Universität Leipzig
Faculty for Mathematics and Computer Science
Institute for Computer Science

Identifying the Human Values behind Arguments

Bachelor’s Thesis

Leipzig, September 22, 2022

handed in by

Handke, Nicolas
Degree Programme Computer Science
Matriculation Number 3734894

1. Referee: Jun.-Prof. Dr. Martin Potthast
Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemässe Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, 22. September 2022

............................................................

Handke, Nicolas
Abstract

Values play a significant role in guiding human behavior, however, they are also related to the way people evaluate situations and how people form their opinions. Priorities on these values (e.g. should we consider *having a world at peace* to be worth more striving for than *having wealth*) form the basis of each person’s value system and the differences between humans are considered an important factor on the formation of opposing sides in controversial argumentation. However, acknowledging another person’s value priorities through the often implicit usage of values in one’s arguments could allow for a better understanding and the possible creation of convincing arguments designed for a specific target audience. The main goal is to open up controversial topics and allow an exchange of opinions beyond the bounds of intercultural understanding and topic-dependent knowledge. For this matter, the thesis at hand contributes a multi-level taxonomy consisting of 54 personal human values derived from various fields of social science. The crowd sourced dataset of 5270 arguments from four different geographical cultures and annotated for each level of the taxonomy is presented alongside. Additionally, this work presents promising baseline results with $F_1$-scores up to 0.81 and 0.25 on average, regarding the first attempt at automated classification of human values behind arguments.
Chapter 1

Introduction

In many cases of argumentative dispute the involved stances or perspectives can’t be directly proven or refuted by either side. In these situations, arguments are instead used to persuade the audience they are addressed to [Bench-Capon, 2003]. Perelman and Olbrechts-Tyteca describe the purpose of such arguments as “to induce the hearer to make certain choices rather than others and, most of all, to justify those choices so that they may be accepted and approved by others.” [Perelman and Olbrechts-Tyteca, 1969, p. 75, italics mine]

This justification of own and others’ actions and attitudes follows the core concept of (human) values [Rokeach, 1968]. Values serve as guiding principles and motivation for human behavior. They represent what people think is generally worth striving for in life and how to do so [Searle, 2003] and act as criteria “for morally judging self and others, and for comparing self with others.” [Rokeach, 1968, p. 160] Some values tend to conflict (e.g., having success vs. being humble) while others seem to align and are sometimes expressed in the same context (e.g., being creative and having freedom of thought). This can cause disagreement on the best course forward, but also the support, if not formation, of political parties that promote the respective highly revered values, suggesting a possible reason for the varying acceptance and strength in persuasion of an argument as each target audience has their own set of value priorities.

Due to their outlined importance, human values are studied both in the social sciences [Schwartz, 1994] and formal argumentation [Bench-Capon, 2003] since decades. According to the former, a “value is a (1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities.” [Schwartz, 1994, p. 20]

As an example on how this value definition relates to arguments, consider the following scenario from Bench-Capon [2003]. The scenario was originally
discussed by Coleman [1992] in an example moral debate and Bench-Capon describes the starting situation as follows:

In the scenario a diabetic, Hal, loses his insulin in an accident through no fault of his own and before collapsing into a coma he hurries to the house of another diabetic, Carla. She is not at home, but Hal enters her house and uses some of her insulin.

As further elaborated by Bench-Capon, one could argue that Hal’s action was justified since

"a person has a privilege to use the property of others to save their life."

To understand the pragmatics of this statement, a reader has to acknowledge the belief (Point 1 in the definition above) that the “end state” (2) of having good health is personally and socially worth striving for (3). To concur with the statement (4), the reader further has to prefer having good health over being compliant (5). This thesis will later discuss this example in more detail with the proposed taxonomy set in place.

"Within computational linguistics, human values thus provide the context to categorize, compare, and evaluate argumentative statements” [Kiesel et al., 2022, p. 4460] which would be beneficial to the assessment of arguments and argumentation with respect to scope and persuasive strength as well as the generation or selection of arguments based on the value system of a target audience [Bench-Capon, 2003]. A major obstacle to the task of identifying the values behind arguments has been the large number of values, their variety and vagueness in definitions, and their mostly implicit use as reasoning behind arguments. However, leveraging advancements in natural language processing (NLP) and understanding, the existence of large argumentation datasets, and the decade-long taxonomization of values by social scientists, a first attempt on an operationalization of human values to classify arguments seems possible.

This thesis’ work surrounds the publication “Identifying the Human Values behind Arguments” by Kiesel et al. at ACL 2022. The core element of the thesis at hand is a consolidated multi-level taxonomy of 54 values taken from four authoritative cross-cultural social science studies (Chapter 3). The taxonomy is aimed to cover the value continuum of human beings as complete as possible and is purposed to be universally applicable in all countries and cultures. In extension of the research presented by Kiesel et al. [2022], this chapter further discusses the steps leading to the selection of human values as categorization aspects, the choices regarding the value schemes used as foundation, and a more detailed explanation of the taxonomy’s formation process. Chapter 4 describes the formation of a crowd-sourced corpus containing 5270 arguments from the US
(most arguments), Africa, China, and India, each of which manually annotated for each level of the taxonomy. In regards to Kiesel et al. [2022], this work goes into more detail about the development of the annotation interface as well as the setup and execution of the crowd sourcing study. In order to provide a baseline on the automated identification of values, Chapter 5 showcases first classification results per taxonomy level both within and across cultures.
Chapter 2

Related Work

As already established, this work focuses on the identification of human values in the context of arguments. Specifically in this field, the work at hand focuses on the collectivity of personal values, emphasized by Rokeach’s fundamental definition of “what it means to say that a person has a value”. [Rokeach, 1973, p. 5] He thereby describes the two concepts of (1) a value as an enduring belief pertaining to desirable modes of conduct or end-states of existence and (2) a value system as prioritization of values based on cultural, social, and personal factors [Rokeach, 1973]. Together with a slightly extended value definition given in the theory by Schwartz [1994], this work leverages these definitions to identify the often implicit usage of values behind arguments.

This work’s proposed multi-level taxonomy (Chapter 3) is based on domain-independent schemes of personal values, as these schemes were considered suitable for the classification of generic and cross-cultural argumentation.

Rokeach [1973] developed a survey of 36 values that distinguishes between values pertaining to desirable end states (e.g. A world at peace) and desirable modes of conduct (e.g. Independent). Brown and Crace [2002] looked at 14 personal values regarding counseling and therapy, such as Health & Activity.

On the prospect of cross-cultural application, Schwartz et al. [2012] proposed 48 value questions derived from the universal needs of individuals and societies. These value questions pertain to 19 separate motivational types which form a circular arrangement listing conflicting values on opposed sides. Regarding the comparison between cultures and the research of values across regions, the World Values Survey [Haerpfer et al., 2022] contains results from 59 countries, analyzing people’s priorities such as the importance of family and the opinion on controversial topics/claims like if it is a child’s duty to take care of ill parents.

Other value schemes are for the most cases strictly more coarse-grained than Schwartz et al.’s theory or even the survey by Rokeach. Cheng and Fleischmann
[2010] consolidated 12 schemes into a “meta-inventory” with 16 values, such as honesty and justice, revealing a large overlap in schemes across fields of research. In addition, some value schemes pertain to specific purposes. These being for example, a scheme towards modeling the influence on management decisions, containing values like social welfare and job satisfaction [England, 1967], a value list designed to measure consumer values, such as warm relationships and self-fulfillment [Kahle et al., 1988], and values addressing technology design, like informed consent and freedom of bias [Friedman et al., 2006].

On the perspective of values in argumentation research and natural language processing (NLP), different approaches consider value systems for real world applications. Extending the definition of the argumentation frameworks [Dung, 1995], Bench-Capon [2003] analyzed argument strength and the persuasion of audiences with his proposed value-based argumentation frameworks which have been manually applied to research the connection between argumentative reasoning and persuasion towards a certain value system [Bench-Capon, 2021]. Using Schwartz’s value theory, Maheshwari et al. [2017] already applied a coarse-grained classification scheme for personality profiling, however, to the best of knowledge, an automated classification of arguments based on human values has not been attempted prior to this work.

There are further concepts established in argumentation research that are closely related to values. The Moral Foundations Theory [Haidt, 2012] analyzes ethical reasoning behind human behavior. A strong connection between values and the Moral Foundations Theory was shown by Feldman [2021]. Improvements in automated value detection are thereby considered beneficial to the classification of Moral Foundations as well. In a similar fashion to values, Moral Foundations have also been considered for argument generation towards a specific target audience [Alshomary and Wachsmuth, 2021].

There also exists a noticeable overlap between values and the concept of framing [Entman, 1993] which emphasizes specific aspects of one’s perceived reality, allowing to measure the cost and benefit of certain actions, “usually measured in terms of common cultural values” [Entman, 1993, p. 2] and guide moral evaluation similarly to human values. The relation between arguments and frames has already been studied in regards of automated classification [Ajjour et al., 2019].

Related tasks concern opinion summarization [Chen et al., 2019; Misra et al., 2016], which aims to extract the most important aspects discussed in a controversial debate, as well as the task on identifying key points in arguments and debates [Bar-Haim et al., 2020; Friedman et al., 2021]. The latter one targets the generation of a small of representative statements from debates and topics. The annotation of arguments resorting to a fixed and universal set of
human values could be beneficial for identifying and analyzing perspectives on controversial topics [Chen et al., 2019].
This chapter elaborates the motivation for constructing a universal classification for arguments and the steps leading to the usage of human values as categorization aspects. An explanation of the choices, regarding the value surveys used as foundation, leads to the description of the taxonomy’s creation process followed by a detailed explanation regarding the taxonomy’s contents. The chapter is concluded by a brief discussion of the resulting taxonomy and a direct application to an example moral debate.

In a first endeavor, a universal classification could allow a much simpler identification of similarities between controversial topics based on their respective arguments. This would not only improve the understanding between cultures and their unique controversial topics but also aid people in general to form an opinion on unknown topics. Especially the latter one, regarding opinion formation, is currently attempted through the use of argument search engines that allow users to get lists of arguments on the selected topics. Such an example is args.me\(^1\) from Wachsmuth et al. [2017] which has arguments associated to ‘aspects’ derived from a Wikipedia list of more than 1000 controversial topics with a respective barycentric visualization of this topic space [Ajjour et al., 2018; Kiesel et al., 2018]. Additionally, the identification of universal or cross-cultural aspects would allow to further analyze and categorize the persuasive strength of arguments, resulting in improvements for strategic gathering and the generation of persuasive arguments towards given target audiences.

Especially with the taken approach of using human values for categorizing arguments, including value-based argument generation and personality profiling [Maheshwari et al., 2017]. At the same time, a universal value taxonomy

\(^1\)https://www.args.me
and value-based persuasion of a certain target audience [Bench-Capon, 2003] includes the risk of manipulative argument generation. In addition, “a value-based analysis could risk to exclude people or arguments based on their values. However, in other cases, for example hate speech, such an exclusion might be desirable.” [Kiesel et al., 2022, p. 4468] The main goal is to open up controversial topics and allow an exchange of opinions beyond the bounds of intercultural understanding and topic-dependent knowledge.

Although the taxonomy proposed in this work is solely based on personal (human) values, multiple classification approaches were taken into consideration. As described earlier (see chapter 2) these namely include the (1) Moral Foundations Theory, (2) frames, (3) opinion summarization and key points, and of course the concept of (4) values.

The (1) Moral Foundations Theory [Haidt, 2012] spans the six moral foundations Care, Liberty, Fairness, Loyalty, Authority, and Sanctity. Kobbe et al. [2020] applied the scheme to classify arguments, however noticed a low human agreement due to the vagueness of the foundations. There also was a significant portion of the arguments (29%, excluding absolute disagreement) which has been considered resorting to none of the moral foundations.

The concept of (2) framing [Entman, 1993] emphasizes specific aspects of controversial debates. Thereby, a set of arguments sharing such an aspect forms a frame. Ajjour et al. [2019] applied this concept using machine learning to identify and extract frames from arguments resulting in generic and topic-specific frames. With this approach they extracted a total of 1623 frames from 465 topics with 80% of the frames occurring in only one topic (topic-specific).

Another topic-dependent classification surrounds the task of (3) opinion summarization which aims to generate overviews of different debates. This includes the identification and comparison of argument facets [Misra et al., 2016], reoccurring propositions in arguments relating to the same topic and approaches to formulate perspectives regarding a certain claim [Chen et al., 2019] or create summaries of debates by extracting pro- and con-points from arguments [Egan et al., 2016]. This is also directly related to the task of key point analysis [Bar-Haim et al., 2020; Friedman et al., 2021] where a small set of key points is used to represent the majority of a topic’s arguments.

In social science the term (4) values has been used for a variety of different socio-psychological constructs and as a result comes with a widespread set of varying definitions [Cheng and Fleischmann, 2010]. To counteract the confusion of terminology, Rokeach conceptualized values, especially human values, as abstract motivation for behavior and formulated a fundamental definition of “what it means to say that a person has a value”. [Rokeach, 1973, p. 5] This allowed to distinguish values from other abstract concepts such as desires or
needs. He also concluded that the number of values a person can be considered to have, denotable as personal values, is likely to be small. [Rokeach, 1968, 1973]

A sophistication of the ‘value’ concept lead to the aforementioned definition by Schwartz that a “value is a belief pertaining to desirable end states or modes of conduct, that transcends specific situations, guides selection or evaluation of behavior, people, and events, and is ordered by importance relative to other values to form a system of value priorities.” [Schwartz, 1994, p. 20] Perelman and Olbrechts-Tyteca [1969] noted the usage of values for audience persuasion and, in an equivalent motivation, Bench-Capon [2003] studied audience persuasion by introducing value-based argumentation frameworks, an extension of the abstract argumentation frameworks of Dung [1995]. By imploring a more general ‘value’ definition, and therefore not only considering human values, this concept has already been manually applied to analyze interactions with reasoning and persuasion subject to a specific value system [Bench-Capon, 2021].

Striving for a fine-grained classification approach with it’s contained aspects not being bound to a specific domain, the decision was made to consolidate a taxonomy of human values and thereby focus on value surveys and lists containing personal values. As a strong connection between personal values and the Moral Foundations Theory has already been shown [Feldman, 2021], there still remains the consideration to include moral foundations into a future version of the taxonomy.

It is important to note that the value taxonomy was not developed through strict psychological studies in social science. The motivation was to create a collection of values suitable to categorize arguments while being as complete as possible in terms of applicability to different controversial topics and different cultures around the world. This was achieved by combining multiple different value surveys and denoting separate values if they contain an arguably distinct enough definition from each other. As such the values in this taxonomy are expected to have a higher correlation in comparison to the original value surveys they were gathered from.

### 3.1 Value Study

Human values have been considered in formal argumentation since about 20 years [Bench-Capon, 2003] and the taxonomization of values by social scientists dates even further back. The value schemes selected for the formation of this work’s proposed taxonomy are the SVS, RVS, and LVI.

The value survey of Rokeach [1973] (RVS) features two lists containing 18 instrumental and 18 terminal values respectively [Rokeach, 1973, p. 28]. With
a similar definition of the value term Schwartz [1992] developed a theory of which principles guide human behavior. His original theory contains 11 motivational types\(^2\) represented by 56 single values, 21 of them being identical to those in the Rokeach list [Schwartz, 1992, p. 17]. Using adjectives and noun phrases the values were also divided into lists of instrumental and terminal values but the usefulness of a terminal-instrumental discrimination was questioned based on empiric findings.

In a later publication, Schwartz [1994] proposed 57 single-value items to represent the supposed 10 motivationally distinct value concepts. The theoretical circular structure of the value concepts was further solidified by empirical data, concluding that values are organized by a common structure of motivational oppositions and congruities for most literate adults across cultures. The application of the Schwartz Value Survey (SVS) [Schwartz, 1994] proved to come with some challenges as it required a high amount of abstract thinking and the values were presented outside of a specific context. This lead to the development of the Portrait Value Questionnaire (PVQ) [Schwartz et al., 2001] with a set of easier understandable questions and a uniform context that people are able to relate to. The research around the PVQ was targeted towards testing the validity and cross-cultural reach of the values theory [Schwartz, 1994]. Therefore the PVQ focused “on the value constructs in [Schwartz’s] theory and the structure of relations among them, not on specific value items.” [Schwartz et al., 2001, p. 520] With the results from previous studies Schwartz et al. [2012] refined the original theory [Schwartz, 1994] proposing 19 distinct value concepts that form a circular motivational continuum and are represented by 48 value items that are phrased as portrait sentences.

Brown and Crace [2002] created the Life Values Inventory (LVI) which features a list of 42 beliefs which aims to help people clarify and prioritize all of the 14 personal values proposed by the authors.

One criterion for the selection of value schemes was the universal applicability, i.e., a value scheme should not be limited to a certain culture or use-case. Some schemes that were taken into consideration were thereby too domain-specific and therefore not suited for a taxonomy of cross-cultural values. As an example, England [1967] studied 66 values related to guiding the decisions of American managers, such as social welfare and job satisfaction.

Another selection criterion, albeit with lesser priority, was the size of the selected value schemes as the following build process of the taxonomy (Section 3.2) relies largely on the overlap between different value schemes to identify values that can be considered universal or cross-cultural. It was observable that in schemes which are not considered domain-specific a respectively low number

\(^2\)The universality of “Spirituality” as a type was doubted but still included, represented by 4 values.
of values often pairs with multiple basic values being comprised into one rather
abstract formulated value. This can be partially exemplified regarding the
names and descriptions of the 19 more abstract value types from Schwartz et al.
[2012] that subsume the original 57 basic human values [Schwartz et al., 2012].
As the task is the aspiration of a taxonomy of basic human values which is as
complete as possible in terms of universal applicability and cross-cultural reach,
this ‘size’ criterion excluded the list of 12 values proposed by Scott [1965], such as
social skills and status, and the List of Values (LOV) by Kahle et al. [1988].
Especially the latter list, spanning a total of 9 values, was not only considered
too coarse-grained for classifying arguments but also pertains to a (domain-)
specific purpose, as it was developed for consumer research regarding values
such as warm relationships and self-fulfillment.

It is, however, worth to note that all mentioned value schemes contain
similar and partially identical value definitions, which has been pointed out by
Cheng and Fleischmann [2010] through the development of a “meta-inventory”
consisting of 16 values that have been consolidated from 12 different value
schemes. The meta-inventory itself is thereby representable through the combi-
nation of the SVS [Schwartz, 1994], RVS [Rokeach, 1973] and LVI [Brown and
Crace, 2002] as all 16 values from the inventory would be semantically included
in the combination of these schemes [Cheng and Fleischmann, 2010].

Finally, the development of the taxonomy also considers the results of
the World Values Survey (WVS) Wave 7 [Haerpfer et al., 2022]. This, again,
ensures the cross-cultural reach and universal applicability of the resulting
value taxonomy. However, this process requires the questions and results from
the WVS to be interpreted in regards of (implicitly) mentioned human values,
inducing a potentially biased perspective, especially for values exclusively
mentioned in the WVS. Therefore, no values were directly extract from the
WVS but instead the WVS was used to confirm the cross-cultural validity of
values from other schemes.

3.2 Build Process of the Value Taxonomy

As elaborated by Cheng and Fleischmann [2010], there are varying definitions
of the ‘value’ concept. The taxonomy proposed in this chapter is mainly based
on the refined theory of Schwartz et al. [2012] with explicit value names taken
from the original theory [Schwartz, 1994]. Therefore, this work adopts the
definition by Schwartz [1994] that a “value is a belief pertaining to desirable end
states or modes of conduct, that transcends specific situations, guides selection
or evaluation of behavior, people, and events, and is ordered by importance
relative to other values to form a system of value priorities.” [Schwartz, 1994,
Additionally, the application of personal values to categorize (possible cross-cultural) arguments emphasizes another characteristic. Extending the example by Bench-Capon [2003] regarding the question whether the taxes should be raised or lowered, some will argue that the taxes should be raised to promote having (social) equality while others will argue that the taxes should be lowered in favor of having a stable society by promoting enterprises. As Bench-Capon points out that both parties can acknowledge the "effects argued by their opponents [...] and both regard greater equality and greater enterprise as good things." [Bench-Capon, 2003, p. 2] Therefore, despite being personal values, if addressed to an audience through an argument’s reasoning they can be understood as values. This characteristic will be further discussed in Chapter 4 when formulating the task of identifying these values behind arguments.

There, however, remains the difficulty of varying interpretations of specific values amongst different people. The problem is well described by van der Weide et al.: “People use their values to evaluate states. However, since values are typically abstract (e.g. fairness or happiness), giving meaning to a value involves interpreting how concrete states relate to abstract values. Concrete interpretations of values are often disputable. For example, although two persons both hold the value of fairness, they may disagree about what they think is fair.” [van der Weide et al., 2009, p. 82]

This work’s approach is founded on the understanding that there only exists a limited number of basic human values [Rokeach, 1973]. These conceptual and abstract values can then be projected onto natural language utterances, i.e., a value name like Be just. The connection behind the concept an it’s utterance is therefore a subject to interpretation. These utterances can also be called abstract words as the concepts, they are referring to, can’t be directly experienced through one’s senses, given that values are beliefs. Abstract words and how people connect them to a specific meaning has been studied by cognitive science for a long time [Zdrazilova et al., 2018]. In an early version of his theory Schwartz et al. [2001] encountered a similar problem which resulted in the development of the PVQ. The presented portrait questions featured a specific context and a less abstract wording. These do not directly relate to single basic values as the focus was mainly on the motivational types as larger value constructs. This work tries to mitigate the challenge by providing a set of definitions or use cases regarding each single value which aims for a universal and cross-audience understanding of each abstract word. Therefore, it is hoped that the task of identifying the values behind a given argument ultimately becomes more coherent and thus also reproducible, as the amount of interpretation from each annotator is greatly reduced.

These specific interpretations of each value still remain controversial and
the problem of biased or even conflicting value definitions is connected to this method as well. However, in order to annotate an argument with a value, a certain form of utterance is necessary to convey any meaning, with this method having a supposedly small bias from individual value understandings of the task’s authors and annotators respectively.

As announced prior, the foundation for this work’s proposed value taxonomy is Schwartz et al.’s refined model, including its hierarchical structure and the circular arrangement regarding the motivational continuum (see Figure 3.1). The 4-Level structure of the taxonomy reflects the same hierarchy. However to achieve a fine-grained naming structure the 48 original value items were identified as *Values* (Level 1) and had their names and descriptions derived using the noun-phrase values from the original theory [Schwartz, 1992, 1994] and the questions from the Portrait Values Questionnaire (PVQ) [Schwartz et al., 2001] as value descriptions. While not all of the 48 items became individual values, which will be discussed further on, this resulted in 45 values with the original 19 ‘values’ (as conceptualized by Schwartz et al. [2012]) now described as *Value categories* (Level 2). The higher-order values (Level 3) and the two dichotomies, *Personal focus/ Social focus* (Level 4a) and *Growth/ Self-protection* (Level 4b), were included without additional changes regarding their contents, structure, or definitions.

The Rokeach Value Survey (RVS, [Rokeach, 1973]) and the Life Values Inventory (LVI, [Brown and Crace, 2002]) focus on personal values as well. Together with the World Values Survey (WVS, [Haerpfer et al., 2022]), these three sources not only provide possible missing values or more narrow value definitions but also support values already included in the base system. Especially the inclusion of the WVS serves this purpose as the values contained in the survey have a high chance to be applicable to value systems in different cultures and countries.

There are 28 out of the 36 values from the RVS and the 14 values from the LVI which were integrated or added to the base system. Not all values from the RVS have been added due to the requirement of being present in at least two of the three additional sources. The only exception is *Be courageous* which is solely present in the RVS. In total 9 values were added to the base system resulting in a final count of 54 values. Furthermore the taxonomy adopted a uniform naming scheme where the value names reflect the distinction made by Rokeach [1973] into instrumental (*be...*) and terminal (*have...*) values that can be easily embedded in sentences, for example, “it is good to *be creative*.”

Two of the added values are not directly related to the universal needs where Schwartz [1994] based the motivational types on, resulting in the addition of a new value category *Universalism: objectivity* (see Figure 3.1). Additionally during the conducted crowd sourcing study (cf. chapter 4) the annotators were
also asked to comment on supposedly missing values. For most of the additional 48 value descriptions (be humane, be fair, be modern, etc.) it was possible to identify values or value combinations in the proposed taxonomy that subsume them, suggesting to extend the value description rather than adding new values.

### 3.2.1 In-depth Category Description

The following will describe the in-depth formation process broken down into each value category (Level 2). Difficulties during the formation process and the reasoning behind made decisions are reported as well. These difficulties mainly occurred regarding the integration of new values in the base system and the classification of values into instrumental and terminal based on Rokeach [1973] definitions.

As mentioned earlier 45 single values had been taken from the Schwartz
### Table 3.1: The 45 values formulated from the refined Schwartz Value Survey (SVS, [Schwartz et al., 2012]) and their correspondence in the original theory [Schwartz, 1992, 1994]. All values are marked in regards of being conceptualized as **instrumental** or **terminal**.

<table>
<thead>
<tr>
<th>Value category</th>
<th>Value</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-direction: thought</td>
<td>Be creative</td>
<td>Creativity/Imagination</td>
</tr>
<tr>
<td></td>
<td>Be curious</td>
<td>Curious/Interested</td>
</tr>
<tr>
<td></td>
<td>Have freedom of thought</td>
<td>Freedom of thought</td>
</tr>
<tr>
<td>Self-direction: action</td>
<td>Be choosing own goals</td>
<td>Choosing own goals/directions</td>
</tr>
<tr>
<td></td>
<td>Be independent</td>
<td>Independent/Self-reliant</td>
</tr>
<tr>
<td></td>
<td>Have freedom of action</td>
<td>Freedom of action</td>
</tr>
<tr>
<td>Stimulation</td>
<td>Have an exiting life</td>
<td>Exciting life/Excitement</td>
</tr>
<tr>
<td></td>
<td>Have a varied life</td>
<td>Varied life/Novelty</td>
</tr>
<tr>
<td></td>
<td>Be daring</td>
<td>Daring/Challenge/Change</td>
</tr>
<tr>
<td>Hedonism</td>
<td>Have pleasure</td>
<td>Pleasure</td>
</tr>
<tr>
<td>Achievement</td>
<td>Be ambitious</td>
<td>Ambitious</td>
</tr>
<tr>
<td></td>
<td>Have success</td>
<td>Successful</td>
</tr>
<tr>
<td></td>
<td>Be capable</td>
<td>Capable</td>
</tr>
<tr>
<td>Power: dominance</td>
<td>Have influence</td>
<td>Social power/Control over others</td>
</tr>
<tr>
<td></td>
<td>Have the right to command</td>
<td>Authority/Right to command</td>
</tr>
<tr>
<td>Power: resources</td>
<td>Have wealth</td>
<td>Wealth/ Material possession</td>
</tr>
<tr>
<td>Face</td>
<td>Have social recognition</td>
<td>Social recognition/respect</td>
</tr>
<tr>
<td></td>
<td>Have a good reputation</td>
<td>Preserving public image/Maintaining face</td>
</tr>
<tr>
<td>Security: personal</td>
<td>Have a sense of belonging</td>
<td>Sense of belonging/feeling others care about me</td>
</tr>
<tr>
<td></td>
<td>Have good health</td>
<td>Healthy</td>
</tr>
<tr>
<td></td>
<td>Have no debts</td>
<td>Reciprocation of favors/avoiding indebtedness</td>
</tr>
<tr>
<td></td>
<td>Be neat and tidy</td>
<td>Clean/Neat, tidy</td>
</tr>
<tr>
<td>Security: societal</td>
<td>Have a safe country</td>
<td>National Security</td>
</tr>
<tr>
<td></td>
<td>Have a stable society</td>
<td>Social order/stability</td>
</tr>
<tr>
<td>Tradition</td>
<td>Be respecting traditions</td>
<td>Respect tradition/Preserve customs</td>
</tr>
<tr>
<td></td>
<td>Be holding religious faith</td>
<td>Devout/Hold religious faith</td>
</tr>
<tr>
<td>Conformity: rules</td>
<td>Be compliant</td>
<td>Obedient</td>
</tr>
<tr>
<td></td>
<td>Be self-disciplined</td>
<td>Self-discipline</td>
</tr>
<tr>
<td>Conformity: interpersonal</td>
<td>Be polite</td>
<td>Politeness</td>
</tr>
<tr>
<td></td>
<td>Be honoring elders</td>
<td>Honor elders/show respect</td>
</tr>
<tr>
<td>Humility</td>
<td>Be humble</td>
<td>Humble/Most</td>
</tr>
<tr>
<td></td>
<td>Have life accepted as is</td>
<td>Accepting my portion in life</td>
</tr>
<tr>
<td>Benevolence: caring</td>
<td>Be helpful</td>
<td>Helpful</td>
</tr>
<tr>
<td></td>
<td>Be honest</td>
<td>Honest</td>
</tr>
<tr>
<td></td>
<td>Be forgiving</td>
<td>Forgiving</td>
</tr>
<tr>
<td>Benevolence: dependability</td>
<td>Be responsible</td>
<td>Responsible/dependable</td>
</tr>
<tr>
<td></td>
<td>Have loyalty towards friends</td>
<td>Loyal/faithful friends</td>
</tr>
<tr>
<td>Universalism: concern</td>
<td>Have equality</td>
<td>Equality</td>
</tr>
<tr>
<td></td>
<td>Be just</td>
<td>Social justice</td>
</tr>
<tr>
<td></td>
<td>Have a world at peace</td>
<td>World at peace</td>
</tr>
<tr>
<td>Universalism: nature</td>
<td>Be protecting the environment</td>
<td>Protecting the environment</td>
</tr>
<tr>
<td></td>
<td>Have harmony with nature</td>
<td>Unity with nature</td>
</tr>
<tr>
<td></td>
<td>Have a world of beauty</td>
<td>World of beauty</td>
</tr>
<tr>
<td>Universalism: tolerance</td>
<td>Be broadminded</td>
<td>Broadminded/Tolerant</td>
</tr>
<tr>
<td></td>
<td>Have the wisdom to accept others</td>
<td>Wisdom/Mature understanding</td>
</tr>
</tbody>
</table>
Value Survey (SVS). They are listed in Table 3.1. For 12 of the values their
instrumental-terminal discrimination varies from the SVS. In this taxonomy, the
maintaining of a desired end-state was not considered to be seen as instrumental
component of a respective value. Therefore, terminal values may contain
instrumental aspects like staying healthy or maintaining a good reputation,
however the focus will be on the fact that there exists a defined reachable
end-state.

**Self-direction: thought**  Creativity is also contained in the LVI and the
WVS [Brown and Crace, 2002; Haerpfer et al., 2022]. The combined concept
focuses on new creations and ideas as an ongoing process of being creatively
expressive and imaginatively active, appearing as an instrumental value. Even
thought the connection was drawn between Creativity and Imagination, the
instrumental value Imaginative [Rokeach, 1973] was not fully considered to
be subsumed by the value Be creative, as the theory around the SVS focuses
mainly on the aspect of creativity [Schwartz, 1994; Schwartz et al., 2012].

**Self-direction: action** The additional value Privacy from the LVI does not
concern a security aspect but instead focuses on the importance of alone time
[Brown and Crace, 2002]. Therefore the value Have privacy has been added
to the freedom-related category Self-direction: action and not to Security:
personal.

**Stimulation** The contained single values as well as the boundaries of this
motivational type remained the same through the revisions of Schwartz’ theory
Schwartz [1992, 1994]; Schwartz et al. [2012]. The definition of the category
consisting solely of excitement, novelty, and change (or challenge in life) was
therefore directly adapted. Schwartz [1992] describes the value Be daring
originally with an adjective (instrumental value). In later revisions of the
theory it is described with the noun change, however, the taxonomy resorts to
the original formulation as an instrumental value.

**Hedonism** In unanimity with the definition by Rokeach, Schwartz et al.
describes that the “conceptual definition and the results of all the analyses
indicate that hedonism has only one component, pleasure.” [Schwartz et al.,
2012, p. 4]

**Achievement** The definitions for the value Have success relates to expressions
as ‘having success according to social standards’ or ‘having others recognize one
as successful’ [Schwartz, 1994; Schwartz et al., 2012], resulting in a categorization
as terminal value. This category also contains the value *Be courageous* which is the only value in the taxonomy originating from just one source, due to its concept being strongly related to *Achievement* and the value was later kept as a result of its still considerable appearance in the crowdsourced dataset.

**Power: dominance & Power: resources** These categories were completely adopted from the SVS. Although the LVI features the value *Financial prosperity* [Brown and Crace, 2002], its conceptual description does not fully coincide with the definition of *Have wealth*, mainly the “power to control events through one’s material assets.” [Schwartz et al., 2012, p. 4]

**Face** While applying a coherent usage of the term ‘maintaining’ towards desirable situations or states in life, the value *Preserving public image/Maintaining face* from the SVS [Schwartz et al., 2012] was changed. The resulting value name *Have a good reputation* was thereby conceptualized as a terminal value.

**Security: personal** In contrast to the conceptualization in the SVS, the value *Have good health* is considered to be terminal. As mentioned earlier, the focus remains on the fact that there exists a defined reachable end-state of having good health while concealing the instrumental aspect of *maintaining* this end-state. Schwartz et al. [2012] also noted that the meaning of health as a value may vary considerably across cultures. This category was extended with the RVS value *Have a comfortable life* as it also resorts to “[s]afety in one’s immediate environment”. [Schwartz et al., 2012, p. 7]

**Security: societal** This category was adopted from the SVS without additional changes.

**Tradition** Regarding the value *Respect tradition/Preserve customs*, the motivation towards an ongoing process of preserving customs as well as family, cultural, or religious traditions lead to the adoption of this value as instrumental.

**Conformity: rules** Schwartz [1992] originally noted the value *Self-discipline* as a noun phrase indicating a terminal value. However the similarity to the instrumental value *Self-controlled* [Rokeach, 1973] lead to the decision to include *Be self-disciplined* as an instrumental value. The LVI also contains the value *Interdependence* stating that it is important to follow the expectations of one’s family, social group, team or organization. With the closest relation being *Conformity: rules*, an adaptation of the value’s description resulted in the
inclusion of the additional value with the slightly less vague value name: *Be behaving properly*.

**Conformity: interpersonal**  Regarding both contained values, *Be polite* and *Be honoring elders*, their motivational definition was understood as a custom that should be followed instead of an end-state of existence.

**Humility**  The state of accepting one’s portion in life is, in opposition to the SVS, understood as a defined end-state. Therefore, the value *Have life accepted as is* has been denoted as terminal.

**Benevolence: caring**  The first three values have been directly adapted from the SVS. The terminal value of *Family security* and the instrumental value of *Loving* from the RVS were added to this category due to their definition being directly related to “caring for the welfare of ingroup members.”[Schwartz et al., 2012, p. 5] Both values were not considered to be subsumed by any of the existing three values as all five of them are contained in the RVS.

**Benevolence: dependability**  Due to the partial overlap in concepts regarding the original SVS values *Loyal* and *Faithful friend*, they were combined into the single value *Have loyalty towards friends*.

**Universalism: concern**  The concept of *Social justice* was originally held as an end-state towards societal concern. However, the PVQ displays the corresponding item as *treat all justly/protect the weak* resorting to a more personal perspective of *contributing* to the society through ones *just-motivated actions* rather the actual state of society-wide justice. This work therefore notes *Be just* as an instrumental value.

**Universalism: nature & Universalism: tolerance**  Both categories were adopted from the SVS without further changes. With the exception of the value *Have harmony with nature*, each value is either present in the RVS or LVI with the same instrumental-terminal discrimination.

**Universalism: objectivity**  The two values *Be logical* (original: *Logical* [Rokeach, 1973]) and *Have an objective view* (original: *Objective Analysis* [Brown and Crace, 2002]) don’t quite fit in any of the existing value categories. Instead a new category *Objectivity* was derived from the overlapping definitions of both values and sorted as subcategory of *Universalism*. The new found category consists of the importance to *use logical principles for understanding*
and solving problems and can therefore be seen between Universalism: tolerance and Self-direction: thought. It was considered combining the two values Be logical and Have an objective view; however, both were kept considering that the latter one mostly describes the result of an objective understanding as motivation for actions and, based on the value differentiation of Rokeach [1973], can be seen as terminal value in contrast to the instrumental value Be logical.

3.3 Taxonomy discussion

The complete proposed value taxonomy can be seen in Table 3.2. Level 1 contains 54 basic human values that are categorized on the more abstract Levels 2–4. Each value has one label per level, with the exception being Have pleasure, as it’s (Level 2) category Hedonism resorts to both Self-enhancement and Openness to change for Level 3, and the values contained in the category Achievement which pertains to both Level 4b labels.

As the values in Level 1 mainly originate from surveys [Rokeach, 1973; Schwartz, 1994] whole taxonomy allows for classification on varying degrees of granularity. The 10 motivational types from the SVS [Schwartz, 1994], being a prior concept of the value categories in Level 2, have already been applied in an approach for the classification of tweets and personality profiling [Maheshwari et al., 2017]. In addition, the promising results on this level during the machine learning experiments (Chapter 5) already motivated the research on improvements for the automated identification of Level 2 labels during the Task on Human Value Detection at SemEval 2023 [Kiesel et al.].

The higher levels of the taxonomy, especially the higher-order values of Level 3, allow for a coarse-grained classification of arguments which can be used to directly reflect different perspectives (e.g. political parties) or directions (Social focus vs. Personal focus) involved in a debate. The circular structure of the taxonomy (cf. Figure 3.1) combined with the different hierarchy levels also provides new opportunities for topic-space visualization [Kiesel et al., 2018], thereby contributing to an improvement for argument search engines.

Regarding the instrumental-terminal discrimination of Level 1, Rokeach already pointed out that the concept of instrumental values can be split in two kinds, competence values and moral values. Together with the strong connection between personal values in general and the Moral Foundations as shown by Feldman [2021], this motivates future research regarding the inclusion of the Moral Foundations Theory [Haidt, 2012]. There also remains a strong connection to the other considered approaches. For example, 14 of the 54 values in this taxonomy are also frames in the dataset of Ajjour et al. [2019]³.

³Per Jaccard similarity of value and frame names ≥ 0.5.
Table 3.2: The 54 values of the taxonomy with sources. The main source taxonomy (•) is the Schwartz Value Survey (SVS, [Schwartz, 1992, 1994; Schwartz et al., 2012]). Additional values are taken from (◦) the Rokeach Value Survey (RVS, Rokeach, 1973), the Life Values Inventory (LVI, Brown and Crace, 2002), and the World Values Survey (WVS, Haerpfer et al., 2022). Table adapted from Kiesel et al. [2022].
3.4 Example moral debate

This section depicts the usage of the taxonomy for categorizing arguments on the exemplary moral debate used by Bench-Capon [2003] to showcase his Value-Based Argumentation Frameworks (VAFs). As mentioned in the beginning, this debate was discussed by Coleman [1992] and further elaborated by Christie [2000]. Bench-Capon describes the initial situation as follows:

In the scenario a diabetic, Hal, loses his insulin in an accident through no fault of his own and before collapsing into a coma he hurries to the house of another diabetic, Carla. She is not at home, but Hal enters her house and uses some of her insulin.

The here considered extend of the moral debate includes the following arguments taken from Bench-Capon [2003]:

(A) A person has a privilege to use the property of others to save their life.

(B) It is wrong to infringe the property rights of another.

(C) If Hal compensates Carla, then Carla’s rights have not been infringed.

(D) If Hal were too poor to compensate Carla, he should none the less be allowed to take the insulin, as no one should die because they are poor.

Regarding their relation in the VAF, the arguments are aligned in an ascending chain of attacks, i.e., (D) attacks (C), (C) attacks (B), and (B) attacks (A). Bench-Capon categorizes the arguments (A) and (D) as resorting to “the value that life is important (life)” and the arguments (B) and (C) promote “the value that property owners should be able to enjoy their property (property)” [Bench-Capon, 2003, p. 443]. However, instead of the persuasion and argument strength towards a certain target audience the focus right now lays on the correspondence of these ‘values’ to the proposed value taxonomy.

The described concept of life has a clear connection to Have good health, as this value is associated with arguments towards avoiding diseases, preserving health, or having physiological and mental well-being. Finding a representation of the property concept proves to be more challenging. As the description speaks about “enjoying their property”, one could argue that it would be similar to Have pleasure as this value is associated with arguments towards making life enjoyable. However, framing the concept and the arguments it is used in specifies property as targeting the rights regarding owned properties and security of named properties. Therefore, the closest relation would be to the value Be compliant, as it is associated with arguments towards abiding to laws or rules.
It is worth to note that, depending on the interpretation, the listed arguments can be associated with a variety of values. For example, one could argue that (D) also resorts to *Have equality* as it concerns the well-being of poor people. This raises the question of how the Value-Based Argumentation Frameworks could model arguments as resorting to more than one value. A theoretical approach could be made using a prioritization of the values behind each argument depending on how strong an argument relates to each individual value. A different approach could be to apply multiple VAFs regarding a selection of subsets of values and their concern towards a certain target audience.

Nevertheless, this moral debate exemplary showcased the classification of arguments using the proposed value taxonomy. However, even for this small selection of arguments it is already notable that the meaning and expression of value concepts depends strongly on interpretation. Therefore, the application of the proposed taxonomy for identifying the often implicit use of values behind arguments presents a difficult task. The following crowd sourcing study (Chapter 4) aims to assess this difficulty on a larger scale and test.
Chapter 4

Crowd Sourcing a Dataset of Values behind Arguments

This chapter reports on the conducted crowd sourcing study, designed to test the taxonomy’s suitability for classifying arguments. The structure of the study’s presentation is based on the extend of a published checklist\(^4\) and begins with an overview of the arguments used for annotation. The chapter continues with a short introduction to the task and the description of the crowd sourcing interface along with its development history. After stating the process of quality control during and after the conducted study, this chapter concludes with the description and discussion of the results from the crowd sourcing study including the aggregated dataset of 5270 arguments. The resulting dataset, a taxonomy description (see Chapter 3) and the annotation interface are published\(^5\) as sources for Kiesel et al. [2022].

4.1 Input Datasets

Following the aspiration of a cross-cultural value taxonomy and using territories as a proxy for cultures, the dataset is composed of four parts: Africa, China, India, and USA. Available argument corpora for non-western countries and cultures are scarce. Therefore, for this work, all non-western arguments had been gathered from online sources associated to a certain region or culture. In order to create a uniform dataset and allow for a better comparison between the cultures, the gathered arguments were paraphrased into a uniform structure. Each argument is thereby composed of three parts: (1) the conclusion an


\(^5\)Identifying the Human Values Behind Arguments at https://github.com/webis-de/ACL-22
argument is referring to, (2) the stance towards that conclusion, and (3) the actual premise as the argument’s main content. Each premise is considered to support (‘pro’ stance) or attack (‘con’ stance) a given conclusion. The following paragraphs describe the sources for all four cultures (taken from Kiesel et al. [2022]):

**Africa** We manually extracted 50 arguments from recent editorials of the debating ideas section of a pan-African news platform, *African Arguments.*

Premises could often be extracted literally, but conclusions were mostly implicit and had to be compiled from several source sentences.

**China** We extracted 100 arguments from the recommendation and hotlist section of a Chinese question-answering website, *Zhihu.* We manually identified key points (premises and conclusions) in the answers and translated them to English.

**India** We extracted 100 arguments from the controversial debate topics 2021 section of *Group Discussion Ideas.* This blog collects pros and cons on various topics from Indian news to support discussions. Premises and conclusions were used as-is.

**USA** We took 5020 arguments with a manual argument quality rating of at least 0.5 from the 30,497 arguments of the IBM-ArgQ-Rank-30kArgs dataset [Gretz et al., 2020]. For the dataset, crowdworkers wrote one pro and one con argument for one of 71 common controversial topics. We rephrased the topics to represent conclusions.

The collection of arguments for the African and Indian part of the dataset was done by Johannes Kiesel whereas the selection and translation of the Chinese arguments was done by Xiaoni Cai, both from Kiesel et al. [2022]. While the effort was made to include texts from different cultures in the final dataset, it is important to note that these samples are not representative of their respective culture, but intended as a first benchmark for measuring the world-wide suitability of the derived value taxonomy and classification robustness across sources (see Chapter 5).

Table 4.1 describes the statistics for each part of the input dataset in regards to the three components of an argument. Token-wise, premises are

---

6https://africanarguments.org
7https://www.zhihu.com
8https://www.groupdiscussionideas.com
### Table 4.1: Numbers of unique conclusions and premises for each part of the contributed dataset, their mean number of space-separated tokens, and stance distribution. Table taken from Kiesel et al. [2022].

<table>
<thead>
<tr>
<th>Part</th>
<th>Conclusions #</th>
<th>Tokens</th>
<th>Premises #</th>
<th>Tokens</th>
<th>Stances Pros</th>
<th>Stances Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>23</td>
<td>10.6</td>
<td>50</td>
<td>28.1</td>
<td>37</td>
<td>13</td>
</tr>
<tr>
<td>China</td>
<td>12</td>
<td>7.3</td>
<td>100</td>
<td>24.5</td>
<td>59</td>
<td>41</td>
</tr>
<tr>
<td>India</td>
<td>40</td>
<td>6.6</td>
<td>100</td>
<td>30.3</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>USA</td>
<td>71</td>
<td>5.6</td>
<td>5020</td>
<td>18.5</td>
<td>2619</td>
<td>2401</td>
</tr>
<tr>
<td>Total</td>
<td>146</td>
<td>5.6</td>
<td>5270</td>
<td>18.9</td>
<td>2775</td>
<td>2495</td>
</tr>
</tbody>
</table>

longer than conclusions with the USA part having the lowest average for both. Additionally, some arguments, especially the (token-wise) longer ones, are suspected to contain more than one premise. However, as the argument part denoted as premise is more precisely seen as argument content in this work, no differentiation will be made regarding the number of premises in each argument. In terms of value annotation this approach is presumed to be without loss of information, as in the cases of multiple premises in one statement, the respective values for all premises are expected to be revealed in the study and therefore all assigned to the argument. It leaves to be seen whether this decision limits the expressive power of the dataset and the connected task of applying machine learning models (see Chapter 5). A future revision of the resulting dataset could attempt to split these arguments and re-assign the respective values, however, this work won’t investigate further on this aspect.

Even though the Indian part of the dataset lists 40 conclusions and 40 premises with negative (con) stance, not every conclusion has an argument containing a negative (con) premise, or positive (pro) premises respectively. For this dataset part in particular the conclusions have between 0 and 3 premises resorting to a pro stance and between 0 and 3 premises resorting to a con stance. This is the same with the African and Chinese part of the dataset where some conclusions have only positive or negative premises.

One exemplary argument from each dataset part can be seen in Table 4.2, which resort to the most frequent value have a stable society.

### 4.2 Crowd Sourcing Setup

This section outlines the concept and the general procedure of the crowd sourcing study. The description is followed by an overview of the used annotation interface and an explanation of the interface’s development process.
CHAPTER 4. CROWD SOURCING

<table>
<thead>
<tr>
<th>Argument</th>
<th>Values</th>
<th>Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>○ Pro “South Africa’s COVID-19 lockdown was too strict”: The economic ramifications of the lockdown have been huge, and have been felt hardest by those who were already most vulnerable.</td>
<td>Have a comfortable life, Have a stable society, Have equality</td>
<td>Africa</td>
</tr>
<tr>
<td>○ Pro “We should protect our privacy in the Internet age.”: The leaked personal information will be defrauded by fraud gangs to gain trust and carry out fraudulent activities.</td>
<td>Have privacy, Have a stable society, Be compliant</td>
<td>China</td>
</tr>
<tr>
<td>○ Con “Rapists should be tortured”: Throughout India, many false rape cases are being registered these days. Torturing all of the accused persons causes torture to innocent persons too.</td>
<td>Have a safe country, Have a stable society, Be just</td>
<td>India</td>
</tr>
<tr>
<td>○ Pro “We should adopt an austerity regime”: An austerity regime will help to reduce the deficit of the country.</td>
<td>Have no debts, Have a stable society, Be responsible</td>
<td>USA</td>
</tr>
</tbody>
</table>

Table 4.2: Four example arguments (stance, conclusion, and premise) and their annotated values. The referenced arguments from the dataset are (top to bottom): B28006, C26030, D27068, and A05074. Table taken from Kiesel et al. [2022].

Starting with the general approach, the task of identifying human values in regards of an argument presents the main challenge of applying a taxonomy derived of personal values onto natural language argumentation. The approach used in this work was based on the observation that a repeated questioning of ‘why’ or, more specific, ‘why something is good’ should eventually reveal the underlying values behind the reasoning for one’s arguments. This approach assumes that, in the regards of arguments, a value can be understood as universally accepted and non-questionable reasoning, i.e., one won’t gain additional information about the motivation behind another person’s argument by questioning this reasoning any further. As an example, consider the following argument against the abolition of zoos:

“Zoos do a lot of conservation work and have successfully bred animals which were on the verge of extinction.”

Questioning the author of this argument, about why they think the named actions should be considered good, could eventually lead to an answer like “These actions are beneficial to the environment” which coincides with the proposed value be protecting the environment. One could further question this answer, however, as values represent universal beliefs [Schwartz, 1994] and

---

9Argument A05064 in the proposed dataset
are criteria for justifying the own and others’ actions \cite{Rokeach1968} it can be assumed that further questioning them yields no additional information regarding the author’s reasoning.

This approach, however, requires that the connection between a given argument and its value-motivated reasoning can always, or at least in the majority of cases, be drawn mentally by humans. Hence the conducted crowdsourcing study also serves as a first assessment regarding the difficulty for trained annotators to draw these connections. The machine learning experiments in Chapter 5 will be used to assess the same problem regarding the approach of automatically modeling possible connections between arguments and values.

The crowdsourcing ran on the MTurk\footnote{https://www.mturk.com} platform. All participating crowd workers have been aware of the study and its data aggregation. However the exact purpose, the annotation of human values, was not communicated in order to minimize the workers bias. Therefore, the concept of ‘values’ was called ‘justifications’ during the complete study and likewise the task for the workers was formulated as to decide which ‘justifications’ could be provided for an argument.

As mandatory for MTurk, annotators were paid on a task basis, which led to an average hourly wage of $8.12, which to the time of the study was above the US federal minimum wage of $7.25. To encourage workers to return for the tasks especially in the early stages of the study and reward annotators who wrote extensive comments, additional bonuses were paid of total $65.65. The annotators were taking on average 2:40 minutes per argument. The total time for the annotations sums up to about 90 days of 8-hour work. No time constraint was given to complete each annotation task.

There were no direct requester-worker interactions as part of the actual crowdsourcing tasks. However the workers were able to comment on each task as well as every single justification in order to indicate problematic arguments or supposedly missing values. There also were additional message exchanges with some workers to clarify certain value descriptions and further address the comments they left.

### 4.2.1 Crowd Sourcing Interface

The applied version of the crowd sourcing interface can be divided into two parts with the top part containing the task’s instructions and examples. The instructions state each workers’ task to “select for each of 5 arguments which of 54 justifications one could provide for it” and workers were asked to leave comments on supposedly missing justifications or if they were unsure about a justification. The instructions also stated an observation made during the
annotation of the check instances (see Section 4.3), where an argument typically had between 1 and 5 suitable justifications that were the most fitting. The examples listed arguments regarding the conclusion “Social media should be banned” which occurred in neither of the arguments used for the study. The arguments were presented with their respective justifications and a small explanatory text to showcase the annotation task and address known difficulties. One such example targeted the three money related values have wealth, have no debts, and have a comfortable life as they had proven to be difficult to distinguish. Additionally, if needed for an easier understanding of the argument in question, the interface was extended with explanations for (domain- or culture-) specific terms, e.g., the “996 overtime system” mentioned in multiple arguments in the Chinese dataset part.

The bottom half forms the main part of the annotation interface, consisting of three panels. The first panel states each argument’s stance, conclusion, and premise while placing them in a uniform scenario for the annotation:

Imagine someone is arguing [in favor of/against] “[conclusion]” by saying: “[premise].”

The scenario is continued in the second panel with a formulation of the annotation task, following the aforementioned approach:

If asked “Why is that good?”, might this be their justification? “Because it is good to [justification].”

This panel also listed exemplary use cases for the selected justification. Below both panels, the third panel provides an overview of all 54 justifications and the corresponding annotation progress on the selected argument.

Finally, the complete annotation process was manageable through keyboard shortcuts in addition to using the cursor. For example, a justification can thereby be annotated as suitable or not-suitable with the left and right arrow-keys respectively which automatically selects the next justification to annotate. This feature allowed faster task completion.

4.2.2 Development Process

The process of developing a suitable annotation interface span a total of 10 versions. Screenshots of the interface during the process of development as well as an example for the top and bottom part of the final interface can be seen in Appendix A. The order of instructions (top), examples (following), and annotation task (bottom) has been decided early on defining the procedure of each crowd sourcing task. However, creating the layout for the actual annotation task presented a greater challenge. The two main issues were the
large number of values and (partially related) the amount of time required for each annotation task.

The first four versions were used for trying out different annotation approaches/concepts, including a hierarchical approach and a list-like sorting inspired by the survey from Rokeach [1973]. In the former, crowd workers would annotate an argument by choosing mainly, also, or not for each of the categories regarding their suitability and for the single category selected as mainly the workers would do same process with the contained values.

The design that was found to be the most intuitive and promised to allow for the most optimizations (keyboard shortcuts) was the 3-panel-layout described earlier. The remaining versions were developed upon this design in a progressive chain, constantly improving the formulation of the instructions and adding examples to preemptively address difficult cases of decision, like an argument resorting to no (listed) values.

One decision that was dropped later on was the usage of an alternative color palette which would allow color discrimination for people with dichromacy or anomalous trichromacy. Albeit being distinguishable, the color scheme was found to be rather distracting and later changed back to the green (suitable) and red (not-suitable) variant while the discrimination of the two options was instead induced through a highlighted check-mark or cross respectively.

4.3 Quality control

An important part for ensuring meaningful results in a crowd sourcing study is the application of quality control. Such process involves excluding data from assignments where workers appear to have ignored the given instructions and preventing the participation of such workers for the majority of the assignments. It is also required to crowd source enough annotations for each item to account for the natural variation in human judgment. Quality control also includes the quality of the task itself, i.e., ensuring that the instructions are easily understandable, and validating the estimated time and work load for each assignment which is necessary for a fair compensation.

The major part of the applied quality control was the split of the crowd sourcing study into a training phase and the main study. The assignments in the training phase were used to train workers on the proposed annotation task and to sort out workers who ignored the provided instructions. As only approved workers from the training phase were allowed in the bulk study, no rejection criteria were applied on these assignments.

During the training phase, submitted assignments with a suitability ratio greater equal 60% (about more than 32 out of 54 justifications selected as
suitable for each argument) over all 5 arguments were automatically excluded and rejected. The remaining assignments were compared to the pre-made annotations of the training arguments as well as to other workers’ annotations on the same task. They only resulted in rejections if the worker selected notably more not-suitable justifications than other workers, indicating that they did not follow the instructions correctly.

A total number of 216 participants was recorded for the study and a minimum amount of 3 annotations per argument was ensured. Workers were required to have an approval rate of at least 98%, at least 100 approved work tasks, and – for language proficiency – being located in the US. No further personal information were gathered. The annotators were first restricted to three annotation tasks. These training tasks contained 5 arguments exclusively from the USA-part of the dataset resulting in 200 arguments used for training. The training tasks have been done before and during the main study to select a total of 27 workers for annotating the bulk of arguments in the main study. Each training argument was annotated beforehand according to the value definitions and used as check instance for the workers performance in the training tasks. These quality checks resulted in 154 work rejections (5% rejection rate) due to ignored instructions, excluding 138 workers in the process. For the remaining 51 workers their submitted tasks were accepted as they followed the instructions, however they often selected not-suitable justifications when they were considered fitting and vice versa, indicating that the value descriptions probably weren’t fully understood.

The check instances used in the training phase were annotated before the crowd sourcing study and therefore have been classified by less than three people. To compensate for an expected bias, the annotations of each check instance given by the selected workers were used to identify their actual annotations for the final corpus using the same aggregation scheme as for the active phase. The first training tasks were also used as a pilot study for the annotation interface regarding the clarity of the provided instructions and value descriptions as well as the average time required for task completion which was essential in calculating an appropriate wage per task. To prevent dropouts and encourage workers, each task in the training phase and main study contained 5 arguments, requiring an average time of 13.3 minutes for completion. Workers were able to revise all given opinions on the suitability of the justifications as well as the optional comments before submitting an assignment. The 5 arguments also didn’t have to be annotated in the presented order. Each worker could submit an assignment only if they annotated each given argument for all 54 values preventing any empty or unfinished annotations. As all workers participating

\[11\] Using the pre-made annotations and aggregated worker selections as expectation for the true labels
in the main study have been manually approved, no additional in-task checks like gold items or attention checks were used. Due to a formatting errors in the uploaded task files, 2 assignments were submitted without annotations. These assignments have not been rejected but were understandably ignored in the final aggregation. Instead the corrected task files for their arguments have been uploaded again for proper annotation. Aside from the defective assignments, all annotations have been completed without any dropouts.

The aggregation process employed MACE Hovy et al. [2013] to fuse the annotations into a single ground truth, applying it label-wise as suggested by the authors for multi-label annotations. However, this approach treats all values independently from each other. As a result the calculated confidence regarding each annotator was not consistent across all values. With the usage of a visual revision tool (see Appendix B) 950 arguments were manually checked in a post-task step by evaluating the selected labels predicted by MACE. Moreover, a manual check was used for the 48 arguments (<1%) to which MACE assigned more than 10 values, reducing their values to the most prevalent 5-7 ones.

4.4 Outcome

During the crowd sourcing study, no harm or inconveniences had been done to the annotators aside from the invested time required for each task which has been fairly compensated. In addition, during the conducted training phase crowd workers had been sufficiently informed regarding the work and compensation for each assignment in the bulk study which was consistent across the entire crowd sourcing study.

Including the 5% rejections during the training phase, a total of 3294 assignments were submitted. As each task required the annotation of 5 arguments for all 54 values, almost 900,000 individual value decisions received over the entire study. The annotations from rejected assignments were discarded from further processing. The remaining data was completely used for aggregating the ground truth. Aside from the 200 arguments used for training, additional 750 arguments from the main study have been manually checked and their labels were partially re-evaluated based solely on the values definitions. These checks were mainly performed for arguments that showed a significantly high disagreement between the workers’ annotations.

Each Worker ID has been replaced by an anonymized hash before processing the annotations. As the final corpus only contains the aggregated data resulting from the annotations, no back-references can be made to the original workers from the publicly available dataset. Regarding the inter-annotator agreement (IAA) the workers reached an average value-wise $\alpha$ of 0.49 [Krippendorff, 2004].
This reflects the expected difficulty of the annotation task. 20 of the 54 values had their respective $\alpha$-value above 0.5 and 10 of them above 0.55 as well. A higher agreement was achieved for some values that occurred more often like Have a safe country (around 18% of USA-corpus) with an $\alpha$-value of 0.61 and Have good health (12%) with 0.70 but also for some fewer occurring values like Be holding religious faith (5%), Be protecting the environment (4%), and Have harmony with nature (6%) that all achieved an $\alpha$-value around 0.7. Considering that the majority of arguments was only annotated by three crowdworkers each, the resulting IAA stays promising in regards to the suitability of the proposed taxonomy.

Workers only annotated each argument for Level 1 of the taxonomy. The annotation for the higher levels was done using the tree-like structure of the taxonomies hierarchy where a value in the ground-truth automatically leads to an assignment of all parent labels in the taxonomy (see Figure 3.1).

For the three non-US parts the dataset was expanded by adding the specific URL reference for each argument and for the Chinese part the original, non-translated premise and conclusion were added as well. From the IBM-ArgQ-Rank-30kArgs dataset [Gretz et al., 2020] the additional quality information provided by the authors was added to the arguments in the US part, because back-references are complicated due to the cleaning and rephrasing of most arguments and the absence of an individual identifier for each argument in the source dataset.

4.4.1 Dataset discussion

The crowd sourcing study aimed to test the two questioned aspects regarding the proposed value taxonomy, namely the actual suitability towards argument classification and the taxonomy’s universal applicability across cultures.

The former appears to be confirmed through the results of the study, as the trained annotators have been able to successfully identify human values behind arguments. Regarding the sample of the 950 manually checked arguments, the annotations formed subsets of values representative for each argument with the selections having an anticipated variance in interpretations regarding the definition of each value as well as the meaning and reasoning behind a given argument. The distribution of the value frequency for each dataset part (see Table 4.1) also closely relates to the topics (or rather conclusions) present in each of the four parts, hinting towards the classification of entire controversial debates based on the values behind their respective arguments.

However, the task of identifying human values behind arguments still proved to be challenging. Especially, the apparent difficulty for human judgment in identifying the values behind arguments surpassed the expectations which is
### CHAPTER 4. CROWD SOURCING

#### Table 4.3: The 54 values of the taxonomy with dataset frequency. Table adapted from Kiesel et al. [2022].

<table>
<thead>
<tr>
<th>Level</th>
<th>1) Value</th>
<th>2) Value category</th>
<th>Dataset frequency (size)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Africa (50)</td>
</tr>
<tr>
<td>Self-direction: thought</td>
<td>Be creative</td>
<td>0.000</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>Be curious</td>
<td>0.000</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Have freedom of thought</td>
<td>0.080</td>
<td>0.000</td>
</tr>
<tr>
<td>Self-direction: action</td>
<td>Be choosing own goals</td>
<td>0.000</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Be independent</td>
<td>0.080</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Have freedom of action</td>
<td>0.080</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Have privacy</td>
<td>0.000</td>
<td>0.040</td>
</tr>
<tr>
<td>Stimulation</td>
<td>Have an exciting life</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Have a varied life</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Be daring</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Hedonism</td>
<td>Have pleasure</td>
<td>0.000</td>
<td>0.020</td>
</tr>
<tr>
<td>Achievement</td>
<td>Be ambitious</td>
<td>0.020</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>Be successful</td>
<td>0.100</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>Be capable</td>
<td>0.040</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>Be intellectual</td>
<td>0.040</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>Be courageous</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>Power: dominance</td>
<td>Have influence</td>
<td>0.040</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Have the right to command</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Power: resources</td>
<td>Have wealth</td>
<td>0.060</td>
<td>0.190</td>
</tr>
<tr>
<td>Face</td>
<td>Have social recognition</td>
<td>0.040</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Have a good reputation</td>
<td>0.020</td>
<td>0.010</td>
</tr>
<tr>
<td>Security: personal</td>
<td>Have a sense of belonging</td>
<td>0.100</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Have good health</td>
<td>0.080</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Have no debts</td>
<td>0.000</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>Be neat and tidy</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Have a comfortable life</td>
<td>0.080</td>
<td>0.260</td>
</tr>
<tr>
<td>Security: societal</td>
<td>Have a safe country</td>
<td>0.160</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Have a stable society</td>
<td>0.420</td>
<td>0.300</td>
</tr>
<tr>
<td>Tradition</td>
<td>Be respecting traditions</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Be holding religious faith</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Conformity: rules</td>
<td>Be compliant</td>
<td>0.040</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>Be self-disciplined</td>
<td>0.000</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Be behaving properly</td>
<td>0.160</td>
<td>0.070</td>
</tr>
<tr>
<td>Conformity: interpersonal</td>
<td>Be polite</td>
<td>0.000</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Be honoring elders</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Humility</td>
<td>Be humble</td>
<td>0.080</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>Have life accepted as is</td>
<td>0.040</td>
<td>0.040</td>
</tr>
<tr>
<td>Benevolence: caring</td>
<td>Be helpful</td>
<td>0.060</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Be honest</td>
<td>0.060</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Be forgiving</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Have the own family secured</td>
<td>0.000</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td>Be loving</td>
<td>0.020</td>
<td>0.020</td>
</tr>
<tr>
<td>Benevolence: dependability</td>
<td>Be responsible</td>
<td>0.060</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Have loyalty towards friends</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Universalism: concern</td>
<td>Have equality</td>
<td>0.240</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>Be just</td>
<td>0.060</td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td>Have a world at peace</td>
<td>0.260</td>
<td>0.000</td>
</tr>
<tr>
<td>Universalism: nature</td>
<td>Be protecting the environment</td>
<td>0.000</td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td>Have harmony with nature</td>
<td>0.000</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>Have a world of beauty</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Universalism: tolerance</td>
<td>Be broadminded</td>
<td>0.100</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Have the wisdom to accept others</td>
<td>0.020</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Have an objective view</td>
<td>0.020</td>
<td>0.120</td>
</tr>
<tr>
<td>Universalism: objectivity</td>
<td>Be logical</td>
<td>0.100</td>
<td>0.160</td>
</tr>
</tbody>
</table>

Table adapted from Kiesel et al. [2022].
directly reflected in the low inter-annotator agreement of 0.49. As a side effect the low IAA also influenced the quality of the dataset during the aggregation process. Together with the problems while applying MACE on multi-label data, the expressive power of the resulting dataset leaves to be questionable.

The further acquisition of the labels for Levels 2–4 presents some issues as well. As seen in Figure 4.1, the fraction of arguments being assigned both labels for Level 4a and 4b, having around 65% and 80% respectively, indicates that these levels might not be dichotomies for arguments. For Level 3, around 8% of the arguments are labeled for all four higher-order values and 30% are labeled as resorting to three of the four higher-order values. Additionally, out of the 2416 arguments that are considered resorting to exactly two of the higher-order values, a total of 408 arguments are labeled to a pair that is considered conflicting (\textit{Openness to change} and \textit{Conservation}, or \textit{Self-enhancement} and \textit{Self-transcendence}). This leads to the conclusion that the bottom-up approach of using the Level 1 labels to annotate all higher levels did not preserve the expected categorization into conflicting/opposing arguments and groups of aligning arguments which was especially expected for the four higher-order values. One way to circumvent this problem in future crowd sourcing tasks could be the usage of a top-down approach where human annotations on each level narrow down the annotation options for the next lower level.

However, the ground-truth labels for Level 1 indicate that the approach used for Levels 2–4 wasn’t the main problem. As an example, the relative co-occurrence of the three money related values (see Table 4.4a) is higher than expected, as the overlap regarding their definitions is fairly small. There is also a noticeable co-occurrence of entire value categories with related definitions. Especially in regards to the prominent categories \textit{Universalism: concern} and \textit{Security: societal} as seen in Table 4.4b, values connected to overall safety and justice were often selected together. In this regard, the expressive power of
CHAPTER 4. CROWD SOURCING

<table>
<thead>
<tr>
<th>Have wealth</th>
<th>Have no debts</th>
<th>Have a comfortable life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have wealth</td>
<td>- 0.51 0.18</td>
<td></td>
</tr>
<tr>
<td>Have no debts</td>
<td>0.23 - 0.10</td>
<td></td>
</tr>
<tr>
<td>Have a comfortable life</td>
<td>0.33 0.39 -</td>
<td></td>
</tr>
</tbody>
</table>

(a) Co-occurrence of the three money related values.

<table>
<thead>
<tr>
<th>Have equality</th>
<th>Be just</th>
<th>Have a world at peace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have a safe country</td>
<td>0.10 0.22 0.76</td>
<td></td>
</tr>
<tr>
<td>Have a stable society</td>
<td>0.21 0.24 0.38</td>
<td></td>
</tr>
</tbody>
</table>

(b) Co-occurrence regarding the categories Universalism: concern and Security: societal.

Table 4.4: Matrices showing the relative co-occurrence of selected values. Cells state the fraction of all arguments labeled with the column’s value that are also labeled with the row’s value. The complete matrix showing the relative co-occurrence of all 54 values can be seen in Appendix C.

the proposed dataset therefore does not reach it’s full potential as arguments often lack a more precise differentiation between values resorting to similar concepts. Future attempts on this crowd sourcing study should revise the value descriptions and further specify the instructions to solely focus on the core values of each argument.

In addition to the pairwise co-occurrence seen in Table 4.4a, the proposed dataset also contains 43 arguments (<1%) that are labeled for all three values Have wealth, Have no debts, and Have a comfortable life. Two examples from the bulk study are arguments against adopting an austerity regime, stating

“austerity regimes can cause widespread unemployment”

and

“an austerity regime can prevent people from having the funds they need to live the life they chose.”

A retrospective of the collected annotations revealed some controversial judgments. Regarding the three money related values, the value that fits both arguments the most is Have a comfortable life which was selected by all annotators. However, on both arguments, one of the three respective annotators

\[12\text{Arguments A22273 and A22305 in the proposed dataset}\]
selected all three values, despite the examples stating that such a case is expected to be very rare. No comments were left on either argument as well.

It thereby appears that the provided instructions and examples as well as the applied quality control were not enough to ensure the annotations to consistently follow the established guidelines. Further crowd sourcing tasks should therefore include specific check-instances in the bulk study as well and require explanatory comments if workers selected certain value combinations or an uncommonly high number of values.

Additionally, MACE selected Have wealth and Have no debts for both of the above arguments, even though only one of the three annotators selected these labels. A revision of the dataset regarding the verification of the ground truth labels is therefore a logical and necessary next step for future work on the proposed dataset. A different aggregation method and a higher number of annotations per argument would also be required for further crowd sourcing tasks, in order to acquire ground truth labels with a greater reliability.

Another issue of the current dataset is the small sample size regarding the three non-US parts. They are sufficiently widespread, especially when combined, to get a first assessment of the taxonomies applicability across cultures. However, in terms of expressive power and the ability to allow direct conclusions regarding the respective culture, the current dataset simply does not contain enough arguments. The cultural variety of the entire dataset is also not large enough to allow for a solidified claim on the universal applicability of the value taxonomy. It is thereby important to point out that the same trained US-American crowdworkers annotated the African, Indian, and Chinese part of the dataset as well. Even though the value taxonomy strives for universalism, a potential risk is that an annotator from a specific culture might fail to correctly interpret the implied values in a text written by people from a different culture. Therefore, the representative power of these three parts and the observed similarities as well as differences in value occurrences still need to be viewed skeptically as the results are only approximations of each respective culture.

For the time being it can be assumed that the proposed value taxonomy is suitable enough to classify arguments in regards of their respective values but further study is required in order to verify this claim.
Chapter 5

Automatically Identifying Values behind Arguments

This chapter reports on the first attempt at the automated identification of values behind arguments. The conducted experiments serve as a difficulty assessment of the task at hand and to provide first baselines for future research. The chapter begins with an introduction of the two experiment types and the used machine learning models, followed by the separate evaluation for each experiment type.

5.1 Experiment setup

Given the small number of arguments for the non-US parts of our dataset (cf. Table 4.1), two machine learning experiments were conducted. The first one evaluated the overall performance of each model on the US arguments as the main part of the dataset. For this matter the 71 conclusions were split\textsuperscript{13} into 60 for training (4240 arguments), 4 for validation (277 arguments), and 7 for testing (503 arguments). Only one very rare value, \textit{be neat and tidy} (0.2\% of arguments in the USA part), does not occur in this test set and was therefore excluded from evaluation. The second experiment tested the robustness of each approach in a cross-cultural setting using the three non-US parts for testing only. No additional re-training was performed on the models in order to achieve better comparison to the results of the first experiment. The non-US parts are considerably smaller and as a result \textasciitilde28\% of the values are lacking arguments (cf. Table 4.3). However, all used machine learning models are equally effected by this lack, thus providing for a comparison with the previous setting.

For both experiments the models predicted the labels for each taxonomy level

\textsuperscript{13}Usage of each argument is noted in the available dataset.
separately and, regarding the low ratio of arguments per conclusion on the non-US parts, only the premise part of each argument was used for the classification. As discussed in Chapter 4, the amount of arguments assigned to each value in the dataset is quite low, averaging about 8% of the US-arguments being assigned to each value. Therefore, the machine learning approaches were expected to struggle on creating accurate models. In addition, the fraction of arguments being assigned both labels for Level 4a and 4b, having around 65% and 80% respectively (cf. Figure 4.1), is the main problem regarding the suitability for automated classification of these levels. The machine learning models were trained and tested on these levels as well to complete the generation of baseline results, however their scores were not expected to be of meaningful impact, as no clear differentiation between the two labels can be learned. With the number of arguments labeled for more than two higher-order values being only slightly lower, the results on Level 3 were also expected to be less representative on the actual task, albeit of more significance than the results on 4a and 4b. However, Level 2 of the taxonomy was considered to contain enough data for meaningful baseline results with the USA-part having on average 17% of the arguments assigned to each value category. There still remained the difficulty of the overall low argument count on the non-US parts, but the results for Level 2 were expected to allow for a first assessment regarding the models’ cross-cultural robustness.

All approaches used out-of-the-box models/concepts which were only slightly fine-tuned on the validation set. The implementation regarding training and testing the used machine learning approaches can be found online.\textsuperscript{14}

\textbf{1-Baseline} As this work is the first attempt at automatically identifying personal values in natural language arguments, a solidified baseline that every following model has to be competitive against, would be a simple function classifying each argument as resorting to all values. Score-wise this classifier would always receive a recall of 1 and its precision equals the actual values’ distribution within the dataset. Especially when employing the $F_1$-score as metric, this model achieves scores that are at least as high if not higher than using label-wise random guessing according to the label frequency. Additionally, as there is no random element involved, the resulting scores for the 1-Baseline model are consistent allowing to replicate and verify the presented results.

\textsuperscript{14}https://github.com/webis-de/ACL-22
SVM To compare against an additional baseline model, we used a linear kernel support vector machine (SVM)\textsuperscript{15} and trained it label-wise with $C = 18$.

Transformer-based model We fine-tuned multi-label bert-base-uncased [Wolf et al., 2020] with a batch size of 8, a learning rate of $2^{-5}$, and 20 epochs. All preparations and executions of the experiments regarding the transformer-based model were done by Milad Alshomary from Kiesel et al. [2022].

5.2 Evaluation

The evaluation focuses on the label-wise $F_1$-score and its mean over all labels (macro-average), as well as its constituents precision and recall. Accuracy is reported for completeness, though the heavily skewed label distribution makes it less suited. The 1-Baseline model always achieves a recall of 1 making it an especially strong model for the $F_1$-score. For calculating the $p$-values when comparing approaches, this work employs the Wilcoxon signed rank significance test [Wilcox, 1996].

5.2.1 Experiment 1 (USA)

Regarding the $F_1$-scores, the BERT model performed better than both Baseline models on Level 1 ($p = 0.007$ vs. SVM and $p = 0.001$ vs. 1-Baseline; $n = 53$) and Level 2 ($p = 0.153$ and $p = 0.117$; $n = 20$). On the higher levels, the number of labels is too small for a significance test. However, the $F_1$-scores for the BERT model on the Levels 3 and 4 are equal to or lower than the scores of the 1-Baseline. Especially, the results for the two base dichotomies (Level 4a and 4b) turned out as expected. For Level 4b the metrics of the BERT model are identical to the 1-Baseline model indicating that the BERT model simply classified each argument as resorting to both labels (cf. Table 5.1).

Not only is the total argument count of the dataset too low in respect of the (on average 4) selected values per argument but also the amount of topics/conclusions and their diversity. Due to the conclusion-based dataset split, the training set contains 180 arguments regarding the value \textit{Be protecting the environment} whereas the test dataset only contains two. Therefore, even though a precision of 0.40 appears too low for practical use, considering the mentioned difficulties of the dataset and the application of out-of-the-box approaches, an average $F_1$-score of 0.25 is promising for future attempts.

\textsuperscript{15}https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
CHAPTER 5. MACHINE LEARNING EXPERIMENTS

<table>
<thead>
<tr>
<th>Model</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4a</th>
<th>Level 4b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>Acc</td>
<td>P</td>
</tr>
<tr>
<td>BERT</td>
<td>0.40</td>
<td>0.19</td>
<td>0.25</td>
<td>0.39</td>
<td>0.30</td>
</tr>
<tr>
<td>SVM</td>
<td>0.21</td>
<td>0.19</td>
<td>0.20</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>1-Baseline</td>
<td>0.08</td>
<td>1.00</td>
<td>0.16</td>
<td>0.08</td>
<td>0.18</td>
</tr>
</tbody>
</table>

**Table 5.1:** Macro precision (P), recall (R), F1-score (F1), and accuracy (Acc) on the USA test set over all labels by level. Best scores per metric and level marked bold. Table taken from Kiesel et al. [2022].

![Values (Level 1)](image1)

**Figure 5.1:** Parallel coordinates plot of F1-scores on the USA test set over the labels by level. The grey bars show the label distribution, which is equal to the F1-score of random guessing as per this distribution. Figure adapted from Kiesel et al. [2022].

In regards to the Levels 1 and 2, BERT reached considerably higher F1-scores for multiple labels (cf. Figure 5.1). The identification performed especially well on the value *Have good health* (F1: 0.81) and the value category *Security: personal* (F1: 0.78) with both having a precision and recall around 0.8. It is also notable that BERT performed on *Have good health* better than on *Be just* (F1: 0.53) even though less arguments are resorting to *Have good health*.

BERT also performed slightly better on the value *Have wealth* (F1: 0.45) than on the category *Power: resources* (F1: 0.39) despite both labels spanning the exact same arguments. This might indicate a beneficial usage of multiple hierarchy levels in a combined classification approach. Future machine learning attempts could thereby achieve higher F1-scores through model stacking or additional feature extraction with convolutional layers.
Table 5.2: Macro F₁-score on each test set over all labels by level. Best scores per part and level marked bold. The scores for USA are the same as in Table 5.1. Table taken from Kiesel et al. [2022].

<table>
<thead>
<tr>
<th>Model</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4a</th>
<th>Level 4b</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>0.20</td>
<td>0.21</td>
<td>0.30</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.37</td>
<td>0.68</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.38</td>
<td>0.37</td>
<td>0.60</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>0.41</td>
<td>0.60</td>
<td>0.68</td>
<td>0.71</td>
</tr>
</tbody>
</table>

5.2.2 Experiment 2 (Cross-cultural)

The BERT model again performed better than both Baselines on Level 1 (p = 0.006 vs. SVM and p < 0.001 vs. 1-Baseline; n = 169) and Level 2 (both p < 0.001; n = 74). For Level 3 (p = 0.179 and p = 0.856; n = 16), the difference

As seen in Table 5.2, each models had a similar performance throughout the four dataset parts. However, it is worth noting that on Level 1 both the SVM and the BERT model achieved a higher F₁-score for the Indian dataset part than for the USA part, despite the lower ratio of annotated values (indicated by the lower F₁-score for 1-Baseline). The BERT model also performed better on this part regarding Level 2 as well.

Regarding the 16 value categories which are actually present in the African dataset part (cf. Table 4.1), an additional 7 categories have less than five
arguments resorting to it. The meaningfulness and representative power of the scores on these labels as well as the resulting averaged F$_1$-score should therefore be taken skeptically. It is still worth to note that the value *Have wealth*, being annotated to 3 arguments in the African dataset part, was predicted with an F$_1$-score of 1 (cf. Figure 5.2). Similarly, the category *Power: resources* achieved an equally high score, as only an additional fourth argument has been misclassified as resorting to this category.

However, a different perspective is shown in regards to the Chinese dataset part, especially comparing the value *Have a good reputation* and its value category *Face*. Even though both are annotated to the exact same single argument (cf. Table 4.1), the category *Face* was predicted with an F$_1$-score of 1 whereas *Have a good reputation* was not predicted correctly (F$_1$: 0.0). As the BERT models accuracy on *Have a good reputation* for this dataset part was 0.99, it certainly classified an argument for this value (false positive). Another example from the Indian dataset part are the values *Be protecting the environment* and *Be humble* annotated to a single argument each and having an F$_1$-score of 1 and 0 respectively. Therefore, the classification for some less represented labels appears more like a lucky guess. For values and value categories with higher argument count like *Have equality* or *Security societal* the BERT model performed quite similar throughout all dataset parts.

Overall, the BERT model appears to be suitable for identifying values behind arguments even across cultures. However, the current size and cultural approximation in the proposed dataset is not sufficient to test this claim further.
Chapter 6

Conclusion

The task of identifying human values behind arguments provides challenges interpretations of value names and concepts. The research in this work contributes (1) a multi-level taxonomy consisting of 54 basic human values, (2) a dataset of 5270 arguments labeled for each taxonomy level and gathered from four different sources, and (2) first baseline results on automated value identification for multiple levels of granularity applied on and compared between different cultures.

Based on the findings, next goals would be to further test the universal applicability of the value taxonomy and to expand the dataset in terms of argument count, cultural variety, and different languages. Especially the argument acquisition process and the expressive power of the ground-truth labels are topics for required improvements in order to more precisely approximate cultures. Further research on improving the machine learning approaches for identifying values behind arguments has already been planned in the Task on Human Value Detection at SemEval 2023 [Kiesel et al.]. As this research is motivated by the promising results of the BERT model on Level 2, improvements on the automated detection could also be beneficial for personality profiling as done by Maheshwari et al. [2017].

A universal taxonomization of values across cultures and domains provides also benefits towards argument strength and value-based argumentation frameworks (VAFs) [Bench-Capon, 2003]. Until now, the VAFs have been mainly applied onto legal arguments [Bench-Capon, 2021] where the respective values have been extracted from domain-specific factors [Chorley and Bench-Capon, 2005]. A taxonomy comprised of universal values extends the range of application outside of law concerning debates, thereby creating further opportunities for the usage of value-based models in practical reasoning across topics.

Finally, a value taxonomy suited for argument classification across topics and cultures provides usage in digital applications as well. Identifying and
precisely stating the values and argument resorts to could assist in avoiding misunderstandings between humans and automated argumentation systems [Kiesel et al., 2021]. Together with the circular arrangement and the different levels of granularity, this classification allows for universal visualizations of a debate’s topic-space which could be used as an improvement for argument search engines [Kiesel et al., 2018; Wachsmuth et al., 2017] and combined with the large amount of data from Internet/Web archives it creates another support to further research on societal challenges [Kiesel, 2022] and even analyze the evolution of human values and their usage in arguments over time.
Bibliography


Appendices
Appendix A

Annotation Interface

The following lists screenshots from various stages regarding the development process of the crowd sourcing interface. Figure A.1 and A.2 show the earliest stages with annotation concepts based on the value surveys which were used for the taxonomy. A following layout test (see Figure A.3) applied an early version of the multi-level taxonomy for the first time. The annotation layout that was eventually decided on can be seen in Figure A.4. The development process also featured an alternative color palette (see Figure A.5) which was later discarded due to it being considered too distracting.

Figure A.6 and A.7 show screenshots of the final annotation interface. Its source code is available online\textsuperscript{16} as part of the published data from Kiesel et al. [2022].

\footnote{https://doi.org/10.5281/zenodo.5657249}
APPENDIX A. ANNOTATION INTERFACE

Figure A.1: Screenshot of the annotation interface based on the SVS [Schwartz, 1994]. Arguments would be annotated firstly in regards to their motivational type (left side) and for the type selected as *mainly* the argument would be annotated in regards to the contained values (right side).

Figure A.2: Screenshot of the annotation interface based on the RVS [Rokeach, 1973]. It uses Sortable.js\(^\text{17}\) to model a prioritization approach similar to the one in the RVS.

\(^{17}\text{https://sortablejs.github.io/Sortable/}\)
Figure A.3: Screenshot of the annotation interface using the Levels 1 and 2 of the proposed taxonomy. It combines the prioritizing approach from Figure A.2 with the two level annotation from Figure A.1. Selecting the i-icon on a label card displays additional information for the respective category or value.
Figure A.4: Screenshot of the first attempt at the new layout for the annotation interface.

Figure A.5: Screenshot of an earlier version of the final annotation interface using an alternative color palette, allowing color discrimination for people with dichromacy or anomalous trichromacy.
Figure A.6: Screenshot of the first part of the annotation interface, containing instructions and examples. Figure taken from Kiesel et al. [2022].
APPENDIX A. ANNOTATION INTERFACE

Figure A.7: Screenshot of the second part of the annotation interface, which consists of three panels: (1) the top left panel places the argument in a scenario ("Imagine"); (2) the top right panel formulates the annotation task for a value (here: *have wealth*) as a yes/no question, describing the value with examples; and (3) the bottom panel shows the annotation progress for the argument and allows for a quick review of selected annotations. Figure taken from Kiesel et al. [2022].
Appendix B

Revision Tool

The revision tool (Figure B.1) was implemented using R [R Core Team, 2021] with the Shiny framework [Chang et al., 2021].

![Screenshot of the interface used for annotation revisions during the crowd sourcing study (see Chapter 4). For check instances (as in this example) the tool highlighted the expected values that are considered definitely suitable (dark green) and possibly suitable (light green) for the respective argument. The multi-label MACE predictions (row name: Multi) are displayed as well.]

Figure B.1: Screenshot of the interface used for annotation revisions during the crowd sourcing study (see Chapter 4). For check instances (as in this example) the tool highlighted the expected values that are considered definitely suitable (dark green) and possibly suitable (light green) for the respective argument. The multi-label MACE predictions (row name: Multi) are displayed as well.
Appendix C

Corpus Statistics

The complete matrix of all 54 values regarding their relative co-occurrence in the crowd sourced dataset (Chapter 4) can be seen in Table C.1. The gradient of red reflects each cells value.
Table C.1: Matrix showing the relative co-occurrence between all 54 values in the final dataset. Cells state the percentage of all arguments labeled with the column's value that are also labeled with the row's value.