guides us in distinguishing what is right from what is wrong). They suggest that morality is not composed of a single scale that ranges from right to wrong, but rather of five innate moral foundations, i.e., five components of morality that range from right to wrong (or, in the words of the authors, from virtue to vice). They compare this approach to how our five taste receptors (sweet, sour, salt, bitter, and umami) combine to yield the tastes we experience.

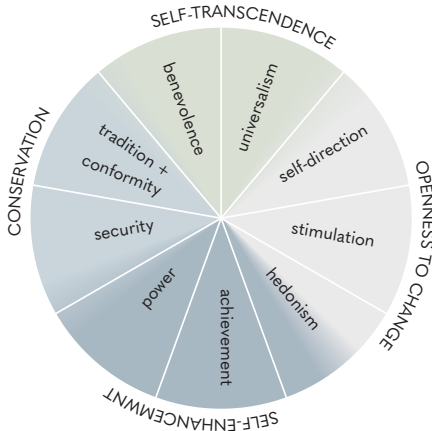


Table 2.1: The moral foundations (virtue/vice) of the Moral Foundation Theory [107].

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

Figure 2.1: The taxonomy of basic human values proposed by Schwartz [268].

Other theories of human values have been proposed over the years [13, 87, 128, 254], and they all roughly share the same basic approach. That is, in practice, multiple values are relevant to each decision-making situation and each of us assigns varying importance

to each of these values (i.e., each of us holds different *value preferences*). The combination of these two aspects determines our individual judgment of the situation. For instance, with a slight simplification, we could argue that the debate on immigration touches on the values of fairness (“Everyone should be given equal opportunities”) and in-group loyalty (“I worry about the preservation of our identity”). The way in which each of us prioritizes fairness vs. loyalty influences our judgment in this debate.

Value theories explain the differences in attitude and behavior at a personal and social level [268]. Extensive experiments have been performed to show that differences in value preferences are predictive of ideological differences [193]. For instance, Graham et al. [107] show that the different prioritizations among the MFT foundations can help explain differences between conservatives and liberals in the US political landscape. Hofstede [128] instead uses his taxonomy of values to explain differences across countries and cultures. Thanks to the powerful capability of explaining deep differences across individuals, in the last decades, values have been operationalized outside of the field of psychology.

Value Sensitive Design Values are central to Value Sensitive Design (VSD) [96], a broad set of methods to design technology that accounts for human values. VSD includes methods for identifying value sources, representing values, and resolving value tensions. The VSD framework includes a general set of values relevant to all design tasks [96]. Then, stakeholders’ value preferences are elicited through techniques such as Value Scenarios [219], Value Dams and Flows [199], and Envisioning Cards [94]. Finally, the elicited values are translated into norms and design requirements by creating a value hierarchy that concretizes and specifies the values into actionable objectives [303]. However, there is an increasing recognition that the instantiation of abstract values in specific contexts is an essential step in the effective realization of VSD [167, 238, 262]. Pommeranz et al. [238] acknowledge the need for self-reflection triggers since reflecting on values is not natural to most people. Our vision of value inference and the method we propose in Chapter 3 to identify context-specific values (Axies) fill the gaps in VSD that Pommeranz et al. [238] recognize. That is, Axies targets the identification of context-specific values by providing concrete triggers to humans (who need not be design experts) for reflecting on values.

Software Engineering Values have also been considered when engineering software [11, 89, 209]. Perera et al. [231] offer an overview of the prevalence of human values in recent Software Engineering (SE) publications. Values of stakeholders are often elicited in the Requirement Engineering (RE) phase. Detweiler and Harbers [73] provide tools to elicit values and embed them in the RE process by collecting value-based user stories. Thew and Sutcliffe [290] elicit stakeholders’ values by linking them to their motivations and emotions. Other works attempt to include values throughout the SE process. For example, Winter et al. [320] propose *Values Q-Sort*, a systematic approach for the elicitation and representation of values across the SE process. Perera et al. [230] introduce *Continual Value(s) Assessment*, a framework that elicits and tracks values throughout the SE process by modeling them as goals. However, such works typically employ existing value taxonomies (e.g., Schwartz’s [268] or Rescher’s [250]) to elicit stakeholders’ values. In our work, we aim to first *identify* a value list relevant to a context. Then, the SE process for applications in a context can use the value list systematically identified for that context instead of general values.

2.2 VALUES IN AI

Values are garnering attention in the AI community, especially since the leading experts in the field have defined the value alignment challenge [40, 100, 261, 283]. As an example, in 2023 the first workshop in Value Engineering in AI (VALE 2023) [224] took place at the 26th European Conference on Artificial Intelligence (ECAI 2023). The workshop contained papers ranging from the importance of values in military decision-making [332] to the identification of value awareness in large language models [1]. Similarly, the SemEval-2023 ValueEval Task: Identification of Human Values Behind Arguments [151] was co-located with the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023). The goal of this challenge was to design the best method to detect which human values support a given sentence. The emergence of such workshops and challenges reflects the affirmation of the concept of values in the AI community. In this section, we survey related works in the Autonomous Agents and MultiAgent Systems (AAMAS) and Natural Language Processing (NLP), as they represent the communities of interest for our work.

2.2.1 VALUES IN NATURAL LANGUAGE PROCESSING

Values may not be explicitly referred to in day-to-day interactions. Often, they are expressed through language, behavior, and customs, and can vary significantly across people, socio-cultural environments, and contexts [62]. Thus, ascertaining values requires extensive personal communication and analysis. The burst of online communication and social media provides an unprecedented opportunity to study value understanding from language. In this section, we provide an overview of the presence of values in the NLP community. First, we survey works aimed at classifying values in natural language. Then, we describe datasets that have been proposed for detecting values in natural language.

Value Classification Value classification has been addressed from both unsupervised and supervised approaches. Unsupervised approaches classify values in language through value lexicons, sets of words descriptive of each value [229]. Value lexicons have been used to detect value-laden text through word count or text similarity [26, 101, 228]. Value lexicons are generated manually [105], via semi-automated methods [18, 131, 251, 318], or expanded from an initial seed via NLP techniques [19, 20, 239]. However, word-level lexicons are limited by the ambiguity of natural language and the restricted range of lemmas, which can be solved by projecting the value lexicon on knowledge graphs that link moral entities and concepts [22, 134]. Supervised methods further approach these limitations through the supervised classification paradigm [142, 150, 164, 173, 206]. A textual dataset is annotated with values belonging to a value taxonomy, and the labels are used to train a supervised model. This paradigm has been especially applied with datasets annotated with the MFT taxonomy [15, 130, 133, 158]. Furthermore, additional experiments have shown that self-supervision is not sufficient for language models to discern value taxonomies such as the MFT, confirming the need for a supervised approach with human labels for the task of value classification [227]. The supervised approach can also be combined with external knowledge, e.g., Lin et al. [173] estimate moral values in tweets by combining textual features and background knowledge (context) from Wikipedia. Experiments have shown that the supervised approach produces better classification results than the unsupervised approach [133, 164, 288]. In our experiments in Part II and Part III,

we employ the supervised approach. However, supervised classifiers often suffer from domain dependency and require fine-tuning in the application context, as we elaborate further in Section 2.3.

Datasets The recent success of NLP models has sparked a surge of research in constructs akin to values, e.g., moral norms, ethical judgments, and social biases. Researchers have collected large datasets annotated with the related implicit components of human language similar to the Schwartz Value Theory [268] and the MFT [107]. Forbes et al. [92] introduced SOCIAL-CHEM-101, a corpus of almost 300,000 rules-of-thumb aimed at learning social and moral norms. Sap et al. [263] collected the Social Bias Inference Corpus with the intent of modeling the way in which people project social biases onto each other. Hendrycks et al. [124] proposed the ETHICS dataset to assess basic knowledge of ethics through well-studied theories of normative ethics (such as deontology and utilitarianism). Lourie et al. [187] introduced SCRUPLES, a dataset composed of 625,000 ethical judgments over 32,000 real-life anecdotes. Emelin et al. [85] presented *Moral Stories*, a crowd-sourced collection of contextualized narratives with the intent of investigating grounded, goal-oriented social reasoning. Jin et al. [141] proposed MoralExceptQA, the novel challenge and dataset on moral exception question answering. Kiesel et al. [150] presented a dataset of 5,270 arguments labeled with the Schwartz theory of basic values and extended it to over 9K arguments for the SemEval-2023 Task 4 [151]. Qiu et al. [244] collected ValueNet, a dataset of dialogues in different social scenarios, also annotated with the Schwartz values. Finally, Hoover et al. [130] and Trager et al. [296] collected the Moral Foundation Twitter Corpus (MFTC) and the Moral Foundation Reddit Corpus (MFRC), respectively, two datasets similarly annotated with the moral values of the MFT. In our experiments in Part II, we use the MFTC, which we introduce more in detail in Section 2.4.2.

2.2.2 VALUES IN ENGINEERING AGENTS AND MULTIAGENT SYSTEMS

Values are garnering increasing attention in engineering intelligent agents [261] and multiagent systems [217]. The AAMAS community is interested in shaping the behavior of autonomous agents and the norms that should govern a sociotechnical system (STS) [217]. In an STS, values can be operationalized at both micro and macro levels [57, 216, 321]. At a micro level, an agent ought to align its actions with an individual's value systems, e.g., by respecting their desire for privacy [6, 215]. For instance, Mosca and Such [208] propose an agent that supports the value of privacy and identifies the optimal data sharing policy by considering the value preferences of users. Mehrotra et al. [196] investigate how human and agent value similarity influences a human's trust in that agent. Chhogyal et al. [54] propose a method to assess trust between agents based on values. At a macro level, values can yield norms to govern the STS [25, 223]. Serramia et al. [270, 271] employ stakeholders' value preferences to select the most value-aligned norm system. Montes and Sierra [204] automate the synthesis of normative systems based on value promotion. Tubella et al. [298] propose the *Glass-Box* approach to evaluate the moral bounds of an AI system by mapping values to norms that constrain inputs and outputs. In Part I, we introduce a methodology that is intended to provide the starting point for such works, by identifying the values that are to be operationalized in the application context. Then, in Part III, we provide methods for estimating individual value systems, which represent the

input for engineering value-aligned AI agents and normative systems.

2.3 OTHER RELEVANT AI CONCEPTS

We review additional related AI concepts and sub-fields. For each one, we indicate the chapter(s) in which they are relevant.

2

Domain Dependency For a language model, the domain from which the training data is sourced represents the relevant decision context. Domain dependency refers to the issue that language models are often specialized or tailored to a specific domain of discourse, and struggle to generalize to data sourced from other domains. Ruder [259] provides an overview of the basic terminology, including generalizability, transferability, and catastrophic forgetting. Domain dependency is a well-known challenge that is gaining attention in the NLP community [3, 39, 192, 221, 255] and is often addressed through domain adaptation, the process of adapting a lexicon or a language model to a novel domain [119, 203, 316, 322]. Domain dependency has been investigated in classification tasks that aim to detect high-level constructs such as sentiment analysis [9, 82, 245], fake news detection [99, 278, 326], and argument mining [8, 67, 292]. However, cross-domain value classification stands out for its multi-label and subjective nature—reasoning about values [238] and thus generating value-annotated datasets is very difficult [130, 296]. In Chapter 4, we evaluate the cross-domain capabilities (generalizability, transferability, and catastrophic forgetting) of a multi-label value classifier. Our goal is to analyze the differences in value expressions across domains, but not to adapt a lexicon or a model to novel domains.

Explainability Explainable AI (XAI) methods aim to explain the decisions taken by an AI system, and are gaining increasing attention in the NLP field [64]. A key distinction is whether an XAI method generates local or global explanations. Local explanations expose the rationale behind an individual prediction of a language model, e.g., by highlighting the most important words in a sentence [191, 252]. Global explanations expose the rationale behind the whole decision-making of the model, e.g., by inducing taxonomies of words that are predictive of the classified labels [184, 242]. In Chapter 5, we propose an XAI method to investigate the source of the differences in cross-domain classification of values. Our method employs a popular XAI technique (SHAP [191]) to investigate how a language model represents value concepts across domains. Precisely, we induce value lexicons to explain the decision-making of the models, as they provide an intuitive global explanation.

Active Learning The key idea behind Active Learning (AL) is that a supervised machine Learning (ML) algorithm can achieve good performance with few training examples if such examples are suitably selected [272]. In a traditional AL setting, a large set of *unlabeled data* is available, and an *oracle* (e.g., human annotators) can be consulted to annotate the unlabeled data. A *sampling strategy* is used to iteratively select the next batch of unlabeled data to be annotated by the oracle, with the intent of rapidly improving the performance of the ML algorithm. A commonly used sampling strategy is uncertainty sampling [249], where at every iteration the ML algorithm is used to predict labels on all the unlabeled

data, and the m unlabeled data with the highest label entropy are selected as the next batch to be annotated (i.e., the data on which the model is least confident about its prediction).

AL has been extensively used in NLP applications [328], with two main strategy approaches. On the one hand, some strategies use the *informativeness* of each unlabeled instance individually, e.g., by measuring the uncertainty of the prediction or the norm of the gradient [327]. The unlabeled instances that are estimated to be most informative are selected to be labeled by the oracle. On the other hand, other strategies focus on the *representativeness* of the data, e.g., by selecting data points that are most representative of the unlabeled set [329] or that are most different from the data that is already labeled [86]. In general, state-of-the-art AL strategies exploit information about the NLP task (i.e., about the NLP model and the available data) with the intent of rapidly improving the performance of the NLP model. However, in our setting, the NLP task of value classification is a means to the end of estimating value preferences. Hence, in Chapter 7, we propose a strategy that is driven by the informativeness of the unlabeled data, but where the informativeness is derived by the downstream task of value preferences estimation.

2.4 DATASETS

We describe the datasets we employed in our experiments, two Participatory Value Evaluations and the Moral Foundation Twitter Corpus.

2.4.1 PARTICIPATORY VALUE EVALUATION (PVE)

A Participatory Value Evaluation (PVE) is a type of survey that elicits citizens' preferences about government policy options [211]. Specifically, a PVE offers a predetermined set of policy options and information about their impacts. Participants are asked for the *choices* of their preferred policy options while respecting a set of constraints (e.g., distributing a maximum amount of points across the options). Then, participants are asked to (optionally) provide textual *justifications* for their choices. Often, these justifications offer valuable insights into the value system of PVE participants. Table 2.2 shows examples of value-laden justifications in a PVE on COVID-19 relaxation measures in the Netherlands [211]. The parallels between the answers to a PVE (i.e., choices and textual justifications) and the composition of the behavioral data that we consider as input for value inference (choices and natural language justifications, see Section 1.1) make PVEs a fertile ground for performing value inference.

Table 2.2: Examples of policy options and corresponding value-laden justifications in a COVID-19 PVE [211].

Policy Option Choice	Justification
Nursing homes allow visitors again	Loneliness and isolation are a bigger killer than Corona.
All restrictions are lifted for persons who are immune	Someone's got to keep the economy going.

In our experiments, we employ data from two PVEs that were conducted in the Netherlands in 2020. The **Covid PVE** was performed to collect Dutch citizens' preferences on *lifting COVID-19 measures in the Netherlands* [211]. The survey was conducted in the

country during 29 April–6 May 2020 when partial lockdown measures were in place in the Netherlands to limit the spread of COVID-19. The government had multiple plans for lifting such measures in the following weeks and months and wanted to gauge the citizens’ opinions on the subject via PVE. The **Energy PVE** was performed to collect residents’ preferences on *future energy policies for the Súdwest-Fryslân municipality* (in the Netherlands) [137]. The survey was conducted during 10 April–3 May 2020 and was aimed at supporting the municipality in co-creating an energy policy, increasing citizen participation, and avoiding public resistance as happened in previous projects related to sustainable energy [111].

In Chapter 3, we collect the justifications provided for the Covid PVE and the Energy PVE as two corpora corresponding to two decision contexts, respectively, and use them as the starting point for identifying context-specific values. In Chapter 6, instead, we use the choices and the justifications provided to the Energy PVE to estimate participants’ value preferences. Finally, in Chapter 7, we extend the value preferences estimation methods by proposing a strategy for disambiguating value conflicts between participants’ choices and justifications. We defer to Chapters 3 and 6 for additional information on how the data is processed for the related experiments.

2.4.2 MORAL FOUNDATION TWITTER CORPUS (MFTC)

The Moral Foundation Twitter Corpus (MFTC) [130] is composed of 35,108 tweets, divided into seven datasets, each corresponding to a topic: All Lives Matter (ALM), Baltimore protests (BLT), Black Lives Matter (BLM), hate speech and offensive language (DAV) [66], 2016 presidential election (ELE), MeToo movement (MT), and Hurricane Sandy (SND). These datasets from complex and diverse socio-political issues allow us to evaluate the transferability by treating each dataset as belonging to a domain. The tweets were annotated by multiple annotators with the MFT taxonomy (see Table 2.1). Hoover et al. [130] provide additional details on the annotation process. They recognize that the vice and the virtue constituting one moral foundation are expressed differently in natural language. For example, an utterance describing a care concern (e.g., “taking care of one’s offspring”) does not necessarily also contain harm expressions. For this reason, each tweet was annotated with all 10 individual moral values plus an additional nonmoral label, resulting in 11 possible labels per tweet. Due to the subjective nature of moral values, different annotators may label the same tweet differently. For this reason, Hoover et al. [130] apply a majority vote to select the definitive label(s) of each tweet. Tweets with no majority label are labeled as nonmoral. Table 2.3 shows three examples of annotated tweets.

Table 2.3: Examples of labeled tweets in three datasets of the MFTC.

Tweet	Dataset	Labels
Police lives matter, all lives matter, peace and love people	ALM	care
Which oppression is worse, sexism or racism?	BLM	harm, cheating
Baltimore Police will deliver an update on the #Fred-dieGray investigation. Listen live on WBAL	BLT	nonmoral

Table 2.4 shows the distribution of labels. The MeanIR is a measure of imbalance in a dataset [50]. MeanIR is the mean of IR_l for each label l , where IR_l is the ratio of the number of instances having the majority (i.e., nonmoral) label and the number of instances having label l . The degree of imbalance varies largely across datasets, which is realistic since different domains are likely to have different distributions of moral content.

Table 2.4: Distribution of labels per dataset of the MFTC, including the MeanIR measure of imbalance per each dataset.

Foundation	ALM	BLT	BLM	DAV	ELE	MT	SND
Care	456	171	321	9	398	206	992
Harm	735	244	1037	138	588	433	793
Fairness	515	133	522	4	560	391	179
Cheating	505	519	876	62	620	685	459
Loyalty	244	373	523	41	207	322	415
Betrayal	40	621	169	41	128	366	146
Authority	244	17	276	20	169	415	443
Subversion	91	257	303	7	165	874	451
Purity	81	40	108	5	409	173	56
Degradation	122	28	186	67	138	941	91
Nonmoral	1744	3826	1583	4509	2501	1565	1313
Total	4424	5593	5257	5358	4961	4591	4891
MeanIR	11.5	51.3	5.4	344.8	9.6	4.0	6.4

The datasets introduced in Section 2.2.1 offer an unprecedented opportunity for studying the social and moral aspects of language. In our research we employ the MFTC as the same moral value theory is used to annotate data in seven different domains, allowing for the cross-domain comparisons that we perform in Part II.

I

2

VALUE IDENTIFICATION

3

IDENTIFYING AND EVALUATING CONTEXT-SPECIFIC VALUES

📖 **Enrico Liscio**, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2022. What Values should an Agent Align with? An Empirical Comparison of General and Context-Specific Values. In *Autonomous Agents and Multi-Agent Systems*, 36, 23.

📖 **Enrico Liscio**, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, Pradeep K. Murukannaiah. 2021. Axes: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '21*, Online, IFAAMAS, 799-808.

📖 **Enrico Liscio**, Michiel van der Meer, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2021. A Collaborative Platform for Identifying Context-Specific Values: Demo Track. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '21*, Online, IFAAMAS, 1773-1775.

The pursuit of values drives human behavior and promotes cooperation. Existing research is focused on general values (e.g., Schwartz) that transcend contexts. However, context-specific values are necessary to (1) understand human decisions, and (2) engineer intelligent agents that can elicit and align with human values. We propose Axies, a hybrid (human and AI) methodology to identify context-specific values. Axies simplifies the abstract task of value identification as a guided value annotation process involving human annotators. Axies exploits the growing availability of value-laden text corpora and Natural Language Processing to assist the annotators in systematically identifying context-specific values. We evaluate Axies in a user study involving 80 human subjects. In our study, six annotators generate value lists for two timely and important contexts: COVID-19 measures and sustainable ENERGY. We employ two policy experts and 72 crowd workers to evaluate Axies value lists and compare them to a list of general (Schwartz) values. We find that Axies yields values that are (1) more context-specific than general values, (2) more suitable for value annotation than general values, and (3) independent of the people applying the methodology.

3.1 INTRODUCTION

Values are abstract ideals and our preferences among relevant and competing values guide our actions and attitude [268]. As agents operate in sociotechnical systems [217] on behalf of and among humans [7], agents' behavior must accord with human values. There is growing recognition [100, 261, 283] that values are central to robust and beneficial AI. In a value-sensitive AI system, an agent must infer the value systems of the stakeholders [29, 282], so that it can reason about aligning its actions with the values of the stakeholders [6, 60, 62, 198]. However, a crucial question that must be answered before these steps is:

What values should an agent elicit, learn, or align with?

Several lists of *general values* have been proposed by psychologists [107, 254, 268], designers [96], and, recently, computer scientists [318]. These value lists aim to be applicable, broadly, across cultures and contexts. However, researchers recognize that not all values are relevant to all contexts [167, 238, 268]. Further, an individual's preferences over general values may not be consistent across contexts [69]. That is, how we perceive and prioritize values is context-dependent. For instance, one might value freedom over safety in general, but prioritize safety over freedom in the context of a global pandemic.

We define *context-specific values* as values that are applicable and defined within a context. For example, in the context of information sharing on social media, privacy is an applicable value, but physical health is likely not (unless we are talking about the health effects of computer use, which is another context). Further, privacy can be interpreted as intruding on one's solitude, or control on information collection, processing, and dissemination [284]. Thus, privacy defined as one's ability to control the extent to which her information is collected, processed, and disseminated is a value specific to the context of social media.

General values give insight into the broad behavioral tendencies of humans, such as openness to immigration and political activism [65]. However, for concrete applications, values must be situated within a context. Consider, for example, the task of value elicit-

tion [167]—identifying individuals’ preferences over competing values—for the purpose of decision-making on green energy transition. Given this concrete task, we can elicit stakeholders’ preferences between two context-specific values such as landscape preservation and energy independence, or between two general values such as security and self-direction. We conjecture that the choice between the context-specific values is both easier for laypeople to express and more insightful for decision-makers than the choice between the general values. Other applications where context-specific values can be beneficial include: (1) communicating values to stakeholders [307], (2) translating values into design requirements [238, 303], (3) reasoning about conflicting values [6, 215], (4) synthesizing normative systems based on values [204, 270, 293], (5) investigating how values influence trust in agents [54, 196], and (6) verifying value adherence of an AI system [298].

How can we identify values specific to a context? Since values are (high-level) cognitive abstractions, human intelligence is necessary to conceptualize a value and reason about its relevance to a context. However, thinking about values is challenging even for humans [167, 238]. Thus, we need to systematically guide and assist humans in the process of identifying context-specific values. To this end, we propose *Axies* (from the Greek word $\alpha\chi\acute{\iota}\epsilon\varsigma$, meaning *values*), a hybrid (human and AI) methodology to engage humans in identifying context-specific values and support the process via Natural Language Processing (NLP) techniques. A key idea behind *Axies* is to simplify the abstract task of value identification to a concrete task of value annotation given a (textual) value-laden opinion. With this approach, *Axies* enables human annotators to (1) learn about a context by exploring opinions about the context, and (2) think about values one opinion at a time.

There is a growing availability of value-laden opinions for many contexts on the Web, e.g., on discussion forums, tweets, and blogs. For example, Figure 3.1 shows examples of value-laden opinions on a Reddit discussion forum. By showing this opinion, *Axies* triggers a value annotator to think about the values of freedom and health in the context of COVID-19 measures. Value-laden opinions can also be collected by explicitly consulting a target population, e.g., [211].



Figure 3.1: Example of value-laden opinions on a Reddit forum. At the top, the topic defines the context under discussion. The three comments explicitly refer to the value of freedom, whereas the third only implicitly refers to the value of health.

Annotating a large opinion corpus is a significant effort. Axies distributes this task among a small group of annotators. Inspired by traditional coding methods such as the grounded theory method [102], the annotators engage in both divergent and convergent thinking by individually exploring the opinion corpus and collaboratively consolidating a value list. Axies employs an active learning strategy [32] to control the order in which opinions are shown to the annotators to reduce the annotation effort.

We conduct three experiments, involving 80 human subjects, to answer five research questions. Our experiments evaluate the characteristics of Axies values (i.e., values generated via Axies) and compare those with general (Schwartz) values [268].

3

Specificity Are Axies values more *context-specific* than general values?

Comprehensibility Are Axies values easier to *comprehend* than general values?

Consistency Does Axies yield a *consistent* set of values, independent of the annotators?

Relationship How do Axies values *relate* to general values?

Application Are Axies values easier to *apply* than general values in an annotation task?

In our first experiment, six annotators (in two groups of three) generate value lists specific to two contexts: COVID-19 relaxation measures, and sustainable ENERGY policies. In the second experiment, two policy experts evaluate the *context-specificity* of Axies and Schwartz value lists. Finally, in the third experiment, 72 crowd workers evaluate the *comprehensibility* of Axies and Schwartz value lists, and perform an annotation task with the value lists. From the crowd annotations, we (1) evaluate the *consistency* between Axies value lists generated by different annotator groups for the same context, (2) empirically study the *relationship* between Axies and Schwartz value lists, and (3) assess the *application* of the value lists by comparing the frequency and inter-rater reliability of value annotations. These experiments provide valuable insights on what values (general vs. context-specific) to choose for engineering a concrete application and the associated trade-offs.

This Chapter is organized as follows. Section 3.2 describes Axies. Section 3.3 describes the experiments. Section 3.4 discusses our results. Section 3.5 concludes the paper. We describe experiment protocols, web platform, and extended results in Appendix A. The code is available online¹.

3.2 AXIES METHODOLOGY

Figure 3.2 shows an overview of the Axies methodology. Given a context-specific opinion corpus, Axies yields a context-specific value list applicable to the *participants* producing the opinion corpus. To do so, Axies (1) exploits NLP techniques and active learning, and (2) engages a group of value *annotators* in the systematic steps of exploration (individual) and consolidation (collaborative).

¹<https://github.com/enricoliscio/axies>

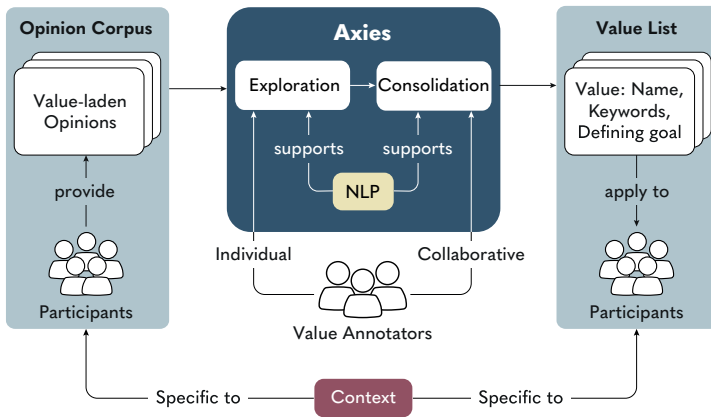


Figure 3.2: Overview of the Axies methodology. Axies takes as input a corpus composed of value-laden opinions and outputs a list of values that are relevant to the participants and the context in which the opinions were provided. Axies is composed of two phases (exploration and consolidation) and is performed by a small group of annotators with the support of NLP techniques.

Opinion Corpus The input to Axies is an *opinion corpus*, a set of participants’ opinions within a context. Axies requires the corpus to include *value-laden* opinions. A value-laden opinion indicates a participant’s value, explicitly or implicitly. For example, in Figure 3.1 the value of freedom is explicitly mentioned but health is an implicit value. We construct the opinion corpora for the Axies evaluation (Section 3.3) using the justifications from the two PVEs described in Section 2.4.1, the Covid PVE and the Energy PVE.

Value List The output of Axies is a *value list* specific to the context in which an opinion corpus is produced, and applicable to the participants producing the corpus. We represent each value in the list by a name, a set of *keywords* that characterize the value in the context, and a *defining goal* [268] that specifies what “holding a value” means in that context. For instance, Table 3.1 shows example context-specific values, applicable to Dutch citizens, produced by executing Axies on the Covid PVE data.

Table 3.1: Examples of Dutch citizens’ values resulting from executing Axies on the Covid PVE data.

Name	Keywords	Defining goal
Mental health	Loneliness, quality of life, stress	The strive towards protecting and improving one’s emotional and psychological well-being.
Economic prosperity	Economy, stability, bankruptcy	Being able to pay and afford what you need.

Value Annotators Axies is intended to be executed by a small group of annotators, who (1) produce individual value lists during *exploration*, and (2) collaboratively merge

the individual lists during *consolidation*. Axies facilitates *inductive reasoning* in that the annotators infer values held by participants (theory) based on the opinions participants express (evidence). A key advantage of this approach is that Axies yields values grounded in data. In addition, the inductive process provides an opportunity to systematically guide the annotators.

Opinion and Value Embeddings Axies represents opinions and values as vectors computed from the Sentence-BERT [248] sentence embedding model M , which takes a word or a sentence as input and returns its vector representation in an n -dimensional space ($n = 768$, in our case). In our experiments, we use the pre-trained bert-base-nli-mean-tokens model. Then, let $M(o)$ be the vector representation of an opinion o . Let n_v be the name and $K_v = \{k_v^1, \dots, k_v^n\}$ be the set of keywords of a value v . Then, Axies computes the value vector $M(v)$ using the Distributed Dictionary Representation [101] as:

$$M(v) = \frac{M(n_v) + \sum_{k \in K_v} M(k)}{\|M(n_v) + \sum_{k \in K_v} M(k)\|}. \quad (3.1)$$

With the vector representations, we can compute the cosine similarity between values and opinions, which we use to guide annotators during the value exploration phase.

3.2.1 VALUE EXPLORATION

In the exploration phase, each annotator independently develops a value list (with name and keywords for each value) by analyzing participants' opinions. Depending on the context, opinion corpora can be quite large. For example, the COVID-19 opinion corpus [211] we evaluate contains about 60,000 opinions. Thus, it is not feasible for an annotator to analyze each opinion in a corpus. Axies seeks to (1) reduce the number of opinions each annotator analyzes to produce a stable value list, and (2) increase the coverage of opinions (with respect to the corpus) the group of annotators analyzes. To this end, Axies employs NLP and active learning techniques to control the order in which the opinions in the corpus are exposed to the annotators. Thus, each annotator analyzes only a subset of the opinions in the corpus. In practice, let A be a set of value annotators that are available for a context. Then, each annotator $a \in A$ follows the exploration steps below.

Opinion selection Axies employs an active learning technique known as *Farthest First Traversal* (FFT) [32, 256]. Using FFT, Axies selects opinions such that an opinion shown to an annotator a is the farthest from the opinions already shown to the annotators in group A and the values already annotated by the annotator a . Algorithm 1 shows the pseudocode for selecting an opinion to show an annotator a . We run one instance of this algorithm to select opinions for all annotators in A to reduce the overlap in opinions shown to different annotators in A (thereby, increasing the coverage of opinions shown to the annotators in A). However, for each annotator $a \in A$, we employ the individual value list, V_a .

Annotation Algorithm 1 shows opinions to an annotator, sequentially. After seeing an opinion, an annotator can add a value (with a name and keywords) or update the name or keywords of an existing value in their value list. The annotator is asked to reason about

Algorithm 1: Fetching the next opinion using FFT.

```

Input:  $O, M$ ; /* Opinions, Embedding model */
Output:  $V_a$ ; /* Value list of  $a$  */
1 initialization:  $\forall o \in O : d_o = \infty; V_a = \emptyset$ ;
2 while  $O \neq \emptyset \ \&\& \neg \text{saturated}(V_a)$  do
3    $o_{\text{next}} = \arg \max_{o \in O} d_o$ ; /* break ties randomly */
4    $O = O - o_{\text{next}}$ ;
5    $V_a^{\text{old}} = V_a$ ;
6    $\text{update\_values}(V_a, o_{\text{next}})$ ;
7    $V_a^\delta = V_a - V_a^{\text{old}}$ ;
8    $\forall o \in O : d_o = \min \left\{ \begin{array}{l} d_o, \\ \text{cosine\_distance}(M(o), M(o_{\text{next}})), \\ \forall v \in V_a^\delta : \text{cosine\_distance}(M(o), M(v)) \end{array} \right\}$ ;

```

the values underlying a participant's opinion. However, the value name or keywords need not explicitly appear in the opinion. Upon adding a value name, Axies shows as keyword suggestions the five most similar words to the value name based on a counter-fitted word embedding model [212], trained to push synonyms closer and antonyms farther.

Termination An annotator must judge when to stop annotating. We suggest the annotators reach *inductive thematic saturation* [264], i.e., to continue annotation until the value list incurs no new changes for several new opinions shown to the annotator. We show a *progress plot* (similar to the example in Figure 3.3) to assist the annotators in deciding on termination. The progress plot shows a bar for each opinion seen by an annotator; the length of the bar is the FFT distance (d_o) at which the opinion was fetched; and the bar color indicates the annotator's action after seeing the opinion. A long sequence of opinions without the addition of value names or keywords is an indicator of a stable value list.

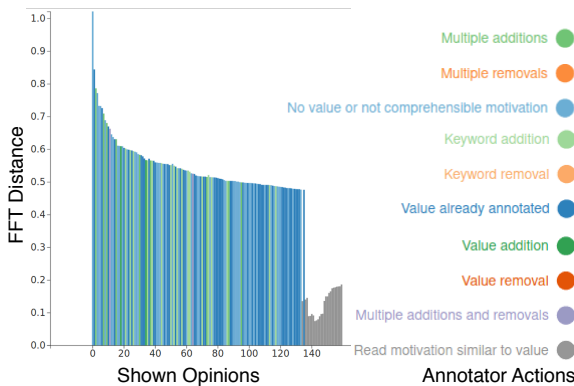


Figure 3.3: Example of a progress plot in the exploration phase. The color of each bar describes the action(s) taken by the annotator upon reading an opinion, whereas the length is the FFT distance to the already-read opinions.

Refinement Finally, Axies can fetch opinions similar to a value by computing the cosine similarity between a value and the opinions not yet shown to an annotator. An annotator can fetch opinions similar to a value to refine the value, especially if it is not well formulated. Such a phase is visible in the final gray bars in Figure 3.3.

3.2.2 VALUE CONSOLIDATION

During consolidation, the annotators in a group collaborate to merge their individual value lists. Exploration and consolidation are complementary in that exploration facilitates divergent thinking whereas consolidation facilitates convergent thinking. To facilitate consolidation, Axies creates a combined value list, $V_A = \bigcup_{a \in A} V_a$ (the union of individual value lists of annotators in group A), and guides the annotators in systematically refining V_A as described next.

Value pairs To simplify the consolidation process, Axies requires the annotators to consolidate only a pair of values at a time. Yet, consolidation is cognitively challenging. If performed naively, the annotators must compare all possible pairs of values in V_A , and repeat that process several times, to arrive at a refined V_A . To reduce the cognitive load, Axies controls the order in which value pairs are presented to the annotators—the most similar value pair from V_A (based on the sentence embeddings model M) is shown first. This approach is beneficial because similar values are likely to be merged, reducing the size of V_A , which in turn, reduces the number of value pairs to consolidate.

Consolidation actions Given a pair of values, the original annotator of each value in the pair describes the value to the other annotators in the group. Axies can fetch the opinions that led to the annotation of a value to assist an annotator in recalling the reasoning behind the annotation. The annotators in the group discuss whether the two values are conceptually similar or distinct. Then, the annotators can take one of the following actions.

- *Merge* the two values, if they are conceptually similar. The annotators may choose one of the two names or a new name for the merged value, and retain or update the keywords.
- *Update* one or both values, if the values are conceptually distinct, but changes in name or keywords make the distinction clearer.
- *Take no action*, if the two values are conceptually distinct, and the distinction is clear as is. If the annotators take no action for a pair of values, that pair is not shown to the annotators again even if that is the most similar value pair in V_A .

Termination Terminating consolidation is subject to annotators' judgment as to whether the value list requires further refinement or not. Axies shows a plot (similar to Figure 3.4) for the annotators to keep track of progress. As shown in the plot, the pairs of similar values shown early in the consolidation process lead to several value updates and merges. However, annotators may also manually choose values to merge or update; the intermittent spikes in Figure 3.4 are due to such manual choices.

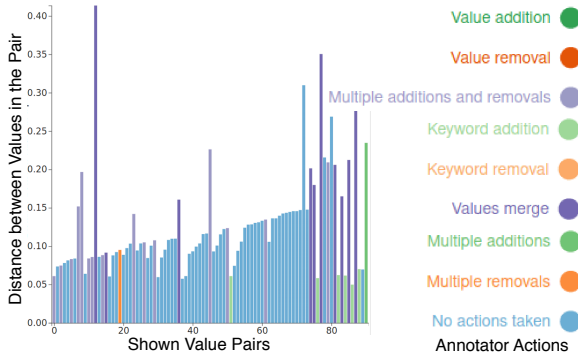


Figure 3.4: Example of a progress plot in the consolidation phase. The color of each bar describes the action(s) taken by the annotators upon reading a value pair, whereas the length is the embedding distance between the values in the pair.

Reflection As the final step, the annotators critically reflect on the consolidated value list. In particular, Axies suggests the annotators analyze each value in the list with respect to the main features of values. Schwartz [268] describes six main features of values; we include five of those, excluding the feature that (basic) values “transcend contexts” since Axies aims for context-specific values. During reflection, Axies also asks the annotators to add a defining goal for each value in the list. The defining goal characterizes what “holding a value” means. That is, a person holding a value in a context is likely to have the corresponding goal in that context. We defer the task of adding defining goals till the end of consolidation so that the task can be performed once for the final list of values.

3.3 EXPERIMENTS

We conducted three experiments with 80 human subjects to evaluate Axies as shown in Figure 3.5. These experiments were approved by the Human Research Ethics Committee of the Delft University of Technology, and we received informed consent from each subject.

In Experiment 1, two groups, G1 and G2, of three annotators each, employ Axies to generate value lists for two contexts (COVID and ENERGY) using a web application we developed [174]. Let the generated value lists be COVID-G1, ENERGY-G1, COVID-G2, and ENERGY-G2. We employ these lists and the full Schwartz list (ten values) [268] in the other two experiments to answer our research questions:

Specificity In Experiment 2, we analyze the context-specificity of COVID (G1 and G2), ENERGY (G1 and G2), and SCHWARTZ values.

Comprehensibility In Experiment 3, we analyze the clarity of each value and the distinguishability between value pairs.

Consistency In Experiment 3, we analyze the consistency between COVID-G1 and COVID-G2, and ENERGY-G1 and ENERGY-G2 using crowdsourced annotations.

Relationship In Experiment 3, we use the annotations on a set of opinions to study the relationship between Axies and Schwartz values.

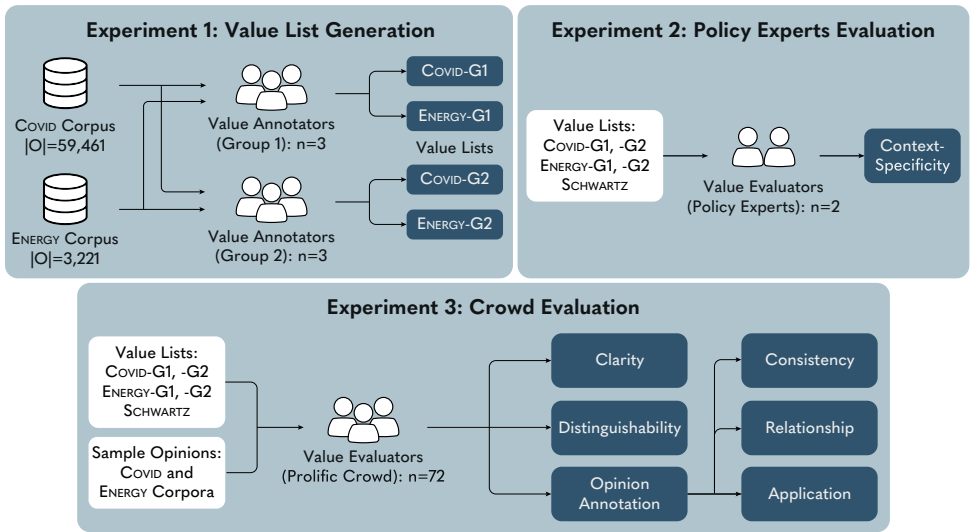


Figure 3.5: Overview of the three experiments that compose our experimental setup. In Experiment 1, two groups of annotators employ Axies to generate value lists in two contexts. In Experiment 2, two experts in policy-making evaluate the context-specificity of the yielded values. In Experiment 3, crowd workers perform a set of annotation tasks to evaluate comprehensibility, consistency, relationship, and application of the value lists.

Application In Experiment 3, we analyze the frequency of annotations and the annotator agreement to study the suitability of a value list for opinion annotation.

Through these experiments, we intend to evaluate the output of the Axies methodology. Thus, we compare the Axies (context-specific) values to the Schwartz list of (general) values due to its high contemporary influence [120]. We do not compare Axies to another value identification methodology since none of the existing methods (to the best of our knowledge) has the same purpose as Axies. Thus, the outputs of existing methods and Axies are not comparable (see Section 1.1.2 and Chapter 2). Most of the existing methods, e.g., [73, 94, 199, 219, 238, 290, 320], perform value estimation, i.e., given an existing list of values, they estimate an individual’s preferences over those values. In contrast, Axies performs context-specific value identification, i.e., given a context, Axies identifies the values relevant to that context. Among the related works, Wilson et al. [318] and Pommeranz et al. [238] are most similar to Axies. However, Wilson et al. [318] specifically pursue the creation of a general list of values. Pommeranz et al. [238] work with context-specific values, but ultimately aim at estimating individuals’ value preferences.

3.3.1 EXPERIMENT 1: VALUE LISTS

Four graduate students and two postdoctoral researchers, each working on a values-related research topic, participated as value annotators in Experiment 1. Two of these participants had a *technology and policy-making* background, and four had a *computer science* background. The two groups, G1 and G2, were constructed to have one member with *technology and policy-making* background and two with a *computer science* background

in each group.

Opinion Corpora We constructed two opinion corpora consisting of Dutch citizens' opinions in two different contexts using data collected via the two PVE surveys described in Section 2.4.1. For each context, we treat all the justifications that have been provided in the corresponding PVE as the opinions composing an opinion corpus. We refer to the PVE about lifting COVID-19 lockdown measures as COVID corpus, and to the PVE about renewable energy policies as ENERGY corpus. The opinions in both corpora were originally in Dutch. Since not all value annotators were fluent in Dutch, the opinions were translated to English using the MarianMT translator [143]. Further, opinions that contained only stop words or punctuation were removed. Then, the COVID corpus contained 59,461 and the ENERGY corpus contained 3,221 opinions.

3.3.2 EXPERIMENT 2: CONTEXT-SPECIFICITY

Two graduate students with *technology and policy-making* background participated in this experiment to evaluate the context-specificity of values. They were familiar with the COVID and ENERGY contexts in which the PVEs were conducted. However, these two participants were not involved in Experiment 1; thus, they did not know which value belonged to which list. We created a value list V_{CES} as the union of COVID-G1, ENERGY-G1, COVID-G2, ENERGY-G2, and SCHWARTZ value lists. Then, for each value $v \in V_{CES}$, we asked each participant the extent to which they agree with the following statement (once per each context) on a Likert scale of 1 (strongly disagree) to 5 (strongly agree):

If I am a policy maker in the COVID (ENERGY) context, knowing citizens' preferences about value v would help me in making a policy decision in that context.

We shuffled the combined value list V_{CES} before asking the questions above so that each participant saw the values in a random order. For each value, we showed its name, keywords, and defining goal. The two participants worked independently. After an initial round of ratings, the Intra-Class Correlation (ICC) between the two raters, an inter-rater reliability (IRR) metric for ordinal data [118], was 0.68. To ensure that the two participants had the same understanding of the task, they discussed their conceptual disagreements. Then, they performed another round of individual ratings, independently. The ICC after the second round was 0.74, which is considered just shy of excellent [118].

3.3.3 EXPERIMENT 3: COMPREHENSIBILITY, CONSISTENCY, RELATIONSHIP, AND APPLICATION

To evaluate the comprehensibility of values in a list, the consistency between Axies value lists for the same context, the relationship between Axies and Schwartz values, and the application of the value lists, we employed 72 Prolific² crowd workers. The crowd workers were directed to the Axies web application to participate in the experiment. Each crowd worker was assigned one value list and the corresponding context (in the case of the workers assigned the SCHWARTZ list, half were assigned the COVID and half the ENERGY

²www.prolific.co

context). First, each worker was asked to read the information provided on the concept of values and the corresponding context. Then, each worker performed three tasks.

CLARITY

For each value in the list assigned to a worker, given the value name, keywords, and defining goal, the worker was asked the extent to which they agree with the following statement on a Likert scale of 1 (strongly disagree) to 5 (strongly agree):

The concept described by the value is clear.

DISTINGUISHABILITY

First, for a value list V , we computed the set P_V of all value pairs: $\forall v_i, v_j \in V : v_i \neq v_j, \{v_i, v_j\} \in P_V$. Then, we showed a subset of value pairs from P_V (along with the respective keywords and defining goals) to each worker assigned to the list V . For each value pair shown, the worker was asked the extent to which they agree with the following statement on a Likert scale of 1 (strongly disagree) to 5 (strongly agree):

The two value concepts are distinguishable.

OPINION ANNOTATION

The final task for the crowd workers was to annotate opinions with values. First, we randomly selected 100 opinions from each opinion corpus. Then, we asked each worker assigned to a value list V to annotate a subset of the opinions selected for V 's context. For each opinion, a worker could select one or more values from V or mark the opinion as not value-laden. We use the annotated opinions to measure the consistency of Axies value lists, the relationship between Axies and Schwartz values, and their application.

Consistency We use the opinion annotations for evaluating the consistency of Axies value lists. Since the same 100 opinions were annotated for both Axies value lists for a context, we can measure the association between values in the two lists based on the opinions annotated with those values. For example, if the same set of opinions is annotated with $v_1 \in \text{COVID-G1}$ and $v_2 \in \text{COVID-G2}$, then we consider v_1 and v_2 as closely associated. Then, we (qualitatively) assess the consistency between COVID-G1 and COVID-G2 (similarly, ENERGY-G1 and ENERGY-G2) based on the extent to which each value in one list (e.g., COVID-G1) is associated with one or more values in another list (e.g., COVID-G2).

Relationship We use opinion annotations to study the relationship between Axies and Schwartz values. Analogous to the procedure described in the previous paragraph, we measure the association between Axies and SCHWARTZ value lists based on the opinions annotated with those value lists.

Application We compute the frequency of annotations (the number of value annotations per opinion) and the inter-rater reliability (IRR) to measure the suitability of a value list for opinion annotation. We measure IRR via Fleiss' Kappa [118] since the annotations were categorical and all opinions were rated by more than two workers.

TASK DISTRIBUTION

Table 3.2 shows the number (#) of workers assigned to each value list, and the numbers of values, value pairs, and opinions assigned to each worker. The value list and the sets of value pairs and opinions were randomly assigned. The number of workers for each list was sufficient to obtain three annotations per opinion and three distinguishability ratings per value pair (one worker in each list annotated fewer than the shown number of pairs since that was sufficient to get three ratings per pair). Each worker rated the clarity of all values in the assigned list.

Table 3.2: Overview of the distribution of the annotation load in the crowd annotation task.

Value List	#Workers	#Values	#Value pairs	#Opinions
COVID-G1	12	11	14	25
COVID-G2	10	9	11	30
ENERGY-G1	15	14	19	20
ENERGY-G2	15	13	16	20
COVID-SCHWARTZ	10	10	7	30
ENERGY-SCHWARTZ	10	10	7	30

QUALITY CONTROL

The crowd workers were required to be fluent in English and have submitted at least 100 tasks with at least a 95% acceptance rate. We included four attention check questions: two in the distinguishability rating and two in the opinion annotation task. A total of 115 workers completed the task. We included a worker's task in our analysis only if the worker (1) passed both attention checks during distinguishability rating; and (2) at least one attention check during opinion annotation (we used one instead of two as the cut-off because there was some room for subjectivity in answering the two attention check questions asked during opinion annotation). These criteria were set before any analysis of crowd work was done. Of the 115 workers, 72 satisfied the criteria above. We suggested the time required for task completion (liberal estimate) as 45 minutes. The mean time spent by a crowd worker on our task was 32 minutes (with 17 17-minute standard deviation). Each worker was paid £5.6 (at the rate of £7.5 per hour).

3.3.4 STATISTICAL ANALYSES

We perform the following statistical analyses on the data we collect. Other types of comparisons (e.g., comparisons of more than two continuous samples) are not applicable to this data.

- (1) To compare two ordinal samples, we employ Wilcoxon's ranksum test (nonparametric) [129] at a 5% significance level.
- (2) To compare two continuous samples, which meet the normality assumption, we employ Welch's t test [71] at 5% significance level. If one of the samples does not meet the normality assumption, we employ Wilcoxon's ranksum test.

- (3) To compare more than two ordinal samples, we employ the Kruskal-Wallis test (nonparametric extension of ANOVA) [129] at 5% significance level. When the Kruskal-Wallis test rejects the null hypothesis, we employ Dunn's multiple comparison test [83] with the Holm-Bonferroni correction to compare pairs of samples.
- (4) To measure the effect sizes (the amount of difference) between pairs of ordinal or continuous samples, we employ Cliff's Delta [58]. The Cliff's Delta is positive when the values in the first sample are greater than the values in the second sample more often, and negative when the values in the first sample are less than the values in the second sample more often. The magnitude of the delta is estimated according to the suggested thresholds: $\delta < 0.147$ is negligible (N); $\delta < 0.33$ is small (S); $\delta < 0.474$ is medium (M); and large (L), otherwise.

3.4 RESULTS AND DISCUSSION

We discuss the main results from our three experiments in this section. Section 3.4.1 introduces the value lists produced in Experiment 1. Sections 3.4.2, 3.4.3, 3.4.4, and 3.4.5 discuss results from Experiments 2 and 3, answering our five research questions.

3.4.1 VALUE LISTS

Two groups of three annotators each performed the two phases of the Axies methodology (exploration and consolidation) in two contexts (COVID and ENERGY).

Exploration A total of 12 explorations (six per context) were performed. In the COVID context, the mean time for exploration was 69.17 minutes (SD = 12.01 minutes), and the mean number of values annotated was 11.17 (SD = 2.64). In the ENERGY context, the mean time for exploration was 67.5 minutes (SD = 10.84 minutes), and the mean number of values annotated was 12.83 (SD = 5.23).

Consolidation A total of four consolidations were performed (two groups of three annotators each; two consolidations, one per context, for each group), producing four value lists. Table 3.3 presents an overview of the four value lists and the SCHWARTZ value list [268] for comparison. The complete lists (including keywords and defining goals) are in Appendix A.3.1. The times spent in consolidating COVID-G1, ENERGY-G1, COVID-G2, and ENERGY-G2 were 105, 110, 115, and 120 minutes, respectively.

3.4.2 CONTEXT-SPECIFICITY

To evaluate the context-specificity of a value list, we measure the extent to which its values can influence policy decisions in the context for which it was produced compared to a list produced for a different context and the SCHWARTZ value list. We compute the specificity of a value v for a context c as the mean of the ratings the two policy experts gave to value v for the context c . The policy experts were not aware of the context for which a value was annotated and spent three hours each to rate the specificity of value lists.

Figure 3.6 (left) compares the specificity of COVID (including G1 and G2), ENERGY (including G1 and G2), and SCHWARTZ values for the COVID context. Figure 3.6 (right)

Table 3.3: The four value lists generated through Axies (by two groups in two contexts) and the SCHWARTZ [268] value list.

Context	List	Value Names
COVID	G1	Well-being, Safety, Economic prosperity, Enjoyment, Fairness, Feasibility, Nuclear family, Autonomy, Care, Control
	G2	Mental health, Safety and health, Economic security, Acceptance of misbehavior, Pleasure, Conformity, Equality, Belonging to a group, Autonomy
ENERGY	G1	Community, Distributional justice, Innovation, Support, Guidance, Landscape preservation, Energy independence, Effectiveness, Sustainability, Planning for rainy days, Equal opportunities, Distrust, Regional benefits, Representation
	G2	Community, Initiative, Freedom, Organizational leadership, Involvement, Nature and landscape, Technical reliability, Technological innovation, Local benefit, Support, Free market economy, Inevitability, Fairness
General	SCHWARTZ	Tradition, Conformity, Security, Power, Achievement, Hedonism, Stimulation, Self-Direction, Universalism, Benevolence

compares the specificity of COVID (including G1 and G2), ENERGY (including G1 and G2), and SCHWARTZ values for the ENERGY context. Since the Kruskal-Wallis test indicated ($p < 0.05$) that one of the three samples is significantly different from the others in both (left and right) comparisons in Figure 3.6, we perform Dunn’s test to compare multiple pairs of samples. The table at the bottom of Figure 3.6 shows the Holm-Bonferroni (H-B) corrected p -values as well as the effect sizes, measured via Cliff’s Delta, for each pairwise comparison. For each cell in the table, the first sample in the comparison is indicated in the column header and the second sample in the comparison is indicated in the row header.

First, we observe that, in the COVID context, COVID values have significantly higher specificity ratings than the ENERGY and SCHWARTZ values with a large effect size. Similarly, in the ENERGY context, ENERGY values have significantly higher specificity ratings than the COVID and SCHWARTZ values with a large effect size. This suggests that Axies values are more context-specific than Schwartz values. This is an important result that demonstrates that Axies serves its purpose of producing context-specific value lists.

Second, the context-specificity varies among the values within the Axies lists. On the one hand, the specificity of a few Axies values is low. Specifically, Control (COVID), Representation, Technological Innovation, and Equal Opportunities (ENERGY) received average ratings lower than 3 for their respective context. We observe that these values are phrased broadly, and they may need refinement. On the other hand, the specificity of some Axies values was high for both contexts. Specifically, the COVID values of Control, Fairness, and Equality were rated higher than 3 for the ENERGY context. Similarly, the ENERGY values of Inevitability, Fairness, and Distrust were rated higher than 3 for the COVID context. Thus, some Axies values can be applicable to more than one context.

Finally, the specificity of SCHWARTZ values varies across contexts. That is, the SCHWARTZ

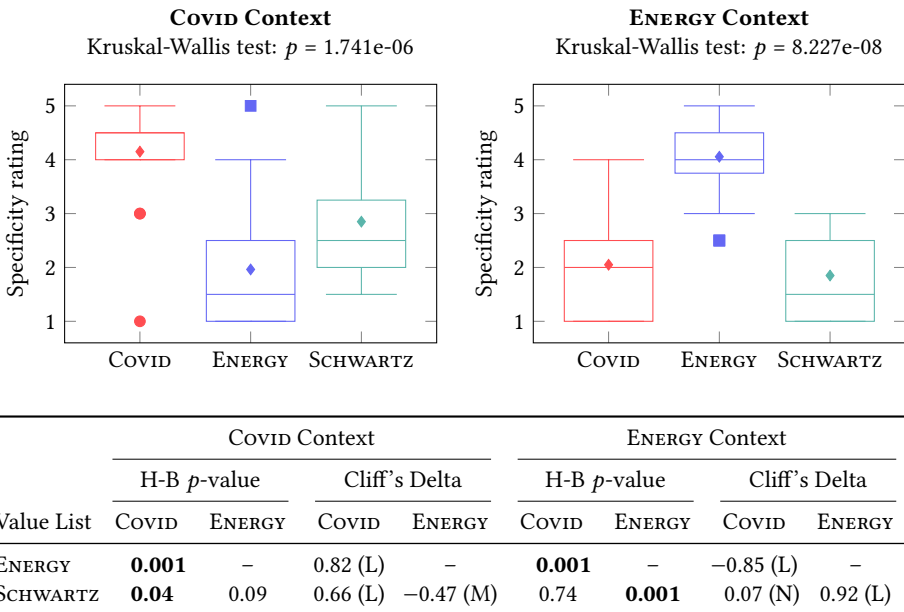


Figure 3.6: The context-specificity of Axios and SCHWARTZ value lists.

values have higher specificity ratings in the COVID context than the ENERGY context. The nature of the two contexts can explain this difference—whereas the COVID context encompasses many aspects of life (at the moment of writing), the ENERGY context is narrower. Hence, in the latter case, the (general) Schwartz values are likely to be less informative.

3.4.3 COMPREHENSIBILITY

We employ crowdsourced data to evaluate the clarity of values and the distinguishability between value pairs in a list.

CLARITY EVALUATION

Recall that the clarity of a value in a list was rated by each crowd worker assigned to that list, yielding at least ten clarity ratings (Table 3.2) per value. Figure 3.7 shows the distribution of mean clarity ratings of COVID, ENERGY, and SCHWARTZ values.

First, the mean clarity rating of all but one Axios value (in all four lists) was at least 3. The ENERGY value of Distrust received a clarity rating of less than 3. The Distrust value has the defining goal “Big players (government, large companies) should not be in charge of solving problems on citizens’ behalf.” We conjecture that the connection between the Distrust value’s name and its defining goal is not obvious, and that is the reason for the value’s low clarity rating. However, a large majority (80.9%) of the Axios values received a mean clarity rating of at least 4, suggesting that Axios value lists are clear to end users.

Second, we observe no significant difference in the clarity of COVID and SCHWARTZ values. However, the COVID and SCHWARTZ values have significantly better clarity than the ENERGY values with a medium and a large effect size, respectively. A potential reason for the

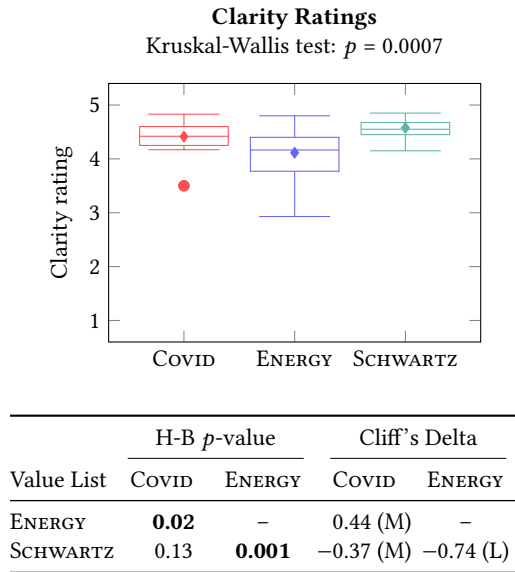


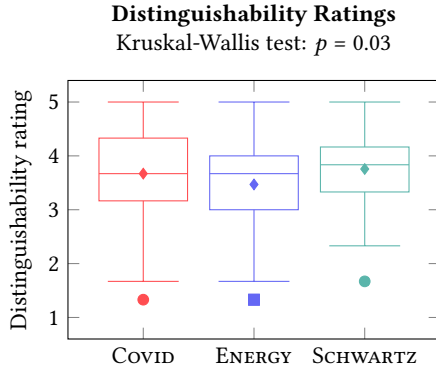
Figure 3.7: Clarity ratings of Axios and SCHWARTZ values.

better clarity of COVID values compared to the ENERGY values is the timeliness of the COVID context. Since people are currently experiencing the pandemic, they can easily understand the values in this context. In contrast, the ENERGY context yields highly specialized values (e.g., Energy Independence) which may appear unclearer to a layperson. A potential reason for the better clarity of SCHWARTZ values compared to ENERGY values (and COVID values although the difference is not statistically significant) is that the Schwartz values, being the result of years of refinement, are polished and easier to understand.

DISTINGUISHABILITY EVALUATION

For each value pair in a value list, three crowd workers indicated how distinguishable the values in the pair were. Figure 3.8 shows the mean distinguishability ratings for pairs of values in the COVID, ENERGY, and SCHWARTZ value lists.

We notice that the distinguishability of value pairs in Axios and SCHWARTZ lists is not significantly different. Further, none of the value pairs have a mean distinguishability rating of 1—that is, no two values in any of the value lists are rated as indistinguishable. However, a good number of Axios value pairs (14.3% in COVID and 22.5% in ENERGY) have a mean distinguishability rating in (1, 3). Thus, although distinguishable, the Axios values within a context have similarities among them. This observation aligns with Schwartz's [268] postulate that values form a continuum of related motivations. In fact, the mean distinguishability rating of a good number (11.1%) of SCHWARTZ value pairs is also in (1, 3). As expected, values that are adjacent in the Schwartz circumplex received low distinguishability scores (such as Conformity and Tradition, rated 1.67), and values at opposite ends received high scores (such as Self-Direction and Conformity, rated 5).



Value List	H-B p -value		Cliff's Delta	
	COVID	ENERGY	COVID	ENERGY
ENERGY	0.08	–	0.15 (S)	–
SCHWARTZ	0.57	0.08	–0.06 (N)	–0.21 (S)

Figure 3.8: Distinguishability ratings of Axies and SCHWARTZ values.

3.4.4 CONSISTENCY

The consistency between two value lists for the same context is measured through the crowdsourced opinion annotations. Each of the 100 opinions selected for each context was annotated by three crowd workers with the Axies value lists generated for that context (Section 3.3.3). We consider an opinion o as annotated with a value v if at least two of the three annotations for o include v . Then, let $v_1 \in \text{COVID-G1}$ and $v_2 \in \text{COVID-G2}$, and O_1 and O_2 be the set of opinions annotated with v_1 and v_2 , respectively. We measure the association between the two values as the Jaccard similarity between their opinion annotations:

$$J(v_1, v_2) = \frac{|O_1 \cap O_2|}{|O_1 \cup O_2|} \quad (3.2)$$

For each value in one value list for a context, Figure 3.9 shows the closest value in the other list for the context, to emphasize the associations between the two lists. Although value lists for the same context differ, we observe that each value in one list for a context is associated (has a non-zero Jaccard similarity) with at least one value in the other list for that context. In some cases, the association is apparent from the value names, e.g., Economic prosperity $\in \text{COVID-G1}$ and Economic security $\in \text{COVID-G2}$. In some cases, despite differences in the names, the values capture similar goals, e.g., Planning for rainy days $\in \text{ENERGY-G1}$ and Technical reliability $\in \text{ENERGY-G2}$, capture the same motivational goal of planning for unforeseen circumstances. In some cases, the motivation behind a value in a list was distributed over more than one value in the other list. For example, Fairness $\in \text{ENERGY-G2}$ is captured by Equal opportunities and Regional benefits $\in \text{ENERGY-G1}$. In essence, no value is conceptually exclusive to one value list within a context.

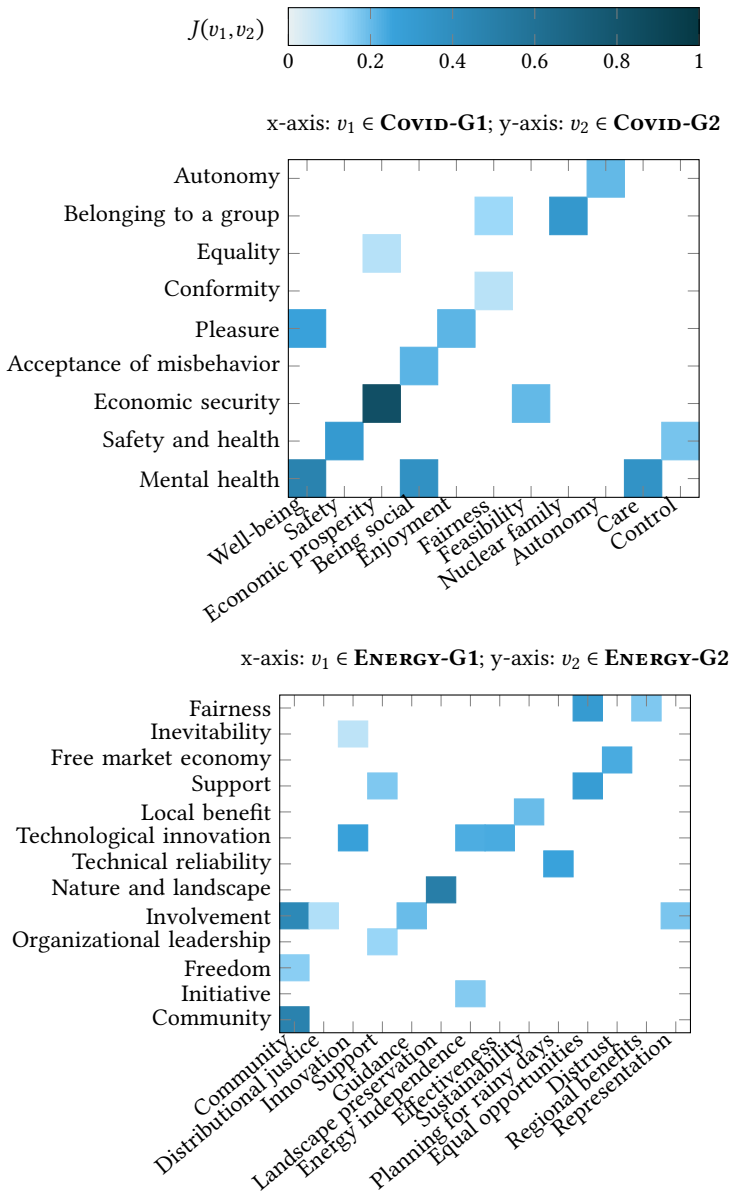


Figure 3.9: Association between G1 and G2 value lists.

3.4.5 RELATIONSHIP

Recall that, similar to Axes value annotations, each of the 100 opinions selected for each context was also annotated by three annotators with the SCHWARTZ value list, resulting in the COVID-SCHWARTZ and ENERGY-SCHWARTZ annotations. To investigate the relationship

between Axes and SCHWARTZ value lists, we employ an approach similar to the consistency evaluation (Section 3.4.4). That is, based on the annotations on the same set of opinions, we compute the Jaccard similarity between two values in different value lists as depicted in Figures 3.10 and 3.11.

3

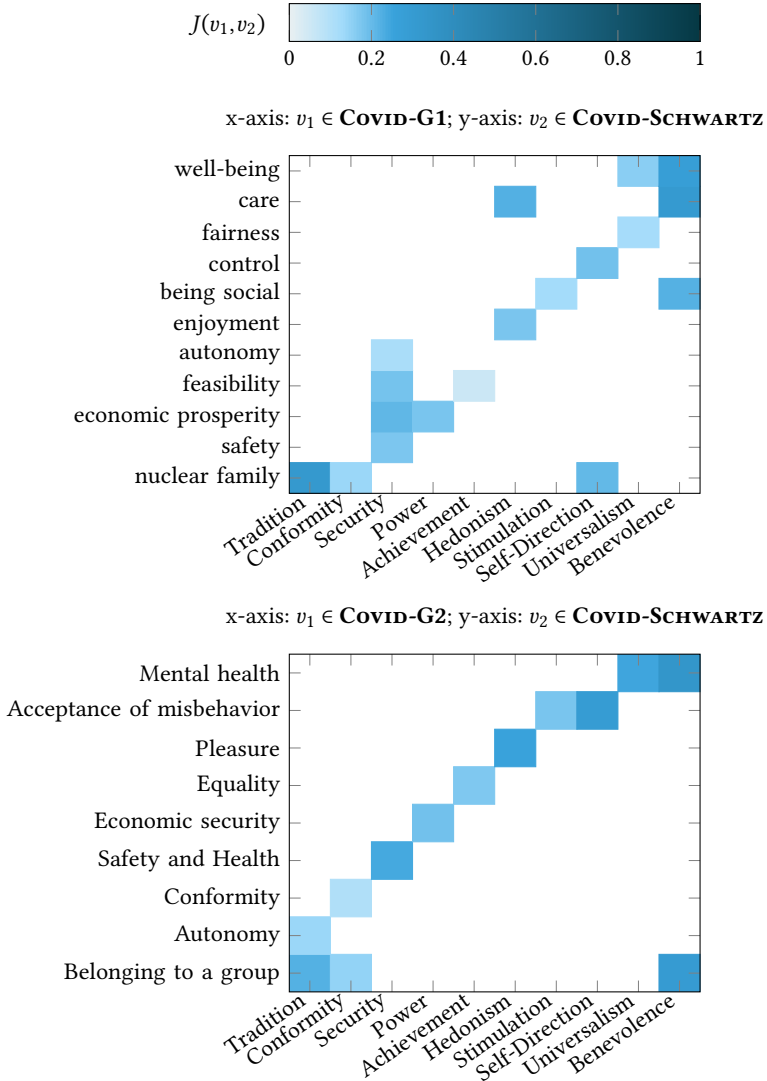


Figure 3.10: Association between Axes and SCHWARTZ values in the COVID context.

First, we observe that each SCHWARTZ value is associated (non-zero Jaccard similarity) with at least one Axes value in each of the four Axes value lists, except for the SCHWARTZ value of Conformity which has no association in the ENERGY-G2 list. However, the intensity of association is low, overall. E.g., the SCHWARTZ values of Achievement and Conformity in

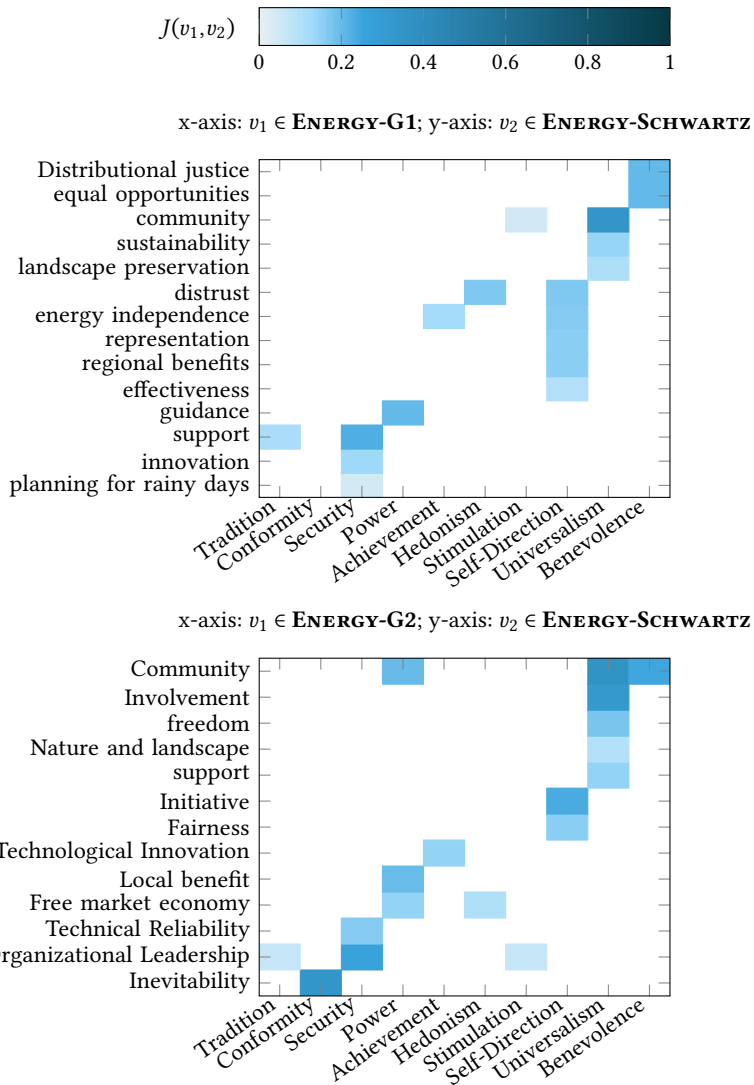


Figure 3.11: Association between Axes and SCHWARTZ values in the ENERGY context.

the COVID context, and Stimulation and Tradition in the ENERGY context have a negligible association with values in both Axies lists generated for the respective contexts.

Second, we notice that some SCHWARTZ values have one-to-many relationships with Axies values. This can be clearly observed in the ENERGY context, where SCHWARTZ values such as Self-Direction and Universalism have multiple matches with both Axies lists. The expected behavior can be also partly observed in the relationship between COVID-G1 and SCHWARTZ value lists (e.g., Security and Benevolence). However, it is less evident in the comparison between COVID-G2 and SCHWARTZ values, where it can only be partially

noticed (e.g., Benevolence).

The results above suggest that the relationship between Schwartz and Axies values depends on the context for which the Axies values are generated. In our case, since ENERGY is a specialized context, only a few general Schwartz values have clear and multiple associations with the context-specific Axies values. In contrast, since the COVID context covers many aspects of life, the Axies values generated for this context have more association with the general Schwartz values.

3

3.4.6 APPLICATION

To assess the application of the value lists, we analyze the opinion annotations. Figure 3.12 shows the number of annotations per opinion with Axies and SCHWARTZ value lists. In both contexts, the Axies values were annotated significantly more often than the SCHWARTZ values. This suggests that the Axies values are easier to recognize than the SCHWARTZ values in the opinions collected in a context.

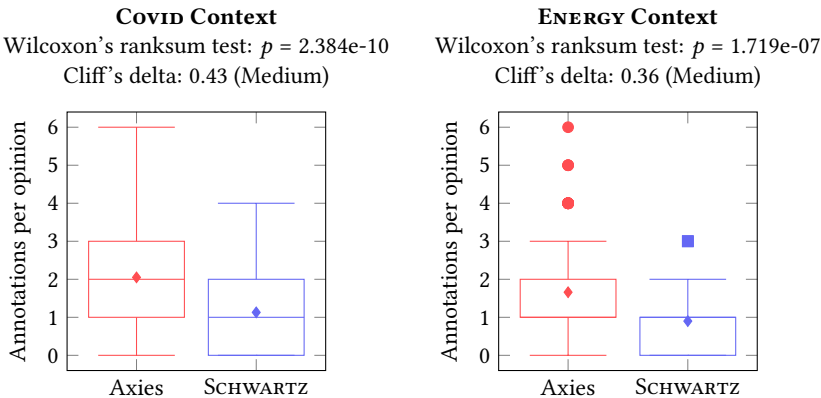


Figure 3.12: Number of annotations with values belonging to the Axies and SCHWARTZ value lists in the two contexts.

Subsequently, we compare the Inter-Rater Reliability (IRR), measured via Fleiss' Kappa, of the annotations with the value lists. Figure 3.13 presents the aggregated IRR [118] for Axies and SCHWARTZ values (Appendix A.3.2 includes IRR for each value).

The IRR is significantly higher for Axies values compared to SCHWARTZ values in both contexts. The average agreement with the SCHWARTZ values is poor, with only two values reaching a fair agreement. In contrast, a large number of Axies values is annotated with a fair agreement and some Axies values reach substantial agreement. This suggests that the annotators interpret Axies values more consistently than the (general) Schwartz values, which is desirable in concrete applications of values. Finally, the IRR is low for all value lists, which can be attributed to the inherent difficulty of annotating values [130], especially for untrained crowd workers. This can be explained by the fact that some values were annotated only a few times, rendering the agreement difficult to evaluate.

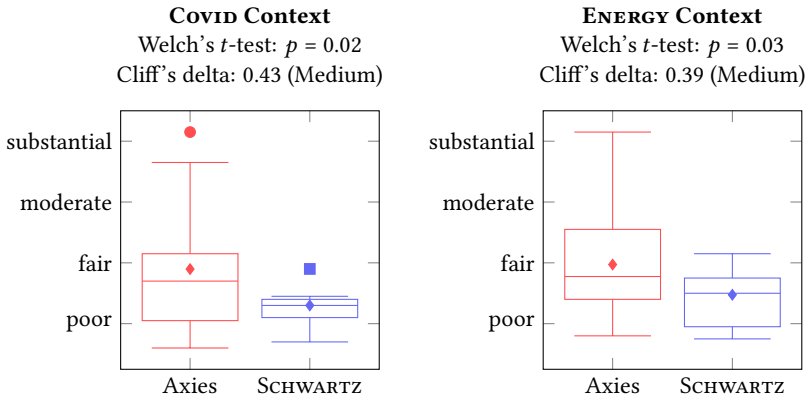


Figure 3.13: Inter-Rater Reliability of annotations with Axies and SCHWARTZ values in the two contexts.

3.4.7 THREATS TO VALIDITY

We identify three main types of threats to the validity of our findings according to the classification by Cook and Campbell [61].

Conclusion Validity concerns the ability to draw correct conclusions from the outcome of an experiment. To answer the RQs on the specificity, comprehensibility, and application of a value list, we employ rigorous statistical methods, validating the underlying assumptions (e.g., normality assumption for t -test) and performing necessary post-hoc analyses (e.g., correcting p -values during multiple comparisons). Thus, the findings on these RQs are robust. However, we could not perform statistical analyses in answering the RQs on the consistency and relationship between the value lists. Although our qualitative analyses yield valuable insights on these RQs, we recognize that these findings must be validated again via better experiment designs.

Internal Validity concerns the influences that may affect the independent variables under study with respect to causality. The subjective interpretation of values is a natural threat to validity in all our experiments. For example, the differences we observe among value lists may be influenced by the differences in the value conceptions of the annotators. The Axies methodology seeks to mitigate this threat by including the consolidation phase, where the annotators discuss their differences in interpretation. Further, in our experiments, we employ two groups of annotators and two contexts to reduce the effect of subjectivity.

External Validity concerns the limits to generalizing the results of our experiment. The small number of annotators who performed the Axies methodology and the limited number of contexts under analysis may reduce the generalizability of our conclusions. First, we required the annotators who performed the Axies methodology (as in Experiment 1) and the policy experts who evaluated context-specificity (as in Experiment 2) to be familiar with the concept of values. Our subjects in these experiments met this requirement but they were all highly educated, living in the Netherlands, and aged between 20 and 35. Thus,

the effects of a larger difference in the value annotators' and policy experts' education, residence, and age on findings in Experiments 1 and 2 remains to be studied. In Experiment 3, we evaluated the features of the values with the help of laypeople, employing a sample of 72 annotators. Although these annotators are from diverse backgrounds (Appendix A.1.3 provides an overview of the annotators' demographics), the sample of annotators is not representative of the real population, e.g., the majority of the annotators in the sample are from Europe. Thus, additional experiments with a more representative set of annotators are necessary to generalize the results to a larger population. Third, the experiments have shown slight variations of outcomes across different contexts (Sections 3.4.2, 3.4.3, and 3.4.5). Further experiments on a varied array of contexts would help in determining the generalizability of our findings. Finally, we compare the Axies value lists with only one list of general values, the Schwartz value list. However, there are other lists of general values, such as Gouveia et al. [104], Hofstede [128], and Inglehart [136]. Although there are similarities and differences among these value lists, empirical data on comparisons of general value lists is sparse [120]. Thus, the generalizability of our findings to general value lists other than the Schwartz value list remains to be studied.

3.5 CONCLUSIONS AND FUTURE DIRECTIONS

Axies combines human and artificial intelligence to yield context-specific values. In a specific context, e.g., driving, context-specific values can be more effective in explaining and predicting human behavior than general values [305]. For instance, an autonomous driving agent can concretely estimate its passengers' preferences over driving-specific values (e.g., safety and efficiency) to tailor the driving experience.

Our experiments highlight important properties of Axies and the trade-offs between context-specific and general values. First, Axies yields values that are *comprehensible* (clear and distinct) to end users. Comprehensibility is important for an agent to (1) estimate value preferences, e.g., by asking whether mental health is more important to a participant than conformity in a context, and (2) explain that the agent made a certain decision because the agent estimated, e.g., fairness as more important to the participant than regional benefits in the decision context. Yet, based on value annotators' feedback and crowd distinguishability results, we observe that values in a context have similarities since they form a motivational continuum. An interesting research direction is to identify and visualize a value continuum (e.g., as a circumplex [268]) from a list of context-specific values. We conjecture that such a visualization would support the process of building a cohesive value list.

Second, as a methodology, we expect Axies to yield reproducible results. Following Axies to annotate an opinion corpus should yield *consistent* value lists independent of the annotators. However, considering the subjective judgments involved, we do not expect a value list produced for a context by one group to be identical to the value list produced by another group. As expected, the value lists generated for the same context by different groups of annotators are not identical but consistent in that each value in one list is associated with one or more values in the other list.

Third, a key result from our experiments is that Axies yields *context-specific* values as it set out to. Specifically, we observe that the values identified for a context are more useful for decision-making in that context than in another context. However, some context-specific values are more broadly applicable than others.

Fourth, we perform an empirical comparison between the context-specific (Axies) values and general (Schwartz) values. Our results indicate that Axies values are indeed more context-specific, but slightly less clear to laypeople than Schwartz values. However, when put to the concrete *application* of value annotation, the same laypeople annotate Axies values more often and with higher agreement. This illustrates the suitability of context-specific values for practical applications.

Finally, we explore the *relationship* between Axies and Schwartz values. Our results show that only a few Schwartz values have a clear correspondence to Axies values (i.e., only the Schwartz values that are relevant to the context) and that values with a clear correspondence are often related to multiple Axies values that describe them in more fine-grained manner in the context. However, we suggest performing more extensive experiments to validate these findings in a varied set of contexts.

Identifying context-specific values is a significant effort. Axies simplifies this process and systematically guides the annotators, who need not be design experts. An interesting future direction is to analyze the benefits of NLP and active learning on the overall process (e.g., by comparing Axies to a baseline without the AI components). Further, in our experiments, the annotators followed the Axies steps one time. In practice, Axies can be used in an agile manner with multiple exploration-consolidation sprints with feedback from evaluations in between the sprints. Axies starts with the assumption that the context for which values are to be identified is already defined. However, defining a context, in itself, is a significant challenge and an essential step in engineering ethical agents [4]. A context may incorporate a variety of spatiotemporal and social elements that influence the interactions among participants and agents [5]. Thus, it is important that the opinion corpus Axies employs is representative of the intended context. For example, in our experiments, the COVID corpus contains the opinions of the residents of a country. Thus, the resulting values are applicable to the residents, but they may not be adequate to capture the values of the healthcare providers (another stakeholder group; thus, a different context). An interesting direction is to employ Axies to compare and contrast contexts. That is, given the Axies value lists for two contexts, the differences between the values in the two lists may indicate the differences between the two contexts.

Value alignment is a long-term research priority for beneficial and robust AI [261]. Our research supports a crucial step in the creation of value-aligned artificial agents—the identification of the values that an agent ought to align with. The values identified via our method can serve as the vocabulary for addressing additional challenges of value alignment such as the translation of values into norms and behaviors [271] and the verification of value adherence to norms [298]. To this end, a repository of values where values are linked with contexts and opinions would be valuable. Given such a repository, designers and developers can reuse values suitable for their contexts and an agent can automatically pick relevant values for a decision context.

II

3

VALUE CLASSIFICATION

4

CROSS-DOMAIN CLASSIFICATION OF MORAL VALUES

4

Moral values influence how we interpret and act upon the information we receive. Identifying human moral values is essential for artificially intelligent agents to co-exist with humans. Recent progress in natural language processing allows the identification of moral values in textual discourse. However, the domain-sensitivity of moral values poses challenges for transferring knowledge from one domain to another. We provide the first extensive investigation of the effects of cross-domain classification of moral values from text. We compare a state-of-the-art deep learning model (BERT) in seven domains and four cross-domain settings. We show that a value classifier can generalize and transfer knowledge to novel domains, but it can introduce catastrophic forgetting. We also highlight the typical classification errors in cross-domain value classification and compare the model predictions to the annotators' agreement. Our results provide insights to computer and social scientists who seek to detect moral values specific to a domain of discourse.

4

4.1 INTRODUCTION

Pluralist moral philosophers argue that human morality can be represented, understood, and explained by a finite number of irreducible basic elements, referred to as *moral values* [107]. The difference in our preferences over moral values explains how and why we think differently. For instance, both conservatives and liberals may agree that individual welfare is important. However, a conservative, who cherishes the values of freedom and independence, may believe that taxes should be decreased to attain more individual welfare. In contrast, a liberal, who cherishes the values of community and care, may believe that taxes should be increased to obtain welfare [105].

It is crucial to understand human morality to develop beneficial AI [261, 283]. To operate among humans, artificial agents must be able to comprehend and recognize the moral values that drive the differences in human behavior [7, 100]. The ability to detect moral values can be instrumental for, e.g., facilitating human-agent trust [54, 196] and engineering value-aligned socitechnical systems [6, 204, 217, 271].

There are survey instruments to estimate individual value profiles [107, 268]. However, reasoning about moral values is challenging for humans [167, 238]. Further, in practical applications, e.g., to conduct meaningful conversations [294] or to identify online trends [206], artificial agents should be able to detect moral values on the fly. The growing capabilities of natural language processing (NLP) enable the detection of moral values from textual discourse [15, 18, 130, 150]. Specifically, a value classifier can be used to identify the moral values underlying a piece of text on the fly. For instance, Mooijman et al. [206] show that detecting moral values from tweets can predict violent protests.

Existing value classifiers are evaluated on a specific dataset, without re-training or testing the classifier on a different dataset. This shows the ability of the classifier to predict values from text, but not the ability to transfer the learned knowledge across datasets. A critical aspect of moral values is that their expression is dependent on *context*, encompassing factors like actors, actions, judges, and values [267]. In the case of a text classifier, the context is defined by the *domain* of the data source, and moral value expressions may take different forms in different domains. For example, in the driving domain, the value of safety concerns speed limits and seat belts, but in the COVID-19 domain, safety concerns social distancing and face masks. Further, a word (broadly, language) may be linked to different moral values in different domains. For example, in a libertarian blog, the word

‘taxes’ may be linked to the authority value, but in a socialist blog, it may be linked to the community value. Thus, a value classifier must be able to recognize the domain-specific connotations of moral values.

Collecting and annotating a sufficient amount of training examples in each domain is expensive and time-consuming. To reduce the need for new annotated examples, we can pre-train classifiers with similar available annotated data and transfer the acquired knowledge to a novel task—a practice known as *transfer learning* [259]. Despite the benefits, transfer learning poses well-known challenges, including: (1) *generalizability*: how well does a classifier perform on novel data? (2) *transferability*: how well is knowledge transferred from one domain to another? and (3) *catastrophic forgetting*: to what extent is knowledge of a previous domain lost after training in a new domain? These challenges are crucial for value classification because of its domain-specific nature.

We perform the first comprehensive cross-domain evaluation of a value classifier. We employ the Moral Foundation Twitter Corpus [130], consisting of seven datasets spanning different socio-political areas, annotated with the value taxonomy of the Moral Foundation Theory [107]. Treating each dataset as a domain, we train a deep learning model, BERT [74], in four training settings to evaluate the value classifier’s generalizability, transferability, and catastrophic forgetting.

Our experiments show that (1) a value classifier can generalize to novel domains, especially when trained on a variety of domains; (2) initializing a classifier with examples from different domains improves performance in novel domains even when little training data is available in the novel domains; (3) catastrophic forgetting occurs even when training on a small portion of data from the novel domain, and its impact must be considered when training on a novel domain; and (4) in the large majority of cases, in all considered training settings, at least one annotator agrees with the model predictions.

Our investigation is significant because moral values are seldom explicit in language, but often lie in subtle domain-dependent cues. Understanding whether a classifier can recognize and transfer such hidden patterns across domains is instrumental for practical use. By unveiling the successes and mistakes of value classifiers in cross-domain settings, we hope to inspire researchers and practitioners to employ value classification responsibly.

This chapter is structured as follows. Section 4.2 describes the experiments we perform to evaluate cross-domain performance. Section 4.3 presents the results. Finally, Section 4.4 concludes the chapter. Appendix B provides additional details on our experimental setup and extended results. The code is available online¹.

4.2 EXPERIMENTAL SETUP

Predicting moral values is a multi-label classification problem. Given a set of textual documents, \mathcal{T} , and a set of moral value labels, $\mathcal{L} = (l_1, l_2, \dots, l_n)$, we wish to learn a mapping $C : \mathcal{T} \mapsto \mathcal{P}(\mathcal{L})$. Each element in $\mathcal{P}(\mathcal{L})$ is a binary vector, $y = (y_1, y_2, \dots, y_n)$, where $y_i = 1$ if the corresponding text is labeled with l_i . The mapping C is learned via BERT [74], a language representation model based on the Transformer architecture [310]. We choose BERT as it represents the state-of-the-art for several NLP tasks, including value classification [15, 150, 158]. We provide additional details, including hyperparameters, in Appendix B.1.

¹<https://github.com/adondera/transferability-of-values>

4.2.1 CROSS-DOMAIN EVALUATION

We perform a cross-domain evaluation of the Moral Foundation Twitter Corpus (MFTC), introduced in Section 2.4. The MFTC is divided into seven datasets, each corresponding to a domain of discourse. To perform our evaluation, we partition the MFTC datasets into \mathcal{T}_{source} and \mathcal{T}_{target} . We treat \mathcal{T}_{source} as available data and \mathcal{T}_{target} as an incoming dataset from a novel domain. In our experiments, \mathcal{T}_{target} is always composed of one MFTC dataset. We experiment with \mathcal{T}_{source} composed of one, three, and six datasets. We present the results for the setting with six datasets as \mathcal{T}_{source} in Section 4.3 and the other settings in Appendix B.2.

For each partition, we train a value classifier, C , in each of the four scenarios shown in Figure 4.1. These scenarios differ in how the classifier is trained. (1) In the *source* scenario, \mathcal{T}_{source} is the training set. (2) In the *target* scenario, \mathcal{T}_{target} is the training set. (3) In the *finetune* scenario, the classifier is first trained on \mathcal{T}_{source} and then continued to train (i.e., finetuned) on \mathcal{T}_{target} . (4) In the *all* scenario, the training set includes both \mathcal{T}_{source} and \mathcal{T}_{target} .

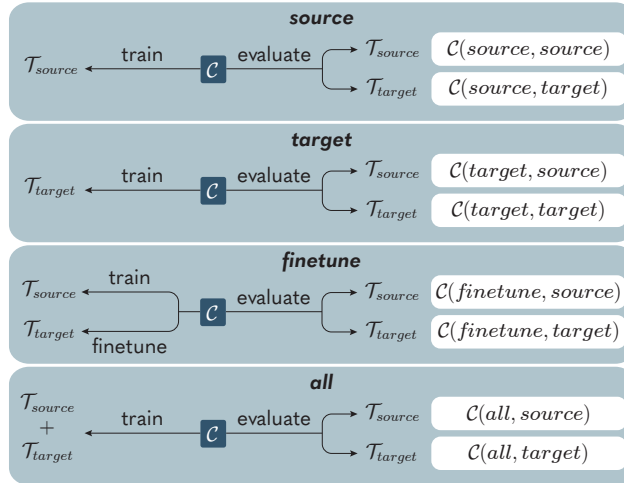


Figure 4.1: The four cross-domain scenarios. E.g., in the *source* scenario, the classifier C is (1) trained on \mathcal{T}_{source} and evaluated on \mathcal{T}_{source} ($C(source, source)$) and (2) trained on \mathcal{T}_{source} and evaluated on \mathcal{T}_{target} ($C(source, target)$).

In each scenario, the classifier is evaluated on both \mathcal{T}_{source} and \mathcal{T}_{target} , resulting in eight settings (combinations of training scenario and evaluation set) as shown in Figure 4.1. For example, $C(source, target)$ indicates that C is trained in the *source* scenario (i.e., on \mathcal{T}_{source}) and evaluated on \mathcal{T}_{target} . As we have seven partitions and four scenarios, we train 28 unique models. We evaluate the models on both \mathcal{T}_{source} and \mathcal{T}_{target} , covering 56 settings.

4.2.2 COMPARISONS

Our experimental setting (partitioning, training scenarios, and evaluation settings) enables a comprehensive cross-domain evaluation of the value classifiers as described below.

Baseline $C(source, source)$ and $C(target, target)$ show the performances of a value classifier on the training domain, when no cross-domain training is performed.

Topline $C(all, source)$ and $C(all, target)$ represent the ideal scenario, where all data is simultaneously available for training.

Generalizability $C(\text{source}, \text{target})$ and $C(\text{target}, \text{source})$ reflect the ability of a value classifier to generalize to a new domain.

Transferability Comparing $C(\text{finetune}, \text{target})$ and $C(\text{target}, \text{target})$ shows whether the knowledge learned by pretraining on $\mathcal{T}_{\text{source}}$ (*finetune* scenario) has an advantage over the absence of pretraining (*target* scenario).

Catastrophic Forgetting Comparing $C(\text{finetune}, \text{source})$ and $C(\text{source}, \text{source})$ shows how the knowledge learned by training on $\mathcal{T}_{\text{source}}$ is lost when finetuned on $\mathcal{T}_{\text{target}}$.

4.2.3 METRICS

Since the imbalance in our datasets varies greatly, we report both the micro F_1 -score and the macro F_1 -score in each setting. The micro F_1 -score, m , is the weighted (by class size) mean of the per-label F_1 -scores. The macro F_1 -score, M , is the unweighted mean of the per-label F_1 -scores.

When training and testing on the same set, we use 10-fold cross-validation with fixed splits into training and test data and report the average F_1 -scores over the 10 runs. For consistency, when testing on a set different from the training set, we test on 10 splits of the set (i.e., ultimately on the whole set) and report average F_1 -scores.

4.3 RESULTS AND DISCUSSION

We evaluate the performance of the model in four training scenarios (*source*, *target*, *finetune*, *all*). Table 4.1 reports the micro and macro F_1 -scores of the eight evaluation settings. The columns indicate the dataset used as $\mathcal{T}_{\text{target}}$ (e.g., in the BLT column, BLT is $\mathcal{T}_{\text{target}}$ and the remaining six datasets compose $\mathcal{T}_{\text{source}}$). The final column reports the average F_1 -scores over the seven datasets. We also report the results of the majority classifier which labels all tweets as nonmoral (the majority class in all datasets), for both $\mathcal{T}_{\text{source}}$ and $\mathcal{T}_{\text{target}}$.

We perform Wilcoxon’s ranksum test [129] to evaluate whether the two results significantly differ. In each column (and in the top half or the bottom half), we choose the setting with the highest F_1 -score and perform a pair-wise comparison with each of the other settings in that (half) column. We highlight, in bold, the best result and the results that are not significantly different ($p > 0.05$) from the best.

4.3.1 GENERAL TRENDS

Before cross-domain analysis, we observe some general trends. First, the topline training scenario (*all*) leads to the best results when evaluating on both $\mathcal{T}_{\text{source}}$ and $\mathcal{T}_{\text{target}}$ (Table 4.1). However, *all* is the ideal scenario. In the top half of the table, $C(\text{source}, \text{source})$ has comparable results to $C(\text{all}, \text{source})$, which is to be expected since the two models are trained on similar data (six out of seven datasets in the *source* scenario, all seven in the *all* scenario). Analogously, in the bottom half of the table, the $C(\text{finetune}, \text{target})$ setting leads to results comparable to $C(\text{all}, \text{target})$. We analyze this result further in Section 4.3.3.

Second, the results are rather consistent across datasets when evaluating on $\mathcal{T}_{\text{source}}$ (top half of Table 4.1), but have large differences when evaluating on $\mathcal{T}_{\text{target}}$ (bottom half of Table 4.1). These differences can be attributed to BLT and DAV, two highly imbalanced datasets (Table 2.4). The class imbalance also justifies the large difference between micro and macro F_1 -scores for these two datasets.

Table 4.1: Results of the four training scenarios evaluated on \mathcal{T}_{source} and \mathcal{T}_{target} . The columns indicate the dataset used as \mathcal{T}_{target} . We report both micro F_1 -score (m , left column) and macro F_1 -score (M , right column).

Classifier Setting	ALM		BLT		BLM		DAV	
	m	M	m	M	m	M	m	M
$C(source, source)$	73.9	65.6	73.9	68.3	71.2	61.8	71.1	66.4
$C(target, source)$	61.6	37.7	43.8	13.1	62.6	43.0	38.8	5.1
$C(finetune, source)$	70.3	57.2	61.2	47.8	69.2	54.9	56.6	41.9
$C(all, source)$	73.7	65.6	73.7	68.0	71.3	62.1	71.0	66.4
Majority (<i>source</i>)	47.0	6.1	42.3	5.6	49.0	6.2	38.8	5.3
$C(source, target)$	63.7	57.9	63.2	29.2	76.1	75.3	83.9	8.7
$C(target, target)$	68.0	56.8	71.4	23.5	84.4	84.6	92.2	9.0
$C(finetune, target)$	69.4	67.0	72.1	37.4	84.6	85.5	92.2	9.2
$C(all, target)$	69.9	67.0	71.2	34.7	83.9	85.2	90.4	9.3
Majority (<i>target</i>)	37.9	5.1	64.8	7.4	28.3	4.2	92.2	8.7

Table 4.1: Results of the four training scenarios evaluated on \mathcal{T}_{source} and \mathcal{T}_{target} . The columns indicate the dataset used as \mathcal{T}_{target} . We report both micro F_1 -score (m , left column) and macro F_1 -score (M , right column). (*continued*)

Classifier Setting	ELE		MT		SND		Average	
	m	M	m	M	m	M	m	M
$C(source, source)$	73.3	66.4	75.7	68.0	74.5	66.5	73.4	66.1
$C(target, source)$	59.3	40.4	52.4	39.1	54.4	36.6	53.3	30.7
$C(finetune, source)$	70.5	61.5	67.7	60.5	68.0	60.8	66.2	54.9
$C(all, source)$	73.6	66.7	75.6	67.7	74.3	66.6	73.3	66.2
Majority (<i>source</i>)	46.1	6.0	49.0	6.2	48.9	6.2	45.9	5.9
$C(source, target)$	63.4	54.8	54.3	51.3	49.2	38.6	64.8	45.1
$C(target, target)$	70.9	52.6	59.4	55.9	65.3	44.6	73.1	46.7
$C(finetune, target)$	72.9	65.2	61.4	59.3	66.7	55.6	74.2	54.2
$C(all, target)$	71.1	62.3	61.4	59.3	66.3	55.6	73.5	53.3
Majority (<i>target</i>)	44.5	5.7	27.9	4.4	26.4	4.0	46.0	5.6

4.3.2 GENERALIZABILITY

We evaluate generalizability through the results for the $C(source, target)$ and $C(target, source)$ settings. In $C(source, target)$, \mathcal{T}_{source} includes six datasets and \mathcal{T}_{target} includes one dataset. In contrast, in $C(target, source)$, \mathcal{T}_{source} includes one dataset and \mathcal{T}_{target} includes six datasets. Thus, $C(target, source)$ is a more challenging setting for generalization than $C(source, target)$.

First, we observe that the model achieves better average F_1 -scores in the $C(source, target)$ setting than the majority (*target*) baseline. This indicates that the value-laden language learned on a varied array of domains is generalizable to a novel domain to some extent, despite the domain-specific nature of moral values. However, the performances in

$C(source, target)$ are not on par with the best results on \mathcal{T}_{target} , as we discuss in Section 4.3.3. Second, we observe that the model achieves better average F_1 -scores in the $C(target, source)$ setting than the majority (*source*) baseline, despite the more challenging setting. However, the results are just marginally better than the majority (*source*) baseline, showing the difficulty in generalizing from one to multiple domains. Finally, in both cases, when we look at the results for individual datasets, the generalizability result does not hold for BLT and DAV, which highlights the challenge of generalizing to domains with a skewed distribution of moral values.

4.3.3 TRANSFERABILITY

Recall that, in the *target* scenario, a model is only trained on \mathcal{T}_{target} , but in the *finetune* scenario, the model is first trained on \mathcal{T}_{source} and then finetuned on \mathcal{T}_{target} . Thus, to evaluate transferability, we compare the $C(finetune, target)$ and $C(target, target)$ settings.

From the average F_1 -scores in Table 4.1, we observe that $C(finetune, target)$ performs better than or on par with $C(target, target)$ —precisely, similar m and 8% increase of M . Thus, the benefits of finetuning are larger for the macro than the micro F_1 -scores. This suggests that pretraining on \mathcal{T}_{source} , which contains a more varied distribution of labels than \mathcal{T}_{target} , improves the prediction of the minority labels in \mathcal{T}_{target} .

To transfer knowledge from \mathcal{T}_{source} to \mathcal{T}_{target} , typically, we need some labeled data in \mathcal{T}_{target} . For the results in Table 4.1, we used 90% of \mathcal{T}_{target} for training and the leftover 10% for evaluating at each fold. However, in practice, such a large amount of training data may not be available in the target domain. Thus, we perform an additional experiment to compare $C(target, target)$ and $C(finetune, target)$, when trained or finetuned, respectively, on a smaller portion of \mathcal{T}_{target} (10%, 25%, and 50%) and tested on a fixed, randomly selected, 10% of \mathcal{T}_{target} . Figure 4.2 shows this comparison. We report the average results of 10-fold cross-validations performed on each of the seven datasets.

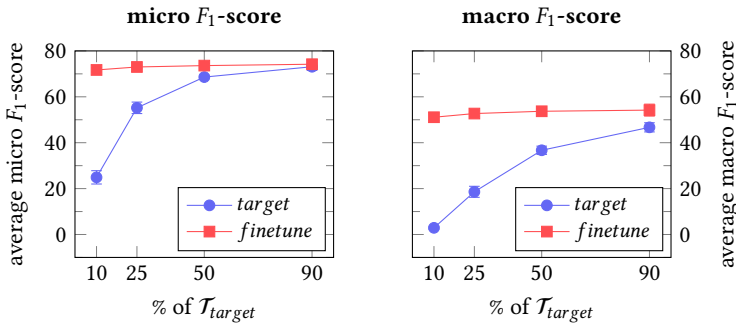


Figure 4.2: $C(target, target)$ and $C(finetune, target)$ results trained with increasing portions of \mathcal{T}_{target} .

We make an important observation from Figure 4.2. The finetuning paradigm does not require a large portion of \mathcal{T}_{target} to perform well in the target domain. In contrast, the performance of $C(target, target)$ increases (but does not surpass $C(finetune, target)$) as training data from \mathcal{T}_{target} increases. Indeed, $C(finetune, target)$ with 10% of \mathcal{T}_{target} performs on par with $C(target, target)$ trained on 90% of \mathcal{T}_{target} . This result shows that transferring

the knowledge of values from source domains to a target domain is valuable especially when the target domain has little training data.

4.3.4 CATASTROPHIC FORGETTING

Recall that, in the *source* scenario, a model is only trained on \mathcal{T}_{source} , but in the *finetune* scenario, the model is first trained on \mathcal{T}_{source} and then finetuned on \mathcal{T}_{target} . Thus, comparing $C(finetune, source)$ and $C(source, source)$ provides insight into the extent to which a model forgot about \mathcal{T}_{source} because of finetuning on \mathcal{T}_{target} .

We observe that the model suffers from catastrophic forgetting as finetuning on \mathcal{T}_{target} reduces the performance on \mathcal{T}_{source} . This is most evident when finetuning on unbalanced datasets such as DAV than balanced datasets such as BLM. In fact, $C(finetune, source)$ has only slightly worse results than $C(source, source)$ in BLM (decrease of 2% in m and 7% in M), with the difference being largest in DAV (decrease of 15% in m and 25% in M).

Figure 4.2 shows that the finetuning paradigm ensures good performances on \mathcal{T}_{target} even when the model is trained on a small portion of \mathcal{T}_{target} . We similarly evaluate catastrophic forgetting comparing $C(source, source)$ and $C(finetune, source)$ when the model is trained with increasing portions of \mathcal{T}_{target} (10%, 25%, and 50%), as shown in Figure 4.3.

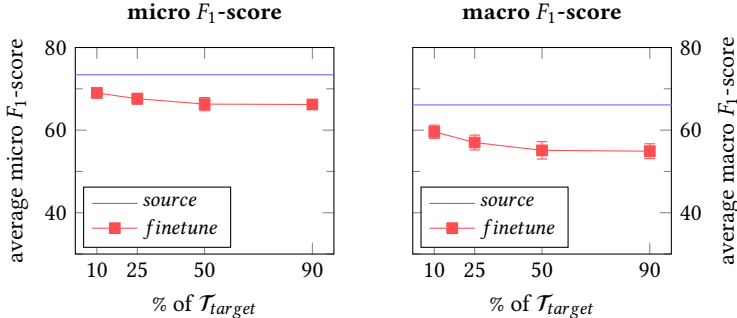


Figure 4.3: $C(source, source)$ and $C(finetune, source)$ results trained with increasing portions of \mathcal{T}_{target} .

Figure 4.3 indicates that catastrophic forgetting worsens as the model is trained with a larger portion of \mathcal{T}_{target} . $C(finetune, source)$ trained with 10% of \mathcal{T}_{target} leads to a decrease of 4% in m and 7% in M compared to $C(source, source)$ (evident by comparing the source flat blue line to the first red *finetune* square in Figure 4.3). Further, $C(finetune, target)$ trained with 10% of \mathcal{T}_{target} leads to an increase of 7% in m and 6% in M compared to $C(source, target)$ (evident by comparing the average $C(source, target)$ in Table 4.1 to the first red *finetune* square in Figure 4.2). These results show the tradeoff between the advantage of transfer learning and the impact of forgetting, even when finetuning with a small portion of \mathcal{T}_{target} .

4.3.5 MISCLASSIFICATION ERRORS

We reported F_1 -scores to provide an overview of the model performance in different training settings. Next, we investigate the behavior of the model through the lens of the MFT. We inspect (1) the confusion between morally loaded and nonmoral tweets, and (2) the mistakes among and within moral foundations since moral foundations are differentially manifested

in language [148]. We highlight the following four types of misclassification errors (which add up to 100%):

Error I A tweet labeled with one (or more) values is classified (by the model) as nonmoral.

Error II A tweet labeled as nonmoral is classified with one (or more) values.

Error III A tweet labeled with a value is classified with values from other foundations.

Error IV A tweet labeled as a vice/virtue is classified as the opposite virtue/vice of the foundation.

Table 4.2 shows the distribution of errors, averaged over the seven datasets.

Table 4.2: Distribution of errors per setting (in percentage).

Setting	Error I	Error II	Error III	Error IV
$C(\text{source}, \text{source})$	25.8	34.3	36.3	3.5
$C(\text{target}, \text{source})$	41.8	24.4	32.0	1.8
$C(\text{finetune}, \text{source})$	38.7	27.5	31.3	2.5
$C(\text{all}, \text{source})$	25.9	34.3	36.3	3.4
$C(\text{source}, \text{target})$	34.7	32.3	30.2	2.8
$C(\text{target}, \text{target})$	31.5	27.6	38.5	2.4
$C(\text{finetune}, \text{target})$	36.0	28.6	32.6	2.8
$C(\text{all}, \text{target})$	30.8	33.0	33.1	3.1

Generalizability In $C(\text{target}, \text{source})$, Error I occurs largely more often than the other errors, indicating that, when generalizing from one to several domains, labeling value-laden tweets as nonmoral is the most common mistake. In contrast, in $C(\text{source}, \text{target})$, when generalizing from several to one domain, Error I is less prominent, indicating that the model attempts to classify moral values in the novel domain.

Transferability Error III is more prevalent in $C(\text{target}, \text{target})$ than $C(\text{finetune}, \text{target})$. Thus, the confusion among moral values reduces when a model is pretrained on the source domain.

Catastrophic Forgetting Error I occurs largely more often in $C(\text{finetune}, \text{source})$ than $C(\text{source}, \text{source})$, indicating that the major type of catastrophic forgetting is missing value-laden language in the source dataset.

Finally, Error IV occurs seldom, suggesting that the models generally learn to not confuse virtues and vices within the same moral foundation.

4.3.6 ANNOTATORS AGREEMENT

We analyze the correspondence between the model predictions and the annotators agreement. Each tweet in the MFTC was annotated by at least three and at most eight different annotators [130, Table 1]. More than 99% of the tweets were annotated by three to five annotators and 84% by three or four annotators. As described in Section 2.4.2, the majority agreement was selected for training and evaluation—that is, only values annotated by at least 50% of the annotators were retained as correct labels. However, given the subjectivity in value annotation, values labeled by a minority of annotators ought to be considered too.

Table 4.3 shows the percentage of annotators that agree with the model predictions considered as errors (Table 4.3a) and accurate (Table 4.3b), averaged over the seven datasets. The columns indicate the percentage of annotators agreeing with the model prediction. For instance, if one out of the four workers who annotated a tweet agrees with the model prediction, we record a 25% agreement.

Table 4.3: Percentage distribution of predictions and annotators agreement percentage.

(a) Percentage distribution of prediction errors and annotators agreement percentage.

Setting	0	(0, 25]	(25, 34]	(34, 50)
$C(\text{source}, \text{source})$	26.1	22.3	45.0	6.6
$C(\text{target}, \text{source})$	49.5	18.0	28.5	3.9
$C(\text{finetune}, \text{source})$	38.5	20.2	36.1	5.2
$C(\text{all}, \text{source})$	26.3	22.2	45.0	6.5
$C(\text{source}, \text{target})$	40.2	23.2	30.4	6.2
$C(\text{target}, \text{target})$	19.7	30.7	40.6	8.9
$C(\text{finetune}, \text{target})$	21.2	30.5	39.9	8.4
$C(\text{all}, \text{target})$	25.6	27.5	39.0	7.9

(b) Percentage distribution of correct predictions and annotators agreement percentage.

Setting	[50, 66)	[66, 75)	[75, 100)	100
$C(\text{source}, \text{source})$	16.9	24.4	20.9	37.7
$C(\text{target}, \text{source})$	16.8	20.0	20.2	43.1
$C(\text{finetune}, \text{source})$	17.0	22.7	20.9	39.4
$C(\text{all}, \text{source})$	17.0	24.5	20.9	37.7
$C(\text{source}, \text{target})$	15.0	27.5	18.5	39.0
$C(\text{target}, \text{target})$	15.0	27.7	18.8	38.5
$C(\text{finetune}, \text{target})$	15.8	28.5	18.7	37.0
$C(\text{all}, \text{target})$	15.7	28.4	18.8	37.2

First, we analyze the classification errors in Table 4.3a. We observe that the sum of the last three columns is always larger than 50%. This indicates that, in all settings, more than half of the model classification errors are not severe in that at least one human annotator agrees with the model prediction. Then, we notice that the settings with the highest incidence of ‘bad’ classification errors (i.e., where no annotators agree with the model prediction) are those employed to evaluate generalizability ($C(\text{target}, \text{source})$ and $C(\text{source}, \text{target})$) and catastrophic forgetting ($C(\text{finetune}, \text{source})$). These results are explained by the harder challenge represented in these settings (refer to Sections 4.3.2 and 4.3.4 for a more in-depth discussion). Finally, we observe that there is a small percentage of errors with agreement between 34% and 50%. For the agreement to be in this range, a tweet must have been annotated by at least 5 annotators. However, 84% of the tweets in the MFTC have been annotated by four annotators or less, thus resulting in a smaller agreement in the last column.

Second, we analyze the correct predictions in Table 4.3b. We notice, in all settings,

a high correspondence between 100% agreement among annotators and correct model predictions—that is, tweets annotated with consistent agreement reliably lead to correct predictions. Further, we observe that the distributions of agreement and correct predictions are consistent across different settings.

4.4 CONCLUSIONS AND DIRECTIONS

We perform a comprehensive cross-domain evaluation of a multi-label value classifier, by comparing a deep learning model (BERT) in seven domains with four cross-domain training scenarios. Our aim is to support practical applications of moral value classification, e.g., the detection of radicalism through the study of moral homogeneity [23], the prediction of violent protests [206], the identification of moral concerns of citizens [211, 276], and the extraction of moral values supporting both stances and arguments [81, 306]. Our findings inform both computer scientists and social scientists on training value classifiers. However, we do not provide a fixed recipe since the right model and approach depend on the time, resources, and data available.

We show that a value classifier generally exhibits the ability to classify moral values across domains. However, the results are highly dependent on the distribution of value-laden language in a domain. Precisely, our experiments support the following key findings. First, a value classifier can generalize to novel domains, especially when trained on multiple domains. However, its performance on the novel domain improves even when trained with a small portion of data from the novel domain. Second, pretraining a value classifier with data from different domains has three benefits when finetuning the classifier. It yields (1) better performances on the novel domain than other settings, (2) good performances even when little training data is available in the novel domain, and (3) smaller confusion among moral values, especially among those less frequent in the novel domain. Third, finetuning on a novel domain causes catastrophic forgetting of the domain it was pretrained with, even when finetuning on a small portion of data from the novel domain. Thus, the tradeoff between the benefits of transferability and the adverse effects of forgetting must be considered in choosing the extent of finetuning. Finally, despite the challenging nature of cross-domain value classification, the majority of classification errors are not severe in that, in all evaluation settings, at least one annotator agrees with the model prediction.

Our investigation opens avenues for additional experiments with advanced methods to improve transfer learning [132, 140, 221] and mitigate catastrophic forgetting [154, 170, 291]. Further, based on the analysis of classification errors, we suggest incorporating the annotators (dis-)agreement into the training of the model, e.g., by employing the full distributions of annotations, as opposed to the current majority approach [301].

5

WHAT DOES A TEXT CLASSIFIER LEARN ABOUT MORALITY? AN EXPLAINABLE METHOD FOR CROSS-DOMAIN COMPARISON OF MORAL VALUES

5

📖 **Enrico Liscio**, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M. Jonker, Kyriaki Kalimeri, Pradeep K. Murukannaiah. 2023. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, Toronto, Canada, ACL, 14113-14132.

Moral values influence our judgment. Although social scientists recognize moral expression as domain specific, there are no systematic methods for analyzing whether a text classifier learns the domain-specific expression of moral language or not. We propose Tomea, a method to compare a supervised classifier’s representation of moral values across domains. Tomea enables quantitative and qualitative comparisons of moral values via an interpretable exploration of similarities and differences across moral values and domains. We apply Tomea on moral narratives in thirty-five thousand tweets from seven domains. We extensively evaluate the method via a crowd study, a series of cross-domain moral classification comparisons, and a qualitative analysis of cross-domain moral expression.

5.1 INTRODUCTION

Moral narratives play a fundamental role in the stance taken on controversial social issues [98]. Recognizing moral narratives helps understand the argumentation around important topics such as vaccine hesitancy [145], violent protests [206], and climate change [75]. Language reveals deep psychological constructs, including *moral values* [107]. Thus, language is an important avenue for analyzing moral expression. In particular, supervised classification models have been showing promising results on morality prediction [15, 124, 188]. These models leverage the wisdom of crowds (via annotations of moral expression) to attain a descriptive representation of morality. However, the supervised learning paradigm can lead to black-box models [64]. Understanding what these models learn about morality is crucial, especially when used in sensitive applications like healthcare [49, 315].

Moral expression is *context* dependent [44, 126, 159], where context refers to factors such as actors, actions, judges, and values [267]. For a text classifier, the *domain* from which the training data is sourced represents the context. For example, in the context of recent Iranian protests, tweets tagged *#mahsaamini* can form the training domain. We expect this domain to have a different moral expression than the training domain of *#prolife* tweets, representing a different context. Recent works [133, 178] analyze the out-of-domain performance of morality classifiers. However, what leads classifiers to perform differently across domains has not been systematically explored. Such an insight is essential for understanding whether classifiers can learn a domain-specific representation of morality.

We propose Tomea (from the Greek *τομέα*, meaning “domain”) to compare a text classifier’s representation of morality across domains. Tomea employs the SHAP method [191] to compile domain-specific *moral lexicons*, composed of the lemmas that the classifier deems most predictive of a moral value in a domain, for each moral value and domain. Through such moral lexicons, Tomea enables a direct comparison of the linguistic cues that the classifier prioritizes for morality prediction across domains.

We employ Tomea to compare moral language across the seven social domains in the Moral Foundation Twitter Corpus (MFTC) [130]. Then, we perform a crowdsourced evaluation to assess the agreement between the human intuition and the automatically obtained results of Tomea. We show that this agreement is consistent across domains but varies across moral values. Further, we find a strong correlation between the results of Tomea and the out-of-domain performance of the models used for obtaining the moral lexicons. In addition, we perform qualitative analyses of the moral impact of specific lemmas, unveiling insightful differences in moral values and domains.

Tomea allows us to inspect and compare the extent to which a supervised classifier

can learn domain-specific moral language from crowdsourced annotations. Tomea can guide computer scientists and practitioners (e.g., social scientists or policy-makers) in the responsible use of transfer learning approaches. In transfer learning, large datasets are used to pre-train language models, which are then finetuned with data collected in the domain of interest. Such pre-training typically helps in improving performance in the finetuning domain. However, increased performance may come at the cost of critical mistakes which may hinder the usage of the model, especially when the finetuning domain concerns minority groups [218]. Tomea can assist in the qualitative comparison of pre-training and finetuning domains by unveiling potential critical differences and guiding practitioners in judging the appropriateness of using a morality prediction model in an application.

This chapter is structured as follows. Section 5.2 introduces the Tomea method. Section 5.3 describes the experiments we perform to evaluate Tomea and Section 5.4 presents the results. Section 5.5 concludes the chapter. Appendix B provides additional details on experiment setup, crowd evaluation, and extended results. The code is available online¹.

5.2 THE TOMEA METHOD

Tomea is a method for comparing a text classifier’s representation of moral values across domains. Tomea takes as input two $\langle \text{dataset, classifier} \rangle$ pairs, where, in each pair, the classifier is trained on the corresponding dataset. Since Tomea intends to compare moral expressions across domains, the two datasets input to it are assumed to be collected in different domains. Tomea’s output is a qualitative and quantitative representation of the differences in moral expressions between the two input domains.

Figure 5.1 shows the two key steps in the method. First, we generate *moral lexicons* capturing the classifiers’ interpretable representations of the moral values specific to their domains. Then, we compare the moral lexicons in two ways. (1) We compare the moral lexicons generated for the same moral values in different domains. (2) We combine the moral lexicons generated for the same domains and provide a single measure of moral value similarity between two domains.

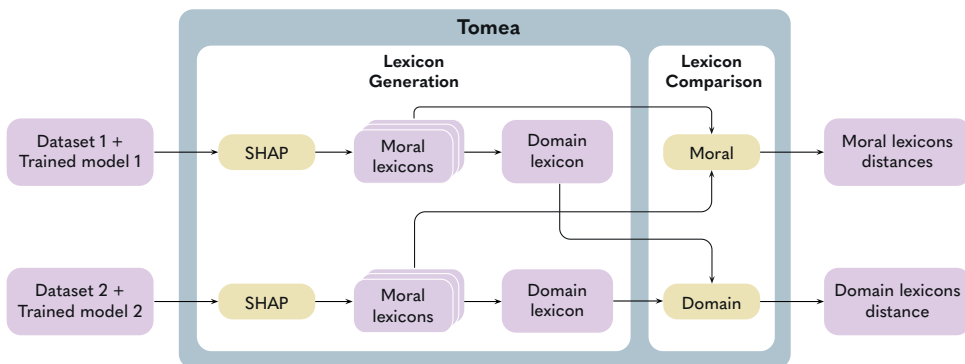


Figure 5.1: Tomea takes as input two $\langle \text{dataset, model} \rangle$ pairs (where the datasets are collected in different domains) and returns the distance in moral expressions across moral values and domains, based on the SHAP XAI technique.

¹<https://github.com/enricoliscio/tomea>

5.2.1 MORAL AND DOMAIN LEXICONS

A *moral lexicon* represents how a morality classifier interprets the expression of a moral value in a domain. We represent the expression of morality by determining the impact that each word has on the classification of a moral value in a domain. Thus, a moral lexicon consists of (w, i) pairs, where w in each pair is a word that the classifier considers relevant for predicting the examined moral value in the domain under analysis and i is its impact. This way, we generate a lexicon for each moral value in each domain. We refer to the union of the moral lexicons generated for all moral values in a domain as the *domain lexicon*.

5.2.2 LEXICON GENERATION

We use Shapley Additive Explanations (SHAP) [191] to generate the lexicons. SHAP uses Shapley values to quantify the extent to which an input component (a word) contributes toward predicting a label (a moral value). The impact of a word is computed as the *marginal contribution* of the word toward a label prediction. Intuitively, the marginal contribution of the word is calculated by removing the word from the sentence and evaluating the difference between the sentence with and without the word. All combinations of words in the sentence (i.e., the *power set* of features) are created to compute the impact of each word. The resulting impact is positive (if the likelihood of predicting a certain label increases when the word is present) or negative (if the likelihood decreases). We aggregate the local explanations to obtain a global ranking of word impact for each moral value. This can be done by adding the local impact of words for each entry of the dataset due to the additive nature of SHAP.

Tomea executes the following steps to obtain moral lexicons from a dataset and a model. (1) Execute SHAP on each entry of the dataset with the related model, resulting in a (w, i) pair for each word that appears in the dataset. (2) Replace each word w with its lemma, if one can be found using NLTK's WordNet-based lemmatizer [38]. (3) Combine words that share the same lemma by adding their impact i together.

5.2.3 LEXICON COMPARISON

Tomea enables the comparisons of (1) moral lexicons across domains, and (2) domain lexicons.

Moral Lexicons First, Tomea normalizes each moral lexicon by substituting each word's impact with its z-score [297] based on the distribution of the impact scores of all words in a moral lexicon. Then, Tomea computes an m -distance (moral value distance) to compare the lexicons of a moral value generated in different domains.

Let $W = \{w_1, \dots, w_n\}$ be the set of n common words between the moral lexicons of a moral value M_i (one of the ten in MFT) in the two domains D_A and D_B (in practice, all words that appear in both lexicons). Then, let the two vectors,

$$\mathbf{i}^{(D_A, M_i)} = [i_1^{(D_A)}, \dots, i_n^{(D_A)}] \quad \text{and} \quad \mathbf{i}^{(D_B, M_i)} = [i_1^{(D_B)}, \dots, i_n^{(D_B)}]$$

represent the impacts of the words belonging to W on M_i in domains D_A and D_B , respectively. Then, the m -distance compares the impacts that the same set of words has in the two

domains D_A and D_B for the moral value M_i as:

$$m\text{-distance}_{M_i}^{(D_A, D_B)} = d(\mathbf{i}^{(D_A, M_i)}, \mathbf{i}^{(D_B, M_i)})/n, \quad (5.1)$$

where d is Euclidean distance. The common set of words W offers a common reference point for measuring the distance between lexicons—however, we employ the full domain vocabulary to perform qualitative comparisons between domains (Section 5.4.4). We normalize the distance by n to reward domains with larger sets of common words. For a domain pair, we compute ten m -distances, one for each M_i .

Domain Lexicons To compare two domain lexicons, Tomea computes a d -distance. The d -distance between two domains D_A and D_B is the Euclidean norm of the vector of all m -distances computed between the two domains. Intuitively, the Euclidean norm represents the length of the vector of m -distances—the larger the m -distances between two domains, the larger the d -distance. For MFT, with ten moral values, d -distance is:

$$d\text{-distance}^{(D_A, D_B)} = \sqrt{\sum_{i=1}^{10} (m\text{-distance}_{M_i}^{(D_A, D_B)})^2} \quad (5.2)$$

5.3 EXPERIMENT DESIGN

We evaluate Tomea on the Moral Foundation Twitter Corpus (MFTC), introduced in Section 2.4. The MFTC is divided into seven datasets, each corresponding to a domain of discourse. Using Tomea, we generate moral and domain lexicons for the seven MFTC domains and perform pairwise comparisons, obtaining 10 m -distances and one d -distance per comparison. The m -distances and d -distances are intended to compare the classifiers' representation of moral values across domains. We perform two types of evaluation to inspect the extent to which these distances capture the differences in moral expression across domains. We also perform a qualitative analysis to find fine-grained differences across domains.

5.3.1 MODEL TRAINING

We treat morality classification as a multi-class multi-label classification with BERT [74], similar to the recent approaches [15, 133, 150, 178]. We create seven models (one per domain) using the *sequential training* paradigm [188], corresponding to the *finetune* approach in Chapter 4. That is, for each domain, the model is first pre-trained on the other six domains and then continued training on the seventh. We choose this paradigm since: (1) it is shown to offer the best performance in transfer learning [178, 188], and (2) it represents a realistic scenario, where it is fair to assume that several annotated datasets are available when a novel dataset is collected. Appendix C.1 includes additional details on the training procedure and the used hyperparameters.

5.3.2 PAIRWISE COMPARISONS

We employ Tomea to perform pairwise comparisons across the seven domains. First, we generate a moral lexicon for each of the ten moral values in each of the seven domains

(we neglect the nonmoral label as it does not expose value-laden language). This yields 70 moral lexicons. For each moral value, we perform pairwise comparisons across the seven domains, resulting in 21 m -distances per value. Finally, we perform pairwise comparisons of the seven domain lexicons to obtain 21 d -distances.

5.3.3 EVALUATION

We evaluate the extent to which m -distances and d -distances are predictive of differences in moral expression across domains. First, we perform a crowd evaluation to compare moral lexicons and their related m -distances. Then, we evaluate domain lexicons and d -distances by correlating them to the out-of-domain performances of the models.

CROWD EVALUATION

We recruited human annotators on the crowdsourcing platform Prolific² to evaluate the comparisons of moral lexicons generated for the same moral value across domains (i.e., the m -distances). We designed our annotation task with the covfee annotation tool [309]. The Ethics Committee of the Delft University of Technology approved this study, and we received informed consent from each subject.

Tomea provides m -distances that indicate the distance between domains for each moral value. We evaluate whether humans reach the same conclusions of domain similarity given the moral lexicons generated by Tomea. However, directly providing a distance or similarity between two domains is a challenging task for humans since it lacks a reference point for comparison. Thus, we re-frame the task as a simpler comparative evaluation.

Crowd task We represent each moral lexicon through a word bubble plot, where the 10 most impactful words are depicted inside bubbles scaled by word impact (Figure 5.2 shows an example). A crowd worker is shown three word bubbles, generated for the same moral value in three domains, D_A , D_B , and D_C . We ask the worker to indicate on a 6-point Likert scale whether D_A is more similar to D_B or D_C based on the shown word bubbles. Appendix C.2 shows a visual example of the task.



Figure 5.2: Word bubble plot used in the crowd evaluation for the moral value *betrayal* in the BLT domain.

²www.prolific.co

We fix one domain as D_A and choose all possible combinations of the other six domains as D_B and D_C , leading to $(6 * 5)/2 = 15$ combinations. We employ each of the seven domains as D_A , leading to 105 combinations. We generate these combinations for each of the ten moral values, resulting in 1050 unique tasks. To account for the subjectivity in the annotation, we ensure that each task is performed by three annotators, pushing the total number of required annotations to 3150. Each annotator performed 20 tasks, resulting in a total of 159 annotators. We included four control tasks in each annotator’s assignment. Appendix C.2 provides additional details on the crowd study.

Evaluation To compare the results of Tomea and the crowd annotations, we compute the correlation between m -distances and crowd answers. Since the Shapiro test showed that the crowd answers are not normally distributed, we choose Spearman correlation in which only the rank order matters.

In the crowd task, workers choose domain similarity on a six-point Likert scale. Given a domain triple (D_A, D_B, D_C) , we represent the three choices indicating D_A to be more similar to D_B than D_C as $[-2.5, -1.5, -0.5]$, and D_A to be more similar to D_C than D_B as $[0.5, 1.5, 2.5]$. For each annotation task, we average the answers received by the three annotators that performed it. In contrast, Tomea computes scores for a domain pair. To compare Tomea’s output with the output of the crowd workers, we transform the results of Tomea into the same triples evaluated in the crowd task. To do so, for a domain triple (D_A, D_B, D_C) and a moral value M_i , we compute:

$$S = m\text{-distance}_{M_i}^{(D_A, D_B)} - m\text{-distance}_{M_i}^{(D_A, D_C)}$$

As m -distances reflect the distance between domains, a negative S indicates that D_A is more similar to D_B than D_C and a positive S indicates that D_A is more similar to D_C than D_B . We correlate S and crowd answers for all 1050 annotated combinations.

OUT-OF-DOMAIN PERFORMANCE

The d -distances computed by Tomea indicate the similarity between two domains. The more similar the two domains are, the better we expect the out-of-domain performance to be. That is, if domains D_A and D_B are similar, we expect a model trained on D_A to have good classification performance on D_B , and vice versa. Thus, we evaluate the d -distances by correlating them to the out-of-domain performances of the models, computed by evaluating each model on the remaining six domains.

5.4 RESULTS AND DISCUSSION

First, we describe the pairwise comparisons resulting from Tomea. Then, we describe the results from the evaluations. Finally, we perform a qualitative analysis to provide fine-grained insights.

5.4.1 CROSS-DOMAIN COMPARISONS

For each moral value, we perform pairwise comparisons across the seven domains, resulting in 21 m -distances per value. We aggregate the moral lexicons obtained for the ten moral values to attain seven domain lexicons. We perform pairwise comparisons across the seven

domain lexicons to obtain 21 d -distances, which we display in Figure 5.1 as a 7x7 symmetric matrix. For readability, we show the scores multiplied by 100.

Table 5.1: d -distances with moral language distance between domains. A darker color depicts a smaller distance.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	6.24	4.64	6.84	5.29	5.38	5.55
BLT	6.24	-	6.23	6.09	5.37	5.50	5.56
BLM	4.64	6.23	-	6.27	4.68	5.14	5.25
DAV	6.84	6.09	6.27	-	5.96	6.54	6.80
ELE	5.29	5.37	4.68	5.96	-	4.72	4.62
MT	5.38	5.50	5.14	6.54	4.72	-	4.96
SND	5.55	5.56	5.25	6.80	4.62	4.96	-

First, we observe that the d -distances have a small magnitude and variation. This is due to the normalization in Equation 5.1 (the length of the shared vocabulary, n , is in the order of thousands). Second, we intuitively expect the moral language in the domains ALM and BLM to be relatively similar compared to other domain pairs involving ALM or BLM. The d -distances support this intuition. Third, the BLT and DAV domains have the largest overall distances from the other domains. This can be explained by their label distribution (Table 2.4), which leads to poor accuracy in predicting moral values [133, 178]. As these two domains contain fewer tweets labeled with moral values, the moral lexicons inferred in these domains are of low quality. This may explain why BLM and BLT, both domains involving protests, do not have a low d -distance. Finally, we caution that the d -distances in Table 5.1 are aggregated across moral values. Although the d -distances provide some intuition, the underlying m -distances provide more fine-grained information (Section 5.4.4 and Appendix C.3).

5

5.4.2 CROWD EVALUATION

Recall that the crowd evaluation consisted of 1050 domain triples and each triple was annotated by three annotators. The resulting Intra-Class Correlation (ICC) between the annotators, an inter-rater reliability (IRR) metric for ordinal data, was 0.66, which can be considered good but not excellent [118]. This shows that crowd workers did not annotate randomly, but can interpret the moral values differently. Such subjectivity is inevitable when annotating constructs such as morality [130, 179].

We compute the Spearman’s rank correlation (ρ) between the crowd annotations and the m -distances as described in Section 5.3.3. Table 5.2 groups the correlations by domains and moral values. The mean correlation (without any grouping) is 0.4.

We make two observations. First, despite the subjectivity and complexity in comparing moral lexicons, Tomea’s results are positively and moderately correlated with human judgment. This shows that Tomea can quantify the differences in how moral values are represented across domains. Second, although the agreement between Tomea and humans is consistent across domains, there are large variations across moral values—spanning strong (e.g., fairness), weak (e.g., authority), and negligible (e.g., purity) correlations. Although the lack of annotations for some moral values in the corpus has likely influenced these results, such variations cannot be solely explained by the label imbalance. In fact, there is

Table 5.2: Correlation between crowd annotations and m -distances, divided by domain and moral value.

(a) Correlation by domain.		(b) Correlation by moral value.	
Domain	ρ	Moral Value	ρ
ALM	0.38	Care	0.34
BLT	0.31	Harm	0.57
BLM	0.43	Fairness	0.74
DAV	0.50	Cheating	0.23
ELE	0.39	Loyalty	0.52
MT	0.42	Betrayal	0.63
SND	0.31	Authority	0.20
Average	0.39 ± 0.07	Subversion	0.51
		Purity	-0.05
		Degradation	0.35
		Average	0.4 ± 0.24

only a weak correlation ($\rho = 0.24$) between the average number of annotations of a moral value across domains (Table 2.4) and the results in Table 5.2b. Thus, we conjecture that other factors influence these variations. On the one hand, some moral values could be more difficult to identify in text than others [18, 148]. On the other hand, a strong correlation for a moral value could suggest clear differences in representing that value across domains, which both humans and Tomea recognize. Instead, a weak correlation indicates that the agreement between Tomea and humans is almost random, which could suggest that the differences across domains are small or hard to identify.

5.4.3 OUT-OF-DOMAIN PERFORMANCE

To compare the domain lexicons, we compare the d -distances to the out-of-domain performance of the models (Section 5.3.3). Table 5.3 shows the out-of-domain macro F_1 -scores of the models. The rows indicate the domain on which the model was trained, and the columns indicate the domain on which the model was evaluated. For each target domain (i.e., each column) we highlight in bold the source domain that performed best.

Table 5.3: Macro F_1 -scores of models trained on the source domain and evaluated on the target domain.

Target →	ALM	BLT	BLM	DAV	ELE	MT	SND
Source ↓							
ALM	-	48.2	83.7	11.0	68.6	61.9	61.2
BLT	58.5	-	71.6	10.7	56.2	52.2	52.7
BLM	74.0	49.9	-	12.8	75.5	64.3	64.9
DAV	49.3	31.7	64.5	-	37.9	40.4	37.1
ELE	73.9	53.6	87.6	11.9	-	67.0	67.5
MT	71.5	56.2	84.4	11.5	72.9	-	72.3
SND	73.4	51.6	88.0	12.7	72.1	67.7	-

We notice that no single domain stands out as the best source for all targets. Thus, the

choice of the source domain influences a model’s out-of-domain performance in a target domain. Hence, we investigate whether the distances Tomea computes are indicative of the out-of-domain performances.

We find a strong negative correlation ($\rho = -0.79$) between the d -distances in Table 5.1 and the out-of-domain F_1 -scores in Table 5.3. Thus, the smaller the d -distance between domains, the higher the out-of-domain performance. This demonstrates that Tomea can provide valuable insights on the out-of-domain performance of a model. To scrutinize this result further, we group the correlations by domain in Table 5.4. There is a moderate to strong negative correlation in all domains except BLT and DAV. We believe that these exceptions are because of the label imbalance and poor model performance in these two domains mentioned in Section 5.4.1.

Table 5.4: Correlation between Tomea results and out-of-domain performance of the models, divided by domain.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ρ	-1.0	0.43	-0.89	0.31	-0.71	-0.83	-0.54

5.4.4 QUALITATIVE ANALYSIS

In addition to quantitative analyses, Tomea enables deep qualitative analyses of the moral expression across domains. In this section, we show examples of (1) words that have a high impact on the same moral value across domains, (2) words that have a largely different impact on the same moral value across domains, and (3) words that have a relatively high impact on two different moral values in two different domains. Then, we show an example procedure for analyzing the differences between two domains. All lexicon values indicated in these analyses are normalized using the z-score.

First, Tomea can detect words that have a high impact on a moral value across domains. For example, the word ‘equality’ has a high impact on fairness in both ALM (21.9) and BLM (27.7) domains; similarly, the word ‘fraudulent’ has a high impact on cheating in both domains (22.6 for ALM and 16.0 for BLM). Such consistencies with a large number of words shared between the domains show consistent moral language across the domains.

Second, Tomea can detect words whose impact on a moral value largely varies across domains. This information offers a qualitative perspective on the domain dependency of moral values. For example, ALM and BLM are two of the most similar domains (Table 5.1). Yet, Tomea indicates that the word ‘treason’ has a relatively low impact on the moral value of betrayal in ALM (2.6) but a considerably higher impact in BLM (24.6); similarly, the word ‘brotherhood’ has a high impact on purity in ALM (26.9) but a comparably lower impact in BLM (8.3). Another interesting comparison can be found between the SND and BLT domains, where the word ‘embarrassing’ has a negligible impact on degradation in SND (-0.1) but a high impact in BLT (27.2). These differences can be explained by anecdotal knowledge—that is, the word ‘embarrassing’ is not relevant for degradation in the Hurricane Sandy relief domain, but it is more relevant in the domain of the Baltimore protests.

Third, Tomea can indicate how a word’s impact can vary across moral values, depending on the domain. For example, the word ‘crook’ has comparable impacts on cheating in the

ELE domain (3.1) and on degradation in the MT domain (3.9); similarly, the word ‘looting’ has a significant impact on harm in ALM (3.5) and on cheating in ELE (6.4). These examples demonstrate why domain is crucial in interpreting the moral meaning of a word.

Finally, Tomea facilitates fine-grained comparisons among specific domains of interest. Take ALM and BLM, two very similar domains according to Table 5.1, for instance. Generally, the m -distances of the moral values are low for these two domains, as shown in Table 5.5. However, the m -distances for authority and subversion are relatively higher than others. We can inspect this further using the moral lexicons generated by Tomea. For example, in subversion, words such as ‘overthrow’ and ‘mayhem’ have a high impact in ALM, whereas words such as ‘encourage’ and ‘defiance’ have a high impact in BLM. This is in line with our intuition that subversion has different connotations in the two domains—whereas subversion is negative in ALM, it is instead encouraged in BLM.

Table 5.5: The m -distances between the ALM and BLM domains, divided by moral value.

Moral Value	m -distance	Moral Value	m -distance
Care	1.62	Harm	1.15
Fairness	1.49	Cheating	1.30
Loyalty	1.54	Betrayal	1.34
Authority	1.80	Subversion	1.85
Purity	1.10	Degradation	1.30

The analyses above are not meant to be exhaustive. We pick examples of moral values, domains, and words to demonstrate the fine-grained analyses Tomea can facilitate. Our observations, considering that we only analyzed a few examples, may not be significant in themselves. Further, these observations may change with more (or other) data.

5.5 CONCLUSIONS AND DIRECTIONS

Tomea is a novel method for comparing a text classifier’s representation of morality across domains. Tomea offers quantitative measures of similarity in moral language across moral values and domains. Further, being an interpretable method, Tomea supports a fine-grained exploration of moral lexicons. Tomea is generalizable over a variety of classification models, domains, and moral constructs.

The similarities computed by Tomea positively correlate with human annotations as well as the out-of-domain performance of morality prediction models. Importantly, Tomea can shed light on how domain-specific language conveys morality, e.g., the word ‘brotherhood’ has a high impact on moral values in the ALM domain, whereas the word ‘treason’ has a high impact in the BLM domain.

Tomea can be a valuable tool for researchers and practitioners. It can be used to study how a text classifier represents moral language across personal, situational, and temporal dimensions, and across different types of moral values [179, 238]. Tomea can support societal applications such as modeling stakeholders’ preferences on societal issues [183, 211, 276], analyzing the impact of events like the COVID-19 pandemic [304], and predicting violent protests [206]. Finally, Tomea can assist NLP researchers in generating morally aligned text [16, 28] that is domain-specific.

A key direction to improve Tomea is incorporating refined explanations, e.g., by rule-based inferences [330]. Additional distance metrics and normalization procedures may also provide a more accurate lexicon comparison. Finally, the qualitative analysis that we performed could be systematized as a methodology for analysts.

III

VALUE ESTIMATION

6

VALUE PREFERENCES ESTIMATION IN HYBRID PARTICIPATORY SYSTEMS

✉ **Enrico Liscio***, Luciano C. Siebert*, Catholijn M. Jonker, Pradeep K. Murukannaiah. Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems. Under review at the *Journal of Artificial Intelligence Research (JAIR)*.

✉ Luciano C. Siebert, **Enrico Liscio**, Pradeep K. Murukannaiah, Lionel Kaptein, Shannon Spruit, Jeroen van den Hoven, Catholijn M. Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect*, IOS Press, 114-127.

Inferring citizens' values in participatory systems is crucial for citizen-centric policy-making. We envision a hybrid participatory system where participants make choices and provide justifications for those choices, and AI agents estimate their value preferences by interacting with them. We focus on situations where a conflict is detected between participants' choices and justifications, and propose methods for estimating value preferences while addressing the detected conflicts by interacting with the participants. We operationalize the philosophical stance that "valuing is deliberatively consequential." That is, if a participant's choice is based on a deliberation of value preferences, the value preferences can be observed in the justification the participant provides for the choice. Thus, we propose and compare value estimation methods that prioritize the values estimated from justifications over the values estimated from choices alone. The results show that explicitly addressing the conflicts between choices and justifications improves the estimation of an individual's value preferences.

6.1 INTRODUCTION

Enhancing citizen participation in decision-making processes is high on the European policy agenda [63]. Initiatives to foster citizens' political power and engagement have been proposed through the use of digital platforms for participatory decision-making [161, 211] and deliberation [97, 135, 275]. To this end, eliciting stakeholders' preferences over competing alternatives only provides superficial information on the debate. Instead, unveiling the trade-offs that stakeholders make among the different *values* that underlie the competing alternatives allows policy-makers to understand stakeholders at a deeper level. Values are the standards or criteria that justify one's opinions and actions, and are intrinsically linked to goals [268]. Values form an ordered system of priorities and the relative importance one ascribes to values (one's *value preferences*) guides action. Yet, how individuals ascribe relative priorities among values can vary significantly across people, socio-cultural environments [78], and decision contexts [126]. Since values preferences tend to be stable over time [268], understanding stakeholders' value preferences on a decision-making subject is crucial for crafting long-term policies on the subject.

In a participatory system, value inference is the challenge of identifying and reasoning about participants' values (Chapter 1). *Value estimation* is the third and final value inference process, and it refers to the task of estimating participants' preferences over the identified values. Value estimation has been traditionally performed based on participants' *choices* over competing alternatives, e.g., from answers to value surveys [107, 268]. However, estimating one's value preferences from both one's choices in a given context and the *justifications* for supporting these choices provides additional insights that could not be achieved considering only one source of information. To this end, the philosopher Samuel Scheffler suggests that "valuing is deliberatively consequential" [266], i.e., if one's choice is based on a deliberation of value preferences, the value preferences can be observed in the justification provided for the choice [77, 149, 233].

We envision a semi-automated approach to value estimation, where AI agents, supported by natural language processing (NLP) techniques, interpret the justifications provided by the participants in support of their choices, and combine the information contained in choices and justifications to estimate their value preferences. But what if the information extracted from the choices is in contrast with the information extracted from the justifications given in support of those choices?

We address choice-justification conflicts by following the mentioned philosophical account that “valuing is deliberatively consequential” [266]. According to this account, valuing something involves a willingness to let the values inform practical reasoning. For example, consider that Alice values online privacy. Then, she will deem reasons related to online privacy as important in related discussions and will consider it a reason for action. For instance, in the context of discussing the indiscriminate use of social media (e.g., sharing potentially sensitive pictures with a large group of participants), we would expect her to explicitly mention privacy. Thus, if Alice mentions privacy as important to her during a related conversation but one of her actions (e.g., sharing a photo of her colleagues) appears to violate the value of privacy, we detect a value conflict and prioritize the value that was explicitly mentioned, following the rationale that it was the result of internal deliberation.

We propose and compare five methods for estimating value preferences from choices and justifications which prioritize values observed in the justifications over values estimated from the choices alone. We employ the proposed methods to estimate the value preferences of the participants of a large-scale survey on energy transition [137]. We evaluate the extent to which our methods’ estimations concur with those of human evaluators. Our results show that addressing the conflicts between choices and justifications improves the estimation of value preferences.

We envision our work as supporting a *hybrid participatory system*, where humans participate in the decision-making process by making choices and providing justifications, and an AI agent supports the decision-making process by estimating the participant’s value preferences, as shown in Figure 6.1. The estimated value preferences can benefit (1) the policy-maker by indicating both what participants prefer and why, and (2) the participant by unveiling the conflicts between their choices and justification, thus helping them to clarify ambiguity and better articulate their value preferences.

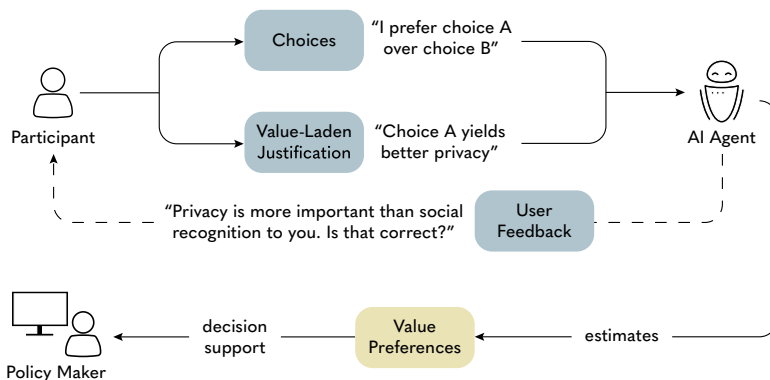


Figure 6.1: A hybrid participatory system, where human participants make choices and motivate those choices, and AI agents estimate the participants’ value preferences to assist in decision-making.

This chapter is organized as follows. Section 6.2 introduces the context behind our analyses. Section 6.3 describes the methods to perform value preferences estimation. Section 6.4 describes our experimental setup and Section 6.5 presents our results. Section 6.6

concludes the chapter. The code will be made public upon publication of the paper.

6.2 BACKGROUND

We introduce a dataset and formalize the key concepts to provide a background for our methods and experiments.

6.2.1 DATA

We estimate individual value preferences from choices and justifications provided to the Energy PVE (see description in Chapter 2.4.1). In this survey, the main question to the citizens was: “What do you find important for future decisions on energy policy?” Six choice options (Table 6.1) were developed in consultation with 45 citizens. These options were presented in the PVE platform and the participants were asked to distribute 100 points among the options. The choice options o_1 and o_2 were preferred more than other options. However, in most cases, participants distributed points to more than one option. After dividing the points, the participants had the chance to motivate each of their choices with a textual justification. 876 participants provided at least one justification for their choices, resulting in a total of 3229 justifications.

Table 6.1: Policy options in the Energy PVE and average amounts of points distributed (out of 100).

Policy option	Description	Avg. points
o_1	The municipality takes the lead and unburdens you	29.05
o_2	Inhabitants do it themselves	21.72
o_3	The market determines what is coming	9.39
o_4	Large-scale energy generation will occur in a small number of places	15.01
o_5	Betting on storage (Súdwest-Fryslân becomes the battery of the Netherlands)	12.96
o_6	Become a major energy supplier in the Netherlands	4.71

Before value preference estimation, the set of values relevant to the decision-making context must be identified [183]. Traditionally, fixed sets of values, e.g., the Schwartz Theory of Values [268] or the Moral Foundation Theory [107], were used for every context. However, in recent years there has been a push toward the identification of context-specific values [95, 175, 179, 238]. In this survey, the values embedded in the textual justifications were identified by a set of four annotators using a grounded theory approach [121]. The annotators were first introduced to foundational concepts [107, 268] and examples of values. Then, they were asked to annotate any keywords from the justifications that relate to values. After a consolidation round, annotators agreed on a list with 18 values. In this paper, we consider only the most frequent values (values mentioned at least 250 times across all project options) to demonstrate our methods. Table 6.2 shows the value list we consider in our experiments.

Table 6.3 shows the number of annotations provided for each of the values we analyze (described in Table 6.2). Although all values have more than 250 annotations (our selection criterion), these values were not annotated equally across the choice options. For example, v_3 was annotated 349 (~76%) times for o_3 , and only 3 times for o_6 .

Table 6.2: Values included in the experiments with the Energy PVE.

Value ID	Value name	Description
v_1	Cost-effectiveness	Money must be well spent and the project must be profitable. No waste. Costs should not be too high
v_2	Nature and landscape	Nature and environment are important. Horizon pollution is often seen as negative. Preserving the Frisian landscape is central
v_3	Leadership	Clarity and control over the sustainability of the energy system. Often about an organization or person that has to take charge
v_4	Cooperation	Working together on a goal. Residents can work together, but also groups and organizations
v_5	Self-determination	The opportunity for residents to make their own decision on renewable energy and to be able to implement it

Table 6.3: Distribution of values annotated for each policy option.

Annotated values	Options							O
	o_1	o_2	o_3	o_4	o_5	o_6		
v_1	90	85	102	85	89	58	509	
v_2	50	29	11	269	27	47	433	
v_3	349	40	42	13	11	3	458	
v_4	80	131	35	17	13	31	307	
v_5	35	305	7	8	20	16	391	
V	604	590	197	392	160	155		

6.2.2 FORMALIZATION

We formalize the concepts associated with the PVE (choices and justifications) and with value preferences estimation (value systems and value-option matrix). These concepts are related as shown in Figure 6.2.

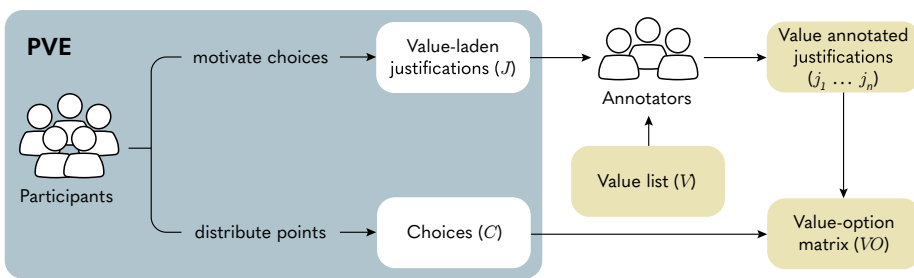


Figure 6.2: Relationship between choices, justifications, and value-option (VO) matrix.

VALUE SYSTEM

Values can be ordered according to their subjective importance as guiding principles [268]. Each person has a *value system* that internally defines the importance the values have to them according to their preference and context. We represent these value preferences via a ranking [331]. Adapting from Serramia et al. [271], we define a value system as follows.

Def 1 A value system is a pair $\langle V, R \rangle$, where V is a non-empty set of values, and R is the ranking of V which represents a person's value preference.

Def 2 A ranking R of V is a reflexive, transitive, and total binary relation, noted as $v_a \succeq v_b$. Given $v_a, v_b \in V$, if $v_a \succeq v_b$, we say v_a is more preferred than v_b . If $v_a \succeq v_b$ and $v_b \succeq v_a$, then we note it as $v_a \sim v_b$ and consider v_a and v_b indifferently preferred. However, if $v_a \succeq v_b$ but it is not true that $v_b \succeq v_a$ (i.e., $v_a \neq v_b$), then we note it as $v_a \succ v_b$.

In this work, we fix the set of values V for all participants (see Table 6.2) and we propose methods to estimate individuals' rankings over V . We refer to this task as *value preferences estimation* in the remainder of the paper. Further, ranking as defined here allows us to know the preferences between any pair of elements (unlike partial orders). We recognize that one's value preferences might not be a total order, since one could consider a given set of values incomparable. Yet, we focus on total orders as an initial step in estimating value preferences, given the challenges of fairly aggregating partial orders [234].

CHOICES AND JUSTIFICATIONS

Our goal is to estimate an individual i 's value preferences via a ranking, R^i , from i 's choices and the justifications provided for these choices. Let $O = \{o_1, \dots, o_n\}$ be a set of options i can choose from in a specific context (for example, the policy options presented in Table 6.1). We assume that i indicates their preferences, C^i , among the choices in O by distributing a certain number of points, p , among the options in O .

$$C^i = \{c_1, \dots, c_n\}, \quad c_i \in [0, p], \quad \sum c_i = p$$

Let J^i be the set of justifications that i provides for their choices:

$$J^i = \{j_1, \dots, j_n\}, \quad \text{where } j_i = \emptyset \text{ if } c_i = 0$$

Following the premise that valuing is deliberatively consequential, if an individual's value system influences their choice c_i , we expect them to mention the values that support choice c_i in the justification provided. Thus, we represent a justification j_i as the set of values (for example, the values in Table 6.2) that are mentioned in the justification:

$$j_i = \{v_1, \dots, v_m\}, \quad \text{if } v_i \text{ influenced } c_i$$

VALUE-OPTION MATRIX

Consider that the values relevant in choosing each option o_i can be determined a priori.

Def 3 A value-option matrix VO is a binary matrix with $|V|$ (number of values) rows and $|O|$ (number of options) columns, where:

$$VO(v, o) = \begin{cases} 1, & \text{if value } v \text{ is relevant for option } o \\ 0, & \text{otherwise.} \end{cases}$$

VO is the starting point for computing individual value rankings, as it represents an initial guess of value preferences in the energy transition context based on the available choices by all participants. Thus, we initialize each individual's VO matrix (VO^i) as:

$$VO^i = VO \tag{6.1}$$

6.3 METHOD

Our goal is to estimate an individual's value ranking from the division of points across a set *choices* and the textual *justifications* provided to each choice. As the choices and justifications were provided within a specific context (energy transition), the resulting value ranking is intended to represent the individual within that context.

Given VO^i , we propose methods to estimate i 's value ranking R^i from (1) choices—method C , (2) justifications—method J , or (3) choices and justifications—methods TB , JC , and JO . We provide the rationale behind each method in the related subsection. The methods C and J , which use either choices or justifications, are used as baselines for evaluating the other three methods (TB , JC , and JO). These methods can be applied sequentially—however, the order in which they are applied can change the final ranking. Figure 6.3 shows the main elements of each method, which are described next.

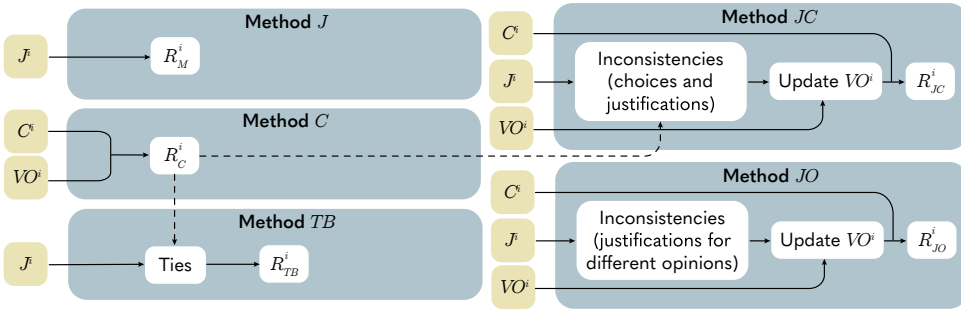


Figure 6.3: Overview of the five proposed value preferences estimation methods. The methods take as input the choices C^i , justifications J^i , and value-option matrix VO^i of an individual i and return a ranking R^i over the set of values V .

METHOD C

To estimate an individual's value ranking R_C^i solely based on their choices C^i (vector of size $|O|$, i.e., number of options), we assume that the individual's choices completely align with their value preferences. First, we compute the importance of values (U^i) for the individual by weighing the values supported by each option with the points (c_i) the individual assigns to the option. Then, we infer a ranking R_C^i from U^i , by ordering the values in V according to their importance score in U^i .

$$U^i = VO^i \times C^{iT} \quad (6.2)$$

$$R_C^i = \text{rank}(U^i) \quad (6.3)$$

METHOD J

To estimate an individual's value ranking R_M^i solely based on the justifications J^i provided to their choices C^i , we first count how many times a given value is mentioned (i.e., annotated) in any of the justifications provided, and attribute one point to each time it is mentioned. Then, we infer the ranking R_M^i by ordering the values accordingly.

METHOD TB: JUSTIFICATIONS AS TIE BREAKERS

We use the justifications J^i as *tie breakers* to reduce indifferent preferences in a value ranking. We start with a given ranking R^i (e.g., R_C^i). Then, let us define that a tie $\tau_{a,b} \in R^i$ between two values $v_a, v_b \in V$ is present when v_a and v_b are indifferently preferred ($v_a \sim v_b$). Due to symmetry, we consider that $\tau_{a,b} = \tau_{b,a}$. If there is a tie $\tau_{a,b}$ and if one of the justifications mentions v_a but none of the justifications mention v_b , then the *TB* method considers $v_a \succ v_b$, and thus breaks the tie. If both values are mentioned in one of the justifications or not mentioned in any justification, the tie remains, as illustrated in Algorithm 2.

Algorithm 2: Method TB

Input: R^i, J^i
Output: R_{TB}^i

```

1  $R_{TB}^i \leftarrow R^i$ 
2 for  $\tau_{a,b} \in R^i$  do
3   if  $(\exists j \in J^i : v_a \in j) \wedge (\nexists j \in J^i : v_b \in j)$  then
4      $\text{set } v_a \succ v_b \text{ in } R_{TB}^i$ ;
5   else if  $(\exists j \in J^i : v_b \in j) \wedge (\nexists j \in J^i : v_a \in j)$  then
6      $\text{set } v_b \succ v_a \text{ in } R_{TB}^i$ ;
7 end
```

6

METHOD JC: JUSTIFICATIONS ARE MORE RELEVANT THAN CHOICES

There may be a conflict between R^i previously estimated for an individual and the values supported by their justifications. That is, R^i indicates $v_b \succ v_a$ but v_a is supported in a justification $j_o \in J^i$, and v_b is not supported in any justification. In this case, the *JC* method prioritizes the value mentioned in the justification over the one not mentioned, assuming that the value not mentioned is not relevant for individual i in option o . When a conflict is detected, we assume that the initial value-option matrix VO^i was inaccurate and update it by setting the cell of VO^i corresponding to v_b for the option o supported by $j_o = \{v_a\}$ to 0. Once VO^i is updated for all conflicts, we compute the value ranking R_{JC}^i as in Algorithm 3.

Algorithm 3: Method JC

Input: R^i, J^i, VO^i, V, C^i
Output: R_{JC}^i

```

1 for  $j_o \in J^i$  do
2   for  $v_a \in j_o$  do
3     for  $v_b \in V \setminus \{v_a\}$  do
4       if  $v_a \prec v_b$  then
5          $VO^i(v_b, o) = 0$ ;
6       end
7     end
8 end
9  $U^i = VO^i \times C^i$ ;
10  $R_{JC}^i = \text{rank}(U^i)$ ;
```

METHOD JO: JUSTIFICATIONS ARE ONLY RELEVANT FOR ONE OPTION

The justifications J^i provided for different options can also bring conflicts. For example, assume options o_1 and o_2 , for which all values $v_i \in V$ are considered relevant. Further, assume that individual i motivated o_1 with value v_3 ($j_1 = \{v_3\}$), and o_2 with value v_5 ($j_2 = \{v_5\}$). From the notion of valuing as a deliberatively consequential process, from j_1 we can infer that $v_3 \succ v_5$, whereas from j_2 we can infer that $v_5 \succ v_3$. As in the JC method, when a conflict is detected, we assume that the initial value-option matrix VO^i was inaccurate and update it. In particular, we set the cell of VO^i corresponding to the value that is part of the conflict but was not mentioned in the provided justification to 0. From our example, the method would set $VO^i(v_5, o_1)$ and $VO^i(v_3, o_2)$ to 0. Once the VO^i matrix is updated for all the justifications \times options conflicts, we compute the value ranking R_{JO}^i . Algorithm 4 illustrates this procedure.

Algorithm 4: Method JO

```

Input:  $J^i, VO^i, C^i, V$ 
Output:  $R_{JO}^i$ 
1  $VO_{JO}^i \leftarrow VO^i$ ; /* Temporary copy, we need information from the
   original  $VO^i$  in the next loops */
2 for  $j_a \in J^i : j_a \neq \emptyset$  do
3   for  $j_b \in J^i \setminus \{j_a\}$  do
4      $V_\alpha = V \setminus \{v : v \in j_a\} : VO^i(v, o_a) == 1$ ; /* Values supporting  $o_a$  in
        $VO^i$ , except values in  $j_a$  */
5     for  $v_x \in V_\alpha$  do
6       if  $v_x \in j_b$  then
7         for  $v_y \in j_a$  do
8            $V_\beta = V \setminus \{v : v \in j_b\} : VO^i(v, o_b) == 1$ ; /* Values supporting
               $o_b$  in  $VO^i$ , except values in  $j_b$  */
9           if  $v_y \in V_\beta$  then
10             $VO_{JO}^i(v_x, o_a) = 0$ ;
11          end
12        end
13      end
14 end
15  $VO^i \leftarrow VO_{JO}^i$ ;
16  $U^i = VO^i \times C^i$ ;
17  $R_{JO}^i = \text{rank}(U^i)$ ;

```

6.4 EXPERIMENTAL SETTING

We describe the experiments we perform to evaluate the five proposed methods. In the context of the Energy PVE described in Section 6.2.1, we consider a value v as relevant for an option o if at least t justifications (in our case, we set $t = 20$) among all participants were annotated with v for o . The resulting VO matrix (as described in Section 6.2.2) is shown in Table 6.4. We use this as a starting point for applying the methods described in Section 6.3.

We analyze each method (C , J , TB , JC , and JO) individually, and a sequential combination of the proposed methods in the following order: $JO \Rightarrow JC \Rightarrow TB$. We choose

Table 6.4: Value-option matrix (VO) for the Energy PVE.

		Options					
		o_1	o_2	o_3	o_4	o_5	o_6
Values	v_1	1	1	1	1	1	1
	v_2	1	1	0	1	1	1
	v_3	1	1	1	0	0	0
	v_4	1	1	1	0	0	1
	v_5	1	1	0	0	1	0

this sequential combination for two reasons: (1) the method TB should be executed last because it does not impact the VO^i matrix directly and thus would not affect the subsequent methods, and (2) we start with JC because it addresses conflicts within the same participant (which happens more frequently), and then continue with JO (less frequent). To combine these methods sequentially we use the ranking resulting from JO as input for JC , and the ranking resulting from JC as input for TB . Finally, for the individual analysis of the methods TB and JC , that require a previously estimated ranking, we start with the ranking estimated from choices alone (method C). We evaluate these methods based on the resulting value preferences rankings, which we refer to as R_C , R_J , R_{TB} , R_{JC} , R_{JO} , and R_{comb} (where R_{comb} is the result of the sequential combination $JO \Rightarrow JC \Rightarrow TB$).

6

6.4.1 EVALUATION PROCEDURE

Two evaluators, with previous knowledge of values and this specific PVE, were asked to independently judge the value preferences of a subset of participants based on their choices C^i and the provided textual justifications (from which J^i was annotated). We did not describe our value preference estimation methods to the evaluators. The evaluators were sequentially presented with a participant's choices and justifications, proposed pairs of values (e.g., v_1 and v_2), and asked to compare the two values with the following options: (1) $v_1 > v_2$; (2) $v_1 < v_2$; (3) $v_1 \sim v_2$; or (4) "I do not know", if they believe there is not enough information to make a proper comparison. This comparison was repeated up to four times per selected participant, with the intent of collecting sufficient information about a participant while increasing the number of analyzed participants.

The values to be compared were randomly selected from a set of value rankings that showed divergence across the methods. Our goal with this procedure is to assess the extent to which the proposed methods estimate value preferences similarly to the human evaluators. Within the application context illustrated in Figure 6.1, we expect that as the methods' rankings mirror human intuition, they might provide meaningful feedback to participants in a participatory system.

6.5 RESULTS AND DISCUSSION

When comparing the five proposed value estimation methods, we aim to answer two questions: (1) How well can each method estimate value preferences compared to humans? (2) How does the estimation of value preferences differ among the methods proposed?

The evaluators performed 1047 comparisons. We discard the responses indicating that there was not enough information to judge values preference ("I do not know"), reducing

the analyzed set to 766 total responses by either one of the evaluators. Figures 6.4a and 6.4b present the performance of each method in terms of matching each evaluator’s responses. These comparisons overlapped 269 times (i.e., the annotators performed the same comparisons). Considering this subset of overlapping comparisons, we find an agreement in 122 (45.35%) and disagreement in 147 (54.65%) comparisons, resulting in a Kappa score of 0.247, which is considered a fair agreement [165]. To mitigate the effect of individual biases, in the remainder of the analysis, we focus on the pairwise comparisons that evaluators agreed on, as presented in Figure 6.4c.

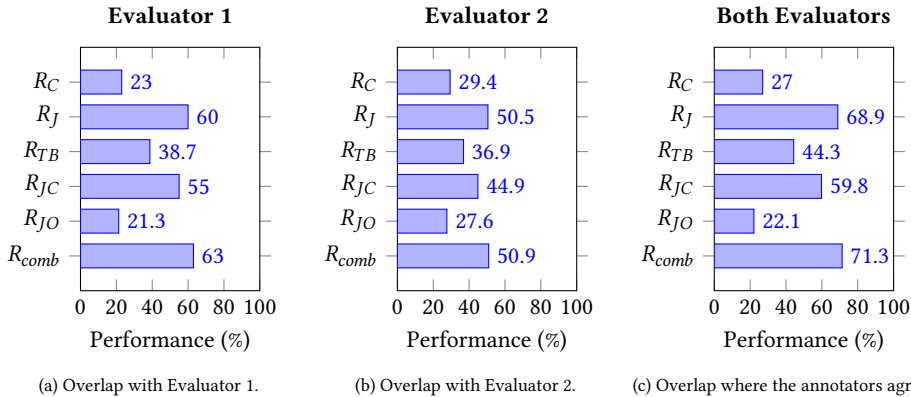


Figure 6.4: Value preferences estimation methods performance, measured as overlap with the evaluators’ answers.

As Figure 6.4 displays, the rankings R_J , R_{JC} , and R_{comb} provide the best performance in terms of human-like value estimation. When compared to R_C , the combined method R_{comb} estimated value preferences 2.64 times more similarly to humans (considering the subset where evaluators agreed). Further, we observe that R_J and R_{JC} also perform better than R_C . The only exception in terms of performance is R_{JO} , which performs slightly worse than R_C . These findings show that combining choices and justifications in estimating value preferences can significantly increase the degree to which an automated method can estimate value preferences similarly to humans, with respect to using only choices.

Finally, we notice that the performance of R_J is similar to the performance of R_{comb} . This is to be expected, as R_{comb} prioritizes justifications over choices, and R_J only employs justifications to estimate value preferences. The visibly better performance of R_J with respect to R_C further motivates the need to consider textual justifications to estimate value preferences that are consistent with human evaluation. With our dataset, combining choices and justifications led to slightly better results than employing just the justifications. Further experiments with other data are needed to confirm this observation.

COMPARATIVE ANALYSIS

For each method, we average the value preference rankings (that is, the position that the values have in the ranking that results after applying the method). We indicate with $>$ the values that have significantly different average rankings ($p \leq 0.05$) and with \geq the values that do not have significantly different averages. The following are the resulting average rankings per each different method:

- $R_C: v_1 \succ v_2 \succ v_4 \succ v_5 \succeq v_3$
- $R_{JC}: v_1 \succ v_2 \succ v_5 \succ v_3 \succ v_4$
- $R_J: v_3 \succ v_1 \succeq v_2 \succ v_5 \succeq v_4$
- $R_{JO}: v_1 \succ v_2 \succ v_4 \succ v_5 \succeq v_3$
- $R_{TB}: v_1 \succ v_2 \succ v_4 \succ v_5 \succ v_3$
- $R_{comb}: v_1 \succeq v_2 \succ v_5 \succeq v_3 \succeq v_4$

Method C ranked the value v_1 as the most important for all individuals, regardless of their choices, due to the characteristics of the value option-matrix (VO) in Table 6.4, which considers v_1 relevant for all choice options. As we attribute the minimum ordinal ranking for the values in case of ties (Def. 2), any choices would lead to R_C^i with v_1 as (one of) the most important value(s), except for method J which does not consider choices.

Let R_C be a baseline for comparison. Figure 6.5 indicates how many positions the final ranking changed across values (we do not consider method J since it did not use R_C as baseline). For example, consider two rankings $R_1 : v_1 \succ v_2 \succ v_3 \succ v_4 \succ v_5$ and $R_2 : v_2 \succ v_3 \succ v_1 \succ v_4 \succ v_5$. We consider four position changes from R_1 to R_2 : v_1 changed from the first to the third position (two changes), v_2 changed from the second to the first position (one change), and v_3 changed from the third to the second position (one change).

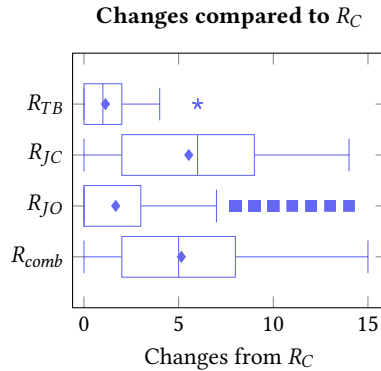


Figure 6.5: Average changes in the value rankings when compared to R_C .

Rankings R_{TB} and R_{JO} barely deviate from the average R_C . Instead, R_{JC} and the combined approach R_{comb} show significant deviation from R_C , indicating a larger difference at an individual value preferences level. The large deviation and the good performance (see Figure 6.4) of these two methods suggest that they estimate individually tailored value preferences that are in line with human intuition.

6.6 CONCLUSION AND FUTURE DIRECTIONS

We propose and compare methods for an AI agent to estimate the value preferences of individuals from one's choices and value-laden justifications, with the goal of generating an ordered value ranking within the analyzed context. We aim to improve the estimation of value preferences by prioritizing value preferences estimated from justifications over value preferences estimated from choices alone. We test our methods in the context of a large-scale survey on energy transition. Through a human evaluation, we show

that incorporating justifications to deal with conflicts in value preferences improves the performance of value estimation by more than two times (in terms of similarity to human evaluators' value estimation) and yields preferences that are more individually tailored.

In future experiments, participants themselves could provide direct feedback to the AI agent, instead of relying on external evaluators. Further, Natural Language Processing algorithms (such as the ones we test in Part II) could be used to scale up experiments by automatically identifying the values supporting the justifications. Finally, we suggest exploring other approaches to associate values with choice options beyond a binary matrix, since values can have different ethical impacts in different contexts.

Our work has the potential to contribute to value alignment between AI and humans in a hybrid participatory system. The estimated individual value preferences can be aggregated at a societal level [168, 169] with the intent of providing policy-makers with an overview of the value preferences of a population. Further, value preferences can serve as a starting point for the operationalization of values, e.g., for the synthesis of value-aligned normative systems [205, 271], as a foundation for international regulatory systems [27], or to formulate ethical principles through a combination of machine learning and logic [153].

IV

CONCLUSIONS

7

CLOSING THE LOOP WITH A HYBRID INTELLIGENCE APPROACH

Value inference cannot be performed based on computational methods alone. In addition to employing computational methods, AI agents ought to foster self-reflection and deliberation among stakeholders. To this end, we propose a Hybrid Intelligence approach that connects two value inference processes—value classification and value preferences estimation—and focuses on promoting self-reflection. This chapter extends Chapter 6, where we propose methods to estimate value preferences based on individuals’ choices and the natural language justifications they provide for their choices in a participatory system setting, and prioritize the values that support the justifications in case of conflicts between the choices and justifications. Here, we investigate the conflicts between the value preferences estimated from participants’ choices and those estimated from their justifications. We propose a strategy to guide the interaction between AI agents and participants to disambiguate these conflicts, by asking the participants to validate the correctness of the value labels predicted for their justifications by a natural language processing algorithm, in an active learning fashion. We compare our method to state-of-the-art active learning methods and find no significant differences. We conclude the chapter with a reflection on the lessons learned by testing our disambiguation strategy.

7.1 INTRODUCTION

Parts I, II, and III contribute to the three value inference processes—identification, classification, and estimation—individually. However, as motivated in Section 1.2, Hybrid Intelligence (HI) [7] approaches are necessary to foster self-reflection and deliberation among stakeholders by closing the loop depicted in Figure 1.2. To do this, in this chapter, we propose an HI approach that connects value classification and value preferences estimation and intends to foster self-reflection in the involved stakeholders (here, our focus is on self-reflection rather than deliberation). This chapter extends the methods and experimental setup described in Chapter 6.

As introduced in Chapter 6, we envision value preferences estimation performed in a participatory sociotechnical system, where AI agents estimate participants’ value preferences on a decision-making subject. In this vision, AI agents, supported by natural language processing (NLP) techniques, interpret the natural language justifications provided by the participants in support of their choices and combine the information contained in choices and justifications to estimate value preferences. In Chapter 6, we propose and compare five methods for estimating value preferences from participants’ choices and justifications. We find that, in case of conflicts between the value preferences estimated from the choices and the value preferences estimated from the justifications, prioritizing the information contained in the justifications results in value preferences that are most aligned with the ones estimated by human annotators.

Nevertheless, the detected choice-justification conflicts should be addressed. Such conflicts may be caused by (1) mistakes in the value inference process (e.g., misclassification of the values supporting the participants’ justifications by an NLP model), or (2) genuine inconsistencies between the participants’ choices and justifications, e.g., due to participants having different assumptions regarding values that drive a choice, or due to the value-action gap [93]. In both cases, addressing the inconsistencies can be beneficial. If the inconsistency is caused by a mistake in the automatic value estimation process, the involved participant should be asked to resolve the mistake, e.g., by correcting a mistake of the NLP model. In case the interpretation is confirmed to be correct and the inconsistency is accurately

detected, the participant can be guided through a process of self-reflection [172, 183] and offered the chance to change their choices or provide additional justifications. We refer to the endeavor of addressing the inconsistencies between the values detected in choices and justifications as *value preferences disambiguation*. We envision a hybrid participatory system (Figure 7.1) where AI agents estimate and disambiguate value preferences to assist in decision-making while fostering self-reflection in participants.

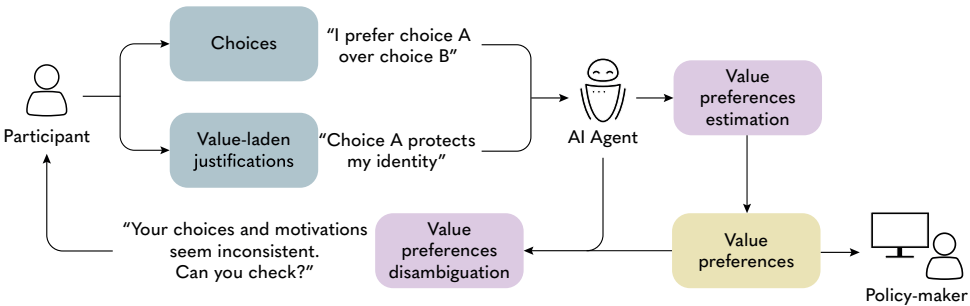


Figure 7.1: A hybrid participatory system where human participants make choices and motivate those choices, and AI agents estimate participants' value preferences through a disambiguation strategy, to assist in decision-making.

In participatory systems, not all participants may be available to take part in such interactions, and the required additional effort may dissuade participants from engaging [275]. In practice, a step-by-step approach may be preferred, as it could be impractical to wait until all participants have been consulted before addressing potential algorithmic mistakes. Inspired by traditional Active Learning (AL) [272] strategies, we propose a *disambiguation strategy* that guides the interactions between AI agents and participants, following the rationale that, by addressing the most informative participants first, the quality of value preferences estimation should rapidly improve for many participants. Precisely, the strategy iteratively selects the participants whose value preferences estimated solely from their choices are most different from the value preferences estimated solely from their justifications. We test our strategy by retrieving the correct interpretation of the justifications provided by the selected participants (i.e., the correct values that support their justifications) to iteratively improve the NLP model tasked to predict the values that support the participants' justifications, which are in turn used to estimate their value preferences.

We evaluate the strategy in an active learning setting with the data from the PVE survey on energy transition and compare it to traditional NLP AL strategies. We show that our method leads to comparable results to the tested baselines, both in NLP performance and value preferences estimation. We discuss these results and elaborate on future directions.

This chapter is organized as follows. Section 7.2 describes the value disambiguation method. Section 7.3 describes our experimental setup to evaluate the disambiguation method and Section 7.4 presents our results. Finally, Section 7.5 concludes the chapter. Appendix D provides additional details on our experimental setup. The code will be made public upon publication of the paper.

7.2 METHOD

The disambiguation strategy is intended to drive the interactions between AI agents and participants by addressing the detected inconsistencies between participants' choices and justifications, to improve the value estimation process. Inspired by popular AL strategies (Section 2.3), the strategy iteratively targets the participants deemed to be most informative. We associate informativeness with the inconsistency between a participant's choices and justifications, assuming that the largest inconsistencies may unveil the biggest mistakes in the value estimation process. By addressing the most informative participants first, we aim to rapidly improve the quality of value preferences estimation for all participants.

Figure 7.2 provides an overview of the proposed strategy. We consider a hybrid participatory setting where the AI agents are equipped with an NLP model tasked to predict the set of values mentioned in each participant's justifications. Then, value preferences are estimated based on the participants' choices and the value labels that are predicted as supporting each justification they provide. We propose that AI agents iteratively interact with the participants with the largest detected inconsistencies between the value preferences estimated from their choices alone and the value preferences estimated from their justifications provided in support of those choices. In our method, the AI agents interact by asking whether the provided justifications have been correctly interpreted (i.e., if the predicted value labels are correct). Other interaction strategies can be implemented (e.g., querying the participants on whether the preference between two values v_a and v_b have been correctly estimated), which we discuss as future work (Section 7.5).

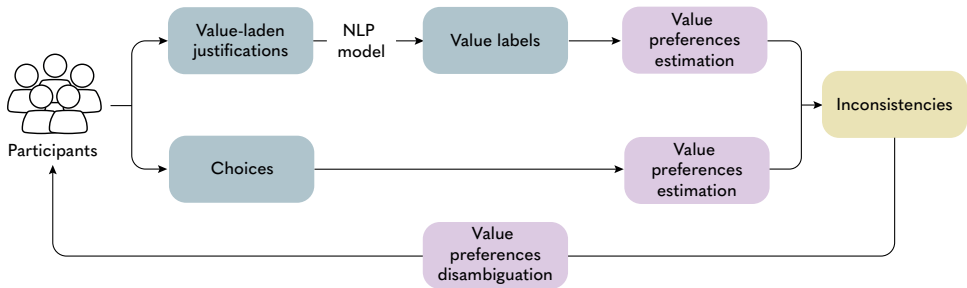


Figure 7.2: Overview of the proposed disambiguation strategy, guided by the detected inconsistencies between value preferences estimated from participants' choices and the value preferences estimated from the value labels classified by an NLP model in the justifications.

Our setting is akin to an AL setting where value labels are iteratively retrieved to train a value classification NLP model. The most informative participants are iteratively selected by the strategy and asked to provide the correct value labels on their justifications, in practice treating the participants themselves as oracles. At every iteration of the AL procedure, we use the current version of the NLP model to predict value labels on all the unlabeled justifications, and use the predicted labels to estimate the value preferences of the participants whose justifications are not yet labeled, with both methods C and method J . Then, for each participant, we calculate the distance between the value ranking estimated with method C and the value ranking estimated with method J . We use the Kemeny distance [123, 147] to measure the distance between rankings, as it accounts for

potential ties between values (see Def. 2 in Chapter 6). The Kemeny distance (d_K) between two value rankings (R_C^i, R_J^i) is defined as:

$$d_K(R_C^i, R_J^i) = \frac{1}{2} \sum_{v=1}^n \sum_{w=1}^n |x_{vw}^{(C)} - x_{vw}^{(J)}|,$$

where n is the number of objects (in our case, $n = 5$ is the number of values), and $x_{vw}^{(C)}$ is equal to 1 if value v is preferred to value w in ranking R_C^i , equal to -1 in the reverse case, and equal to 0 if the two values are equally preferred. Finally, we choose as the next batch the p participants with the largest Kemeny distance between the value rankings estimated with method C and method J , and retrieve value labels for the justifications they provided. The NLP model is trained with the newly collected annotated justifications, and the AL strategy is re-iterated with the updated version of the NLP model.

7.3 EXPERIMENTAL SETTING

We test the disambiguation strategy as a sampling strategy in an AL setting, where the justifications' annotations are iteratively retrieved and used to train an NLP model tasked to classify the values that support each justification. We treat value classification as a multi-label classification task, where each justification is annotated with zero or more value labels. Since not all provided justifications ought to be value-laden, a justification may have zero labels in case none of the values in Table 6.2 is deemed relevant. Multi-label BERT [74] has been shown to produce state-of-the-art performances on similar value classification tasks [133, 150, 178, 244]. Thus, we use RobBERT [72], a RoBERTa variant [186] which is considered state-of-the-art for text in Dutch. We have also translated the data to English and tested equivalent English models, obtaining similar performances. We show results for the original Dutch data in the main body of the paper. Hyperparameters tuning and comparison with the English models can be found in Appendix D.1.

We employ the annotations described in Section 6.2.1 to simulate the AL procedure. At every iteration of the AL procedure, we have a set of labeled justifications (whose labels have been retrieved and that are used to train the NLP model), a set of unlabeled justifications (whose labels can be retrieved if selected by the sampling strategy), and a set of test justifications (that are only used for evaluation). Analogously, we have a set of labeled participants, unlabeled participants, and test participants, who have provided the justifications in the corresponding sets. At every iteration, the model is trained with the labeled justifications, and then used to predict labels on the unlabeled justifications. With the predicted labels, the value preferences of the unlabeled participants are estimated. The disambiguation strategy is then used to select the p unlabeled participants with the most inconsistent value preferences estimated from choices and justifications alone. The p participants are added to the set of labeled participants, and the labels of the justifications provided by the participants are retrieved and the justifications added to the set of labeled justifications.

As is common in AL settings, we warm up the NLP model by initializing the set of labeled participants with 10% of the available participants, and the set of labeled justifications with the justifications provided by those participants. At each iteration, we train the NLP model with the labeled justifications. We use the trained model to predict labels

on the test justifications and use these labels to (1) estimate the value preferences of the test participants with the best-performing value estimation method, and (2) evaluate the performance of the NLP model. We then use the disambiguation strategy to select $p = 39$ participants, to add 5% of the available participants to the labeled participants set at each iteration. We iterate the procedure for 5 iteration steps and repeat it in a 10-fold cross-validation.

7.3.1 EVALUATION PROCEDURE

We evaluate how the disambiguation strategy drives the NLP model performance and the estimation of value preferences, comparing it to the respective topline and baselines.

We perform 10-fold cross-validation to measure the performance of the NLP model trained on all available data and use the result as an NLP topline during the AL procedure. We use a model trained on all data to predict labels on all the justifications and use the predicted labels to estimate all participants' value preferences with the best-performing value estimation method. We treat the resulting value rankings as value preferences topline during the AL procedure, as they represent the best possible value rankings that can be estimated with the mistakes introduced by using the labels predicted by an NLP model instead of the ground truth annotations. At every iteration of the AL procedure, we compare the NLP performance on the test set to the NLP topline, and the estimated value preferences of the test participants to the value preferences topline. For the NLP performance, we report the F_1 -score. As the label distribution is balanced (Table 6.3), there is a small difference between micro and macro F_1 -scores—we report the micro F_1 -score as it accounts for the label distribution. Finally, for the value estimation performance, we report the Kemeny distance between the estimated value preferences of the test participants and the corresponding value preferences topline.

We compare the results to two baselines. First, we employ the uncertainty sampling strategy (Section 2.3) to select 5% of justifications ($j = 145$ justifications) at each iteration, similarly to the evaluated disambiguation strategy. This strategy is solely driven by justification informativeness, ignoring the connection between the justifications and their authors. We choose this strategy as a baseline since traditional NLP AL strategies are solely driven by information about the NLP task, as described in Section 2.3. Second, we employ a random baseline, where at each iteration 5% random participants ($p = 39$ participants, similarly to the proposed disambiguation strategy) and their justifications are added to the labeled set. With both our proposed strategy and the baselines, we plot the trend of the NLP and value estimation performances throughout the progressive iterations. We compare them with each other and to the corresponding toplines.

7.4 RESULTS AND DISCUSSION

We present and discuss the evaluation of the disambiguation strategy. First, we report the results of the toplines. The NLP topline resulted in an average micro F_1 -score of 0.64, which is slightly lower than similar value classification tasks [133, 178], likely due to the smaller dataset size. For the value preferences topline, we use the predicted justification labels to estimate value rankings through the R_{comb} method (the best-performing value estimation method). The value preferences topline resulted in an average Kemeny distance

of 1.88 (with 2.88 standard deviation) from the value rankings estimated with the combined method (with the resulting value preferences ranking R_{comb}) by using the ground truth annotations on the justifications. We use these toplines to measure the trend of the results throughout the AL iterations.

We report the results of our experiments in Figures 7.3 and 7.4. In all experiments, at every iteration we used the tested strategy to select 5% of the data to be added to the set of labeled data. However, since different participants provided different numbers of justifications, selecting the justifications provided by 5% of the participants may not correspond to 5% of all available justifications. In Figures 7.3 and 7.4, we show on the x-axis the number of justifications used for training the NLP model at the corresponding iteration. While that corresponds to exactly 5% increments in the case of the uncertainty strategy (which selects 5% of the justifications at every iteration), it is not the case for the random and disambiguation strategies (which selects 5% of the participants at every iteration).

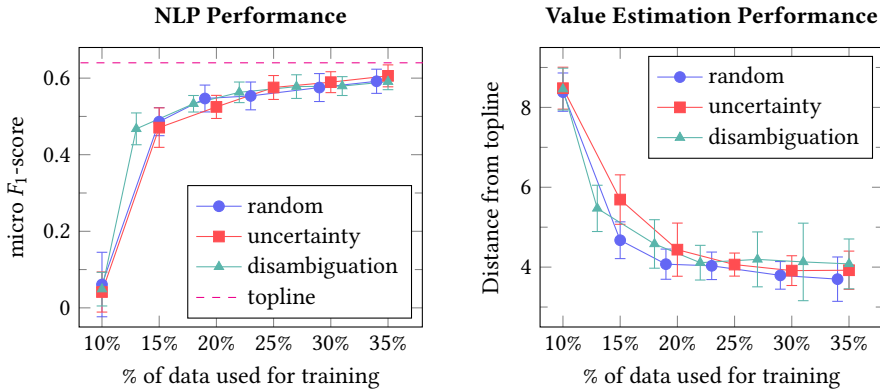


Figure 7.3: NLP performance (micro F_1 -score), compared to the NLP topline (dashed horizontal line). Figure 7.4: Value estimation performance (average Kemeny distance from the value preferences topline).

The random strategy has a varying step size that roughly averages to 5%, as expected by a strategy that randomly selects participants. Instead, the step size of the disambiguation strategy is consistently smaller than the other two (for this strategy we plot six steps, as opposed to five for the other strategies), meaning that at every iteration the strategy chooses participants who have provided fewer justifications than the average participant. This empirically matches the intuition behind the strategy—participants who have provided few justifications have a R_J (value ranking calculated from justifications alone) that is mostly composed of ties between values. Such undetermined R_J have a large distance from the corresponding R_C , which instead considers all the choices provided by the participants.

The NLP performances of the model trained with the disambiguation strategy and with the two baseline strategies (uncertainty and random) are illustrated in Figure 7.3. No significant difference between the compared methods is visible, as all three strategies lead to a rapid improvement in performance that approaches the NLP topline when roughly 30% of the available justifications are used for training. In line with these results, experimental findings [84] show that there is no single AL strategy that outperforms all others across different datasets, and, in some cases, no significant difference is observable with the

random strategy. Ultimately, these results demonstrate that the proposed disambiguation strategy, despite being guided by the downstream task of estimating value preferences, does not significantly affect the NLP performance.

Figure 7.4 presents the value estimation performance of the three compared strategies, measured as the average Kemeny distance between the value preferences estimated with the labels predicted by the current iteration of the NLP model and the value preferences topline. First, we remind that the topline has been calculated with the label predictions resulting from a model trained with all available data. However, the training process with the tested strategies is performed as 10-fold validation, thus a different subset of the dataset is used for training in each fold. Consequently, we do not expect the Kemeny distance to approach zero, as different data was used during the training process (thus resulting in different individual value preferences). Still, the topline reference allows us to compare the value estimation performance trend of the three strategies.

We observe that the value estimation performance trend is similar for all three strategies, leading to a rapid decrease in distance from the topline that mirrors the rapid improvement in the F_1 -score. While the results are comparable when 20% or more justifications are used for training, the results with ~15% of the training data show small differences—while the F_1 -score performances at this stage are almost identical, there is a small difference in value preferences estimations. In particular, the uncertainty strategy (which ignores the link between participants and justifications) is worse than the other two tested strategies, which motivates the usage of a participant-driven strategy instead of a justification-driven strategy. However, the differences are not sufficiently large to draw a definitive conclusion.

Overall, we notice no significant difference between the proposed strategy and the baselines. We discuss two possible reasons. First, the NLP performance is the biggest driver of value estimation performance—in practice, the more justifications are correctly labeled, the more accurate the value estimation is. With the analyzed data and the relatively small dimension of the dataset, no significant difference is noticeable between the tested strategies in NLP performance, including the random strategy, resulting in a similar trend in the value estimation performance. Second, the distance between R_J and R_C may not be the best indicator for the informativeness of a participant. Considering the annotations from Section 6.2.1, there is a distance of 8.0 (with 3.5 standard deviation) between R_J and R_C estimated for the same participants. Thus, large distances between the two rankings may not be particularly informative in this dataset. However, we believe that a strategy driven by the downstream application may be particularly useful in similar settings, as we elaborate as Future Work.

7.5 CONCLUSION AND FUTURE DIRECTIONS

We propose a disambiguation strategy to drive the interactions between AI agents and participants, with the intent of improving the value estimation performance. Our strategy prioritizes the interaction with the participants whose value preferences estimated from choices alone are most different from the value preferences estimated from justifications alone, following the rationale that such participants would be the most informative for rapidly adjusting and improving the value estimation process. However, our results show no significant difference with compared baseline strategies, including a strategy where interactions with participants are randomly determined.

Despite the inconclusive results, we believe that our proposed disambiguation strategy opens novel research avenues. Such a hybrid approach to an interaction strategy for value preferences disambiguation can help iteratively address algorithmic mistakes while fostering self-reflection in participants [172]. A strategy driven by the downstream task of value preferences estimation helps in integrating the different components involved in the value estimation process (value label classification and aggregation of one's choices and justifications). Further, different disambiguation approaches could be tested. For instance, the strategy could target the participants with the most different choice distribution from the average, or with the largest amount of ties in their estimated value rankings.

Our approach is intended to be used in an online setting in a participatory system. The proposed strategy addresses inconsistencies between choices and justifications by retrieving the correct value label that supports the justifications. In future experiments, participants themselves could provide direct feedback to the AI agent, instead of relying on external evaluators. Additionally, participants may be offered the option to adjust the estimated value preferences directly, instead of being limited to providing the correct value label supporting their justifications. Machine learning methods could then be employed for value estimation, learning directly from the feedback provided by the participants.

This work represents the first attempt to close the value inference loop through an HI strategy. Other works could investigate the combination of different value inference processes. For instance, future research could investigate the effectiveness of zero-shot classification of pre-trained language models when predicting a novel value label, in case the identified value list is updated during a participatory process. Another avenue is to foster deliberation among participants by guiding them in the discussion of detected value preferences inconsistencies.

8

CONTRIBUTIONS AND FUTURE WORK

Values are central to the construction of ethical societies where human and AI agents co-exist. This thesis contributes to *value inference*, the endeavor of identifying relevant values and estimating individuals' value preferences. Value inference is a prerequisite for value alignment, as AI agents ought to infer human values before aligning their behavior with those values. In **Chapter 1**, we break down the value inference challenge into three processes (value identification, value classification, and value preferences estimation). We discuss how Natural Language Processing (NLP) is a key component of value inference, as natural language is often the means through which humans reveal their value preferences. Then, we identify two core challenges for value inference—namely, value preferences are (1) dependent on context, and (2) often implicit even to the human holding those values. To this end, this thesis investigates how AI agents can identify, classify, and estimate context-specific values. Then, we motivate how a Hybrid Intelligence (HI) approach is key to guiding humans in reflecting on their value systems, and provide the first example of a HI approach to involve stakeholders in the value inference procedure.

This Chapter is structured as follows. Section 8.1 reviews the contributions of this thesis concerning the individual value inference processes and identifies future work avenues for these processes. In Section 8.2, we examine the need for HI approaches to value inference. Section 8.3 presents challenges and opportunities that relate to value inference as a whole. Finally, Section 8.4 discusses the limitations of our experiments, and Section 8.5 reflects on the societal implications of our research.

8.1 THE VALUE INFERENCE FRAMEWORK

Our vision of the value inference challenge is divided into three processes: value identification, value classification, and value preferences estimation. For each of these processes, we discuss our contributions and identify future work avenues.

8.1.1 VALUE IDENTIFICATION

Not all values are relevant to all contexts and the interpretation of a value may change across contexts. To this end, **Chapter 3** introduces Axies, a HI method for identifying and defining the values that are relevant to a decision context. We employed Axies to identify context-specific values with two groups of annotators in two decision contexts and evaluated its results with a study involving 80 crowd workers. Our experiments show that Axies yields context-specific values that are comprehensible to laypeople and consistent across different groups of annotators. Then, we compared the values yielded by Axies to the Schwartz set of basic values [268]. We found that only a few Schwartz values are related to Axies values (i.e., only the Schwartz values that are relevant to the context) and Schwartz values with a distinct correspondence are frequently associated with multiple Axies values, which provide a more detailed description in the given context. Finally, our evaluation shows that laypeople annotate Axies values with a higher agreement than Schwartz values, showing the suitability of context-specific values for practical applications.

Basic and context-specific values are complementary. Basic values help explain human behavior across contexts. However, context-specific values are necessary for concrete applications, such as the value-sensitive design of a chicken husbandry system or to support energy-related policy-making [137, 303]. Furthermore, our experiments show

clear correspondences between context-specific and basic values. Understanding how context-specific values vary and how they relate to basic values across contexts is yet to be explored. To this end, Axioms can be employed to create a database of values that are linked with contexts. AI agents could automatically select from this database the set of values relevant to the decision context. However, this database would only represent a starting point for value identification, as it may not be shared by all relevant stakeholders. AI agents need to be able to dynamically adjust the set and the interpretation of the context-specific values. HI approaches must be devised to actively engage relevant stakeholders in the value identification process, e.g., by asking them whether the set of values is representative of the full breadth of their preferences, or by asking them to compare and discuss their interpretation of the values with other stakeholders (thus fostering self-reflection and deliberation).

8.1.2 VALUE CLASSIFICATION

Since natural language is our preferred means for communicating value preferences [261, 266], AI agents need to be equipped with NLP algorithms that can classify value-laden content in natural language. Furthermore, in sociotechnical systems (STSs), AI agents interact with stakeholders across a multitude of decision contexts, requiring the ability to classify value-laden content across different contexts. To this end, in **Chapter 4** we perform an evaluation of value classifiers across contexts, and in **Chapter 5** we investigate the difference in how the language models represent the different value concepts across contexts. We perform our experiments with a fixed set of basic values to focus the investigation on the influence of context. Our experiments show that value classifiers can generalize to different contexts, recognizing value-laden words and expressions that are shared across contexts. However, on closer inspection, we find that language models also learn to recognize context-specific expressions that do not generalize across contexts. These context-specific value representations may introduce biases that could lead AI agents to critical mistakes when generalizing to novel contexts.

The state-of-the-art language models have demonstrated impressive zero-shot generalization to previously unseen tasks and contexts [34, 240, 313]. However, value classification is subjective by nature, since different individuals may have a different interpretation of (1) the meaning of a value, or (2) what are the value(s) that support a natural language statement. To reflect this diversity, NLP-based AI agents need to take a *perspectivist* approach [47, 301]. That is, they need to learn from a diversity of annotations rather than from one consensus annotation (often obtained through majority voting [47, 213]). Employing consensus labels to train and evaluate language models could lead to the exclusion of minority opinions, as we explore in Section 4.3.5. In our experimental setting, we use the labels obtained through majority voting to train and evaluate the value classifiers. However, we show that in the majority of mistakes that the classifiers make (i.e. the cases in which the model's prediction differs from the majority label), there is at least one (minority) annotator that agrees with the model's prediction. Thus, language models ought to be instead trained to represent the societal distribution of value interpretations. This challenge is twofold. On the one hand, language models ought to learn from a distribution of opinions rather than from a consensus label [301]. On the other hand, it must be ensured that the distribution of opinions is collected fairly and is representative of the relevant stakeholders, e.g., involving

enough and heterogeneous annotators [21, 46].

8.1.3 VALUE ESTIMATION

Estimating stakeholders' value preferences is the last step of our value inference approach. Knowing how individuals prioritize relevant values in a decision context is necessary for AI agents to align their behavior with stakeholders' values. To this end, in **Chapter 6**, we propose and compare different approaches to estimating stakeholders' value preferences based on the choices and textual justifications provided in a survey about energy transition. We follow a philosophical account [266] suggesting that, in case of conflicts between the values that support the choices and those that support the justifications, the values that support the justifications should be prioritized. Our experiments show that this method produces results that are more aligned with the value preferences that humans estimate compared to the value preferences estimated from choices or justifications alone.

The estimation of value preferences has typically been approached through questionnaires, criticized for not including contextual factors in the estimation process (see Sections 1.1 and 2.1). Our approach takes a step forward by estimating value preferences based on the choices and natural language justifications provided within a decision context. However, similar to value questionnaires, our approach offers a static glimpse into stakeholders' value preferences, based on one-time choices and justifications. The advancement in state-of-the-art language models has demonstrated that AI agents can be equipped with the ability to converse about high-level concepts such as emotions and values [51, 52, 190]. Thus, we envision an *interactive* value preferences estimation method (which can be incorporated within our HI approach in Chapter 7), where AI agents ask for feedback and adjust the estimated value preferences online. This approach would provide several advantages. First, stakeholders would be able to describe the dependency of their value preferences on context, indicating which contextual element lead them to a change in value preferences. Second, stakeholders would be able to indicate the extent to which AI agents should prioritize the value preferences estimated from their choices over the value preferences estimated from their justifications, as that is dependent on individuals' preferences and decision context. Third, AI agents would be able to ask for additional justifications to disambiguate undefined value preferences. We further elaborate on the idea of interactive value preferences estimation in the next section within our hybrid value inference vision.

8.2 HYBRID VALUE INFERENCE

As motivated in Section 1.2, AI agents ought to interact with stakeholders to foster self-reflection and deliberation. To effectively do so, agents ought to situate their interaction strategy in concrete observations, e.g., by asking questions that spur from the interpretation of stakeholders' observed behavior. To this end, in **Chapter 7** we propose a strategy for guiding the interaction between AI agents and stakeholders. This chapter builds upon Chapter 6, where we introduce methods for estimating value preferences based on individuals' choices and the natural language justifications accompanying those choices. In instances of conflicts between choices and justifications, we prioritize values aligned with the justifications. Here, we delve into conflicts arising between value preferences

derived from individuals' choices and those inferred from their justifications. We propose a strategy facilitating AI agent-stakeholder interaction to resolve these conflicts. In an active learning approach, stakeholders validate the correctness of value labels classified in their justifications by an NLP algorithm. Our method does not reveal significant differences when compared to state-of-the-art active learning approaches. Yet, it introduces a new paradigm for guiding an active learning strategy to foster self-reflection in stakeholders.

Our strategy guides the interactions between AI agents and stakeholders. However, values are high-level and abstract motivations, and reasoning about them is difficult for humans [167, 237]. Thus, the design of such interactions is a challenging task. To the best of our knowledge, Chen et al. [52] propose the first and only attempt to value-aligned conversational agents by designing dialogues that aim at correcting potential misalignments between the estimated value preferences and the true stakeholders' preferences, which may vary over time or due to contextual factors. Similar approaches to conversational agents have been proposed to elicit other abstract concepts such as emotions and engagement [51, 190], or in sensitive applications such as the healthcare domain [166]. Furthermore, conversational agents have been used to moderate and foster deliberation among stakeholders [55, 112, 185, 295]. The design of the interactions between AI agents and stakeholders is critical for the success of value inference, as it promotes engagement and trust in the process [189, 197].

8.3 CHALLENGES AND OPPORTUNITIES

So far, we have addressed the challenges related to the individual aspects of value inference. Here, we introduce the computational and human-centered research challenges and opportunities associated with hybrid value inference as a whole. These challenges show that value inference is a cross-cutting topic that can contribute to and benefit from interdisciplinary research.

Behavior Observation The observation of value-laden human behavior constitutes a field of study on its own, as motivated in Section 1.1.1. We identify three main challenges. First, the sensing of human activity, including, among others, video and audio capture, biometric sensors, and location tracking [14, 289]. This ought to be performed while respecting stakeholders' privacy and under informed consent, as elaborated in Section 8.5. Second, the interpretation of sensory data to detect which sensorial input corresponds to which activity [2, 110]. This enables the construction of a map of stakeholders' behavior where a particular justification can be linked to a specific action. Lastly, the distinction between value-laden and non-value-laden behavior [311]. There is no unanimous agreement among individuals regarding which actions and justifications are driven by values (e.g., brushing teeth or explaining a train delay). AI agents ought to discern what constitutes value-laden behavior, with input from the concerned stakeholders.

Identifying Context Shifts Value systems have been recognized to be context-specific, as situational factors affect our priorities [44, 126, 159, 180]. This thesis investigates the impact of contextual factors and shifts on the value inference processes. Besides our work, the dependency on context has been investigated in engineering AI agents and

multiagent systems [214, 215, 293] and in NLP applications (see domain dependency in Section 2.3). However, further research is required to allow AI agents to *identify* relevant context shifts, i.e., recognizing a change in contextual factors that may lead to a change in value preferences. To this end, Pyatkin et al. [243] propose ClarifyDelphi, the first example of a conversational agent that learns to ask questions to elicit additional relevant context when judging the morality of a situation. The success of this approach shows that language models have an understanding of how context influences the perceived morality of a situation, and a similar approach can be used to recognize the factors that may influence stakeholders' value systems. This method can be extended by populating a knowledge graph [139] that connects contextual factors to value systems shifts. This graph would be populated through HI approaches similar to ClarifyDelphi, with clarifications on contextual influences that result from the interactions between AI agents and stakeholders.

Explainability We identify three connections between explainable AI (XAI) and value inference. First, we emphasize the importance of *interactive* explanations [31, 200, 257], as AI users find a single explanation insufficient [162]. Dialogue-based interactive explanations are a key research challenge for realizing the hybrid value inference framework we envision. Second, explanations are crucial for validating the value inference processes. We envision an AI system that provides explanations for each value inference process with the intent of improving the process itself. To this end, *actionable* explanations (i.e., explanations that humans or agents can act upon) constitute an essential research avenue [31, 146, 241]. Third, we point to the novel challenge of generating *value-based* explanations [319], i.e., natural language clarifications that expose an underlying value reasoning. Such explanations are necessary for communicating the results of value inference to stakeholders.

Bias, Fairness, and Quality of Data Value inference is crucial for sensitive AI applications, e.g., to make life-changing decisions in a healthcare STS. Therefore, it is essential to guarantee that these decisions do not reflect discriminatory behavior. This amounts to ensuring that the value inference processes are fair and free of bias [163, 194]. This is part of the broader challenge of ensuring the *quality* of the data employed by the value inference processes. To this end, strategies must be devised to *curate* (build, maintain, and evaluate) the datasets involved in value inference. For example, qualitative and quantitative relationships between value datasets can be modeled in a knowledge graph [139] to describe the links between the (context-specific) values inferred in the associated contexts. This is in line with the emerging trend in Data-Centric AI [285], which recommends a focus shift from the models to the underlying data used to train and evaluate models.

Resilience Value inference is sensitive to misbehavior, as humans may misreport or maliciously misguide their agents when providing feedback. We envision two related research challenges. On the one hand, we can consider how to deter manipulation, which is challenging because it calls for the detection of individual and collective misbehaviors [12]. This would require collaboration with social scientists and economists to design mechanisms for encouraging truthful reporting and feedback that prevent manipulation. On the other hand, we ought to analyze and empirically quantify the *resilience* of the value inference processes when coping with varying populations of misbehaving humans (e.g., by

investigating the robustness of the system [207, 269]). Importantly, given the compositional nature of the proposed value inference framework, resilience should be quantified both for individual processes and for the framework as a whole.

Verification and Validation The results of value inference need to be both verified (i.e., checking whether the processes operate as intended) and validated (i.e., checking whether the system operates to the satisfaction of the users) [30]. Both verification and validation can be performed via HI approaches as described in Section 1.2. Although value inference can be incrementally verified and validated throughout the lifecycle of an STS, it is necessary to define a moment in which the results are sufficiently satisfactory for being operationalized (e.g., to design policies). Identifying such *satisfaction criterion* is a significant research challenge. This investigation is akin to measuring saturation in qualitative analysis [264], which considers, among other, stakeholders' validation of each value inference process, time and effort required by stakeholders, and evolution of the results (e.g., by identifying asymptotic trends).

Responsible Autonomy Designing autonomous agents that align with their human users' values is an important step toward trustworthy AI [279, 280]. To this end, the value inference processes must be legitimate [33, 108]. The involvement of stakeholders in the hybrid value inference processes is key to legitimacy, as stakeholders ought to believe that the overall procedure is fair [225]. In particular, consent and dissent are important aspects for ensuring legitimacy [80, 280]. On the one hand, for value inference to be legitimate, the stakeholders must consent to the inference processes. In addition, there must be explicit dissent channels for the stakeholders to question the outcomes of the inference processes. On the other hand, value inference enables a broader understanding of consent, as advocated by Pitkin [235, 236], as not merely seeking a stakeholder's permission but evaluating whether the consented action aligns with the stakeholder's values. Although the concepts of consent and dissent are well-studied in the legal literature [17], computational modeling of these abstractions is an open challenge.

8.4 LIMITATIONS

We discuss three main limitations of our work.

First, the subjective nature of values affects the replicability of our experiments. Throughout the thesis, we employ humans to annotate and/or evaluate our methods. We attempt to mitigate the effect of subjectivity by employing different sets of annotators (Chapter 3) or a large set of evaluators (Chapters 3 and 5). However, we required the involved annotators to be fluent in English, and their demographic distribution (Appendices A.1.3 and C.2.3) is skewed towards Europe. These factors could lead to the perpetuation of Western values and biases [194] in our analyses. Additional experiments could verify whether annotators with different backgrounds would reach the same conclusions. Nevertheless, the inherent subjectivity of the task may impede replicability even with the same annotators. For this reason, additional evaluation metrics are required to evaluate the value inference processes, especially when used in HI fashion, such as stakeholders' satisfaction [324] or trust [196] in the system.

Second, we acknowledge limitations related to the background of the corpora we use, the Participatory Value Evaluations and the Moral Foundation Twitter Corpus (Section 2.4). Both corpora are composed of Western-centric data in Western languages (Dutch and English) and we employ language models pre-trained in the respective languages. Our proposed methods and experiments are effective under these conditions. However, the effectiveness with different datasets, e.g., datasets in morphologically richer languages or diverse cultures, remains to be investigated. Further, the scalability to longer text formats (e.g., news articles) is yet to be explored. However, our methods and evaluation procedures can be applied to larger and culturally diverse datasets as well.

Lastly, the data we use consists of discrete observations of a stakeholder's behavior, i.e., tweets and choices and justifications provided in a survey. This type of data offers limited insight into the context of the topic under discussion and the background of the stakeholder. For example, a language model may misclassify values due to the lack of background information about a mentioned event, or misinterpret slang expressions. In our HI vision, value inference is performed through interactions between AI agents and stakeholders. In this setting, conversations represent the main means through which values are expressed and inferred. Through conversations, stakeholders can provide additional clarifications behind their statements, and AI agents are offered more data to learn stakeholder-specific value expressions. This setting allows to perform a more accurate value inference but introduces additional challenges. Among others, AI agents ought to have mechanisms for adapting their language models to a specific stakeholder, or strategies for updating the inferred value systems based on new clarifications provided by the stakeholder.

8.5 SOCIETAL IMPLICATIONS

Value inference ought to be deployed with care. As discussed in Section 8.3, the value inference processes ought to be thoroughly validated and stakeholders ought to actively indicate their consent to them. The privacy of the involved stakeholders ought to be respected at all stages [91, 280]. Striking a balance between the benefits of value inference and protecting sensitive information is crucial. It necessitates robust regulatory frameworks and transparent data-handling practices [202]. For instance, regulations may determine that individual value systems can only be stored at the individual agent level, with only the aggregated value systems being shared with authorities or external parties. Finally, heightened awareness must be fostered among stakeholders. Thoughtful reflection and proactive measures are essential to uphold the privacy rights of individuals. To this end, our envisioned HI approach is suited to fostering awareness of all value inference processes by situating self-reflection in specific contexts and behaviors.

Value inference could be misused by malicious agents, especially targeting sensitive features including ethnicity and political orientation [144, 286]. For instance, authorities in non-liberal countries could use value inference to identify repressed minorities by detecting value preferences that diverge from the expected preferences. Mechanisms must be devised during the development of AI agents to avoid misuse. To this end, ongoing research is investigating methods that mitigate bias and unfairness by design [79, 157, 308]. Furthermore, as suggested by Russell [260], uncertainty can be built in AI agents by design, so that the AI agents can learn to consult the relevant stakeholders when doubting whether their actions may cause undesired consequences.

Our work is concerned with descriptive ethics—we use AI systems to understand how humans reason about values. However, the use of systems trained to discern descriptive ethics for normative ethics (i.e., to make value-based judgments such as religious prescriptions and medical advice) can be problematic [286]. Deriving a normative framework from descriptive datasets implicitly associates the average view with moral correctness. The perspectivist approach we discuss in Section 8.1.2 only partially addresses the problem—even when modeling a representative sample of the population, the majority view emerges as correct. To this end, we envision a HI approach where AI systems do not take moral stances, but rather provide humans with all the necessary contextual factors to allow them to make normative decisions.

The systems we develop will likely be used in applications considered high-risk under the European AI Act [59] (e.g., medical applications) and thus should be under strict control. As advocated by Ferrari et al. [88], regulators ought to be able to observe, inspect, and modify AI systems used in sensitive applications. That is, regulators ought to be informed about when, where, and how AI systems are used. They ought to be granted access to information such as training data, model architectures, and hyperparameters. Finally, they ought to be in a position to mandate changes to e.g., training data and infrastructure to ensure compliance with regulations.

If used cautiously, value inference has the potential to drastically improve our society. As motivated throughout the thesis, value inference serves as the starting point to guide AI agents in aligning their behavior with our value systems. Furthermore, it can allow us humans to better understand each other. Policy-makers can employ value inference to gauge citizens' value systems at scale over divisive issues. Doctors can use it to tailor rehabilitation paths at an individual level. In the next chapter, we discuss these and other beneficial applications that are being developed based on value inference.

9

BEYOND VALUE INFERENCE

Value inference is the first step toward value alignment. In practice, value inference is followed by the *operationalization* of values, both at the agent and sociotechnical system levels. By operationalization, we refer to the usage of the inferred (or pre-loaded) value systems to model an AI agent’s behavior [6, 208, 215, 300]. Values have been used for eliciting appropriate trust [195], plan selection [62], negotiation [25], social simulation [122], and engineering normative systems [204, 207, 270, 281]. We envision value inference and operationalization as actively influencing each other throughout the lifecycle of an STS. An example of such a connection is the evaluation of norm compliance [70, 298], i.e., assessing whether the implemented norms align with the inferred values.

Value inference facilitates other various practical applications. In Section 9.1, we describe a project that expands upon the methods introduced in Chapter 6 by aggregating individual value systems at a societal level to support decision-making. Then, Section 9.2 offers an overview of other concrete applications that are being developed based on our work on value inference.

9.1 AGGREGATING VALUE SYSTEMS FOR DECISION SUPPORT

Value inference results in the estimation of the value systems of individual stakeholders, allowing AI agents to align their behavior with the estimated value systems. However, several concrete applications require the *aggregation* of individual systems at a group level. For instance, making policy decisions that align with stakeholders having a variety of value systems (e.g., when deciding over water governance [232, 233]) or designing human-agent teams involving humans with differing ethical perspectives (e.g., in the medical field, where teams are constantly tasked with scenarios that require ethical consideration [90]). To this end, value systems ought to be aggregated to yield a consensus value system. However, from a social choice perspective, value systems can be aggregated following different ethical principles (e.g., utilitarian or egalitarian).

We propose a method for aggregating value systems at a group level that considers a range of ethical principles, from utilitarian to egalitarian. In our experimental setting, we assume to have a fixed set of values over which different stakeholders have indicated their preferences, and aim to aggregate the individual value systems at a group level. We test our proposed method by aggregating the individual value systems obtained in Chapter 6. Our experimental evaluation shows how different consensus value systems are obtained depending on the ethical principle of choice, leading to practical insights for a decision-maker on how to perform value system aggregation.

9.1.1 METHOD

To aggregate value systems, we employ the approach proposed by Lera-Leri et al. [168] (adapted from González-Pachón and Romero [103]), who introduce a distance function to aggregate rankings over objects (values, in our case). Adapting the formula to the formalization of value systems introduced in Section 6.2.2, the distance function is:

$$U_p = \left[\sum_{i=1}^N \sum_{j=1}^{|V|} |R_i[j] - R_s[j]|^p \right]^{1/p}, \quad (9.1)$$

where $R_i[j]$ is the rank position provided by the i -th member of the society for the j -th object (in our case, the j -th value belonging to the value list V) within the ranking. $R_s[j]$ is

the *consensus* position assigned by the society as a whole to the j -th value, i.e., the unknown consensus ranking that we seek to obtain. The value system aggregation problem consists of computing the consensus ranking R_S that minimizes the distance U_p . Lera-Leri et al. [168] propose an ℓ_p -regression approach to address this optimization task, which is out of the scope of this thesis. We defer to their work for additional insight.

From the U_p distance function, Lera-Leri et al. [168] derive two cases of interest. First, by setting $p = 1$, the general distance in Eq. 9.1 yields:

$$U_B = \sum_{i=1}^N \sum_{j=1}^{|V|} |R_i[j] - R_S[j]|. \tag{9.2}$$

The consensus that minimizes U_B provides the social optimum from the point of view of the majority, i.e., the *utilitarian* solution (or Benthamite solution [36]) that maximizes the total welfare. Instead, by setting $p = \infty$, the distance function in Eq. 9.1 yields:

$$U_R = \max_{i,j} [|R_i[j] - R_S[j]|]. \tag{9.3}$$

Finding the consensus, in this case, implies the minimization of the disagreement of the member of the society most displaced with respect to the majority solution defined by the utilitarian case above (Eq. 9.2). This solution is *egalitarian* [53] since it represents the social optimum from the point of view of the minority (from the perspective of the worst-off member of the society according to the Rawls' principle [247]).

In addition to the utilitarian and egalitarian cases, we can use $p \in [2, \infty)$ for computing different consensus. To illustrate the semantics of the ethical principle p and its impact on the consensus, we show a test case (with fabricated data) in Figure 9.1, which plots the judgments of 25 individuals on two objects. The circles represent the individuals' judgements $R_i[1]$ and $R_i[2]$ about objects 1 and 2 within the x and y axis, respectively.

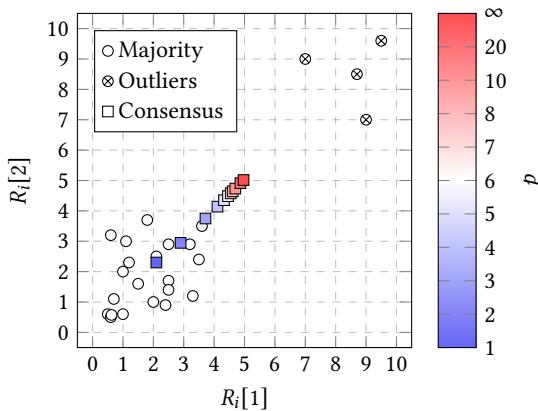


Figure 9.1: The impact of the ethical principle p on the computed consensus of a set of individuals' judgment of two objects: $R_i[1]$ and $R_i[2]$. Circles show the individuals' judgment and squares represent the consensus computed. Squares are filled with the color of the ethical principle used to compute the consensus.

In this example, we distinguish two groups of individuals: (1) a clustered set of individuals that represent the majority (with values for x and y smaller than 4), and (2) a

few individuals who represent outliers distant from the majority (shown with a crossed circle). The squares represent the position of the consensus computed with different ethical principles p 's, represented with a color scale from blue ($p = 1$) to red ($p = \infty$). As we can observe, the utilitarian consensus ($p = 1$), is at the center of the majority. As p increases, the consensus moves towards the outliers, converging to the egalitarian solution ($p = \infty$) which reduces the distance of the consensus to the worst-off member of the society.

9.1.2 RESULTS

We employ the described method to aggregate the individual value systems obtained in Chapter 6 with the best-performing value estimation method (R_{comb}). Table 9.1 shows the consensus rankings resulting from the aggregation with different ethical principles (p): from 1 (utilitarian) to 10, and ∞ (egalitarian). Each column $R_S[v_i]$ indicates the position of value v_i in the ranking as computed by our aggregation. Note that we obtained a partial ranking as a consensus ranking for each ethical principle. That is, the order (preferences) between values in each consensus ranking can contain ties between values. For instance, in the first row, value v_1 is equally preferred to value v_2 (because $R_S[v_1] = R_S[v_2] = 2$).

Table 9.1: Computed consensus ranking for different ethical principles p .

p	$R_S[v_1]$	$R_S[v_2]$	$R_S[v_3]$	$R_S[v_4]$	$R_S[v_5]$	Consensus ranking
1	2.00	2.00	3.00	3.00	3.00	$v_1 \sim v_2 \succ v_3 \sim v_4 \sim v_5$
2	2.14	2.53	2.90	2.90	2.91	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
3	2.31	2.65	2.95	2.96	2.94	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
5	2.51	2.78	2.98	2.99	2.97	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
7	2.63	2.84	2.99	3.00	2.98	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
10	2.74	2.89	2.99	3.00	2.99	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
∞	3.00	3.00	3.00	3.00	3.00	$v_1 \sim v_2 \sim v_3 \sim v_4 \sim v_5$

From Table 9.1, we distinguish three types of consensus rankings.

- $p = 1$ (utilitarian): v_1 and v_2 are equally preferred, and they are both preferred over the others (v_3, v_4, v_5), which in turn are equally preferred.
- $p \in [2..10]$ (intermediate): v_1 is more preferred than v_2 , and both are more preferred than the other values (v_3, v_4, v_5). The indifference between v_3, v_4 and v_5 holds.
- $p = \infty$ (egalitarian): all values are equally preferred.

We make two observations from Table 9.1. First, Lera-Leri et al. [168] show that for $p = 2$ the consensus ranking results correspond to computing the mean of the individual rankings to aggregate. Accordingly, the consensus ranking for $p = 2$ is the same as the one obtained in Chapter 6, where we employ the mean of individual rankings to compute a consensus ranking. Second, the consensus position R_S for all moral values converges to 3 (central position in the ranking) as the value of parameter p increases.

CHARACTERIZING THE SPACE OF ETHICAL PRINCIPLES

We characterize the whole space of ethical principles (from utilitarian to egalitarian) that are available to a decision-maker when computing a consensus value system, as introduced

by Lera-Leri et al. [168]. Our goal is to determine whether an ethical principle p produces a consensus leaning towards utilitarian ($p = 1$) or egalitarian ($p = \infty$). To achieve our objective, we compute the consensus ranking R_S considering a given p (denoted as $R_S^{(p)}$) and we measure the distance between $R_S^{(p)}$ and the one corresponding to $p = 1$ and $p = \infty$, denoted as $R_S^{(1)}$ and $R_S^{(\infty)}$ respectively. We refer to these two distances as $\|R_S^{(1)} - R_S^{(p)}\|_p$ and $\|R_S^{(\infty)} - R_S^{(p)}\|_p$. These distances allow us to measure a *transition point* (\bar{p}) that is the equidistant ethical principle whose computed consensus is between the utilitarian and the egalitarian consensus. As a result, we can characterize different zones within the space of ethical principles as Figure 9.2 illustrates. The *utilitarian zone* is composed of the ethical principles p that lean towards the utilitarian consensus ($p < \bar{p}$). The *egalitarian zone* is composed of the ethical principles that lean towards the egalitarian consensus ($p > \bar{p}$).

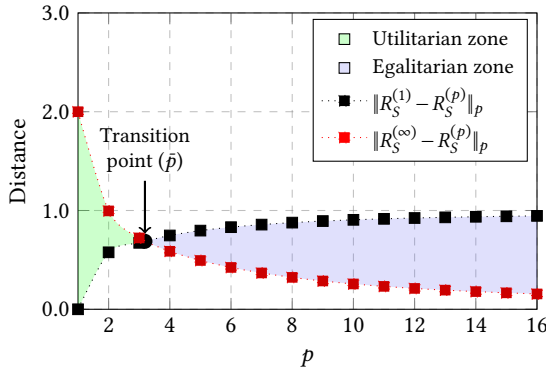


Figure 9.2: Distance between the consensus computed according to ethical principle p and the consensus computed according to $p = 1$ (utilitarian, black line) and $p = \infty$ (egalitarian, red line). \bar{p} is the transition point that yields a consensus that is equidistant from the utilitarian and egalitarian ones. Thus, \bar{p} divides the space of ethical principles into a *utilitarian zone* (more similar to the utilitarian consensus, green) and an *egalitarian zone* (more similar to the egalitarian consensus, blue).

We observe that the transition point is near 3 ($\bar{p} \sim 3$). Therefore, a given ethical principle p can be interpreted as “more utilitarian” or “more egalitarian” depending on its relative position with respect to the transition point. However, notice that this transition point \bar{p} differs from the one obtained by Lera-Leri et al. [168] in their experiments, showing its dependency on the data distribution.

9.1.3 TAKEAWAYS

Individual value systems can be aggregated differently according to different ethical principles. We employ a method for aggregating value systems with an ethical principle that ranges from utilitarian to egalitarian. Our experiments show that the choice of the ethical principle leads to different consensus value systems. The visual analysis displayed in Figure 9.2 intends to provide useful guidance for decision-makers concerned with obtaining a consensus on different value systems following an ethical principle of choice. In particular, we show how to (1) quantify the impact of choosing different ethical principles on the resulting consensus value system, and (2) determine whether a given ethical principle

produces a consensus leaning towards the utilitarian or the egalitarian aggregation. As the transition between ethical principles is dependent on the distribution of the data, we envision our proposed approach and visualization to be used by decision-makers when deciding how to aggregate individual value systems.

9.2 REAL-WORLD APPLICATIONS (BEING) DEVELOPED

Value inference and value alignment are priorities for beneficial AI [261, 283], yet their development is in the early stages. While much of the related research is carried out in laboratories, real-world applications are starting to surface. Numerous concrete applications are on the horizon where the importance of value inference can be demonstrated.

First, as exemplified in the previous section, the inferred value systems can be invaluable to policy-makers, as they them to understand the deeper motivations of citizens on divisive issues. Our work with the PVE surveys shows how to estimate citizens' value preferences based on participatory democratic tools. To this end, we have collaborated with the creators of the PVE. We helped them analyze the Covid PVE data (Section 2.4) with NLP tools, as the Dutch government required quick insight into the stance of Dutch citizens on COVID-19 policies [210]. Furthermore, we have collaborated with Populytics, a TU Delft spinoff that branched out to commercially perform PVEs, on the analysis of a survey on renewable energy in the Foodvalley region in the Netherlands¹.

Second, value inference can help us better understand each other. Experiments have shown that AI agents can help improve the quality and depth of online conversations [112, 113]. We envision AI agents that provide value-based explanations of the stances held by participants to an online conversation or deliberation, intending to deepen the mutual understanding of participants. Our collaboration with the The Hague University of Applied Sciences (THUAS) points in this direction. Researchers at THUAS have developed a deliberation platform² where citizens can deliberate on a topic of relevance to the municipality of the Hague. We intend to perform value inference based on the conversations on the platform, investigating whether unveiling participants' value systems helps in achieving a better common understanding of the discussion and the different stances.

Third, value inference can help in creating tailored solutions for sensitive applications such as healthcare. Researchers in the Netherlands Organisation for Applied Scientific Research (TNO) are developing an AI agent aimed at supporting lifestyle changes in diabetes patients [68]. In their vision, the AI agent ought to infer a patient's values to learn what is important to them, and to adapt its behavior and strategy at an individual level to follow the most effective recovery path. Similarly, De Kindertelefoon is a Dutch online platform that allows children and adolescent to anonymously and safely discuss their problems with peers and expert moderators. De Kindertelefoon intends to employ AI techniques to offer better support to children in need. For instance, Al Owayyed [10] developed a platform for training moderators through a conversational AI agent. We are currently developing HI approaches to value inference, with the involvement of experts and moderators. The estimation of the value systems and needs of the participating children can help moderators in tailoring their support and experts in understanding the shared concerns of participants.

¹<https://populytics.nl/en/cases/foodvalley/>

²<https://civictchnology.nl/project/public-dialogues-goodtalk/>

V

APPENDICES



IDENTIFYING AND EVALUATING CONTEXT-SPECIFIC VALUES

A.1 EXPERIMENTS PROTOCOL

We provide additional information on the three experiments outlined in Section 3.3.

A.1.1 EXPERIMENT 1: VALUE LISTS

As Section 3.3.1 describes, six annotators were invited to generate value lists with the use of Axies. A brief survey revealed that the annotators consisted of one graduate student, three doctoral students, and two postdoctoral researchers, aged between 20 and 35, and with previous experience with personal values.

24h before each of their participation to the first exploration, we sent an email to each participant asking to create a user on the web application and accept the Informed Consent Form. Upon acceptance, they were shown general information about personal values and the analyzed contexts (COVID and ENERGY). At the beginning of the first exploration and first consolidation sessions, each annotator was shown instructions and goals of the respective phases. Informed Consent Form, introductory information, and phases instructions are detailed in the supplemental material [176].

A.1.2 EXPERIMENT 2: SPECIFICITY

As described in Section 3.3.2, two policy experts were invited to perform Experiment 2. Upon giving informed consent, the experts reported to be graduate students in the *technology and policy making field*, with experience with the two participatory value evaluations (PVEs) at the base of the analyzed contexts (COVID and ENERGY) through previous projects. Before starting the individual phase of the experiment, they were provided with instruction for the evaluation task (including information about the contexts). Informed Consent Form and instructions are detailed in the supplemental material [176].

Both annotators individually gave specificity ratings to all 57 values (including all Axies and Schwartz values). Afterwards, they were invited to deliberate about the values for which their ratings differed more than two points on Likert scale (2 values out of 57).

Following their discussion, they were offered the option to change the rating for these values, provided they noted down why they changed it. Table A.1 presents the Intraclass Correlation (ICC) coefficients illustrating the agreement between annotators, highlighting the differences between Axies and Schwartz values.

Table A.1: ICC of context-specificity ratings.

	Axies values	Schwartz values	All values
ICC before discussion	0.69 (good)	0.51 (fair)	0.68 (good)
ICC after discussion	0.76 (excellent)	0.51 (fair)	0.74 (good)

A.1.3 EXPERIMENT 3: COMPREHENSION AND CONSISTENCY

Section 3.3.3 presents an overview of Experiment 3. We initially opened a pilot annotation task on Prolific for four user, and set the expected completion time to 50 minutes. Results encouraged us to proceed. Although we expected the completion time to be lower, we preferred to keep the expected completion time to 45 minutes to encourage users to spend more effort on the task.

Upon taking the task on Prolific, workers were redirected to the web application hosted on our servers. Here, after accepting the Informed Consent Form, they were given a small introduction to the annotation task and the assigned context (COVID or ENERGY). Then, they were guided sequentially through the three steps (clarity evaluation, distinguishability evaluation, opinion annotation), while being shown instructions at the beginning of the respective step. Informed Consent Form, introductory information, and steps instructions are all detailed in the supplemental material [176].

USER DEMOGRAPHICS

Upon giving informed consent, workers were asked the following demographic information:

- What is your age?
- What gender do you identify as?
- Where is your home located?
- What is the highest degree or level of education you have completed?

Figure A.1 presents the aggregated results of the demographics of the 72 users whose submissions were considered in the study.

QUALITY CONTROL

As mentioned in Section 3.3.3, four attention checks were included in each task. In distinguishability evaluation, we showed an extra pair composed of twice the same value, which the users were intended to label as not distinguishable (1 out of 5). Then, we showed a pair consisting of two values taken from opposite ends of the Schwartz circumplex (Tradition and Self-Direction), which the users were expected to rate with a distinguishability score of at least 3 out of 5. In the opinion annotation task, each worker was shown two artificial

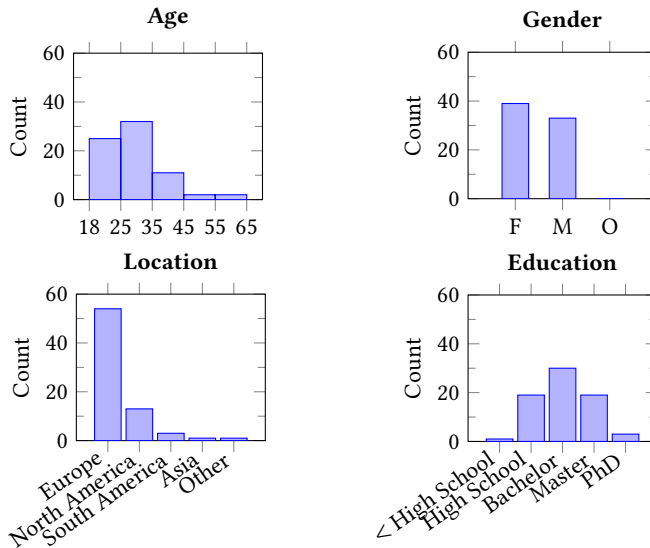


Figure A.1: Demographics of crowd workers.

opinions explicitly highlighting one of the values present in the value list (for instance, an opinion such as “Safety comes above all.”, with Safety being one of the values present in the list). The check would be considered successfully passed if the related value is among the values chosen to annotate the opinion.

Workers’ submissions were accepted if they would spend at least 20 minutes on the task, or if they would pass more than half of the attention checks. Of the 115 workers who completed the task, 107 fulfilled these requirements and were paid. The submissions were then considered in our results only if both distinguishability evaluation attention checks and at least one of the completeness evaluation attention checks (due to the more subjective nature of the latter task) would be successfully passed. 72 submissions were finally considered in our analysis.

A.2 WEB PLATFORM

A computational platform is necessary to support the annotators in applying Axies without exposing them to the underlying technical mechanisms. To enable these features, we develop an intuitive and reusable web platform with an AI backend. The platform is implemented in Python on the Flask micro web framework [109]. The backend is also implemented in Python to provide seamless integration with state-of-the-art NLP models. All data is stored in an SQLite database [127]. Further, we developed functionalities to import the opinion corpus in a csv or yaml format. Finally, the responsive web interface is implemented in JavaScript. The interface can be used on small (e.g., smartphone) and large screens, and it utilizes the de facto standards in modern web applications. The modular setup of the two phases enables easy extension to new annotation tasks. The source code



Figure A.2: The exploration phase in the web application. Annotators add values and/or keywords based on the shown motivation. Next, they fetch a new motivation as the farthest to the currently displayed (via FFT), or as the most similar to an annotated value.

is available on GitHub¹ and a video demonstration on YouTube².

Annotators are required to register with a username and a password. Operations can be performed asynchronously. Data is stored in the SQL database upon input, allowing the annotators to leave and return to the platform without losing progress. A top navigation bar is accessible from any page (as shown in Figure A.2), permitting users to switch between the two phases of Axies (Explore and Consolidate) and different contexts (e.g., COVID and ENERGY in our experiments).

¹Code: <https://github.com/enricoliscio/axies>

²Demonstration: <https://youtu.be/s7nJPr2Z80w>

A.3 EXTENDED RESULTS

We offer additional details on the results presented in Section 3.4. Raw results are in the supplemental material [176].

A.3.1 VALUE LISTS

We provide further details on the results of Experiment 1 (Section 3.3.1 and 3.4.1).

EXPLORATION

Six annotators performed exploration on two contexts, resulting in 12 exploration sessions. The main results are presented in Section 3.4.1. An overview of the sessions is presented in Table A.2, highlighting the number of values generated in each exploration.

Table A.2: Overview of the duration of the exploration phase.

Annotator ID	Group	Context	#Values	Duration
1	1	COVID	8	55 min
1	1	ENERGY	8	70 min
2	2	COVID	11	80 min
2	2	ENERGY	12	80 min
3	1	COVID	13	60 min
3	1	ENERGY	18	75 min
4	2	COVID	13	80 min
4	2	ENERGY	19	60 min
5	2	COVID	8	60 min
5	2	ENERGY	6	50 min
6	1	COVID	14	80 min
6	1	ENERGY	14	70 min

CONSOLIDATION

Two groups (of three annotators each) performed consolidation on two contexts, resulting in four consolidation sessions. The main results are presented in Section 3.4.1. Table A.3 shows the number of values at the start and at the end of each consolidation phase, and its duration. The four complete value lists (including value names, keywords, and defining goals) are in Tables A.5, A.6, A.7, and A.8. We retain all keywords as originally annotated, removing one keyword we consider inappropriate.

Table A.3: Overview of the consolidation phases.

Group	Context	#Start values	#End values	Duration
1	COVID	35	11	105 min
1	ENERGY	40	14	110 min
2	COVID	32	9	115 min
2	ENERGY	37	13	120 min

A.3.2 COMPREHENSIBILITY

Here we present a detailed picture of the comprehensibility evaluation results obtain with Experiment 3, described in Section 3.3.3. The main results are presented in Section 3.4.3.

CLARITY EVALUATION

Section 3.4.3 presents the clarity evaluation results per each context. Figure A.3 presents the average clarity ratings given to the values of the five value lists.

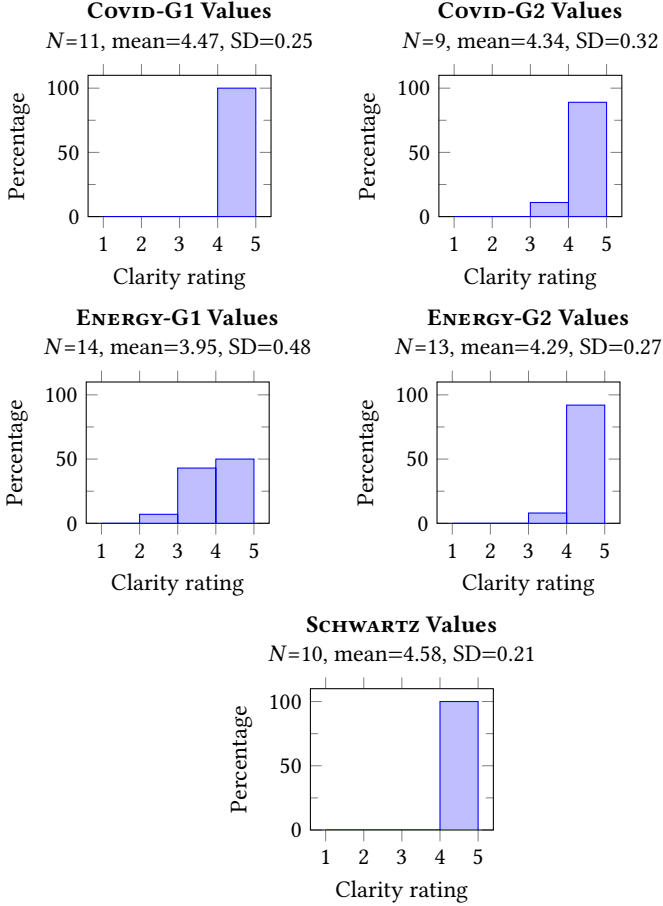


Figure A.3: Histograms of value clarity ratings.



DISTINGUISHABILITY EVALUATION

Section 3.4.3 presents the distinguishability evaluation results per each context. Figure A.4 presents the average distinguishability ratings divided in the five value lists.

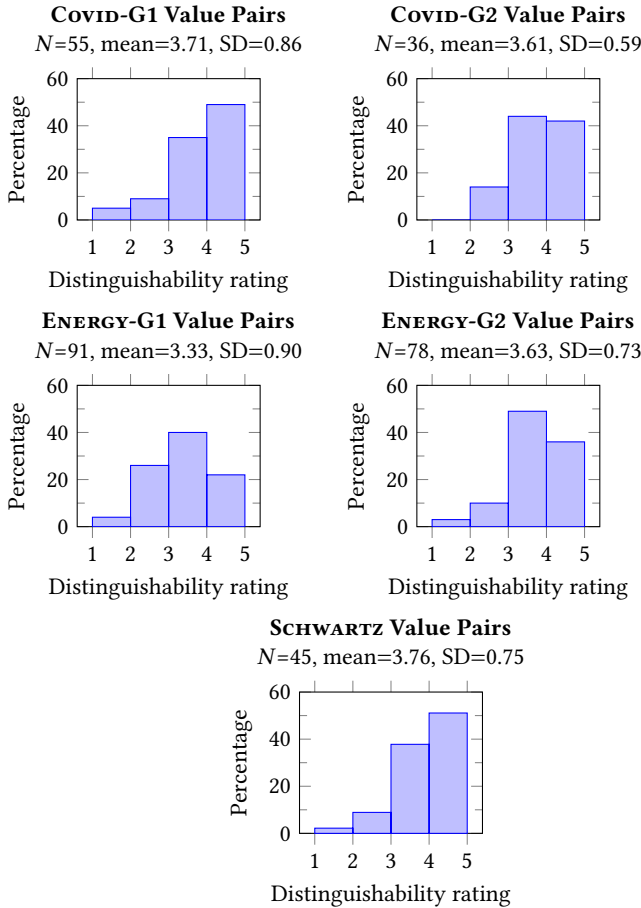


Figure A.4: Histograms of value distinguishability ratings.

A

CROWD ANNOTATION TASK

Section 3.4.6 describes the results of the crowdsourced annotations. Figures A.5 and A.6 illustrate, per value list, the number of opinions that were annotated with each value belonging to the list. Recall that each value list was used to annotate 100 opinions. Table A.4 presents the Inter-Rater Reliability (and its interpretation) for each value in the value lists.

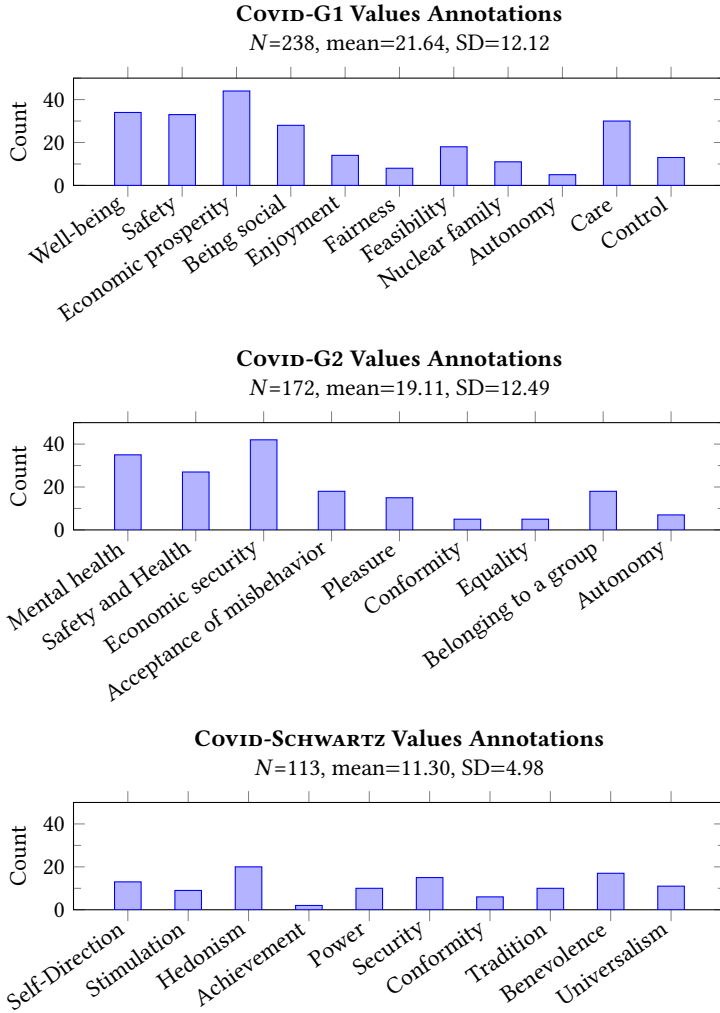


Figure A.5: Histogram of annotated opinions per value in context COVID.

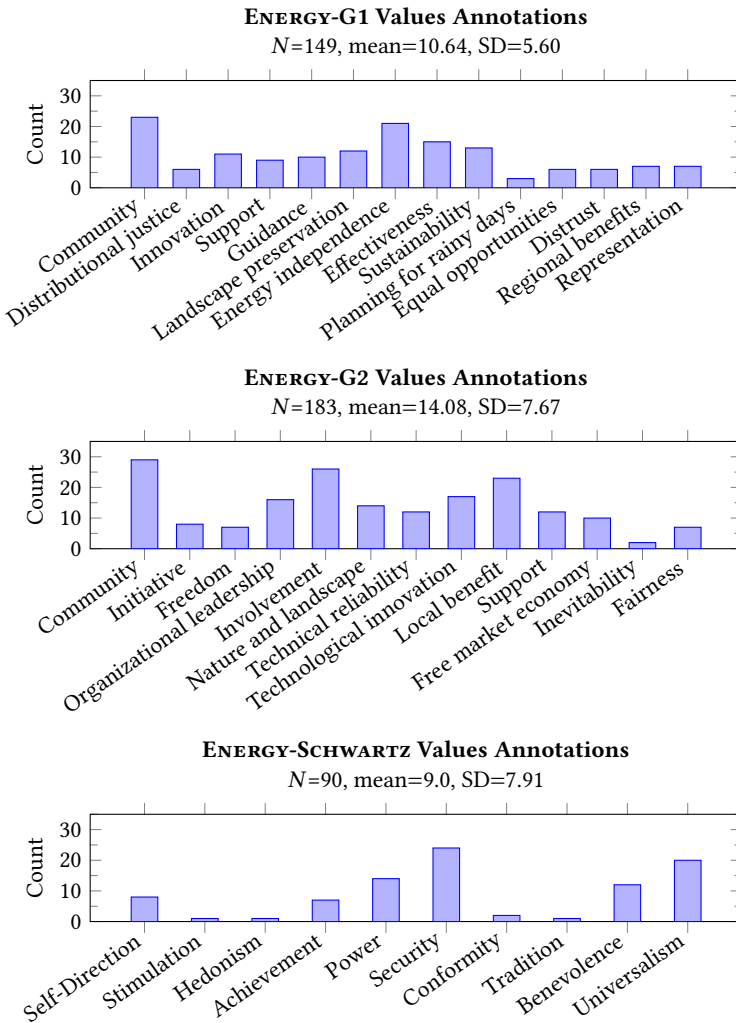


Figure A.6: Histogram of annotated opinions per value in context ENERGY.

Table A.4: Inter-Rater Reliability with Axes and Schwartz value lists.

COVID-G1		COVID-G2	
Value	Fleiss' Kappa	Value	Fleiss' Kappa
fairness	0.14 (poor)	Mental health	0.51 (moderate)
care	0.1 (poor)	Safety and Health	0.23 (fair)
being social	0.22 (fair)	Economic security	0.63 (substantial)
enjoyment	0.1 (poor)	Acceptance of misbehavior	0.39 (fair)
economic prosperity	0.53 (moderate)	Pleasure	0.22 (fair)
nuclear family	0.27 (fair)	Conformity	-0.02 (poor)
control	-0.08 (poor)	Equality	0.08 (poor)
safety	0.08 (poor)	Belonging to a group	0.16 (poor)
autonomy	-0.07 (poor)	Autonomy	0.01 (poor)
well-being	0.15 (poor)		
feasibility	-0.06 (poor)		

ENERGY-G1		ENERGY-G2	
Value	Fleiss' Kappa	Value	Fleiss' Kappa
Distributional justice	-0.03 (poor)	Community	0.31 (fair)
innovation	0.22 (fair)	Initiative	0.08 (poor)
guidance	0.1 (poor)	freedom	0.34 (fair)
energy independence	0.28 (fair)	Organizational Leadership	0.26 (fair)
effectiveness	0.05 (poor)	Involvement	0.32 (fair)
support	0.09 (poor)	Nature and landscape	0.32 (fair)
sustainability	0.03 (poor)	Technical Reliability	0.16 (poor)
planning for rainy days	0.08 (poor)	Technological Innovation	0.31 (fair)
distrust	0.27 (fair)	Local benefit	0.08 (poor)
landscape preservation	0.63 (substantial)	support	0.12 (poor)
equal opportunities	0.15 (poor)	Free market economy	0.31 (fair)
regional benefits	-0.04 (poor)	Inevitability	0.23 (fair)
representation	0.03 (poor)	Fairness	0.14 (poor)
community	0.41 (moderate)		

COVID-SCHWARTZ		ENERGY-SCHWARTZ	
Value	Fleiss' Kappa	Value	Fleiss' Kappa
Self-Direction	0.06 (poor)	Self-Direction	0.11 (poor)
Stimulation	0.06 (poor)	Stimulation	0.02 (poor)
Hedonism	0.07 (poor)	Hedonism	0.2 (fair)
Achievement	0 (poor)	Achievement	-0.05 (poor)
Power	0.04 (poor)	Power	0.23 (fair)
Security	0.08 (poor)	Security	0.13 (poor)
Conformity	-0.06 (poor)	Conformity	-0.04 (poor)
Tradition	0.08 (poor)	Tradition	0.08 (poor)
Benevolence	0.18 (poor)	Benevolence	0.17 (poor)
Universalism	0.09 (poor)	Universalism	0.1 (poor)

Value name	Keywords	Defining goal
fairness	apply to everyone, differences, equality, distribution, discrimination	Measures should apply to everyone, without discriminating among groups (age and region especially).
care	caregiving, care for each others, help, support	Make sure that everyone is taken care of and looked after.
being social	contact, physical, oil of society, social isolation, neighbors, hug, weighs very heavily, fellow believers, religion, friends, each others	Being in contact with friends, neighbors and other acquaintances in your social circle.
enjoyment	beer, getaway, recreational sex, sport, entertainment, drinking, fun, pleasure, celebrate, festivals, weekends	Being able to enjoy life at full and indulge in fun activities.
economic prosperity	unemployment, poverty, bankruptcy, economy, companies, entrepreneur, money, interests, restart, must be kept running, running, income, provide, jobs, industry, work, worried, stability	Being able to pay and afford what you need.
nuclear family	couch, wedding, parents, sister, birth, family members, foundation, social education, unnatural, child, relatives	Being together with your parents, kids and siblings.
control	cross-border traffic, traffic, necessary adjustments, protective equipment, distance, necessary, obliging mouth caps, can maintain, tests, impossible, protectors, solutions, masks, disinfect, app, protocol, monitored, humidity monitored, difficult, distance, sly, secretive, obey	Having and respecting regulations in order to avoid the spread of the disease.
safety	safety measures, mouth caps, gloves, precaution, distance, wash hands, low risk, not transferring, many strangers come very close, immunity temporary, infection hazard, transfer limited, below 60, risk, building immunity, group immunity, tests, dangers to the elderly, monitoring our health, health	Staying healthy and not infected.
autonomy	everything open, restrictions, enforcement, choice, own, liberty, individualist, self-concern	Being able to determine by yourself what you do and are allowed to do.
well-being	alone, loneliness, psychological influence, relaxation, depression, mental health, emotional, beautiful moments, mental health, psychically, suicide, pressure, needs, normal, loneliness, isolation, restless, apprehensiveness, expectation, hope	Being content, doing well, without worries.

feasibility	test case, research, possibilities, demonstrated, charts, statistics, science, rvm, try out, unlikely, logical, practical, reasonable, common sense	Having policies that are doable and effective.
-------------	---	--

Table A.5: Complete COVID-G1 value list.

Value name	Keywords	Defining goal
Mental health	mentally, stress, Mental damage, positivity, living alone, psychic injury, stress, physical proximity, mental complaints, tension, spiritual relaxation, quality of life, psychiatry, loneliness, walk around, visit ill people in nursery home, cheer up, last days, empathy, compassion, last phase of life, psychologist, personal care	The strive towards protecting and improving one's emotional and psychological well-being
Safety and Health	hygiene, facial protection, screen, immune, small scale, masks, protected, keep distance, distance, Mouth caps, gloves, busy, crowded, health, groups, personal, safe, mortality, fatality, death, dying, diabets, die, immune system, temperature	Personal protection against the health-related impacts caused by the coronavirus
Economic security	money, shops, recession, company, jobs, entrepreneurs, self-employed, survival, companies, financial impact, fall, restart activity, poverty, bankrupt, restaurants, entrepreneurship, economic, taxation, productivity, unemployment, resume work, heavy economic times, economy, zzp, zzper, financial damage, companies fall down, earn money, Relaunching economics	Mitigating the economic downsides of the situation caused by the pandemic and its countermeasures.
Acceptance of misbehavior	working black, empathy, illegal, ignorance, misbehavior, People are already doing it, We see it every day, abuse, illusion, lying, unfair	Acceptance to the fact that people might not conform to measures.
Pleasure	pleasure, fun, activity, entertainment, cozy, Drink, Balance, social interaction, festivals, dates, food, recreational, leisure, liveliness, kiss, hug	Being able to undertake activities that promotes personal satisfaction and pleasure.

Conformity	official, government, behavior, citizen, control, measures, controlled protection, 1.5 meter, working undercover, RIVM, keep distance, strict regulations, mouth caps, rules, fines, norms, regulations, stay home, responsible, unless you're ill, no busy places, limit risk, work at home, trust	Striving to comply to the guidelines and rules imposed by the authorities
Equality	equal, the same, unfair, Distinction, discrimination, doesn't apply only to certain groups of subjects, help other people, Human side, humanity	Ensuring that all people are given the same treatment and act for the common good
Belonging to a group	friends, miss friends, social contact, Zoom, Skype, meet friends, social, cozy, chill, friends, buddies, youth, together, contact, church, sing, pray, sports, family, school, peers, daughter, cuddling, grandparent, relatives	Being able to closely interact with the people that you care about or enjoy spending time with
Autonomy	walk around, own decision, choice, proximity, do something useful, police state, personal, choice, freedom, self, autonomy, companies decide, decide for themselves, responsibility for individual, you may visit or not, own responsibility, choose for themselves	Being able to make your own decisions and take the responsibility for your actions

Table A.6: Complete COVID-G2 value list.

Value name	Keywords	Defining goal
Distributional justice	Everywhere, not just in Friesland, spread, across the netherlands, they pay	Fair distribution of burdens and benefits.
innovation	alternatives, stimulate, under development, future, bet on, biogas, new technologies, creativity, progress, invention	Keep on producing new and better technologies.
guidance	direction, Obliges companies, control, central, disagree, coordinated, overview, centrally regulated, distribute, take the lead, government, municipality, monitoring	Having a central entity that decides and regulates energy policies.
energy independence	themselves, own backyard, self, self-doing, close to home, private, reserve, need, storage	Having an independent source of energy, without relying on external providers.

A

effectiveness	success, feasible, effective, appropriate, optimal, in order to get away from , inevitable, necessary, does its job, very busy, well-led, needed, Prevent things from being done twice, Many different housing situations, possible, small, fitting	Creating tailor-made, doable policies to reach the renewable energy target.
support	help, care, possibilities, don't know, knowledge, unable, weakest	People receive advice and assistance.
sustainability	pollution, renewable energy, care for environment	Having energy policies that increase renewable energy generation.
planning for rainy days	expectation, storage, seasonal, weather conditions, unforeseen	Having plans for unforeseen circumstances.
distrust	only revolves around money, fill his own cases, not leave it to the market, small part, delivered to the gods, serving, repugnance, economy, savings, anti-politics, many beautiful words, few deeds, anti-market, mistrust	Big players (government, large companies) should not be in charge of solving problems on citizens' behalf.
landscape preservation	Billiard towel with holes, beautiful, landscape, messy, few places as possible, not stand out, landscape pollution, inconspicuous, opposed to large-scale, beauty, nature is affected, interference, for nature, surroundings	Leave landscape untouched.
equal opportunities	benefits everyone, rich and poor, paid by everyone, strongest win from the weakest, possible for everyone, fairness	Everyone should be given a chance to participate and speak up.
regional benefits	jobs, own gain, profit, investment, opportunities	Bring advantages to job market and economy of South-West Friesland.
representation	Approach all residents, stand up for, accountability, responsibility	Every member of society should be accounted for when taking decisions.
community	decide for themselves, determining people, people determining, free, own steps, willingness, leave it to the people, from the bottom, choice, self-management, independence, autonomy, Local needs, own community, small-scale, own initiative, involvement, with residents, participation, each others, solidarity, engagement	Creation and ownership by and of the community.

Table A.7: Complete ENERGY-G1 value list.

Value name	Keywords	Defining goal
------------	----------	---------------

Community	cooperation, Encouraging residents, ideas, involve, involvement, local binder, limits and conditions of government and residents , mei elkoar, mien-skip, think along, each other, care, contribute, balance, protection , everyone, Solidarity	Preserving the feeling of doing it together and taking care of each other
Initiative	Involvement, Do something themselves, empty roofs, local entrepreneur, residents, buying solar panels, self-doing, conscious behaviour, heat pumps, solar panels, enthusiasm, regulate its own energy, opportunities, Encouraging residents, Stimulates, stick behind the door, Initiating	Participants value acting towards their own plans
freedom	own direction, my choice, freedom, independence, autonomy, responsibility for themselves, private, liberty, themselves, voice, residents, small scale	Participants value the ability to freely speak, think or make their own choices in the energy transition.
Organizational Leadership	organized, mess, conflict, central management, delays, decision-making, Prevent things from being done twice, oversee, higher level, cooperation, director, supervises, compliance, regulate, energy co-operations, director, leader, control, supervision, protection, direction, conditions, central point, democratic, Consults, municipality, Approach all residents, decision, coordinate, lead, expertise, nation, municipality, Europe, politicians, official	A single organization is in charge of supervision and organizing the process towards reaching the energy goals.
Involvement	involve people, involve citizens, democratically elected, voting, veto, election, financial participation, think along, participation, public, public evaluation, survey, opinion	People have a say in the process of reaching the energy goals
Nature and landscape	Nature, landscape, few places as possible, preserve, nature conservation, find a place, nature protection, small number of places, disruption of landscape, view in nature, ugly windmills, putting trees, trees, bushes, clustering, cluttering of the landscape, landscape, living, surrounding, scenic, ecology, not stand out, flat building, further from the inhabited world, not in my backyard, underground, minimal nuisance, less burdened, minimize burdening, liveability, noise, in front of my nose	Preserving nature and the aesthetic of the landscape

A

Technical Reliability	Transport losses, Overproduction, help with industry, severe winter, retention of electricity, spikes, excess generated energy, stored, later use, always electricity, stable grid, control, stability, direction of the municipality, electricity grid, energy security	Ensuring that people can rely on energy solutions and have a stable energy grid without hampering in their way of life
Technological Innovation	Hydrogen storage, Frontrunner, water treatment, biogas, wave energy, most recent products, newest technology, H2	Capacity to come up with new and better solutions to energy-related problems
Local benefit	Investment, Profits, revolves around money, earning model, profitable, rewarding, an extra penny, earn money, local profit, labour, financial risk reduction, no big investors, local labour, jobs for citizens, mercy of wealthy companies	To try and steer the (financial) benefits from a solutions towards one that is best for the participant and its peers.
support	municipality can help me with that, professional expertise, help me decide, older people, elderly, support, help, assist, aid, facilitation, legal, permit, subsidy	Ensuring that all participants can rely on the expertise of the decision makers, and are all assisted during the organizational procedures.
Free market economy	profitable, efficient, lucrative, most profit per area, optimize space, scalability, large scale, companies are effective, company, income, players, capitalist , profit	The belief that a free, self-regulated market economy will result in the best gains for all participants
Inevitability	no choice, energy security, necessity, needed, required, necessary, important, uncertainty , responsibility	The realization that actions need to be taken even if your preferences are not aligned with that action
Fairness	Same playing field, social approach, social, Each province must take a share, weaker, share profit, Divide the burden, strongest wins, neighbourhoods differ, local decision, local solution, Keep it local	The strive towards a proper division of benefits and responsibilities.

Table A.8: Complete ENERGY-G2 value list.

B

B

CROSS-DOMAIN CLASSIFICATION OF MORAL VALUES

B.1 EXPERIMENTAL DETAILS

As we train deep learning models, reproducibility is an issue due to the inherent randomness of the training procedure. Nevertheless, we seek to provide all possible tools for reproducing our experimental results. To do so, we attach our code and the complete set of results. Furthermore, the following sections describe our data preprocessing, the hyperparameters, the computing infrastructure, and the random seeds used in our experiments.

B.1.1 DATA PREPROCESSING

We choose to use the datasets as they are, despite their imbalanced label distribution (Table 2.4), since such imbalance is representative of realistic applications. We preprocess the tweets by removing URLs, emails, usernames and mentions. Next, we employ the Ekphrasis package¹ to correct common spelling mistakes and unpack contractions. Finally, emojis are transformed into their respective words using the Python Emoji package².

B.1.2 HYPERPARAMETERS

To select the hyperparameters, we trained and evaluated the model on the entire MFTC corpus with 10-fold cross-validation. Table B.1 shows the hyperparameters that were compared in this setting, highlighting in bold the best performing option that we then used in the experiments described in the paper. If a parameter is not present in the table, the default value supplied by the framework was used.

B.1.3 COMPUTING INFRASTRUCTURE

The following are the main libraries and computing environment used in our experiments.

- PyTorch: 1.8.1
- TensorFlow: 2.5.0
- FastText: 0.8.22

¹<https://github.com/cbaziotis/ekphrasis>

²<https://pypi.org/project/emoji/>

Table B.1: Hyperparameters tested and selected (in bold).

Hyperparameters	Options
Model name	bert-base-uncased
Number of parameters	110M
Max sequence length	64
Epochs	2, 3 , 5
Batch size	16 , 32, 64
Dropout	0.05, 0.1 , 0.02
Optimizer	AdamW
Learning Rate	$5 \cdot 10^{-5}$
Loss function	Binary Cross Entropy

B

- Huggingface’s Transformers: 4.6.0
- NVIDIA GeForce RTX 2080 Ti GPU
- CUDA: 11.2
- cuDNN: 8.1.1.33

Refer to the code base for a detailed list of the libraries we used, and their versions.

The following list details the amount of GPU hours spent for obtaining our results:

- Tables 4.1, B.2, and B.3: 44 hours
- Figures 4.2 and 4.3: 33 hours
- Tables B.4, B.5, and B.6: 24 hours
- Table B.8: 26 hours

The error analysis (Tables 4.2, 4.3a, and 4.3b) did not require additional GPU time.

B.1.4 RANDOM SEEDS

In our experiments, we ensured that the same train-test splits are used across different runs of each experiment. Further, to control for randomness, we fixed the random seeds in the following libraries:

- Python (`random.seed`);
- NumPy (`numpy.random.seed`);
- PyTorch (`torch.manual_seed`);
- Tensorflow (`tensorflow.random.set_seed`).

B.1.5 ARTIFACTS USAGE

We have mainly used two artifacts in this research: the MFTC and BERT. The MFTC was collected with the intent of facilitating NLP research on moral values [130]. It can be downloaded³ and used under the Creative Commons Attribution 4.0 license. BERT [74] was created with the intent of performing, among others, text classification. Thus, we are using it as originally intended, under its Apache 2.0 distribution license⁴.

³<https://osf.io/k5n7y/>

⁴<https://github.com/google-research/bert/blob/master/LICENSE>

B.2 EXTENDED RESULTS

In this Section, we extend the results presented in the paper. Raw results are available online [177].

B.2.1 MODEL COMPARISON

We have presented the results of the transferability analysis with the BERT model. In order to evaluate whether our conclusions generalize to other model architectures, we repeat the experiment conducted in the paper (see Sections 3.3 and 3.4) with the following two additional models:

- Long Short Term Memory (**LSTM**), a category of Recurrent Neural Networks (RNN). We choose LSTM as a baseline model since it is commonly used in moral value classification [130, 173, 206, 251].
- **fastText**, a machine learning approach that learns character-level information, in contrast to the whole word representations LSTM employs. This flexibility makes fastText a good candidate for transfer learning. Further, we choose fastText as it attains performances on par with state-of-the-art deep learning methods, but is considerably faster.

Tables B.2 and B.3 present the results of the transferability analysis, performed and presented analogously to Table 4.1, for LSTM, fastText, and BERT. We observe that BERT outperforms fastText and LSTM in most settings. This is not surprising, since BERT is state-of-the-art for text classification. Both BERT and fastText outperform LSTM, the model extensively used for predicting moral values. Further, we notice that the general trends observed in Section 4.3.1 hold for all three models.

Table B.2: Results of the four training scenarios and three models evaluated on \mathcal{T}_{source} . The columns indicate the dataset used as \mathcal{T}_{target} . For each experiment we report micro F_1 -score (m , left-hand column) and macro F_1 -score (M , right-hand column).

Classifier Setting	ALM		BLT		BLM		DAV		ELE		MT		SND		Average	
	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M
<i>LSTM</i>																
$C(source, source)$	64.1	45.7	64.0	52.1	61.1	39.6	59.2	48.0	63.5	46.5	66.4	47.1	65.6	46.8	63.4	46.5
$C(target, source)$	47.8	19.3	41.0	6.1	53.5	25.6	38.8	5.1	51.1	20.2	39.1	11.9	35.1	16.1	43.8	14.9
$C(finetime, source)$	61.4	37.4	48.3	25.1	60.0	39.6	41.6	11.0	60.7	40.5	55.1	39.1	52.3	36.6	54.2	32.8
$C(all, source)$	64.5	46.7	63.2	49.2	62.3	41.4	59.3	47.7	64.2	48.6	66.4	48.7	65.8	48.1	63.7	47.2
<i>fastText</i>																
$C(source, source)$	66.8	56.0	65.9	57.8	64.4	51.5	63.1	56.9	66.6	56.7	69.5	59.5	67.8	56.8	66.3	56.5
$C(target, source)$	54.5	30.9	42.7	8.5	56.4	33.1	38.7	5.1	52.2	30.0	48.9	22.0	41.3	20.3	47.8	21.4
$C(finetime, source)$	62.1	48.8	54.4	39.5	62.6	46.4	52.9	39.9	61.4	50.8	57.3	45.7	56.7	49.7	58.2	45.8
$C(all, source)$	66.9	56.3	66.0	57.5	64.8	52.1	63.1	56.7	66.9	57.0	68.7	58.2	67.5	56.4	66.3	56.3
<i>BERT</i>																
$C(source, source)$	73.9	65.6	73.9	68.3	71.2	61.8	71.1	66.4	73.3	66.4	75.7	68.0	74.5	66.5	73.4	66.1
$C(target, source)$	61.6	37.7	43.8	13.1	62.6	43.0	38.8	5.1	59.3	40.4	52.4	39.1	54.4	36.6	53.3	30.7
$C(finetime, source)$	70.3	57.2	61.2	47.8	69.2	54.9	56.6	41.9	70.5	61.5	67.7	60.5	68.0	60.8	66.2	54.9
$C(all, source)$	73.7	65.6	73.7	68.0	71.3	62.1	71.0	66.4	73.6	66.7	75.6	67.7	74.3	66.6	73.3	66.2
Majority (<i>source</i>)	47.0	6.1	42.3	5.6	49.0	6.2	38.8	5.3	46.1	6.0	49.0	6.2	48.9	6.2	45.9	5.9

Generalizability All three models achieve better average F_1 -scores in the $C(source, target)$ setting than the majority (*target*) baseline. However, compared to the majority (*source*) baseline, $C(target, source)$ performs worse with LSTM, comparably with fastText, and much better with BERT. This suggests that a contextualized representation, as in BERT, is necessary for value classification in novel domains, especially for the novel domains with a large moral vocabulary as is the case in $C(target, source)$.

Table B.3: Results of the four training scenarios and three models evaluated on \mathcal{T}_{target} . The columns indicate the dataset used as \mathcal{T}_{target} . For each experiment we report micro F_1 -score (m , left-hand column) and macro F_1 -score (M , right-hand column).

Classifier Setting	ALM		BLT		BLM		DAV		ELE		MT		SND		Average	
	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M
<i>LSTM</i>																
$C(source, target)$	52.5	40.2	61.7	19.3	59.6	43.2	85.9	8.5	52.7	35.7	43.3	33.3	36.9	21.8	56.1	28.9
$C(target, target)$	47.2	25.7	64.1	8.2	71.6	55.8	92.2	9.0	56.4	24.5	37.2	18.3	50.1	26.4	59.8	24.0
$C(finetune, target)$	61.4	51.2	69.0	23.2	78.2	77.2	92.2	9.0	64.7	44.6	49.6	43.3	54.7	36.8	67.1	40.8
$C(all, target)$	57.6	48.7	65.2	20.3	71.1	64.4	90.3	9.1	60.3	42.3	47.8	41.2	51.1	35.3	63.3	37.3
<i>fastText</i>																
$C(source, target)$	57.5	46.8	57.1	23.1	62.9	54.6	83.5	8.9	54.1	39.5	49.2	45.5	38.5	24.9	57.5	34.8
$C(target, target)$	62.4	50.4	69.2	18.3	77.6	74.2	92.1	9.0	63.8	39.5	49.4	40.8	57.4	34.0	67.4	38.0
$C(finetune, target)$	62.5	57.5	68.6	30.1	77.8	78.6	88.6	9.7	65.8	53.3	51.4	47.6	59.0	46.7	67.7	46.2
$C(all, target)$	61.8	55.3	66.8	30.4	75.2	75.3	88.1	9.8	63.1	51.6	52.5	49.2	57.1	45.1	66.4	45.2
<i>BERT</i>																
$C(source, target)$	63.7	57.9	63.2	29.2	76.1	75.3	83.9	8.7	63.4	54.8	54.3	51.3	49.2	38.6	64.8	45.1
$C(target, target)$	68.0	56.8	71.4	23.5	84.4	84.6	92.2	9.0	70.9	52.6	59.4	55.9	65.3	44.6	73.1	46.7
$C(finetune, target)$	69.4	67.0	72.1	37.4	84.6	85.5	92.2	9.2	72.9	65.2	61.4	59.3	66.7	55.6	74.2	54.2
$C(all, target)$	69.9	67.0	71.2	34.7	83.9	85.2	90.4	9.3	71.1	62.3	61.4	59.3	66.3	55.6	73.5	53.3
Majority ($target$)	37.9	5.1	64.8	7.4	28.3	4.2	92.2	8.7	44.5	5.7	27.9	4.4	26.4	4.0	46.0	5.6

Transferability From the average F_1 -scores in Table B.3, we observe that $C(finetune, target)$ performs better than or on par with $C(target, target)$ across all three models. The benefits of finetuning are most evident for LSTM (7% increase in the average m and 17% increase in M). The benefits can also be observed for fastText (similar m and 8% increase of M) and BERT (similar m and 8% increase of M), but to a lesser degree than LSTM.

Catastrophic Forgetting We observe that all three models suffer from catastrophic forgetting since finetuning on \mathcal{T}_{target} reduces the performance on \mathcal{T}_{source} . As mentioned in the paper, the degree of catastrophic forgetting is most evident when finetuning on unbalanced datasets such as DAV than balanced datasets such as BLM.

TRAINING TIME

In some applications, e.g., estimating value trends on Twitter, value classifiers need to be re-trained frequently since the trends can shift fast. Similarly, to employ techniques such as active learning for value annotation requires training a classifier at every iteration to prompt for new labels. In such cases, training time is an important factor for selecting an approach and model. Figure B.1 shows the average training time in logarithmic scale, for different models and scenarios (Appendix B.1.3 describes our computing infrastructure).

Two considerations are evident. First, fastText trains significantly faster than the other two models. Second, for all three models, the training time is approximately proportional to the amount of data in the training set—the $target$ and $finetune$ scenarios employ a similar amount of data, which is roughly six times smaller than in the $source$ and all scenarios.

B.2.2 COMPOSITION OF THE SOURCE DATASET

In Section 4.2.1, we mention that in our experiments \mathcal{T}_{target} is always composed of one dataset of the MFTC, while we test with \mathcal{T}_{source} being composed of one, three, or six datasets. In the main paper we present the results where \mathcal{T}_{source} is composed of six datasets. Here, we present the results where it is composed of one or three datasets, using BERT.

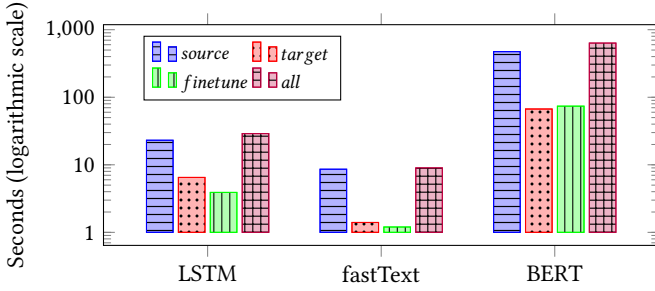


Figure B.1: Average training time (seconds) per model and scenario.

ONE DATASET AS \mathcal{T}_{source}

Not all the settings described in Section 4.2.1 can be meaningfully replicated when \mathcal{T}_{source} is composed of just one dataset. For instance, $C(source, source)$ and $C(target, target)$ would coincide, as well as $C(source, target)$ and $C(target, source)$. Thus, in Tables B.4, B.5, and B.6 we present the results along the lines of generalizability, transferability, and catastrophic forgetting, respectively. When possible, we compare the results to the results presented in the paper (where \mathcal{T}_{source} is composed of six datasets). As in the paper, we highlight in bold the best result and the results that are not significantly different from it.

Table B.4: Generalizability: the model is trained on \mathcal{T}_{source} and evaluated on \mathcal{T}_{target} .

$\mathcal{T}_{target} \rightarrow$	ALM		BLT		BLM		DAV		ELE		MT		SND	
$\mathcal{T}_{source} \downarrow$	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>
ALM	-	-	65.6	21.3	72.0	55.4	87.2	8.5	58.4	30.3	45.1	33.1	44.8	24.2
BLT	33.4	11.4	-	-	36.0	17.6	90.9	8.6	44.9	8.4	26.9	9.2	30.3	7.3
BLM	64.1	53.6	64.2	21.6	-	-	86.4	8.4	65.2	49.7	49.7	43.3	44.5	30.4
DAV	35.8	4.9	63.0	7.3	25.3	3.9	-	-	46.6	6.0	27.8	4.5	25.2	3.9
ELE	53.7	35.2	63.5	22.7	60.8	49.8	85.8	9.6	-	-	48.4	41.3	47.3	30.8
MT	47.9	43.8	58.8	20.5	54.9	48.3	49.9	6.0	54.7	41.9	-	-	41.5	29.2
SND	47.7	33.5	54.8	22.6	50.6	37.2	79.1	8.6	48.9	33.6	42.8	35.1	-	-
Six	63.7	57.9	63.2	29.2	76.1	75.3	83.9	8.7	63.4	54.8	54.3	51.3	49.2	38.6

Generalizability To evaluate generalizability (Table B.4), the model is trained on \mathcal{T}_{source} and evaluated on \mathcal{T}_{target} , akin to the $C(source, target)$ setting described in the paper. Thus, at the end of the table, we append the results of $C(source, target)$ from Table 4.1 (where \mathcal{T}_{source} is composed of six datasets). First, we notice that the results are generally better when \mathcal{T}_{source} is composed of six datasets. Further, there is no dataset that stands out as clearly better than the other six in generalizability.

Transferability To evaluate transferability (Table B.5), the model is trained on \mathcal{T}_{source} , retrained on \mathcal{T}_{target} , and evaluated on \mathcal{T}_{target} , akin to the $C(finetune, target)$ setting described in the paper. Thus, at the end of the table, we append the results of $C(finetune, target)$ from Table 4.1 (where \mathcal{T}_{source} is composed of six datasets). First, we notice that the results are generally better or on par to the results where \mathcal{T}_{source} is composed of six datasets. Further, there is no dataset that stands out as clearly better than the other six in transferability. These two aspects suggest that a combination of the six datasets as \mathcal{T}_{source} consistently leads to better transferability results.

Table B.5: Transferability: the model is trained on \mathcal{T}_{source} , retrained on \mathcal{T}_{target} , and evaluated on \mathcal{T}_{target} .

$\mathcal{T}_{target} \rightarrow$	ALM		BLT		BLM		DAV		ELE		MT		SND	
$\mathcal{T}_{source} \downarrow$	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>
ALM	-	-	74.3	31.8	85.3	86.0	89.8	8.6	72.4	62.7	61.1	58.8	67.4	54.5
BLT	69.4	58.0	-	-	82.9	83.6	91.7	8.7	72.1	62.7	58.4	55.4	65.2	47.2
BLM	66.9	60.8	72.6	33.4	-	-	92.5	8.8	72.4	66.9	61.0	59.1	68.8	62.6
DAV	23.7	13.8	68.1	16.9	56.2	43.9	-	-	46.9	33.1	29.6	16.2	46.6	25.5
ELE	68.6	61.1	72.1	36.2	82.9	83.5	92.5	8.8	-	-	60.0	58.7	66.9	53.6
MT	66.7	60.2	72.9	36.4	83.8	84.1	90.1	8.6	73.4	61.1	-	-	65.7	52.5
SND	69.6	66.7	73.9	34.7	83.6	85.1	91.9	8.7	68.7	58.8	60.7	56.4	-	-
Six	69.4	67.0	72.1	37.4	84.6	85.5	92.2	9.2	72.9	65.2	61.4	59.3	66.7	55.6

Table B.6: Catastrophic forgetting: the model is trained on \mathcal{T}_{source} , retrained on \mathcal{T}_{target} , and evaluated on \mathcal{T}_{source} .

$\mathcal{T}_{target} \rightarrow$	ALM		BLT		BLM		DAV		ELE		MT		SND		No retrain	
$\mathcal{T}_{source} \downarrow$	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>	<i>m</i>	<i>M</i>
ALM	-	-	48.4	34.6	67.0	64.3	49.2	24.3	60.6	55.2	57.2	52.5	60.1	57.7	68.0	56.8
BLT	66.0	24.0	-	-	65.7	25.8	67.6	12.7	64.8	28.6	62.5	28.9	57.6	25.7	71.4	23.5
BLM	79.4	79.8	60.2	55.4	-	-	52.2	40.5	77.7	78.1	74.9	74.5	74.5	76.6	84.4	84.6
DAV	45.1	4.3	91.5	8.7	70.3	6.9	-	-	59.9	6.3	45.0	4.9	63.1	6.6	92.2	9.0
ELE	67.6	48.0	57.8	33.1	70.0	55.3	46.5	8.4	-	-	63.8	56.7	59.8	52.8	70.9	52.6
MT	51.3	45.0	40.2	28.2	55.4	50.8	28.4	5.1	55.8	52.2	-	-	54.3	51.0	59.4	55.9
SND	54.0	37.5	39.9	20.4	55.8	41.7	26.9	4.4	55.0	43.3	57.4	47.3	-	-	65.3	44.6

Catastrophic Forgetting To evaluate catastrophic forgetting (Table B.6), the model is trained on \mathcal{T}_{source} , retrained on \mathcal{T}_{target} , and evaluated on \mathcal{T}_{source} , akin to the $C(\text{finetune}, \text{source})$ setting described in the paper. However, we cannot compare the results with the $C(\text{finetune}, \text{source})$ setting, as the evaluation sets differ (one dataset in Table B.6, six datasets in $C(\text{finetune}, \text{source})$ in Table 4.1). However, we compare to the case where the model is only trained on \mathcal{T}_{source} . Differently from the previous tables, the evaluation sets (i.e., \mathcal{T}_{source}) are consistent in every row, not in every column. Thus, we highlight the best results per row. It is evident that catastrophic forgetting happens even when \mathcal{T}_{source} is composed of one dataset. Further, there is no dataset that stands out as better than the other six in mitigating forgetting.

THREE DATASETS AS \mathcal{T}_{source}

When employing three datasets as \mathcal{T}_{source} , the settings described in Section 4.2.1 can be meaningfully reproduced. However, the selection of the three datasets (out of the six available at each experiment) that compose \mathcal{T}_{source} is not trivial. Experimenting with all possible combinations would result in $\frac{6!}{3!(6-3)!} = 20$ experiments per setting. In order to simplify the experiments, we decide to test with only one combination of three datasets, selected as the best performing combination from the experiments in Section B.2.2. We average the results of Tables B.4, B.5, and B.6, and for each dataset used as \mathcal{T}_{target} , we select the three datasets that led to the best average performance. Due to the class imbalance of all datasets, one of the biggest challenges is to achieve good performances across all values. Thus, we decide to consider only the average macro F_1 -scores. We report the best resulting datasets in Table B.7—for each dataset that we use as \mathcal{T}_{target} in the following experiments, we use the indicated three datasets as \mathcal{T}_{source} .

Table B.8 reports the complete cross-domain evaluation results, analogously to Table 4.1. For

Table B.7: The three datasets used as \mathcal{T}_{source} in Table B.8.

\mathcal{T}_{target}	\mathcal{T}_{source}
ALM	BLM, MT, SND
BLT	ELE, MT, SND
BLM	ALM, ELE, MT
DAV	BLT, BLM, ELE
ELE	BLM, MT, SND
MT	BLM, ELE, SND
SND	BLM, ELE, MT

further comparison, we add the results from Table 4.1 (where \mathcal{T}_{source} is composed of six datasets). The results in the bottom half of the table can be directly compared, as in each column the model is evaluated on the same test set. However, the results on the top half cannot be directly compared, as the model is evaluated on different test sets (three and six datasets, respectively).

It is evident that the results are consistent with the results presented in the main paper. In the top half of the table, the best performing settings are $C(source, source)$ and $C(all, source)$, both when \mathcal{T}_{source} is composed of three and six datasets. In the bottom half, where the results can be directly compared, we notice that the best performing settings are consistent, and lead to comparable results.

We conclude that selecting the three best performing datasets as \mathcal{T}_{source} has neither advantage nor disadvantage over selecting all six datasets. However, selecting all six allows for a consistent evaluation, where all MFTC datasets are used in all evaluation settings, thus avoiding the arbitrary choice of datasets to be used as \mathcal{T}_{source} that we described at the beginning of this section.

Table B.8: Results of the four training scenarios evaluated on \mathcal{T}_{source} and \mathcal{T}_{target} , when \mathcal{T}_{source} is composed of three or six datasets. The columns indicate the dataset used as \mathcal{T}_{target} . We report both micro F_1 -score (m , left column) and macro F_1 -score (M , right column).

Classifier Setting	ALM		BLT		BLM		DAV		ELE		MT		SND		Average	
	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M
Three datasets as \mathcal{T}_{source}																
$C(source, source)$	70.9	68.8	66.1	62.5	67.2	63.4	76.1	70.3	71.2	69.1	75.0	71.8	72.4	69.6	71.3	67.9
$C(target, source)$	52.8	40.7	34.2	8.8	59.1	49.4	46.3	6.0	52.9	44.3	50.6	43.8	48.3	37.3	49.2	32.9
$C(finetime, source)$	64.1	59.4	50.9	38.5	65.0	58.8	58.7	34.6	66.9	63.7	68.2	65.7	65.0	62.7	62.7	54.8
$C(all, source)$	70.9	69.1	66.3	62.6	67.1	63.4	75.9	69.8	70.9	68.9	74.6	71.5	72.5	69.7	71.2	67.9
Six datasets as \mathcal{T}_{source}																
$C(source, source)$	73.9	65.6	73.9	68.3	71.2	61.8	71.1	66.4	73.3	66.4	75.7	68.0	74.5	66.5	73.4	66.1
$C(target, source)$	61.6	37.7	43.8	13.1	62.6	43.0	38.8	5.1	59.3	40.4	52.4	39.1	54.4	36.6	53.3	30.7
$C(finetime, source)$	70.3	57.2	61.2	47.8	69.2	54.9	56.6	41.9	70.5	61.5	67.7	60.5	68.0	60.8	66.2	54.9
$C(all, source)$	73.7	65.6	73.7	68.0	71.3	62.1	71.0	66.4	73.6	66.7	75.6	67.7	74.3	66.6	73.3	66.2
Three datasets as \mathcal{T}_{target}																
$C(source, target)$	64.8	58.9	61.4	26.6	77.1	74.5	85.3	8.8	60.0	54.7	54.9	51.7	51.3	41.1	65.0	45.2
$C(target, target)$	68.1	56.8	71.1	23.3	83.8	84.2	92.2	8.7	71.0	53.6	59.1	54.9	65.2	44.7	72.9	46.6
$C(finetime, target)$	70.1	67.4	72.6	37.4	84.9	85.4	92.2	8.7	72.9	64.7	61.2	59.6	68.0	58.3	74.5	54.5
$C(all, target)$	69.6	66.2	71.2	35.0	84.0	85.1	91.0	9.3	71.7	64.2	61.0	59.2	67.8	58.3	73.7	53.9
Six datasets as \mathcal{T}_{target}																
$C(source, target)$	63.7	57.9	63.2	29.2	76.1	75.3	83.9	8.7	63.4	54.8	54.3	51.3	49.2	38.6	64.8	45.1
$C(target, target)$	68.0	56.8	71.4	23.5	84.4	84.6	92.2	9.0	70.9	52.6	59.4	55.9	65.3	44.6	73.1	46.7
$C(finetime, target)$	69.4	67.0	72.1	37.4	84.6	85.5	92.2	9.2	72.9	65.2	61.4	59.3	66.7	55.6	74.2	54.2
$C(all, target)$	69.9	67.0	71.2	34.7	83.9	85.2	90.4	9.3	71.1	62.3	61.4	59.3	66.3	55.6	73.5	53.3
Majority (target)	37.9	5.1	64.8	7.4	28.3	4.2	92.2	8.7	44.5	5.7	27.9	4.4	26.4	4.0	46.0	5.6

C

C

AN EXPLAINABLE METHOD FOR CROSS-DOMAIN COMPARISON OF MORAL VALUES

C.1 EXPERIMENTAL DETAILS

We provide here all the information needed for reproducing our experimental results.

C.1.1 DATA PREPROCESSING

We preprocess the tweets by removing URLs, emails, usernames and mentions. Next, we employ the Ekphrasis package¹ to correct common spelling mistakes and unpack contractions. Finally, emojis are transformed into their respective words using the Python Emoji package².

C.1.2 HYPERPARAMETERS

To select the hyperparameters, we trained and evaluated the model on the entire MFTC corpus with 10-fold cross-validation. Table C.1 shows the hyperparameters compared in this setting, highlighting in bold the best-performing option that we then used in the experiments described in the paper. If a parameter is not present in the table, the default value supplied by the framework is used.

C.1.3 MODEL TRAINING

As introduced in Section 5.3.1, we trained seven models on the seven domains of the MFTC, respectively. Each model was first trained on the remaining six domains and then continued training on the domain under analysis. The training on the seventh domain was performed on 90% of the domain, leaving 10% out for evaluation. Table C.2 shows the performances of the models on the portions of the domains left out for evaluation. The trained models are available online [181].

¹<https://github.com/cbaziotis/ekphrasis>

²<https://pypi.org/project/emoji/>

Table C.1: Hyperparameters tested and selected (in bold).

Hyperparameters	Options
Model name	bert-base-uncased
Number of parameters	110M
Max sequence length	64
Epochs	2, 3 , 5
Batch size	16 , 32, 64
Dropout	0.05, 0.1 , 0.02
Optimizer	AdamW
Learning Rate	$5 \cdot 10^{-5}$
Loss function	Binary Cross Entropy

Table C.2: Models performance (macro F_1 -score).

	ALM	BLT	BLM	DAV	ELE	MT	SND
F_1 -score	70.3	32.1	85.3	8.7	64.8	62.3	53.9

C.1.4 COMPUTING INFRASTRUCTURE

The following are the main libraries and computing environments used in our experiments.

- PyTorch: 1.8.1
- Huggingface’s Transformers: 4.6.0
- NVIDIA GeForce RTX 2080 Ti GPU
- CUDA: 11.2
- cuDNN: 8.1.1.33
- SHAP: 0.40.0

We spent 7 GPU hours to train the models and 70 CPU hours to generate the moral lexicons.

C.1.5 RANDOM SEEDS

In our experiments, to control for randomness, we fixed the random seeds in the following libraries:

- Python (`random.seed`)
- NumPy (`numpy.random.seed`)
- PyTorch (`torch.manual_seed`)
- CUDA (`torch.cuda.manual_seed_all`)

C.1.6 ARTIFACTS USAGE

We have mainly used three artifacts in this research: the MFTC [130], SHAP [191], and BERT [74]. The MFTC was collected with the intent of facilitating NLP research on morality. It can be downloaded³ and used under the Creative Commons Attribution 4.0 license. SHAP was intended to explain the output of any machine learning model. Thus, we are using it as originally intended, under its MIT

³<https://osf.io/k5n7y/>

license⁴. BERT was created with the intent of performing, among others, text classification. Thus, we are using it as originally intended, under its Apache 2.0 distribution license⁵.

C.2 CROWD EVALUATION

Section 3.3.3 introduces the crowd experiment. We first opened a pilot annotation job on Prolific for nine users with an expected completion time of 25 minutes. The average completion time was 21 minutes and the average ICC was 0.61. These results encouraged us to proceed with the rest of the experiment. Ultimately, the average time spent by a crowd worker on a job was 22 minutes (\pm 12 minutes SD). Each worker was paid £3.75 (at the rate of £9/h as per Prolific suggestion of fair retribution).

C

C.2.1 ANNOTATION JOB LAYOUT

Upon taking the annotation job on Prolific, workers were redirected to a web application hosted on our servers. Here, after accepting the informed consent form, they were asked demographic questions and then were given a brief introduction to the annotation tasks and the moral elements involved. Informed consent forms, instructions, and all word bubbles are available online [182].

Figure C.1 shows an example of an annotation task. In each individual task, annotators needed to indicate whether the word bubble describing domain D_A was more similar to the one describing domain D_B or D_C . The annotators were given the following six options on a Likert scale:

1. A is clearly more similar to B (than to C)
2. A is more similar to B (than to C)
3. A is slightly more similar to B (than to C)
4. A is slightly more similar to C (than to B)
5. A is more similar to C (than to B)
6. A is clearly more similar to C (than to B)

After the initial instructions, each annotator was guided through four sections. Each section contained five tasks where all word bubbles were generated for the same moral element (but multiple different domains), plus one control task (as described in Section C.2.2). Before each section, the annotator was introduced to the moral element concerned in the following section. Thus, each annotator was introduced to four different moral elements. These elements were chosen from two different moral foundations, for a total of two moral foundations per annotator. For instance, one annotation job could be composed of four annotation sections corresponding to the moral elements of *care*, *harm*, *authority*, and *subversion*, resulting in 24 annotations tasks (including four control tasks).

C.2.2 QUALITY CONTROL

The crowd workers were required to be fluent in English and have submitted at least 100 Prolific jobs with at least 95% acceptance rate. We included four control tasks, one per section. In each, the word bubbles describing D_A and D_B were identical, and different from the word bubble describing D_C . A total of 186 workers completed the job. Using the Likert options enumeration introduced in Section C.2.1, we included a worker's job in our analysis only if (1) all four control tasks were answered with options 1, 2, or 3; and (2) at least two control tasks were answered with options 1 or 2. These criteria were set before any analysis of crowd work was done. Of the 186 workers, 159 satisfied the criteria above.

⁴<https://github.com/slundberg/shap/blob/master/LICENSE>

⁵<https://github.com/google-research/bert/blob/master/LICENSE>

The following word bubbles describe the moral concept of care. Please indicate whether the word bubble A is more similar to the word bubble B or C. Please make sure to read all the words in the bubbles.

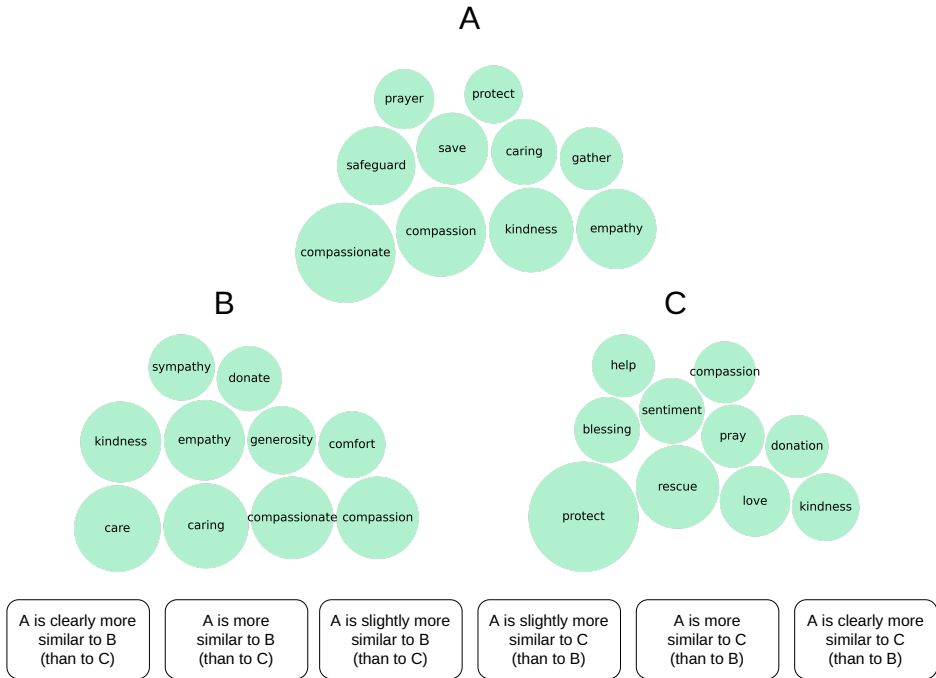


Figure C.1: The annotator is asked to take a choice on a 6-points Likert scale based on the shown word bubbles.

C.2.3 USER DEMOGRAPHICS

Upon giving informed consent, workers were asked the following demographic information:

- What is your age?
- What gender do you identify as?
- Where is your home located?
- What is the highest degree or level of education you have completed?

Figure C.2 shows the demographics of the 159 users whose submissions were considered in the study.

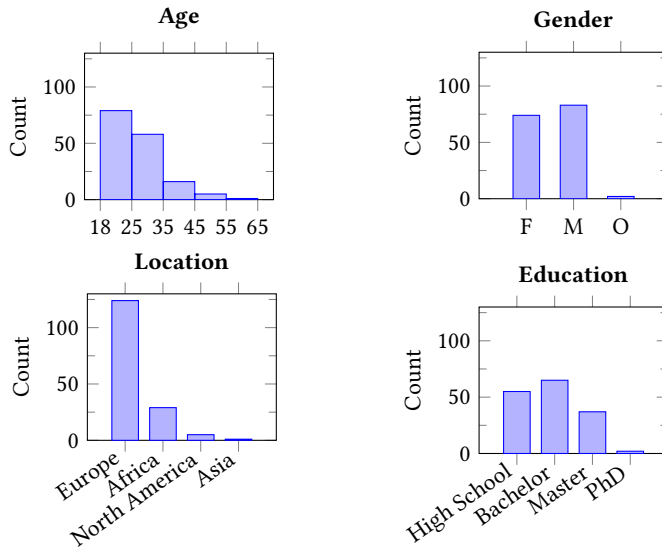


Figure C.2: Demographics of crowd workers.

C.3 EXTENDED RESULTS

We describe additional results. The raw results are available online [182].

C.3.1 *M*-DISTANCES

In Table 5.1 we show the d -distances describing the distance between domains. In tables C.3a to C.3j we display the m -distances describing the distance between domains for each moral element. For readability, we show the scores multiplied by 100.

The most apparent consideration is that moral expression similarity is not consistent across domains, but rather depends on the moral element under analysis. In Section 5.4.4 we provide examples on how to explore such fine-grained differences across domains. On top of the explored cases, another insightful example is represented by two domains that ranked with a higher distance, ALM and SND. Nevertheless, the domains ranked relatively more similar in the *care* element. Let us inspect closely the moral lexicons generated for *care* for ALM and SND. At first, we notice some differences, such as the words ‘rescue’ and ‘donation’ that are specific to the SND domain, being especially relevant in a hurricane relief domain. However, we also notice many similarities, such as the words ‘protect’ and ‘compassion’, typical for describing in-group care.

C.3.2 CORRELATION BY DOMAIN AND ELEMENT

Table C.4 shows the Spearman correlation (ρ) by moral element and domain. We notice that ρ is generally consistent across moral elements—for instance, the elements of *fairness* and *betrayal* have the highest ρ , while *purity* have the lowest. However, there are some exceptions. SND has a comparatively low ρ for *harm*, and MT for *subversion*, despite having a large number of annotations (Table 2.4). A possible reason is that the expression of these elements in these domains is less domain-specific than in other domains, leading to lower ρ with crowd intuition. Instead, DAV has a high ρ for *harm* and *betrayal*. This can be explained by the nature of the domain (hate speech), which would lead to highly specific lexicons for these elements.

Table C.3: m -distances for the ten moral elements. A darker color indicates a smaller distance between domains.

(a) m -distances for the *care* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.66	1.62	2.28	1.72	1.51	1.43
BLT	1.66	-	1.68	1.13	1.70	1.62	1.53
BLM	1.62	1.68	-	1.28	1.41	1.98	1.80
DAV	2.28	1.13	1.28	-	1.67	1.96	2.26
ELE	1.72	1.70	1.41	1.67	-	1.82	1.64
MT	1.51	1.62	1.98	1.96	1.82	-	1.61
SND	1.43	1.53	1.80	2.26	1.64	1.61	-

(b) m -distances for the *harm* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.45	1.15	2.48	1.26	1.23	1.12
BLT	1.45	-	1.44	1.85	1.34	1.33	1.38
BLM	1.15	1.44	-	2.19	1.17	1.14	1.06
DAV	2.48	1.85	2.19	-	1.69	2.15	2.11
ELE	1.26	1.34	1.17	1.69	-	1.11	1.11
MT	1.23	1.33	1.14	2.15	1.11	-	1.02
SND	1.12	1.38	1.06	2.11	1.11	1.02	-

(c) m -distances for the *fairness* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.17	1.49	2.21	1.65	1.66	1.86
BLT	2.17	-	2.34	2.24	1.96	1.98	2.09
BLM	1.49	2.34	-	2.22	1.67	1.82	1.93
DAV	2.21	2.24	2.22	-	2.14	2.17	2.49
ELE	1.65	1.96	1.67	2.14	-	1.58	1.66
MT	1.66	1.98	1.82	2.17	1.58	-	1.73
SND	1.86	2.09	1.93	2.49	1.66	1.73	-

(d) m -distances for the *cheating* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.82	1.30	2.06	1.34	1.60	1.62
BLT	1.82	-	1.84	1.79	1.63	1.62	1.75
BLM	1.30	1.84	-	2.09	1.24	1.35	1.44
DAV	2.06	1.79	2.09	-	2.06	1.98	2.31
ELE	1.34	1.63	1.24	2.06	-	1.23	1.35
MT	1.60	1.62	1.35	1.98	1.23	-	1.47
SND	1.62	1.75	1.44	2.31	1.35	1.47	-

(e) m -distances for the *loyalty* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.58	1.54	2.46	1.93	2.01	1.96
BLT	1.58	-	1.82	1.36	1.65	1.91	1.73
BLM	1.54	1.82	-	2.35	1.60	1.55	1.99
DAV	2.46	1.36	2.35	-	2.40	2.40	2.75
ELE	1.93	1.65	1.60	2.40	-	1.30	1.68
MT	2.01	1.91	1.55	2.40	1.30	-	1.59
SND	1.96	1.73	1.99	2.75	1.68	1.59	-

(f) m -distances for the *betrayal* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.02	1.34	1.75	1.19	1.21	1.13
BLT	2.02	-	1.92	2.04	1.56	1.84	1.73
BLM	1.34	1.92	-	1.69	0.85	1.12	0.90
DAV	1.75	2.04	1.69	-	1.56	1.73	1.61
ELE	1.19	1.56	0.85	1.56	-	1.05	0.87
MT	1.21	1.84	1.12	1.73	1.05	-	0.88
SND	1.13	1.73	0.90	1.61	0.87	0.88	-

(g) m -distances for the *authority* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.18	1.80	2.21	2.02	1.87	2.00
BLT	2.18	-	2.20	2.31	1.67	1.75	1.65
BLM	1.80	2.20	-	1.81	1.80	1.62	1.79
DAV	2.21	2.31	1.81	-	1.61	2.06	1.82
ELE	2.02	1.67	1.80	1.61	-	1.77	1.63
MT	1.87	1.75	1.62	2.06	1.77	-	1.58
SND	2.00	1.65	1.79	1.82	1.63	1.58	-

(h) m -distances for the *subversion* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.10	1.85	2.48	1.84	2.17	2.30
BLT	2.10	-	1.98	2.12	1.87	1.78	1.66
BLM	1.85	1.98	-	2.30	1.61	2.05	2.05
DAV	2.48	2.12	2.30	-	2.11	2.00	2.35
ELE	1.84	1.87	1.61	2.11	-	1.72	1.63
MT	2.17	1.78	2.05	2.00	1.72	-	1.84
SND	2.30	1.66	2.05	2.35	1.63	1.84	-

(i) m -distances for the *purity* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.86	1.10	1.85	2.14	1.56	2.44
BLT	2.86	-	2.78	2.29	2.24	1.98	2.40
BLM	1.10	2.78	-	1.75	1.79	1.72	1.94
DAV	1.85	2.29	1.75	-	1.61	1.71	2.00
ELE	2.14	2.24	1.79	1.61	-	1.51	1.67
MT	1.56	1.98	1.72	1.71	1.51	-	1.87
SND	2.44	2.40	1.94	2.00	1.67	1.87	-

(j) m -distances for the *degradation* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.44	1.30	1.65	1.34	1.94	1.03
BLT	1.44	-	1.27	1.77	1.11	1.47	1.40
BLM	1.30	1.27	-	1.89	1.38	1.61	1.21
DAV	1.65	1.77	1.89	-	1.77	2.40	1.44
ELE	1.34	1.11	1.38	1.77	-	1.60	1.09
MT	1.94	1.47	1.61	2.40	1.60	-	1.76
SND	1.03	1.40	1.21	1.44	1.09	1.76	-

Table C.4: Spearman correlation (ρ) between m -distances and crowd results, divided by domain and moral element. A darker color indicates a higher correlation.

	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
ALM	0.49	0.53	0.65	0.34	0.49	0.63	0.11	0.47	0.03	0.25
BLT	0.10	0.46	0.73	0.15	0.17	0.59	0.38	0.37	-0.01	0.29
BLM	0.20	0.54	0.66	0.27	0.60	0.67	0.27	0.61	0.16	0.36
DAV	0.43	0.84	0.80	0.18	0.63	0.75	0.39	0.65	-0.26	0.45
ELE	0.41	0.58	0.69	0.43	0.48	0.55	-0.11	0.70	-0.19	0.42
MT	0.36	0.50	0.76	0.24	0.51	0.53	0.25	0.30	0.08	0.44
SND	0.37	0.25	0.73	0.05	0.58	0.69	-0.01	0.47	-0.13	0.21

C

C.3.3 QUALITATIVE ANALYSIS

In Section 5.4.4 we suggest methods for qualitatively comparing moral rhetoric across domains. In particular, we show similarities and differences between two domains, ALM and BLM. These are among the most similar domains for the moral elements of *fairness* (Table C.3c) and *cheating* (Table C.3d). For both domains, the words ‘equality’ and ‘fraud’ are among the most impactful words for the two elements, respectively. In Table C.5 we show examples of tweets where these words are used, to provide additional context on their usage.

Table C.5: Examples of tweets with similar moral rhetoric in the ALM and BLM domains.

Tweet	Domain	Label
<i>Equality</i> is key. #AllLivesMatter pray over everyone. Cherish your life cause today you never know	ALM	fairness
Praying for Justice and <i>equality</i>	BLM	fairness
Of course #AllLivesMatter Shep, you self righteous, dangerously politically correct <i>fraud</i> posing as a fair journalist.	ALM	cheating
Shaun King is/was a <i>fraud</i> and a liar and deserved to be outed as such. #BlackLivesMatter deserves better.	BLM	cheating

On the other hand, ALM and BLM differ in the moral element of *subversion* (Table C.3h). Here, words such as ‘overthrow’ and ‘mayhem’ have a high impact in ALM, whereas words such as ‘encourage’ and ‘defiance’ have a high impact in BLM. In Table C.6 we show examples of tweets where these words are used, to provide additional context on their usage.

Table C.6: Examples of tweets with different moral rhetoric in the ALM and BLM domains.

Tweet	Domain	Label
I am a proponent of civil disobedience and logic driven protest only; not non irrational violence, pillage & <i>mayhem</i> !	ALM	subversion
For those who try to confuse acts of <i>defiance</i> with deliberate acts of racist terrorism, we pray	BLM	subversion

D

CLOSING THE LOOP WITH A HYBRID INTELLIGENCE APPROACH

D

D.1 NLP MODEL TRAINING

We provide additional details on the choice and the training of the NLP model used in the active learning setting to evaluate the disambiguation strategy (Section 7.3 offers an overview of the experimental settings).

The PVE corpus was originally collected in Dutch. Thus, we chose to test the state-of-the-art model in Dutch, `pdelobelle/robert-v2-dutch-base`¹. However, due to the more widespread usage of the English language in NLP models, we also decided to translate the corpus to English with the Microsoft Azure Text Translation service² and test two models trained in English—a RoBERTa model (similar to the tested Dutch model) trained on a sentiment analysis task on tweets (`cardiffnlp/twitter-roberta-base-sentiment`³) and a comparable model with a different architecture, XLNet (`xlnet-base-cased`⁴).

To select the hyperparameters, we trained and evaluated each model on the entire PVE corpus with 10-fold cross-validation. Table D.1 shows the hyperparameters that were compared in this setting, highlighting in bold the best-performing option and reporting the micro and macro F_1 -scores resulting in the best hyperparameters. If a parameter is not present in the tables, the default value supplied by the framework is used.

As noticeable, the difference between the three tested models is minimal. Thus, we decided to use the RoBERTa Dutch model to employ the original data.

¹<https://huggingface.co/pdelobelle/robert-v2-dutch-base>

²<https://azure.microsoft.com/en-us/services/cognitive-services/translator/>

³<https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

⁴<https://huggingface.co/xlnet-base-cased>

Table D.1: Hyperparameters tested and selected (in bold) and F_1 -scores resulting with the selected hyperparameters.

(a) Results with the Dutch RoBERTa model.

robert-v2-dutch-base	
Hyperparameter	Options
Model type	RoBERTa
# of parameters	125M
Max seq. length	64 , 128
Epochs	3, 4
Batch size	8 , 16, 32
Dropout	0.05, 0.1 , 0.2
F_1-score	Best Result
micro F_1 -score	0.64
macro F_1 -score	0.63

(b) Results with the English RoBERTa model.

twitter-roberta-base-sentiment	
Hyperparameter	Options
Model type	RoBERTa
# of parameters	125M
Max seq. length	64, 128
Epochs	3, 4
Batch size	8 , 16, 32
Dropout	0.05, 0.1, 0.2
F_1-score	Best Result
micro F_1 -score	0.65
macro F_1 -score	0.64

(c) Results with the English XLNet model.

xlnet-base-cased	
Hyperparameter	Options
Model type	XLNet
# of parameters	110M
Max seq. length	64 , 128
Epochs	3, 4
Batch size	8, 16 , 32
Dropout	0.05, 0.1 , 0.2
F_1-score	Best Result
micro F_1 -score	0.65
macro F_1 -score	0.64

BIBLIOGRAPHY

REFERENCES

- [1] Giulio Antonio Abboa, Serena Marchesi, Agnieszka Wykowska, and Tony Belpaeme. Do LLMs Show Traits of Value Awareness? In *Preproceedings of the Value Engineering in AI Workshop, at 26th European Conference on Artificial Intelligence (ECAI 2023)*, 2023.
- [2] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):1–43, 2011.
- [3] Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. In Neural Machine Translation, What Does Transfer Learning Transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL '20*, pages 7701–7710, Online, 2020. ACL.
- [4] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Designing ethical personal agents. *IEEE Internet Computing*, 22(2):16–22, 2018.
- [5] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI '18*, pages 28–34, Stockholm, Sweden, 2018.
- [6] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, pages 16–24, Auckland, New Zealand, 2020. IFAAMAS.
- [7] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Gusztai Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn J. M. Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8):18–28, 2020.
- [8] Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '16*, pages 1395–1404. ACL, 2016.
- [9] Tareq Al-Moslimi, Nazlia Omar, Salwani Abdullah, and Mohammed Albared. Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review. *IEEE Access*, 5:16173–16192, 2017.
- [10] Mohammed Al Owayyed, Sharon Grundmann, Merijn Bruijnes, and Willem-Paul Brinkman. Training child helpline counsellors with a BDI-based conversational agent. In *BNAIC/BeNe-Learn 2023*, Delft, the Netherlands, 2023.

- [11] Huib Aldewereld, Virginia Dignum, and Yao-Hua Tan. Design for Values in Software Development. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, pages 831–845. Springer Netherlands, 2015.
- [12] Shani Alkobi, David Sarne, Erel Segal-Halevi, and Tomer Sharbat. Eliciting Truthful Unverifiable Information. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '18, pages 1850–1852, Stockholm, Sweden, 2018. IFAAMAS.
- [13] Gordon W Allport. *Pattern and growth in personality*. Holt, Reinhart & Winston, 1961.
- [14] Wael Alsafery, Omer Rana, and Charith Perera. Sensing within smart buildings: A survey. *ACM Computing Surveys*, 55(13s):1–35, 2023.
- [15] Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL '22, pages 8782–8797, Dublin, Ireland, 2022. ACL.
- [16] Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. Aligning to Social Norms and Values in Interactive Narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '22, pages 5994–6017, Seattle, USA, 2022. ACL.
- [17] Elizabeth Anderson. The Epistemology of Democracy. *Episteme: A Journal of Social Epistemology*, 3(1-2):8–22, 2006.
- [18] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. *Knowledge-Based Systems*, 191: 1–29, 2020.
- [19] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. The Language of Liberty: A preliminary study. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 623–626, Ljubljana, Slovenia, 2021. ACM.
- [20] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. LibertyMFD: A Lexicon to Assess the Moral Foundation of Liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, GoodIT '22, page 154–160, New York, NY, USA, 2022. ACM.
- [21] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.
- [22] Luigi Asprino, Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci, and Misael Mongiovi. Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, DeeLIO '22, pages 33–41, Dublin, Ireland, 2022. ACL.
- [23] Mohammad Atari, Aida Mostafazadeh Davani, Drew Kogon, Brendan Kennedy, Nripsuta Ani Saxena, Ian Anderson, and Morteza Dehghani. Morally Homogeneous Networks and Radicalism. *Social Psychological and Personality Science*, 12:1–11, 2021.
- [24] Various Authors. Mahabharata, 300.

- [25] Reyhan Aydogan, Özgür Kafali, Furkan Arslan, Catholijn M. Jonker, and Munindar P. Singh. Nova: Value-based Negotiation of Norms. *ACM Transactions on Intelligent Systems and Technology*, 12(4):1–29, 2021.
- [26] Mohamed Bahgat, Steven R. Wilson, and Walid Magdy. Towards Using Word Embedding Vector Space for Better Cohort Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '20, pages 919–923, Atlanta, Georgia, 2020. AAAI Press.
- [27] Ondrej Bajgar and Jan Horenovsky. Negative Human Rights as a Basis for Long-term AI Safety and Regulation. *Journal of Artificial Intelligence Research*, 76:1043–1075, 2023.
- [28] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, NeurIPS '22, pages 38176–38189. Curran Associates, Inc., 2022.
- [29] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online ai systems. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI '19, pages 3–11, Honolulu, Hawaii, USA, 2019. AAAI Press.
- [30] Jerry Banks. *Handbook of Simulation*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1998.
- [31] Gagan Bansal. Explanatory dialogs: Towards actionable, interactive explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 356–357, New Orleans, LA, USA, 2018. ACM.
- [32] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, SDM '04, pages 333–344, Orlando, Florida, USA, 2004. Society for Industrial and Applied Mathematics.
- [33] David Beetham. Political legitimacy. In *The Blackwell Companion to Political Sociology*, pages 107–116. Malden and Oxford: Blackwell, New York City, NY, USA, 2001.
- [34] Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, and Sameer Singh. Zero-and few-shot nlp with pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 32–37, 2022.
- [35] Roland Benabou, Armin Falk, Luca Henkel, and Jean Tirole. Eliciting Moral Preferences: Theory and Experiment. Technical report, Princeton University, 2020.
- [36] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. Payne and Son, London, 1789.
- [37] Shruti Bhargava, Anand Dhoot, Ing-marie Jonsson, Hoang Long Nguyen, Alkesh Patel, Hong Yu, and Vincent Renkens. Referring to screen texts with voice assistants. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, ACL '23, pages 752–762, Toronto, Canada, 2023. ACL.
- [38] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

- [39] Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. Multilingual Transfer Learning for QA Using Translation as Data Augmentation. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI '21*, pages 12583–12591, Online, 2021.
- [40] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 2014.
- [41] Dries H. Bostyn, Sybren Sevenhant, and Arne Roets. Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas. *Psychological Science*, 29(7):1084–1093, 2018.
- [42] Ryan L. Boyd, Steven R. Wilson, James W. Pennebaker, Michal Kosinski, David J. Stillwell, and Rada Mihalcea. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media, ICWSM '15*, pages 31–40, Okford, UK, 2015. The AAAI Press.
- [43] Engin Bozdag and Jeroen van den Hoven. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265, 2015.
- [44] Johan Brännmark. Moral disunitarianism. *The Philosophical Quarterly*, 66(264):481–499, 11 2015.
- [45] Samuel Butler. *Erewhon*. Trubner & Co., London, 1872.
- [46] Federico Cabitza, Andrea Campagner, Domenico Albano, Alberto Aliprandi, Alberto Bruno, Vito Chianca, Angelo Corazza, Francesco Di Pietto, Angelo Gambino, Salvatore Gitto, et al. The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Applied Sciences*, 10(11):4014, 2020.
- [47] Federico Cabitza, Andrea Campagner, and Valerio Basile. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI '23*, pages 6860–6868, 2023.
- [48] James Cameron. *The terminator*, 1984.
- [49] Jay Carriere, Hareem Shafi, Katelyn Brehon, Kiran Pohar Manhas, Katie Churchill, Chester Ho, and Mahdi Tavakoli. Case Report: Utilizing AI and NLP to Assist with Healthcare and Rehabilitation During the COVID-19 Pandemic. *Frontiers in Artificial Intelligence*, 4(2):1–7, 2021.
- [50] Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. Addressing Imbalance in Multilabel Classification: Measures and Random Resampling Algorithms. *Neurocomputing*, 163:3–16, 2015.
- [51] Kushal Chawla, Rene Clever, Jaysa Ramirez, Gale M Lucas, and Jonathan Gratch. Towards emotion-aware agents for improved user satisfaction and partner perception in negotiation dialogues. *IEEE Transactions on Affective Computing*, 2023.
- [52] Pei-Yu Chen, Myrthe L Tielman, Dirk KJ Heylen, Catholijn M Jonker, and M Birna Van Riemsdijk. Acquiring semantic knowledge for user model updates via human-agent alignment dialogues. In *HHAI 2023: Augmenting Human Intellect*, pages 93–107. IOS Press, 2023.

- [53] Yann Chevalere, Paul Dunne, Ulle Endriss, Jerome Lang, Michel Lemaitre, Nicolas Maudet, Julian Padget, Steve Phelps, Juan A Rodriguez-Aguilar, and Paulo Sousa. Issues in multiagent resource allocation. *Informatica*, 30:3–31, 2006.
- [54] Kinzang Chhogyal, Abhaya Nayak, Aditya Ghose, and Hoa K. Dam. A Value-based Trust Assessment Model for Multi-agent Systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI '19, pages 194–200, 2019.
- [55] Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrera, et al. Can language model moderators improve the health of online discourse? *arXiv preprint arXiv:2311.10781*, 2023.
- [56] Yejin Choi. Common sense: The dark matter of language and intelligence. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 2. IFAAMAS, 2023.
- [57] Amit K. Chopra and Munindar P. Singh. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 48–53, New Orleans, LA, 2018. ACM.
- [58] Norman Cliff. *Ordinal methods for behavioral data analysis*. Psychology Press, Hove, East Sussex, UK, 2014.
- [59] European Commission. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Technical report, European Union, 2023.
- [60] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI '17, pages 4831–4835, San Francisco, California, USA, 2017. AAAI Press.
- [61] T.D. Cook and D.T. Campbell. *Quasi-experimentation – Design and Analysis Issues for Field Settings*. Houghton Mifflin Company, Boston, USA, 1979.
- [62] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in bdi agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI '17, pages 178–184, Melbourne, Australia, 2017. The AAAI Press.
- [63] Erich Dallhammer, Roland Gaugitsch, Wolfgang Neugebauer, and Kai Böhme. Spatial planning and governance within eu policies and legislation and their relevance to the new urban agenda. Technical report, European Committee of the Regions: Bruxelles, Belgium, 2018.
- [64] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A Survey of the State of Explainable AI for Natural Language Processing. In *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*, ACL '20, page 447–459, Suzhou, China, 2020. URL <https://aclanthology.org/2020.aacl-main.46.pdf>.
- [65] Georg Datler, Wolfgang Jagodzinski, and Peter Schmidt. Two theories on the test bench: Internal and external validity of the theories of Ronald Inglehart and Shalom Schwartz. *Social Science Research*, 42(3):906–925, 2013.

- [66] Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM '17*, pages 512–515, 2017.
- [67] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP '17*, pages 2055–2066. ACL, 2017.
- [68] Maike H. de Boer, Jasper van der Waa, Sophie van Gent, Quirine TS Smit, Wouter Korteling, Robin M van Stokkum, and Mark Neerincx. A contextual Hybrid Intelligent System Design for Diabetes Lifestyle Management. In *International Workshop Modelling and Representing Context, ECAI '23, Krakow, Poland, 2023*.
- [69] Jacques de Wet, Daniela Wetzelhütter, and Johann Bacher. Revisiting the trans-situationality of values in Schwartz’s Portrait Values Questionnaire. *Quality and Quantity*, 53(2):685–711, 2018.
- [70] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: Values, norms and culture in multi-agent systems. *Artificial Intelligence and Law*, 21(1): 79–107, 2013.
- [71] Marie Delacre, Daniël Lakens, and Christophe Leys. Why psychologists should by default use Welch’s t-Test instead of Student’s t-Test. *International Review of Social Psychology*, 30(1): 92–101, 2017.
- [72] Pieter Delobelle, Thomas Winters, and Bettina Berendt. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online, 2020. ACL.
- [73] Christian Detweiler and Maaïke Harbers. Value stories: Putting human values into requirements engineering. *CEUR Workshop Proceedings*, 1138:2–11, 2014.
- [74] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '19*, page 4171–4186, 2019.
- [75] Janis L Dickinson, Poppy McLeod, Robert Bloomfield, and Shorna Allred. Which moral foundations predict willingness to make lifestyle changes to avert climate change in the USA? *PLoS ONE*, 11(10):1–11, 2016.
- [76] Thomas Dietz. Bringing values and deliberation to science communication. *Proceedings of the National Academy of Sciences of the United States of America*, 110(3):14081–14087, 2013.
- [77] Thomas Dietz and Paul C. Stern. Toward a theory of choice: Socially embedded preference construction. *Journal of Socio-Economics*, 24(2):261–279, 1995. doi: 10.1016/1053-5357(95)90022-5.
- [78] Virginia Dignum. Responsible Autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI '17*, pages 4698–4704, Melbourne, Australia, 2017. AAAI Press.

- [79] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 314–331, Online, 2020. ACL.
- [80] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:1–17, 2021.
- [81] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 135–145, Regensburg, Germany, 2022. ACM.
- [82] Chunnging Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 4019–4028, 2020.
- [83] Olive Jean Dunn. Multiple Comparisons Using Rank Sums. *Technometrics*, 6(3):241–252, 1964.
- [84] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 7949–7962, Online, 2020. ACL.
- [85] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 698–718, Online and Punta Cana, Dominican Republic, 2021. ACL.
- [86] Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsnér, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie Catherine de Marneffe. Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '19, pages 2223–2234, Minneapolis, Minnesota, USA, 2019. ACL.
- [87] Norman T Feather. Values, valences, and choice: The influences of values on the perceived attractiveness and choice of alternatives. *Journal of personality and social psychology*, 68(6): 1135, 1995.
- [88] Fabian Ferrari, José van Dijck, and Antal van den Bosch. Observe, inspect, modify: Three conditions for generative ai governance. *New Media & Society*, pages 1–19, 2023.
- [89] Maria Angela Ferrario, Will Simm, Stephen Forshaw, Adrian Gradinar, Marcia Tavares Smith, and Ian Smith. Values-first SE: Research principles in practice. *Proceedings of the 38th International Conference on Software Engineering*, pages 553–562, 2016.
- [90] Christopher Flathmann, Beau G Schelble, Rui Zhang, and Nathan J McNeese. Modeling and guiding the creation of ethical human-AI teams. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 469–479, 2021.
- [91] Luciano Floridi and Mariarosaria Taddeo. What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160360, 2016.

- [92] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 653–670, Online, 2020. ACL.
- [93] Chiara Franco and Claudia Ghisetti. What shapes the “value-action” gap? the role of time perception reconsidered. *Economia Politica*, 39(3):1023–1053, 2022.
- [94] Batya Friedman and David G. Hendry. The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1145–1148, 2012.
- [95] Batya Friedman and David G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press, 2019.
- [96] Batya Friedman, Peter H. Kahn, and Alan Borning. Value sensitive design and information systems. In *The Handbook of Information and Computer Ethics*, pages 69–101. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2008.
- [97] Dennis Friess and Christiane Eilders. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339, 2015.
- [98] Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoțiu-Pietro. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC '16, pages 3730–3736, 2016.
- [99] Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL '21, Online, 2021. ACL.
- [100] Jason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437, 2020.
- [101] Justin Garten, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Deghani. Dictionaries and distributions: Combining Expert Knowledge and Large Scale Textual Data Content Analysis: Distributed Dictionary Representation. *Behavior Research Methods*, 50(1):344–361, 2018.
- [102] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory*. Aldine Publishing, Chicago, Illinois, USA, 1967.
- [103] Jacinto González-Pachón and Carlos Romero. Distance-based consensus methods: a goal programming approach. *Omega*, 27(3):341–347, 1999.
- [104] Valdiney V. Gouveia, Taciano L. Milfont, and Valeschka M. Guerra. Functional theory of human values: Testing its content and structure hypotheses. *Personality and Individual Differences*, 60: 41–47, 2014.
- [105] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology*, 96(5): 1029–1046, 2009.

- [106] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366, 2011.
- [107] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, 47:55–130, 2013.
- [108] Stephan Grimmelikhuijsen and Albert Meijer. Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, 5(3):232–242, 2022.
- [109] Miguel Grinberg. *Flask Web Development: Developing Web Applications with Python*. O’Reilly Media, Inc., Boston, Massachusetts, USA, 2018.
- [110] Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. A survey on deep learning for human activity recognition. *ACM Computing Surveys*, 54(8): 1–34, 2021.
- [111] Den Haag. National Climate Agreement-The Netherlands. Technical report, Dutch Ministry of Economic Affairs and Climate, 2019.
- [112] Rafik Hadfi and Takayuki Ito. Augmented Democratic Deliberation: Can Conversational Agents Boost Deliberation in Social Media? In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’22, pages 1794–1798, Online, 2022. IFAAMAS.
- [113] Rafik Hadfi, Jawad Haqbeen, Sofia Sahab, and Takayuki Ito. Argumentative Conversational Agents for Online Discussions. *Journal of Systems Science and Systems Engineering*, 30(4): 450–464, 2021.
- [114] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart J. Russell. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, NeurIPS ’16, pages 3916–3924, Barcelona, Spain, 2016. Curran Associates, Inc.
- [115] Catherine Hafer and Dimitri Landa. Deliberation as self-discovery and institutions for political speech. *Journal of Theoretical Politics*, 19(3):329–360, 2007.
- [116] Meera Hahn, Amit Raj, and James M. Rehg. Which way is ’right’?: Uncovering limitations of vision-and-language navigation models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’23, page 2415–2417. IFAAMAS, 2023.
- [117] Johanna Hall, Mark Gaved, and Julia Sargent. Participatory Research Approaches in Times of Covid-19: A Narrative Literature Review. *International Journal of Qualitative Methods*, 20: 1–15, 2021.
- [118] Kevin A. Hallgren. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol*, 8(1):23–34, 2012.
- [119] William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP ’16, pages 595–605, Austin, TX, USA, 2016. ACL.

- [120] Paul H.P. Hanel, Lukas F. Litzellachner, and Gregory R. Maio. An empirical comparison of human value models. *Frontiers in Psychology*, 9(9):1–14, 2018.
- [121] Helen Heath and Sarah Cowley. Developing a grounded theory approach: a comparison of Glaser and Strauss. *International Journal of Nursing Studies*, 41(2):141–150, 2 2004.
- [122] Samaneh Heidari, Maarten Jensen, and Frank Dignum. Simulations with values. In Harko Verhagen, Melania Borit, Giangiacomo Bravo, and Nanda Wijermans, editors, *Advances in Social Simulation*, pages 201–215. Springer International Publishing, Cham, 2020.
- [123] Willem J. Heiser and Antonio D’Ambrosio. Clustering and prediction of rankings within a kemeny distance framework. In *Algorithms from and for Nature and Life*, pages 19–31. Springer International Publishing, Cham, 2013.
- [124] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI With Shared Human Values. In *Proceedings of the Ninth International Conference on Learning Representations, ICLR ’21*, pages 1–29, Online, 2021.
- [125] Mireille Hildebrandt. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law*, 20(1):83–121, 2019.
- [126] Patrick L. Hill and Daniel K. Lapsley. Persons and situations in the moral domain. *Journal of Research in Personality*, 43(2):245–246, 2009.
- [127] Richard D. Hipp. SQLite, 2020.
- [128] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1):1–26, 2011.
- [129] Myles Hollander and Douglas A. Wolfe. *Nonparametric Statistical Methods*. Wiley, New York, USA, 1999.
- [130] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.
- [131] Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. The extended Moral Foundations Dictionary (eMFD): Development and Applications of a Crowd-Sourced Approach to Extracting Moral Intuitions from Text. *Behavior Research Methods*, 53: 232–246, 2021.
- [132] Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-Tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL ’18, pages 328–339, 2018.
- [133] Xiaolei Huang, Alexandra Wormley, and Adam Cohen. Learning to Adapt Domain Shifts of Moral Values via Instance Weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT ’22)*, pages 121–131, Barcelona, Spain, 2022. ACM.

- [134] Ioana Hulpus, Jonathan Kobbe, Heiner Stuckenschmidt, and Graeme Hirst. Knowledge Graphs meet Moral Values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, *SEM '20, pages 71–80, Barcelona, Spain (Online), 2020. ACL.
- [135] Luca Iandoli, Ivana Quinto, Anna De Liddo, and Simon Buckingham Shum. On online collaboration and construction of shared knowledge: Assessing mediation capability in computer supported argument visualization tools. *Journal of the Association for Information Science and Technology*, 67(5):1052–1067, 2016.
- [136] Ronald Inglehart. Modernization and postmodernization in 43 societies. In *Modernization and Postmodernization*, pages 67–107. Princeton University Press, 1997.
- [137] Anatol Itten and Niek Mouter. When digital mass participation meets citizen deliberation: combining mini-and maxi-publics in climate policy-making. *Sustainability*, 14(8):4656, 2022.
- [138] Ole Sejer Iversen, Kim Halskov, and Tuck Wah Leong. Rekindling values in Participatory Design. In *Proceedings of the 11th Biennial Participatory Design Conference*, pages 91–100, Sydney, Australia, 2010. ACM.
- [139] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.
- [140] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 2177–2190. ACL, 2020.
- [141] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. When to make exceptions: Exploring language models as accounts of human moral judgment. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35 of *NeurIPS '22*, pages 28458–28473. Curran Associates, Inc., 2022.
- [142] Kristen Johnson and Dan Goldwasser. Classification of Moral Foundations in Microblog Political Discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 720–730, Melbourne, Australia, 2018. ACL.
- [143] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, ACL '18, pages 116–121, Melbourne, Australia, 2018. ACL.
- [144] Kyriaki Kalimeri, Mariano G. Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, 92:428–445, 2019.
- [145] Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. Human values and attitudes towards vaccination in social media. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 248–254, 2019.

- [146] Amir Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pages 353–362, New York, NY, USA, 2021. ACM.
- [147] John G. Kemeny and L. J. Snell. Preference ranking: an axiomatic approach. In *Mathematical models in the social sciences*, pages 9–23. Ginn New York, 1962.
- [148] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. Moral Concerns are Differentially Observable in Language. *Cognition*, 212:104696, 2021.
- [149] Jasper O Kenter, Mark S Reed, and Ioan Fazey. The deliberative value formation model. *Ecosystem Services*, 21:194–207, 2016.
- [150] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL '22*, pages 4459–4471, Dublin, Ireland, 2022. ACL.
- [151] Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. SemEval-2023 Task 4: ValueEval: Identification of Human Values behind Arguments. In *17th International Workshop on Semantic Evaluation, SemEval '23*, pages 2290–2306, Toronto, Canada, July 2023. ACL.
- [152] Hyunwoo Kim, Eun Young Ko, Donghoon Han, Sung Chul Lee, Simon T. Perrault, Jihee Kim, and Juho Kim. Crowdsourcing perspectives on public policy from stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pages 1–6, Glasgow, UK, 2019. ACM.
- [153] Tae Wan Kim, John Hooker, and Thomas Donaldson. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, 70:871–890, 2021.
- [154] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017.
- [155] Mark Klein. Enabling Large-Scale Deliberation Using Attention-Mediation Metrics. *Computer Supported Cooperative Work (CSCW)*, 21(4-5):449–473, 2012.
- [156] Mark Klein. How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium. Technical report, Center for Collective Intelligence, 2012.
- [157] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27, 2018.
- [158] Jonathan Kobbe, Ines Rehbein, Ioana Hulpus, and Heiner Stuckenschmidt. Exploring Morality in Argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, online, 2020. ACL.

- [159] Ilir Kola, Ralvi Isufaj, and Catholijn M. Jonker. Does Personalization Help? Predicting How Social Situations Affect Personal Values. In *HAI2022: Augmenting Human Intellect*, pages 157–169, 2022.
- [160] Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, ACL '23, pages 707–719, Toronto, Canada, 2023. ACL.
- [161] Cristina Lafont. Deliberation, participation, and democratic legitimacy: Should deliberative mini-publics shape public policy? *Journal of political philosophy*, 23(1):40–63, 2015.
- [162] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking Explainability as a Dialogue: A Practitioner’s Perspective. In *Proceedings of the Workshop on Human-Centered AI @ NeurIPS*, HCAI '22, pages 1–23, Online, 2022. Curran Associates, Inc.
- [163] Alexander Lam. Balancing fairness, efficiency and strategy-proofness in voting and facility location problems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 1806–1807, Online, 2021. IFAAMAS.
- [164] Alex Gwo Jen Lan and Ivandr e Paraboni. Text- and author-dependent moral foundations classification. *New Review of Hypermedia and Multimedia*, 0(0):1–21, 2022.
- [165] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, 1977.
- [166] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.
- [167] Christopher A. Le Dantec, Erika Shehan Poole, and Susan P. Wyche. Values as Lived Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1141–1150. ACM Press, 2009.
- [168] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar. Towards Pluralistic Value Alignment: Aggregating Value Systems through ℓ_p -Regression. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pages 780–788, Online, 2022. IFAAMAS.
- [169] Roger X. Lera-Leri, Enrico Liscio, Filippo Bistaffa, Catholijn M. Jonker, Maite Lopez-Sanchez, Pradeep K. Murukannaiah, Juan A. Rodriguez-Aguilar, and Francisco Salas-Molina. Aggregating value systems for decision support. *Knowledge-Based Systems*, 287:111453, 2024.
- [170] Zhizhong Li and Derek Hoiem. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.
- [171] Q. Vera Liao and Michael Muller. Enabling Value Sensitive AI Systems through Participatory Design Fictions, 2019.
- [172] Catherine Y. Lim, Andrew B.L. Berry, Andrea L. Hartzler, Tad Hirsch, David S. Carrell, Zo e A. Bermet, and James D. Ralston. Facilitating Self-reflection about Values and Self-care among Individuals with Chronic Conditions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, UK, 2019. ACM.

- [173] Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. Acquiring Background Knowledge to Improve Moral Value Prediction. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '18, pages 552–559, Barcelona, Spain, 8 2018. IEEE.
- [174] Enrico Liscio, Michiel van der Meer, Catholijn M. Jonker, and Pradeep K. Murukannaiah. A Collaborative Platform for Identifying Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 1773–1775, Online, 2021. IFAAMAS.
- [175] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. Axes: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 799–808, Online, 2021. IFAAMAS.
- [176] Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep Murukannaiah. Axes: Identifying and Evaluating Context Specific Values - supplemental material, 2021. <https://doi.org/10.4121/13705423>.
- [177] Enrico Liscio, Alin E. Dondera, Andrei Geadau, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Cross-Domain Classification of Moral Values - supplemental material, 2022. <https://doi.org/10.4121/518aac3c-27ef-4309-8523-805580c35035>.
- [178] Enrico Liscio, Alin E. Dondera, Andrei Geadau, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Cross-Domain Classification of Moral Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, USA, 2022. ACL.
- [179] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. What Values Should an Agent Align With? *Autonomous Agents and Multi-Agent Systems*, 36(23):32, 2022.
- [180] Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M. Jonker, Kyriaki Kalimeri, and Pradeep K. Murukannaiah. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, ACL '23, pages 14113–14132, Toronto, Canada, 2023. ACL.
- [181] Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M. Jonker, Kyriaki Kalimeri, and Pradeep K. Murukannaiah. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric - models, 2023. <https://doi.org/10.4121/646b20e3-e24f-452d-938a-bcb6ce30913c>.
- [182] Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M. Jonker, Kyriaki Kalimeri, and Pradeep K. Murukannaiah. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric - supplemental material, 2023. <https://doi.org/10.4121/cd43b76a-850e-4222-ab81-5a20b7b1b93d>.
- [183] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. Value Inference in Sociotechnical Systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1774–1780, London, United Kingdom, 2023. IFAAMAS.

- [184] Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. On interpretation of network embedding via taxonomy induction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pages 1812–1820. ACM, 2018.
- [185] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023.
- [186] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.
- [187] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 13470–13479, 2021.
- [188] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 13480–13488, 2021.
- [189] Kate Loveys, Gabrielle Sebaratnam, Mark Sagar, and Elizabeth Broadbent. The effect of design features on relationship quality with embodied conversational agents: a systematic review. *International Journal of Social Robotics*, 12(6):1293–1312, 2020.
- [190] Birgit Lugrin, Catherine Pelachaud, and David Traum. *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*. ACM, 2022.
- [191] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, NeurIPS '17, pages 1208–1217, Long Beach, CA, USA, 2017.
- [192] Iliia Markov and Walter Daelemans. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online, 2021. ACL.
- [193] Marissa McNeace and Jeffrey Sinn. Moral Foundations Theory vs. Schwartz Value Theory: Which Theory Best Explains Ideological Differences? *The Winthrop McNair Research Bulletin*, 4(6), 2018.
- [194] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 2021.
- [195] Siddharth Mehrotra. Modelling Trust in Human-AI Interaction. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 1814–1816, Online, 2021. IFAAMAS.
- [196] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. More Similar Values, More Trust? The Effect of Value Similarity on Trust in Human-Agent Interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 777–783. Association for Computing Machinery, 2021.

- [197] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. Integrity Based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Transactions on Interactive Intelligent Systems*, 2023.
- [198] Rijk Mercurur, Virginia Dignum, and Catholijn M. Jonker. The value of values and norms in social simulation. *Journal of Artificial Societies and Social Simulation*, 22(1):1–9, 2019.
- [199] Jessica K. Miller, Batya Friedman, Gavin Jancke, and Brian Gill. Value tensions in design: The value sensitive design, development, and appropriation of a corporation’s groupware system. In *Proceedings of the International ACM Conference on Supporting Group Work*, GROUP ’07, pages 281–290, 2007.
- [200] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
- [201] Sören Mindermann and Stuart Armstrong. Occam’s razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems*, NeurIPS ’18, pages 5598–5609, Montreal, Canada, 2018. Curran Associates, Inc.
- [202] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2): 2053951716679679, 2016.
- [203] Omid Mohamad Beigi and Mohammad H. Moattar. Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification. *Knowledge-Based Systems*, 213:106423, 2021.
- [204] Nieves Montes and Carles Sierra. Value-Guided Synthesis of Parametric Normative Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’21, pages 907–915, Online, 2021. IFAAMAS.
- [205] Nieves Montes and Carles Sierra. Synthesis and Properties of Optimally Value-Aligned Normative Systems. *Journal of Artificial Intelligence Research*, 74:1739–1774, 2022.
- [206] Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, 2(6): 389–396, 2018.
- [207] Javier Morales, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Michael J. Wooldridge, and Wamberto Vasconcelos. Automated synthesis of normative systems. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’13, pages 483–490, Saint Paul, Minnesota, USA, 2013. IFAAMAS.
- [208] Francesca Mosca and Jose M Such. ELVIRA: an Explainable Agent for Value and Utility-driven Multiuser Privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’21, pages 916–924, Online, 2021. IFAAMAS.
- [209] Davoud Mougouei, Harsha Perera, Waqar Hussain, Rifat Shams, and Jon Whittle. Operationalizing human values in software: A research roadmap. *ESEC/FSE 2018 - Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 780–784, 2018.

- [210] Niek Mouter, Shannon Spruit, Anatol Itten, Jose Ignacio Hernandez, Lisa Volberda, and Sjoerd Jennings. Resultaten van een raadpleging onder 30.000 nederlanders over de versoepeling van coronamaatregelen. Technical report, Delft University of Technology, 2020.
- [211] Niek Mouter, Jose Ignacio Hernandez, and Anatol Valerian Itten. Public participation in crisis policymaking. How 30,000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures. *PLoS ONE*, 16(5):1–42, 2021.
- [212] Nikola Mrkšić, Diarmuid Séaghdha, Blaise Thomson, Milica Gasić, Lina Rojas-Barahona, Pei Hao Su, David Vandyke, Tsung Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '16*, pages 142–148, San Diego, California, USA, 2016. ACL.
- [213] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI conference on human factors in computing systems, CHI '23*, pages 1–16, 2021.
- [214] Pradeep K Murukannaiah. Reasoning about context and engineering context-aware agents. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '14*, pages 1733–1734, Paris, France, 2014. IFAAMAS.
- [215] Pradeep K. Murukannaiah and Munindar P. Singh. Xipho: Extending tropos to engineer context-aware personal agents. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '14*, pages 309–316, Paris, France, 2014. IFAAMAS.
- [216] Pradeep K. Murukannaiah and Munindar P. Singh. From machine ethics to internet ethics: Broadening the horizon. *IEEE Internet Computing*, 24(3):51–57, 2020.
- [217] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn J. M. Jonker, and Munindar P. Singh. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, pages 1706–1710, Auckland, 2020. IFAAMAS.
- [218] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL '21*, pages 5356–5371, Online, 2021. ACL.
- [219] Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. Value scenarios: A technique for envisioning systemic effects of new technologies. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, pages 2585–2590, 2007.
- [220] Andrew Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00*, page 663–670, Stanford, CA, USA, 2000. Cambridge University Press.
- [221] Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP '21*, pages 5414–5426, Online and Punta Cana, Dominican Republic, 2021. ACL.

- [222] Richard E Nisbett and Timothy D Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231, 1977.
- [223] Pablo Noriega, Harko Verhagen, Julian Padget, and Mark D’Inverno. Ethical Online AI Systems Through Conscientious Design. *IEEE Internet Computing*, 25(6):58–64, 2021.
- [224] Nardine Osman, Luc Steels, and Katarzyna Budzynska. The first workshop in Value Engineering in AI (VALE 2023). In *26th European Conference on Artificial Intelligence, ECAI ’23*, Krakow, Poland, 2023.
- [225] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK, 1990.
- [226] Ovid. *Metamorphoses (Book 11)*. Various editions, 8.
- [227] Jeongwoo Park, Enrico Liscio, and Pradeep K. Murukannaiah. Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1–20, St. Julian’s, Malta, 2024. ACL.
- [228] Matheus C. Pavan, Vitor G. Santos, Alex G. J. Lan, Joao Martins, Wesley Ramos Santos, Caio Deutsch, Pablo B. Costa, Fernando C. Hsieh, and Ivandre Paraboni. Morality Classification in Natural Language Text. *IEEE Transactions on Affective Computing*, 3045(c):1–8, 2020.
- [229] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic Inquiry and Word Count (LIWC). *Mahway: Lawrence Erlbaum Associates*, 71, 2001.
- [230] Harsha Perera, Waqar Hussain, Jon Whittle, Arif Nurwidiantoro, Davoud Mougouei, Rifat Ara Shams, and Gillian Oliver. A study on the prevalence of human values in software engineering publications, 2015 - 2018. In *Proceedings of the 42nd International Conference on Software Engineering*, pages 409–420, 2020.
- [231] Harsha Perera, Gunter Mussbacher, Waqar Hussain, Rifat Ara Shams, Arif Nurwidiantoro, and Jon Whittle. Continual human value analysis in software development: A goal model based approach. In *Proceedings of the IEEE International Conference on Requirements Engineering*, pages 192–203, 2020.
- [232] Klara Pigmans, Neelke Doorn, Huib Aldewereld, and Virginia Dignum. Decision-Making in Water Governance: From Conflicting Interests to Shared Values. In *Responsible Innovation*, pages 165–178. Springer, 2017.
- [233] Klara Pigmans, Huib Aldewereld, Virginia Dignum, and Neelke Doorn. The Role of Value Deliberation to Improve Stakeholder Participation in Issues of Water Governance. *Water Resources Management*, 33(12):4067–4085, 2019.
- [234] Maria Silvia Pini, Francesca Rossi, Kristen Brent Venable, and Toby Walsh. Aggregating partially ordered preferences: impossibility and possibility results. In *Proceedings of the 10th conference on Theoretical aspects of rationality and knowledge*, pages 193–206, 2005.
- [235] Hanna Pitkin. Obligation and Consent–I. *The American Political Science Review*, 59(4):990–999, 1965.
- [236] Hanna Pitkin. Obligation and Consent–II. *The American Political Science Review*, 60(1):39–52, 1966.

- [237] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn M. Jonker. Self-Reflection on Personal Values to Support Value-Sensitive Design. In *Proceedings of the 25th BCS Conference on Human Computer Interaction*, HCI '11, pages 491–496, Newcastle-upon-Tyne, UK, 2011. BCS Learning & Development.
- [238] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn M. Jonker. Elicitation of Situated Values: Need for Tools to Help Stakeholders and Designers to Reflect and Communicate. *Ethics and Information Technology*, 14(4):285–303, 2012.
- [239] Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. Development and Validation of the Personal Values Dictionary: A Theory-Driven Tool for Investigating References to Basic Human Values in Text. *European Journal of Personality*, 34(5):885–902, 2020.
- [240] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 2022.
- [241] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijn De Bie, and Peter Flach. Face-feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 344–350, New York, NY, USA, 2020. ACM.
- [242] Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wager. Deconfounded Lexicon Induction for Interpretable Social Science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '18, pages 1615–1625, New Orleans, Louisiana, USA, 2018.
- [243] Valentina Pyatkin, Jena D Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. Clarifydelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 11253–11271, 2023.
- [244] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. ValueNet: A New Dataset for Human Value Driven Dialogue System. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, AAAI '22, pages 11183–11191, 2022.
- [245] Xiaoye Qu, Zhikang Zou, Yu Cheng, Yang Yang, and Pan Zhou. Adversarial category alignment network for cross-domain sentiment classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '19, pages 2496–2508, Minneapolis, Minnesota, USA, 2019. ACL.
- [246] Chirag Raman. *Towards Artificial Social Intelligence in the Wild: Sensing, Synthesizing, Modeling, and Perceiving Nonverbal Social Human Behavior*. PhD thesis, Delft University of Technology, 2023.
- [247] John Rawls. *A Theory of Justice*. Oxford University Press, Oxford, 1973.
- [248] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP '19, pages 3973–3983, Hong Kong, China, 2019. ACL.

- [249] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9):1–40, 2021.
- [250] Nicholas Rescher. *Introduction to Value Theory*. Prentice-Hall, 1969.
- [251] Rezvaneh Rezapour, Saamil H. Shah, and Jana Diesner. Enhancing the Measurement of Social Effects by Capturing Morality. In *Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Minneapolis, Minnesota, USA, 2019.
- [252] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.
- [253] David A. Robb, Xingkun Liu, and Helen Hastie. Explanation styles for trustworthy autonomous systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 2298–2300. IFAAMAS, 2023.
- [254] Milton Rokeach. *The Nature of Human Values*. Free Press, New York, USA, 1973.
- [255] Subendhu Rongali, Beiye Liu, Liwei Cai, Konstantine Arkoudas, Chengwei Su, and Wael Hamza. Exploring Transfer Learning For End-to-End Spoken Language Understanding. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 13754–13761, Online, 2021.
- [256] Daniel J. Rosenkrantz, Richard E. Stearns, and Philip M. Lewis II. An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing*, 6(3):563–581, 1977.
- [257] Nicholas A Roy, Junkyung Kim, and Neil Rabinowitz. Explainability via causal self-talk. In *Advances in Neural Information Processing Systems*, NeurIPS '22, pages 1–16, New Orleans, LA, USA, 2022. Curran Associates, Inc.
- [258] Jonathan Rubin, Jason Crowley, George Leung, Morteza Ziyadi, and Maria Minakova. Entity contrastive learning in a large-scale virtual assistant system. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, ACL '23, pages 159–171, Toronto, Canada, 2023. ACL.
- [259] Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, NUI Galway, 2019.
- [260] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, New York, NY, USA, 2019.
- [261] Stuart J. Russell, Daniel Dewey, and Max Tegmark. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, 36(4):105–114, 2015.
- [262] Mostafa Saket. Putting Values in Context: an augmentation of Value Sensitive Design (VSD). *Journal of Ethics and Emerging Technologies*, 31(2):1–9, 2021.

- [263] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 5477–5490, Online, 2020. ACL.
- [264] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality and Quantity*, 52(4):1893–1907, 2018.
- [265] M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, R. Pinho Dos Santos, H. Vina-greiro Alves, E. Vecchione, and Scheunemann L. Values and Identities - a policymaker's guide – Executive summary. Technical report, Publications Office of the European Union, 2021.
- [266] Samuel Scheffler. Valuing. In *Equality and Tradition: Questions of Value in Moral and Political Theory*, chapter 7, page 352. Oxford University Press, Oxford, UK, 1st edition, 2012.
- [267] Chelsea Schein. The Importance of Context in Moral Judgments. *Perspectives on Psychological Science*, 15(2):207–215, 2020.
- [268] Shalom H. Schwartz. An Overview of the Schwartz Theory of Basic Values. *Online readings in Psychology and Culture*, 2(1):1–20, 2012.
- [269] Nicolas Schwind, Emir Demirovic, Katsumi Inoue, and Jean Marie Lagniez. Partial robustness in team formation: Bridging the gap between robustness and resilience. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 1142–1150, Online, 2021. IFAAMAS.
- [270] Marc Serramia, Maite Lopez-Sanchez, and Juan A. Rodríguez-Aguilar. A Qualitative Approach to Composing Value-Aligned Norm Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '20, pages 1233–1241, 2020.
- [271] Marc Serramia, Maite Lopez-Sanchez, Stefano Moretti, and Juan A. Rodríguez-Aguilar. On the dominant set selection problem and its application to value alignment. *Autonomous Agents and Multi-Agent Systems*, 35(42), 2021.
- [272] Burr Settles. *Active Learning*. Morgan & Claypool, 2012.
- [273] Yiting Shen, Steven R. Wilson, and Rada Mihalcea. Measuring Personal Values in Cross-Cultural User-Generated Content. In *Proceedings of the 10th International Conference on Social Informatics*, SocInfo '19, pages 143–156. Springer, 2019.
- [274] Shuming Shi, Enbo Zhao, Wei Bi, Deng Cai, Leyang Cui, Xinting Huang, Haiyun Jiang, Duyu Tang, Kaiqiang Song, Longyue Wang, Chenyan Huang, Guoping Huang, Yan Wang, and Piji Li. Effdit: An assistant for improving writing efficiency. In Danushka Bollegala, Ruihong Huang, and Alan Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, ACL '23, pages 508–515, Toronto, Canada, 2023. ACL.
- [275] Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep K. Murukannaiah, and Catholijn M. Jonker. Reason against the machine? Future directions for mass online deliberation. *Frontiers in Political Science*, 4:1–17, 10 2022.

- [276] Luciano C. Siebert, Enrico Liscio, Pradeep K. Murukannaiah, Lionel Kaptein, Shannon L. Spruit, Jeroen van den Hoven, and Catholijn M. Jonker. Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect*, pages 114–127, Amsterdam, the Netherlands, 2022. IOS Press.
- [277] Luciano C. Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker, Jeroen van den Hoven, Deborah Forster, and Reginald L. Legendijk. Meaningful human control: actionable properties for AI system development. *AI and Ethics*, 5(1):1–15, 2022.
- [278] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 557–565, 2021.
- [279] Amika M. Singh and Munindar P. Singh. Wasabi: A conceptual model for trustworthy artificial intelligence. *Computer*, 56(2):20–28, 2023.
- [280] Munindar P. Singh. Consent as a Foundation for Responsible Autonomy. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, AAAI '22, pages 12301–12306, Online, 2022. The AAAI Press.
- [281] Munindar P. Singh and Pradeep K. Murukannaiah. Toward an Ethical Framework for Smart Cities and the Internet of Things. *IEEE Internet Computing*, 27(2):51–56, 2023.
- [282] Nate Soares. The value learning problem. Technical report, Machine Intelligence Research Institute, Berkeley, California, USA, 2014.
- [283] Nate Soares and Benya Fallenstein. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity: Managing the Journey*, pages 103–125. Springer, Berlin, 2017.
- [284] Daniel J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3): 477–560, 2006.
- [285] Eliza Strickland. Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big. *IEEE Spectrum*, 59(4):22–50, 2022.
- [286] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '22, pages 769–779, Seattle, USA, 2022.
- [287] Yue Tan, Bo Wang, Anqi Liu, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. Guiding dialogue agents to complex semantic targets by dynamically completing knowledge graph. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6506–6518, Toronto, Canada, 2023. ACL.
- [288] Livia Teernstra, Peter van der Putten, Liesbeth Noordegraaf-Eelens, and Fons Verbeek. The morality machine: Tracking moral values in tweets. In *Advances in Intelligent Data Analysis XV: 15th International Symposium*, IDA '16, pages 26–37, Stockholm, Sweden, 2016. Springer.

- [289] Thiago Teixeira, Gershon Dublon, and Andreas Savvides. A survey of human-sensing: Methods for detecting presence, count, location, track, and identity. *ACM Computing Surveys*, 5(1): 59–69, 2010.
- [290] Sarah Thew and Alistair Sutcliffe. Value-based requirements engineering: method and experience. *Requirements Engineering*, 23(4):443–464, 2018.
- [291] Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '19*, pages 2062–2068, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.
- [292] Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaaard. Spurious correlations in cross-topic argument mining. In *Proceedings of the Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021*, pages 263–277, Online, 2021. ACL.
- [293] Myrthe L. Tielman, Catholijn M. Jonker, and M. Birna Van Riemsdijk. Deriving Norms from Actions, Values, and Context. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '19*, pages 2223–2225, 2019.
- [294] Anna Tiginova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. Listening Between the Lines: Learning Personal Attributes from Conversations. In *Proceedings of the 2019 World Wide Web Conference, WWW '19*, pages 1818–1828, 2019.
- [295] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15, 2023.
- [296] Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*, 2022.
- [297] Mario Triola. *Elementary Statistics*. Pearsons, 13th edition, 2017.
- [298] Andrea Aler Tubella, Andreas Theodorou, Frank Dignum, and Virginia Dignum. Governance by glass-box: Implementing transparent moral bounds for AI behaviour. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI '19*, pages 5787–5793, 2019.
- [299] Sanna Tuomela, Netta Iivari, and Rauli Svento. User values of smart home energy management system: sensory ethnography in vsd empirical investigation. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia, MUM '19*, pages 1–12, Pisa, Italy, 2019. ACM.
- [300] Sz-Ting Tzeng. Engineering Normative and Cognitive Agents with Emotions and Values. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS '22*, pages 1878–1880, Online, 2022. IFAAMAS.
- [301] Alexandra N Uma, Dirk Hovy, Barbara Plank, and Massimo Poesio. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.
- [302] Tilburg University. European Value Study, 2021. <https://europeanvaluesstudy.eu>.

- [303] Ibo van de Poel. Translating values into design requirements. In *Philosophy and Engineering: Reflections on Practice, Principles and Process*, pages 253–266. Springer Netherlands, Dordrecht, Netherlands, 2013.
- [304] Ibo van de Poel, Tristan de Wildt, and Dyami van Kooten Pássaro. COVID-19 and Changing Values. In *Values for a Post-Pandemic Future*, pages 23–58. Springer International Publishing, 2022.
- [305] Tom G.C. van den Berg, Maarten Kroesen, and Caspar G. Chorus. Does morality predict aggressive driving? A conceptual analysis and exploratory empirical investigation. *Transportation Research Part F: Traffic Psychology and Behaviour*, 74(1):259–271, 2020.
- [306] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *HHAI2022: Augmenting Human Intellect*, pages 17–31, Amsterdam, the Netherlands, 2022. IOS Press.
- [307] W. Fred van Raaij and Theo M. M. Verhallen. Domain-specific market segmentation. *European Journal of Marketing*, 28(10):49–66, 1994.
- [308] Francisco Vargas and Ryan Cotterell. Exploring the Linear Subspace Hypothesis in Gender Bias Mitigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP '20*, pages 2902–2913, 2020.
- [309] Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, Proceedings of Machine Learning Research, pages 265–293. PMLR, 2022.
- [310] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems, NeurIPS '17*, pages 5998–6008, Long Beach, CA, USA, 2017.
- [311] Zina B. Ward. On value-laden science. *Studies in History and Philosophy of Science Part A*, 85: 54–62, 2021.
- [312] Caleb Warren, A. Peter McGraw, and Leaf Van Boven. Values and preferences: Defining preference construction. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(2):193–205, 2011.
- [313] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations, ICLR '22*, 2022.
- [314] Nathaniel Weir, Xingdi Yuan, Marc-Alexandre Côté, Matthew Hausknecht, Romain Laroche, Ida Momennejad, Harm Van Seijen, and Benjamin Van Durme. One-Shot Learning from a Demonstration with Hierarchical Latent Language. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23*, page 2388–2390. IFAAMAS, 2023.
- [315] Andrew Wen, Sunyang Fu, Sungrim Moon, Mohamed El Wazir, Andrew Rosenbaum, Vinod C. Kaggal, Sijia Liu, Sunghwan Sohn, Hongfang Liu, and Jungwei Fan. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *npj Digital Medicine*, 2(130):1–7, 2019.

- [316] Garrett Wilson and Diane J. Cook. A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5), 2020.
- [317] Kelly G. Wilson, Emily K. Sandoz, Jennifer Kitchens, and Miguel Roberts. The valued living questionnaire: Defining and measuring valued action within a behavioral framework. *Psychological Record*, 60(2):249–272, 2010.
- [318] Steven R. Wilson, Yiting Shen, and Rada Mihalcea. Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In *Proceedings of the 10th International Conference on Social Informatics*, SocInfo '18, pages 455–470, St. Petersburg, Russia, 2018. Springer.
- [319] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artificial Intelligence*, 300(1):103554, 2021.
- [320] Emily Winter, Steve Forshaw, and Maria Angela Ferrario. Measuring human values in software engineering. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 10–13, 2018.
- [321] Jessica Woodgate and Nirav Ajmeri. Macro Ethics for Governing Equitable Sociotechnical Systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pages 1824–1828, Online, 2022. IFAAMAS.
- [322] Fangzhao Wu and Yongfeng Huang. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 301–310, Berlin, Germany, 2016. Association for Computational Linguistics.
- [323] WVSA. World Value Survey, 2022. <https://www.worldvaluessurvey.org/wvs.jsp>.
- [324] Chenxing Xie, Yanding Wang, and Yang Cheng. Does Artificial Intelligence Satisfy You? A Meta-Analysis of User Gratification and User Satisfaction with AI-Powered Chatbots. *International Journal of Human-Computer Interaction*, 40(3):613–623, 2022.
- [325] Weilai Xu, Fred Charles, and Charlie Hargood. Generating stylistic and personalized dialogues for virtual agents in narratives. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 737–746. IFAAMAS, 2023.
- [326] Hua Yuan, Jie Zheng, Qiongwei Ye, Yu Qian, and Yan Zhang. Improving Fake News Detection with Domain-Adversarial and Graph-Attention Neural Network. *Decision Support Systems*, 53: 113633, 2021.
- [327] Ye Zhang, Matthew Lease, and Byron C. Wallace. Active Discriminative Text Representation Learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, AAAI '17, pages 3386–3392, San Francisco, California, USA, 2017.
- [328] Zhisong Zhang, Emma Strubell, and Eduard Hovy. A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, pages 6166–6190. ACL, 2022.
- [329] Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. Active Learning Approaches to Enhancing Neural Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online, 2020. ACL.

-
- [330] Yilun Zhou, Marco Tulio Ribeiro, and Julie Shah. Exsum: From local explanations to model understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '22*, pages 5359–5378, Seattle, USA, 2022. ACL.
- [331] Luisa M. Zintgraf, Diederik M. Roijers, Sjoerd Linders, Catholijn M. Jonker, and Ann Nowé. Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '18*, pages 1477–1485, Stockholm, Sweden, 2018. IFAAMAS.
- [332] Tomasz Zurek, Tom van Engers, and Jonathan Kwik. Values, proportionality, and uncertainty in military autonomous devices. In *Preproceedings of the Value Engineering in AI Workshop, at 26th European Conference on Artificial Intelligence (ECAI 2023)*, 2023.

SIKS DISSERTATIONS

- 2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
- 02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
- 03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
- 04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
- 05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
- 06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
- 07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
- 08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
- 09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
- 10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
- 11 Anne Schuth (UvA), Search Engines that Learn from Their Users
- 12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
- 13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
- 14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
- 15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
- 16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
- 17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
- 18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
- 19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
- 20 Daan Odijk (UvA), Context & Semantics in News & Web Search
- 21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
- 22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
- 23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
- 24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
- 25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
- 26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
- 27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
- 28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

-
- 29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
 - 30 Ruud Mattheij (TiU), The Eyes Have It
 - 31 Mohammad Khelghati (UT), Deep web content monitoring
 - 32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
 - 33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
 - 34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
 - 35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
 - 36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
 - 37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
 - 38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
 - 39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
 - 40 Christian Detweiler (TUD), Accounting for Values in Design
 - 41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
 - 42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
 - 43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
 - 44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
 - 45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
 - 46 Jorge Gallego Perez (UT), Robots to Make you Happy
 - 47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
 - 48 Tanja Buttler (TUD), Collecting Lessons Learned
 - 49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
 - 50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
-
- 2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
 - 02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
 - 03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
 - 04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
 - 05 Mahdiah Shadi (UvA), Collaboration Behavior
 - 06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
 - 07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
 - 08 Rob Konijn (VUA), Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
 - 09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
 - 10 Robby van Delden (UT), (Steering) Interactive Play Behavior
 - 11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

- 12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
- 13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
- 14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
- 15 Peter Berck (RUN), Memory-Based Text Correction
- 16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
- 17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
- 18 Ridho Reinanda (UvA), Entity Associations for Search
- 19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
- 20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
- 21 Jeroen Linszen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
- 22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
- 23 David Graus (UvA), Entities of Interest – Discovery in Digital Traces
- 24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
- 25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
- 26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
- 27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
- 28 John Klein (VUA), Architecture Practices for Complex Contexts
- 29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
- 30 Wilma Latuny (TiU), The Power of Facial Expressions
- 31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
- 32 Thaer Samar (RUN), Access to and Retrieval of Content in Web Archives
- 33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
- 34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
- 35 Martine de Vos (VUA), Interpreting natural science spreadsheets
- 36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
- 37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
- 38 Alex Kayal (TUD), Normative Social Applications
- 39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
- 40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
- 41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
- 42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
- 43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

-
- 44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
 - 45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
 - 46 Jan Schneider (OU), Sensor-based Learning Support
 - 47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
 - 48 Angel Suarez (OU), Collaborative inquiry-based learning
-
- 2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
 - 02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
 - 03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
 - 04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
 - 05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
 - 06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
 - 07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
 - 08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
 - 09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
 - 10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
 - 11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
 - 12 Xixi Lu (TU/e), Using behavioral context in process mining
 - 13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
 - 14 Bart Joosten (TIU), Detecting Social Signals with Spatiotemporal Gabor Filters
 - 15 Naser Davarzani (UM), Biomarker discovery in heart failure
 - 16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
 - 17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
 - 18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
 - 19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
 - 20 Manxia Liu (RUN), Time and Bayesian Networks
 - 21 Aad Sloomaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
 - 22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
 - 23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
 - 24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
 - 25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
 - 26 Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
 - 27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
 - 28 Christian Willemsse (UT), Social Touch Technologies: How they feel and how they make you feel
 - 29 Yu Gu (TIU), Emotion Recognition from Mandarin Speech
 - 30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
-

- 2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
- 02 Emmanuelle Beaux Ausalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
- 03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
- 04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
- 05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
- 06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
- 07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
- 08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
- 09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
- 10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
- 11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
- 12 Jacqueline Heinerman (VUA), Better Together
- 13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
- 14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
- 15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
- 16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
- 17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
- 18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
- 19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
- 20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
- 21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
- 22 Martin van den Berg (VUA), Improving IT Decisions with Enterprise Architecture
- 23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
- 24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
- 25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
- 26 Prince Singh (UT), An Integration Platform for Sychromodal Transport
- 27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
- 28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
- 29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
- 30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
- 31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
- 32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
- 33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks

-
- 34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
- 35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
- 36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
- 37 Jian Fang (TUD), Database Acceleration on FPGAs
- 38 Akos Kadar (OU), Learning visually grounded and multilingual representations
-
- 2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
- 02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
- 03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
- 04 Maarten van Gompel (RUN), Context as Linguistic Bridges
- 05 Yulong Pei (TU/e), On local and global structure mining
- 06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
- 07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
- 08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search
- 09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
- 10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
- 11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation Methods for Long-Tail Entity Recognition Models
- 12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
- 13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
- 14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
- 15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
- 16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
- 17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
- 18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
- 19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
- 20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
- 21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
- 22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
- 23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
- 24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
- 25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
- 26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer Optimization
- 27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
- 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
- 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

- 30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
 - 31 Gongjin Lan (VUA), Learning better – From Baby to Better
 - 32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
 - 33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
 - 34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
 - 35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
-
- 2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
 - 02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
 - 03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
 - 04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
 - 05 Davide Dell’Anna (UU), Data-Driven Supervision of Autonomous Systems
 - 06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
 - 07 Armel Lefebvre (UU), Research data management for open science
 - 08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
 - 09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children’s Collaboration Through Play
 - 10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
 - 11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
 - 12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
 - 13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
 - 14 Negin Samaemofrad (UL), Business Incubators: The Impact of Their Support
 - 15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
 - 16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
 - 17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
 - 18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
 - 19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
 - 20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
 - 21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
 - 22 Sihang Qiu (TUD), Conversational Crowdsourcing
 - 23 Hugo Manuel Proença (UL), Robust rules for prediction and description
 - 24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
 - 25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
 - 26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

-
- 27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
- 28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
-
- 2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
- 02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
- 03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
- 04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
- 05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
- 06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
- 07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
- 08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
- 09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
- 10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
- 11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
- 12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
- 13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
- 14 Michiel Overeem (UU), Evolution of Low-Code Platforms
- 15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
- 16 Pieter Gijbbers (TU/e), Systems for AutoML Research
- 17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
- 18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
- 19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
- 20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
- 21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
- 22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
- 23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
- 24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
- 25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
- 26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
- 27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
- 28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

-
- 29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
 - 30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
 - 31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
 - 32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
 - 33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
 - 34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
 - 35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
-
- 2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
 - 02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
 - 03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
 - 04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
 - 05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
 - 06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
 - 07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
 - 08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
 - 09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
 - 10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
 - 11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
 - 12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
 - 13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
 - 14 Selma Čaušević (TUD), Energy resilience through self-organization
 - 15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
 - 16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
 - 17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
 - 18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
 - 19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
 - 20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
 - 21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
 - 22 Alireza Shojaifar (UU), Volitional Cybersecurity
 - 23 Theo Theunissen (UU), Documentation in Continuous Software Development
 - 24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
 - 25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs

-
- 26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
 - 27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
 - 28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
 - 29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
-
- 2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
 - 02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
 - 03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
 - 04 Mike Huisman (UL), Understanding Deep Meta-Learning
 - 05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
 - 06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
 - 07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
 - 08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
 - 09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
 - 10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
 - 11 Mahmoud Shokrollahi-Far (TiU), Computational Reliability of Quranic Grammar
 - 12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
 - 13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence

ACKNOWLEDGMENTS

I have reached the end of my journey as a Ph.D. candidate, and I could not have made it without the support of all of you.

I start by thanking the external defense committee members: Dr. Ibo van de Poel, Dr. Piek Vossen, Dr. Pinar Yolum, and Dr. Munindar Singh. Thank you for your feedback and especially for the availability in such a short time frame.

Pradeep, you've been the best supervisor I could ask for. We are similar enough that we could easily get along both from a personal and a professional side, but also different enough that I could learn so much from you. You let me explore my ideas and follow my intuitions but taught me where I needed most teaching. It is easy to work with you and you always know how to contribute. I am glad we will continue to work together.

Catholijn, you've been an excellent guide since the day I suddenly walked into your office. You have inspired me to become a scientist. You have taught me the value of collaborating across fields and of patient diplomacy. Through you, I have come to understand the true essence of a passion for science.

I want to thank all the collaborators that I had the pleasure to work with in these years. Michiel, our countless hours on Skype have been a source of comfort and fun during the challenges of beginning a Ph.D. during a pandemic. You are my first collaborator and a true friend. Luciano, you're the one who (unintentionally) pulled me into this Ph.D., thank you! You've been my first mentor outside of my supervisors. The amazing group that hosted me in Barcelona: Roger, Filippo, Jar, and Maite. Being there with you has been a truly formative experience from a personal and professional level, thank you for having me. The unforeseen collaboration with a Spanish (Oscar), an Italian (Lorenzo), and a Greek (Kyriaki), which sounds like the beginning of a bad joke, was instead lovely and successful. One day we will eventually meet in person! Tristan and Francesca, even though our proposal was unsuccessful, I enjoyed writing it with you. Neil, thank you for the guidance when I most needed it. Next, I want to thank all my students, starting with the students with whom I wrote papers (Alin, Andrei, Ionut, Jeongwoo), followed by the other students that I have worked with and am still working with (Dragos, Florentin, Mei Lan, Zhiheng, Bianca, Kenzo, Kirsten, Nathaniël, Rob, Luca, Jahson, Antonio). Finally, all the other collaborators that I had the pleasure to work with: Roel (a fellow Systems & Control student), Niek, Shannon, Aske, Lionel, Jeroen, and Francisco. I have learned something from every one of you.

I couldn't have made this without my colleagues from the Interactive Intelligence group. Starting a Ph.D. during a pandemic is not easy, but we eventually carved out great relationships. Thank you Mani (it's so much fun hanging out with you, man), Carolina (a friend always ready to give the best advice), Nele (the bland food specialist), Mo (my favorite nerd), Masha (the best ceramist), Sid (we started and we'll finish together), Zuzanna (the party organizer), Amir (thank you for the help with the BNA presentation), Laxmi & Edgar (to complete the best office ever), Ruben (my London roommate), Elena (l'unica in ufficio durante la pandemia), Davide (incredibile avere un collega dei Romiti), Pei-Yu (it's not my fault, it's Chirag that started it), Paul, Deborah, Emma, Micha, Sietze, Antonio, Morita, Rolf, Miguel, Ilir, Willem-Paul, Mark (thank you for the availability as reserve member), Myrthe, Frans, Catha, Frank, Yanzhe, Jinke, Tina, Stephanie, Luuk, Agnes, Sandy, Yu-Wen, Joanna, Aleks, Fran, Merijn, Eric, and all the other present and past II members. Thanks also to the support staff, Ruud, Bart, Wouter, and especially Anita (the most invaluable member of the group).

The Hybrid Intelligence consortium has been my extended academic family. So many meet-ups and so much quality time spent together, especially in Vlieland. Thank you Bram (please don't tell anyone that a Dutch knows more about wine than me), Selene (the sweetest one), Tae (the craziest one), Chirag (I think Pei-Yu still hates us, I blamed it on you), Davide (it was great organizing the SIG with you), Jasper (thank you for the help with the propositions), Urja (it was fun in Seattle!), Tiffany (I hope you enjoyed my tips for New York), Bernd (yes, I'm still this tall), Lea, Íñigo, Andreas, Davide, Merle, Anna, Loan, Emre, Annet, Wijnand, Davide, Mark, Niklas, Kata, Maria, Sharvaree, Cor, Nicole, J.D., Ludi, Putra, Delaram, Johanna, and all other present and past HI members. And thank you Frank and Wendy for all the patient organization.

A special thank you to my closest friends in the Netherlands: Barbara, Davide, Silvia, Simone, and Valerio. You are like a second family to me. Even though we have reached the season finale, you've been the best sitcom partners I could ask for. I also want to thank all the other friends in Delft (thank you for having been and still being like an extended family), in Fizyr (leaving was a hard decision, but I am glad I got to meet you), in Romagna (sono stra contento di essere ancora in contatto con tutti voi nonostante la distanza), in Bologna (è stato super divertente studiare assieme), in DIG (jams were so much fun), in the Netherlands, in Italy, and all over the world. The list is too long to be spelled out here, but each of you has made my journey special with your support and the shared moments we've had together.

Finally, I want to thank my family. Mamma, babbo, è grazie a voi che sono la persona che sono oggi. So che vi chiedete ancora da dove sia scappato fuori un ingegnere, ma è tutto merito vostro. Mi avete insegnato la perseveranza e il senso del dovere. Grazie per il vostro costante supporto. E grazie al resto della famiglia per essere sempre stati al mio fianco.

Kamila, you are more than a friend, you are more than family. I am so happy to have you in my life. You are my first supporter, motivator, and manager. I am so looking forward to seeing where life will bring us, together. Kocham cię. Bardzo.

CURRICULUM VITÆ

Enrico LISCIO

20/09/1993 Born in Forlì, Italy.

EDUCATION

- 2020–2024 **Ph.D. in Computer Science**
Delft University of Technology, the Netherlands
- 2022 **Visiting Researcher**
IIIA-CSIC, Barcelona, Spain
- 2015–2017 **Master of Science in Systems & Control**
Delft University of Technology, the Netherlands
- 2012–2015 **Bachelor of Science in Automation Engineering**
Università di Bologna, Italy

EXPERIENCE

- 2017–2020 **Fizyr**, Delft, the Netherlands
Project Lead (Oct '18 – Mar '20)
Deep Learning Developer (Sep '17 – Sep '18)
- 2016–2017 **Heemskerk Innovative Technology**, Delft, the Netherlands
Graduate Intern

LIST OF PUBLICATIONS

2024

- 1. Roger Lera-Leri, **Enrico Liscio**, Filippo Bistaffa, Catholijn M. Jonker, Maite Lopez-Sanchez, Pradeep K. Murukannaiah, Juan A. Rodriguez-Aguilar, Francisco Salas-Molina. 2024. Aggregating Value Systems for Decision Support. In *Knowledge-Based Systems*, 287, 111453.
- 2. Jeongwoo Park*, **Enrico Liscio***, Pradeep K. Murukannaiah. 2024. Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning. In *Findings of the Association for Computational Linguistics: EACL 2024*, St Julian's, Malta, ACL, 654-673.
- 3. Michiel van der Meer, **Enrico Liscio**, Catholijn M. Jonker, Aske Plaat, Piek Vossen, Pradeep K. Murukannaiah. A Hybrid Intelligence Method for Argument Mining. To appear in *Journal of Artificial Intelligence Research (JAIR)*.

2023

- 1. **Enrico Liscio**, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M. Jonker, Kyriaki Kalimeri, Pradeep K. Murukannaiah. 2023. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL '23*, Toronto, Canada, ACL, 14113-14132.
- 2. **Enrico Liscio**, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Pradeep K. Murukannaiah. 2023. Value Inference in Sociotechnical Systems: Blue Sky Ideas Track. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23*, London, United Kingdom, IFAAMAS, 1774-1780.
- 3. **Enrico Liscio**, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Pradeep K. Murukannaiah. 2023. Inferring Values via Hybrid Intelligence: Poster Track. In *HHAI2023: Augmenting Human Intellect*, Munich, Germany, IOS Press, 373-378.
- 4. **Enrico Liscio**, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M. Jonker, Kyriaki Kalimeri, Pradeep K. Murukannaiah. 2023. Tomea: an Explainable Method for Comparing Morality Classifiers across Domains. In *35th Benelux Conference on Artificial Intelligence and 32nd Belgian-Dutch Conference on Machine Learning, BNAIC '23*, Delft, the Netherlands, 1-4.


2022


- 1. **Enrico Liscio**, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2022. What Values should an Agent Align with? An Empirical Comparison of General and Context-Specific Values. In *Autonomous Agents and Multi-Agent Systems*, 36, 23.
- 2. **Enrico Liscio**, Alin E. Dondera, Andrei Geadau, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2022. Cross-Domain Classification of Moral Values. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '22*, Seattle, WA, USA, ACL, 2727-2745. [**Reproducibility Badge (top 3%)**]

- 3. Luciano C. Siebert, **Enrico Liscio**, Pradeep K. Murukannaiah, Lionel Kaptein, Shannon Spruit, Jeroen van den Hoven, Catholijn M. Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect*, Amsterdam, the Netherlands, IOS Press, 114-127. **[Best paper award finalist]**
- 4. **Enrico Liscio**, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2022. Identifying Context-Specific Values via Hybrid Intelligence: Poster Track. In *HHAI2022: Augmenting Human Intellect*, Amsterdam, the Netherlands, IOS Press, 298-301. **[Best poster award]**
- 5. Michiel van der Meer, **Enrico Liscio**, Catholijn M. Jonker, Aske Plaat, Piek Vossen, Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *HHAI2022: Augmenting Human Intellect*, Amsterdam, the Netherlands, IOS Press, 17-31. **[Best paper award]**

2021

- 1. **Enrico Liscio**, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, Pradeep K. Murukannaiah. 2021. Axes: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '21*, Online, IFAAMAS, 799-808.
- 2. **Enrico Liscio**, Michiel van der Meer, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2021. A Collaborative Platform for Identifying Context-Specific Values: Demo Track. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '21*, Online, IFAAMAS, 1773-1775.

 Included in this Thesis.

 Won a best paper, tool demonstration, or proposal award.

* Equal contribution.