A Survey on Modelling Morality for Text Analysis

Ines Reinig*♠, Maria Becker,*♣ Ines Rehbein*♠, Simone Paolo Ponzetto♠

♠Data and Web Science Group
University of Mannheim, Germany
{reinig,rehbein,ponzetto}@uni-mannheim.de

♣Germanistisches Seminar
University of Heidelberg, Germany
maria.becker@gs.uni-heidelberg.de

Abstract

In this survey, we provide a systematic review of recent work on modelling morality in text, an area of research that has garnered increasing attention in recent years. Our survey is motivated by the importance of modelling decisions on the created resources, the models trained on these resources and the analyses that result from the models' predictions. We review work at the interface of NLP, Computational Social Science and Psychology and give an overview of the different goals and research questions addressed in the papers, their underlying theoretical backgrounds and the methods that have been applied to pursue these goals. We then identify and discuss challenges and research gaps, such as the lack of a theoretical framework underlying the operationalisation of morality in text, the low IAA reported for many human-annotated resulting resources and the lack of validation of newly proposed resources and analyses.

1 Introduction

With the rise of large language models, research goals in NLP have also become more ambitious, tackling new challenges like the prediction of complex psychological constructs. A case in point is the modelling of morality in text, a task that requires a deep and comprehensive understanding of natural language. More and more studies at the interface of NLP and Computational Social Science (CSS) have addressed this task, based on different theoretical frameworks, and a variety of methods have been applied to predict moral values from text. These studies aim at modelling the moral sentiment that a person or group holds toward a certain target in order to investigate research questions from the political or social sciences. Others model morality in the context of AI applications, e.g., to study the inherent moral biases learned by language models. Many resources have been created, but are often difficult to find due to their heterogeneous, interdisciplinary research backgrounds.

Modelling morality using NLP techniques has also been at the center of recent events such as workshops or coding competitions. For example, Kiesel et al. (2023) identify human values behind arguments in a SemEval shared task (ValueEval'23)[1] while the MP^2 workshop[2] at NeuRIPs 2023 looks at the application of theories from moral philosophy and psychology to AI practices, demonstrating the growing interest in exploring morality with computational methods in interdisciplinary settings.

This survey aims at providing a systematic overview of recent work on modelling morality in text, the resources that are available, and the methods that have been applied. We argue that the operationalisation of morality has an immense impact on (i) the annotated resources that are created, (ii) the models that are trained, based on these resources, (iii) the predictions obtained from these models, and (iv) the analyses that result from the predictions. Our survey therefore focusses on the theoretical basis and modelling part, as well as on the validation of the concept.

This is *not* a survey on ethics in AI, which was the main focus of Vida et al. (2023). Motivated by the confusion regarding concepts from philosophical ethics in NLP research, Vida et al. (2023) focus on concepts from philosophy and analyse literature on moral NLP with respect to their philosophical foundation and terminology. We, instead, focus on the *theoretical modelling and operationalisation of morality for text analysis* and the challenges that arise for applications in CSS and Cultural Analytics. These different backgrounds are also reflected in the paper selection method and the set of surveyed papers, which we briefly discuss in the

*Equal contribution between the first three authors.

[1] https://touche.webis.de/semeval23/touche23-web/index.html

[2] https://aipsychphil.github.io/

Findings of the Association for Computational Linguistics ACL 2024, pages 4136–4155

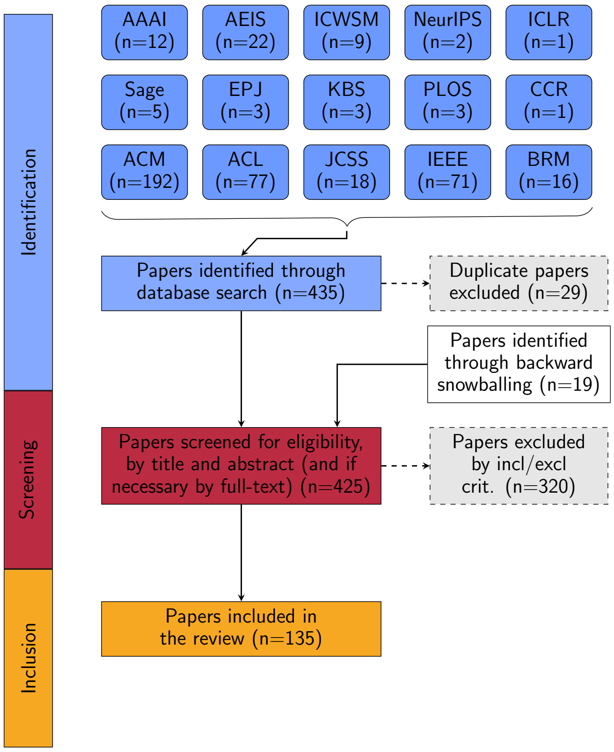

Figure 1: PRISMA-inspired flow diagram describing our paper sampling method.

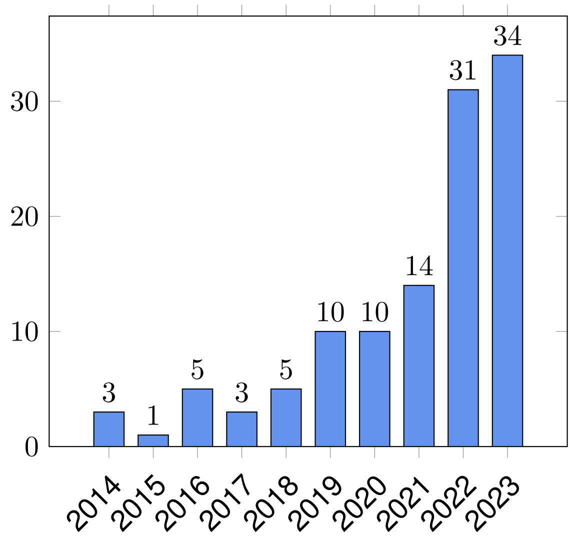

Figure 2: Distribution of publication years of the 116 reviewed papers.

Appendix, §A.2.

The survey is structured as follows. We first describe our survey methodology (§2) and outline the research objectives and background of the papers included in the survey (§3). Then we describe different operationalisations of morality in text (§4) and their impact on applications in the social sciences (§5). We discuss trends and research gaps in Section 6 before we conclude and outline some recommendations (§7). The supplementary materials together with our code can be found in our GitHub repository: https://github.com/umanlp/survey_morality.

2 Survey methodology

In the following, we briefly describe our methodology for selecting and reviewing the papers for this survey. Our paper selection method is inspired by Alturayeif et al. (2023) and also follows recommendations from Moher et al. (2009). The different steps are summarized in Figure 1, more details can be found in the appendix.

Paper sampling (see §A.1.1 and §A.1.2) We first semi-automatically identify potentially relevant papers. We search for specific keywords in 15 selected journals and venues (blue boxes in Figure 1). After this paper sampling step, we obtain 435 papers, which we then deduplicate, amounting to 406 papers. We identify 19 more papers that might be relevant for the survey using backward snowballing.

Screening and reviewing (see §A.1.4 to §A.1.6) The set of papers resulting from paper sampling and snowballing is then screened following six selection criteria (cf. §A.1.4) that must all be satisfied for a paper to be included in the survey. After screening, 135 papers are kept and distributed amongst the authors for reviewing. To increase consistency across reviewers, we use a survey form with an accompanying codebook for reviewing, both refined during an initial pilot study based on eight selected papers. After the reviewing process, we identify 4 demo papers, 1 shared task paper and 14 papers that, according to our selection criteria, were not included in the survey. This leaves us with 116 papers as the basis of the analyses presented in the remainder of this work. Figure 2 shows the distribution of publication years for the papers, demonstrating the growing interest in the field (also see Table 7 in the appendix).

3 Survey overview

To get an overview of the work included in the survey, we first classify papers according to their main research objectives, as listed below (note that a paper can have more than one objective).

1. **Values, Stance, Framing:** investigate the moral values of a person, group, or culture;

Research interest	# papers
Value/Stance/Framing	87
Morality in AI	30
Comparison	21
Moral Theories	10
Other Theories	2

Table 1: Objectives for modelling morality.

explore the moral sentiment towards a target; identify moral rhetoric and framing.

2. **Morality in AI**: investigate morality in the context of AI systems or applications.
3. **Comparison**: compare moral values to other concepts (e.g., stance, emotions).
4. **Moral Theories**: evaluate or improve a moral theory.
5. **Other Theories**: evaluate or improve another theory (not related to morality).

The majority of the papers model morality to analyse the moral values and stances held by an individual or group, and to investigate moral framing (87 papers, see Table 1). The next frequent class focusses on morality in AI (30 papers), often with the goal of improving an LLMs understanding of moral values or to investigate moral bias encoded in LLMs. 21 papers compare moral values to other concepts, such as stances and emotions, and 12 papers aim at evaluating or improving a theoretical framework.

According to their main contributions, the papers can be broadly categorised into experimental papers, analysis papers and resource papers.[3]

Experimental papers This category includes work that focusses on the development and evaluation of methods for the identification of moral language in text. With 81 papers, it is the largest of the three categories. Regarding the machine learning methods used, we observe the following trends (see Table 2): Not surprisingly, most works use fine-tuned transformers (29 papers), followed by dictionary- or rule-based approaches (25 papers). Feature-based ML has mostly been applied in older papers, while more recent work also uses zero- and few-shot learning based on LLMs.

Analysis papers The second largest category includes 65 papers that present an analysis based on the application of NLP methods to predict moral values in text. Papers in this category address research questions from the CSS, for example, Zhang and Counts (2016) investigate moral values in the context of anti versus pro-abortion policies, Islam and Goldwasser (2022b) study moral messages used in COVID-19 vaccine campaigns while Ertugrul et al. (2019) predict social protest activities from social media discussions. Most analysis papers are from the political and social sciences (46), followed by media and communication studies (11), psychology (5) and other fields (3).

Method	# papers
Fine-tuned transformers	29
Dictionary-/rule-based	25
Feature-based ML	18
LLMs (zero-/few-shot)	14
NNs with static embeddings	8
Semi-supervised ML	8
Logic-based ML	6
Unsupervised ML	10
Reinforcement Learning	2

Table 2: Methods employed in the papers (ignoring baselines).

Resource papers The last and smallest category includes 56 papers that release annotated datasets (e.g., Trager et al. (2022); Mooijman et al. (2018)), dictionaries (e.g., Zúquete et al. (2023); Araque et al. (2022)) or ontologies (e.g., De Giorgis et al. (2022); Hulpuș et al. (2020)). Most of the 56 papers present a new, annotated dataset (38 papers), 11 papers create or expand a moral dictionary, while 9 others create, link or augment ontologies or knowledge graphs with moral vocabulary.[4] The majority of the papers employ trained annotators (20 papers) while 14 papers use crowdsourcing. To control for annotation bias, 8 of the 14 crowdsourcing papers collect information regarding the coders' demographics or their political or moral views.

In summary, the majority of the papers in our survey use text-based models of morality to study moral sentiment and moral framing, typically applied to research questions from the CSS. Given the strong focus on real-world applications, the question arises as to the validity of the various approaches used to model morality in text. We address this issue in the next section.

4 Operationalisations of morality

Next, we study how the different papers in the survey operationalise the concept of morality. We first give an overview over the task of moral value

[3]Note that the contributions are not mutually exclusive.

[4]These contributions are, again, not mutually exclusive.

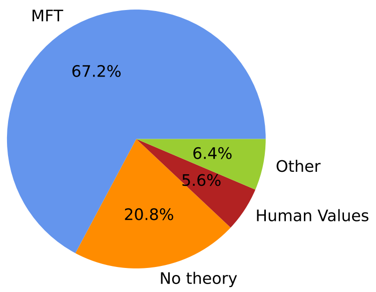

Figure 3: Theories used in the papers.

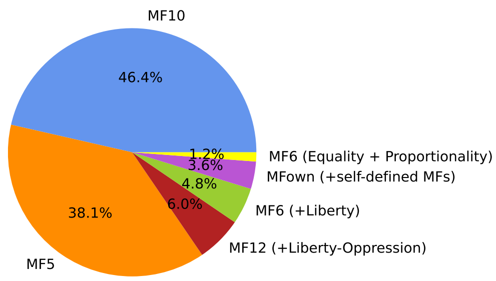

Figure 4: Number of foundations encoded in papers based on MFT

prediction. Then we look at the theoretical frameworks that have been employed, and turn to the question on which level morality is investigated in the papers. Finally, we discuss the issue of underspecification in modelling morality and its negative impact on the reliability and validity of the analyses.

4.1 Task description

Most papers in the survey model morality by predicting the moral attitudes, beliefs, sentiment or emotions expressed in a text. Let us consider Example 4.1 below, taken from the Twitter corpus of Roy et al. (2022).

Example 4.1. "Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first. "

The task in those papers is then to assign one or more labels that describe the moral values expressed in the text, where the labels depend on the theoretical framework used to model morality (see §4.2). The label could simply predict whether the text includes moral language or not (morality: yes/no), or it could include more fine-grained information describing the moral values or beliefs expressed in the text, for example, the moral foundations described in MFT (see §4.2 below) or the human values defined in Schwartz and Bilsky (1987). Some works additionally encode the strength of the moral message by means of a continuous score (Araque et al., 2020), others additionally try to identify moral roles, such as the entity causing harm or the target of the moral message (Roy et al., 2022), as illustrated in the example below.

Example 4.2.

"Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first."

patients: Target of Care/Harm

Another difference we found with regard to the operationalisation of morality in the survey relates to the level of analysis, i.e., whether moral values are predicted on the level of documents, sentences or tokens or whether the work also tries to identify moral frames together with their roles expressed in the text (see §4.3).

4.2 Overview of theoretical frameworks

Figure 3 provides an overview of the theoretical frameworks used in the papers. The vast majority (84 papers) uses Moral Foundations Theory (MFT) (Graham et al., 2013), which we describe below. Remarkably, 25 papers do not rely on a specific theory or framework. Out of those, 8 papers aim at modelling social norms by extracting unspoken commonsense rules, often referred to as Rules-of-Thumb (RoT) (Forbes et al., 2020).[5]

Other theories and concepts that have been used to model morality include Schwartz' Theory of Human Values (Schwartz and Bilsky, 1987), Hofstede's Cultural Dimensions (Hofstede, 2001), the Theory of Contractualist Moral Decision-Making (Levine et al., 2018; Awad et al., 2022), Moral Disengagement (Bandura, 1999, 2016), the Path Model of Blame (Malle et al., 2014) and the Economics of Convention Framework (Boltanski and Thévenot, 2006).

Since MFT is the predominant theoretical framework for modelling morality in texts, below we provide a brief introduction to the main concepts of this theory.

Moral Foundations Theory (MFT) is a descriptive, pluralist theory of morality that has been highly influential within the field of moral psychology (Haidt et al., 2009; Graham et al., 2013).

[5]Please note that while we decided to categorise this work as not being based on any specific theory, the authors point out relations to work on descriptive and applied ethics (Hare, 1981; Kohlberg, 1976).

While monists believe that one single dimension is sufficient to understand and explain morality, pluralists argue that the concept of morality is based on more than one such dimension, or foundation.

In MFT, moral foundations (MFs) are conceptualised as intuitions, i.e., as an "evaluative feeling (like–dislike, good–bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion" (Haidt and Bjorklund, 2008). In short, moral foundations are intuitions or "gut feelings" that often drive moral reasoning and turn it into rationalisation.[6] MFT assumes that the foundations have been developed during evolution as responses to several adaptive challenges. The foundations that have been proposed so far can be classified into *binding* foundations (ingroup LOYALTY, respect for AUTHORITY, and PURITY) and *individualising* foundations (CARE and FAIRNESS).[7]

MFT does not claim to know how many of these foundations exist, instead it proposes a set of criteria for *foundationhood* (see Table A.4) and encourages researchers to revise and extend the set of moral foundations. Another basic assumption of MFT states the existance of an innate draft of the moral mind that is later revised by experience and cultural influences (Graham et al., 2013, p. 9), thus making it an interesting basis for cross-cultural investigations of morality.

MFT comes in different flavours We observe that the MFT-based papers differ with respect to the number of foundations considered for analysis. As shown in Figure 4, many papers either apply a set of five moral foundations (MF5: Care, Fairness, Loyalty, Authority, Purity) or use the five foundations, but with separate classes for the vice–virtue scale (MF10: Care vs. Harm, Fairness vs. Cheating etc.). This is probably due to the fact that many papers rely on existing resources that utilise the MF5 or MF10 schema (e.g. the English Moral Foundations Dictionary (MFD)).

Only three papers include new, self-defined moral foundations. One of them is Cheng and Zhang (2023) which we describe in more detail below, as the same validation method has also been applied for the creation of the English and Japanese Moral Foundations dictionaries. The authors adapt the English MFD v2.0 to Chinese and propose six new candidates for moral foundations (Liberty, Altruism, Diligence, Waste, Resilience, and Modesty).[8] They start with a translation of the MFD2.0 to Chinese and add over 1,200 Chinese words related to morality. Then they use several rounds of expert coding to assign the dictionary entries to MFs and ask crowdworkers to write short essays about moral issues related to each foundation. This resulted in a benchmark with over 2,200 texts labelled for the different MFs that are then used to validate the dictionary, based on the average word-frequencies for each class and a word density analysis. The authors emphasise the importance of testing MFT in different cultural settings, in order to advance and refine the theory.

While this approach is very much in line with the theoretical assumptions of MFT, such as the five criteria for *foundationhood* mentioned above (also see Table A.4), other work seems less aware of the theoretical constraints imposed by MFT, adding arbitrary new candidates to the list without providing proper validation. A case in point are candidates like *Feminism–Maleness; Sustainability–Climate change; Peace–War* (González-Santos et al., 2023) that clearly ignore the criteria of *foundationhood*.

4.3 Level of analysis

One perspective from which we can investigate how papers operationalise the analysis of morality in text is the *level* of analysis, i.e., whether morality has been analysed on the *document, segment, sentence* or *token* level, or whether the analysis is based on *frames*. By *segment* we refer to any text span that is a substring of the document, longer than a token, and not a sentence (this includes both text spans that are shorter and longer than sentences). As opposed to *segment*, the *document* level refers to an entire text; this includes tweets and other social media posts or messages, since they contain the whole text and not just a substring of it. We define *frames* as units that additionally encode the relations between text spans describing moral situations, actions, goals or values and their roles, such as the target or holder of a moral sentiment.

Table 3 shows that the majority of the papers analyse morality on the document level. The high

[6] Note that this social intuitionist view is crucially different from other views that consider moral intutions as "strong, stable, immediate moral beliefs" (Sinnott-Armstrong et al., 2010) or as moral judgments (McMahan, 2000).

[7] In newer versions of MFT, FAIRNESS has been further divided into the EQUALITY and PROPORTIONALITY MFs.

[8] Some of these had already been proposed in the literature while others are new.

Level	# papers
Document	77
Segment	18
Token	14
Sentence	9
Frame	9

Table 3: Levels of analysis for annotating morality.

number can be traced back to the fact that most papers base their analysis on tweets and Reddit posts. Only a small number of papers (9) provide a frame-based analysis of morality that explicitly encodes the holder and target of the moral sentiment.

We argue that the practise of annotating morality at the document or sentence level instead of explicitly encoding the participants and their roles leads to *underspecification* in the coding scheme which not only results in low inter-annotator agreement (IAA) but also fails to capture *perspective*, i.e., who is the holder of the moral sentiment. We discuss these issues below.

4.4 The problem of low IAA

A number of the survey papers have created manually annotated resources, however, less than half of the papers make their annotation guidelines publicly available (22 out of 54 resource papers). Looking at the available guidelines, we find that about half of them have a length of less than or up to one page only. These short guidelines often present the coders with a highly underspecified task description, instructing them to assign labels to sentences or documents without an operationalisable definition of the concept of morality.

Our survey includes a number of papers dedicated to the annotation of morality. Many of them report rather low scores for inter-annotator agreement (see Table 4). Note that the scores are not comparable, given that the projects use different schemas and annotation settings. However, we can see that about half of the projects report an IAA below 0.5. Three projects report medium scores (>0.6) while only four of the projects achieved an IAA above 0.7 (κ/α).

We now look at studies that report high IAA in order to determine the underlying reasons. The highest scores are reported in Cheng and Zhang (2023) who asked trained coders to classify moral words from the Chinese Moral Dictionary into MFs. Weinzierl and Harabagiu (2022) report high IAA for COVID-19 Vaccine Hesistancy Frames (VHFs). They identify and aggregate VHFs in tweets and annotate the aggregated and cleaned frame representations. Islam and Goldwasser (2022a) also analyse COVID-19 Vaccine Hesistancy, focussing on Facebook ads. They report a high IAA (0.74 Cohen's κ) for the annotation of MFs, based on a small set of 110 instances. Johnson and Goldwasser (2018) collect tweets by US congress members for a number of controversial, morally charged topics (Abortion, ACA, Guns, Immigration, LGBTQ, Terrorism) and report a κ of 0.79 for the annotation of MFs for those topics. Shahid et al. (2020) annotate moral foundations in news articles on the sentence level. They report negative IAA for the binary distinction of whether a sentence includes moral language or not. However, for sentences where the annotators agreed on the presence of moral language, IAA for the MF type was high (0.85 Krippendorff's α).

IAA	Schema	Source
	Krippendorff's alpha	
-0.03	moral (yes/no)	Shahid et al. (2020)
0.2–0.46	MF6+Liberty	Ziems et al. (2022)
0.61	MF12+Liberty	Pacheco et al. (2022)
0.91	MF+own	Cheng and Zhang (2023)
	Cohen's kappa	
0.28-0.53	MF5	Kobbe et al. (2020)
0.32	moral values	Alshomary et al. (2022)
0.32	MF12+Liberty	Beiró et al. (2023)
0.62	Blame Theory	Orizu and He (2016)
0.65	MF10	Rezapour et al. (2019)
0.74	MF6+Liberty	Islam and Goldwasser (2022a)
0.79	MF10	Johnson and Goldwasser (2018)
0.85	MF10	Weinzierl and Harabagiu (2022)
	Fleiss's kappa	
0.46	MF+own	Karami et al. (0)
0.16-0.46	MF10	Hoover et al. (2020)

Table 4: IAA for the annotation of morality reported in the papers. Note that the scores are not comparable, as the annotations use different label sets and IAA metrics.

In sum, all studies that obtained high agreement coded moral foundations in texts that were already filtered for morally charged language. Thus, the difficult part of the annotation is *not* to classify moral language into MFs but to decide whether or not a text includes moral language. In other words, it seems as if the low IAA reported in many papers can be traced back to the question of what counts as moral language. This finding suggests that full-text annotation, i.e., assigning labels to each sentence in a news article or to each document in a collection of social media messages, might not be a good approach. This is further illustrated by the following example from the Moral Foundations Reddit Corpus (Trager et al., 2022):

Example 4.3. "Dude I think Bernie is racist. No shit I'm gonna think Le Pen supporters are racist."
(MFRC, Subreddit: neoliberal)

Coder1: EQUALITY
Coder2: CARE, PURITY, EQUALITY
Coder3: LOYALTY, AUTHORITY

Here, three trained coders assigned 5 out of 6 possible MF labels to this post, with only one label chosen by more than one coder. This inconsistency is not necessarily evidence that the coders have different moral values but could simply show that, when left without specific instructions, the coders focus on different aspects in the text and, instead of annotating the text author's moral values, assign labels based on free word associations like *racist* → EQUALITY, *supporters* → LOYALTY.

4.5 The problem of perspective

Example 4.5 shows that the annotators can be tempted to code different perspectives in the same text, as illustrated below. In this case, coder 1 has chosen the AUTHORITY label, probably referring to Wilder's moral foundations, while the other annotators use EQUALITY to encode the moral values of the text author.

Example 4.4. "Don't know about Le Pen but I know Wilders was worryingly on the fascist side of things" (MFRC, Subreddit: europe)
Coder1: AUTHORITY
Coder2: CARE, PURITY, EQUALITY
Coder3: EQUALITY

This underspecification and mix of perspectives casts considerable doubt on the construct validity of the annotations and their reliability for text analysis. As a remedy to this problem, we recommend the use of a frame-based approach, as in Roy et al. (2022). We argue that encoding moral roles like the holder and target of a moral sentiment is crucial for the analysis of morality in text. This is illustrated in Roy et al. (2022) who show that in discussions about abortion in the US both conservatives and liberals use the CARE-HARM moral foundation, however, with different targets. While conservatives focus on the protection of unborn life, liberals prioritise the well-being of the women. This example shows that the MF label on its own is not sufficient to capture the differences between conservatives and liberals. We therefore find the lack of studies that code morality on the level of frames surprising and hope that this research gap will be addressed in the near future.

5 Applications of moral value prediction

We now focus on the application of computational methods for moral value prediction in real-world analyses. Our findings show that the thoroughness of the research methodology leaves something to be desired. Only few of the 65 analysis papers in the survey explicitly state their research questions (13 papers), and an equally small number of papers (10) formulates hypotheses and tests for statistical significance, while the vast majority of the analysis papers use data exploration or visualisations to support their findings. As dictionaries were among the most often applied method for predicting moral values in the analysis papers (also see Table 2), we next look at the available resources and their validity in real-world applications.

5.1 Validity of dictionary-based approaches

The disadvantages of dictionary-based text analysis are well known and have been discussed at length, e.g., in the context of sentiment analysis. Those drawbacks also apply to the analysis of moral values, most importantly the insensitivity of dictionaries to word meaning in context and their failure to handle compositionality, such as negation (Wiegand et al., 2010). Another crucial issue of dictionary-based methods is their failure to capture perspective, i.e., to identify the holder of the moral sentiment expressed in the text.

The most often used resources for dictionary-based moral value prediction are the English Moral Foundations Dictionary (MFD) (Graham et al., 2009) and extensions thereof. The MFD was originally developed for comparing Christian sermons from liberal and conservative churches in the US, as sermons often include moral narratives. The resulting scores were mostly in line with the predictions made by the theory, which Graham et al. (2009) take as a validation of the approach. However, the authors report that they also tried to use the MFD on a corpus of Republican and Democratic candidates' convention speeches but were not able to extract distinctive moral values from the data, thus calling into question the general applicability of the approach to texts from other domains where moral narratives might not be as omnipresent as in the sermons.

Hopp et al. (2021) question the representativeness of the original, expert-created MFD and propose using crowdworkers to identify morally relevant passages in text, and then use the crowdsourced annotations to extract terms for their expanded version of the dictionary, the eMFD. They argue that this procedure is more apt to treat MFs as "the products of fast, spontaneous intuitions", thus being superior to an expert-curated word list.

To validate their approach, Hopp et al. (2021) apply the eMFD to news articles from far-left, center-left and far-right news outlets. As the eMDF yielded more distinctive scores for the different partisan news than the MFD and MFD2.0, the authors conclude that their method is superior to the other two dictionaries. This, however, is no conclusive proof that the above dictionaries are reliable and valid approximations of measurement tools like the MFT Questionnaire (Graham et al., 2011) which has been successfully tested for *internal and external validity* and *test-retest reliability* (see §A.5), using confirmatory factor analysis.

Another question that arose from the survey concerns the assumptions and prerequisites that must be fulfilled to guarantee the reliability and validity of the results. In the following, we discuss one of the most important prerequisite for empirical analyses, namely the *representativeness* of the data for the population studied.

5.2 Representativity of the data

The question of when a corpus is representative enough to answer research questions about a certain population has been discussed at length in the field of corpus linguistics (see, e.g., Biber (1993); Egbert et al. (2022)) and these findings and best practices should be taken into account.

Moral values across cultures Representativity is particularly important for research questions that investigate moral values *across cultures*, such as Wu et al. (2023), who compare folk tales from 27 cultural backgrounds, based on a crowdsourced dictionary. In their study, they use an opportunistic collection of folk tales translated into English. The European cultures are represented at fine-grained levels in the corpus, with small countries like Denmark, Norway and Sweden regarded as separate cultures. The African continent, on the other hand, is categorised as one culture only, represented by folk tales from West Africa. This design seems rather imbalanced and Eurocentric and casts doubts on the validity of the analysis.

We therefore argue that while it is necessary to formulate hypotheses and carry out significance tests (as done by only 10 out of 65 analysis papers), it is crucial to also ensure the representativeness of the data for the respective research question. We therefore release a checklist that addresses some of the design decisions relevant for ensuring the representativity of the data.[9]

6 Trends, research gaps and recommendations

After discussing issues regarding the modelling of morality in text and limitations of current applications to research questions in the Computational Social Sciences, we now outline some trends emerging from the survey, as well as research gaps and challenges that we would like to see addressed in future work.

Resources for languages other than English While several corpora and dictionaries are available for the analysis of morality in English text, only few resources exist for other languages. This is in line with the findings in Vida et al. (2023). Out of the 18 studies that work with languages other than English, only 6 release an annotated dataset for a new language. 5 papers use dictionary-based approaches, typically based on translations of one or more of the English MFDs without proper validation for the new language. A notable exception is Cheng and Zhang (2023) who introduce new moral foundations for Chinese and also test for validity (see §4.2). The remaining papers either use survey-derived measures of morality, annotate stances on moral topics, or present an annotation tool. Thus, the creation of annotated datasets and tools for languages other than English remains an important research objective.

Modelling grounded in theory A high percentage of the papers included in the survey (20%) are not based on any theoretical framework. We argue that grounding models of morality in theory has several advantages. First, it provides a link to previous research and can thus inform our research design and modelling decisions. Second, a sound and well-defined theoretical framework can help address the problem of underspecification in modelling. Finally, the theory can provide us with testable hypotheses and, vice versa, our results can help improve theory development. We therefore recommend researchers that aim at building new

[9]The checklist is included in our GitHub repository: https://github.com/umanlp/survey_morality.

resources to choose an appropriate framework as the theoretical basis for their work.

Underspecification Another problem we found concerns the creation of annotated resources as training data. As mentioned above, annotation procedures and guidelines lack (i) *transparency and reproducibility*, as only half of the survey papers release their guidelines, and (ii) *specificity*, as the guidelines are often shorter than one page. We therefore recommend the creation and publication of more specific and detailed guidelines for the annotation of morality that provide coders with sufficient information on how to deal with the already challenging task.

In Section 4.5, we argued that it is crucial for the annotation of morality to also encode the *perspective*, not only to reduce inconsistencies in the annotations, but also to make the resulting resources more useful for different types of analyses. We therefore recommend annotating morality at the level of frames and roles, to explicitly capture this information.

Representativity, reliability, validity We highlight the importance of ensuring the representativity of the research data for the population relevant for the respective research questions, especially in comparative studies. To increase the reliability and validity of the analyses, we advise researchers to formulate their research questions, form hypotheses and test for statistical significance. In addition, more work is needed on how to test for different types of validity concerning the construct of morality.

Morality in AI systems While not in the center of our interest, we find 20 papers that probe large language models (LLMs), either to search for biases or to investigate what LLMs have learned about morality. Some papers explore whether language models such as BERT or ChatGPT capture moral norms or include a "moral compass" (cf. e.g. Hendrycks et al. (2021); Schramowski et al. (2022)). Most of the approaches try to assess a language model's knowledge of moral concepts by analyzing the semantic space of the underlying word or sentence representations or use prompts to provide the models with morality related questions or scenarios. While these papers mainly try to uncover the moral and ethical biases of AI systems, we expect to see more research in the future which goes one step further, by not only making the (mostly western centric) moral biases in LLMs transparent but by developing methods for removing the undesirable properties of these models.

7 Conclusion

In the survey, we present a systematic review of work on modelling morality in text. We highlight problems that need to be addressed, one of them being the lack of resources for languages other than English. Another issue concerns the low IAA for the annotation of morality in text. We argue that one reason for this, besides the inherent subjectivity of the task, results from underspecified task instructions, witnessed by a) very short or unavailable guidelines and b) the attempt to code morality on the sentence or document level, making it hard for the annotators to know which aspects of the text they are supposed to code. To address this issue, we recommend annotating morality at the level of frames, to make it clear what is being annotated and from whose perspective.

Another problem we found is the lack of validation of both the resources and the analyses. While many papers use NLP methods to investigate research questions in CSS, only few studies formulate hypotheses or use significance testing. Equally important is the representativeness of the data, especially in comparative studies, and the extent to which it is justified to replace carefully validated methods such as questionnaires with automated dictionary-based text analysis procedures. This should be investigated in future studies. With the paper, we release a checklist that addresses crucial design decisions for text-based analysis that we hope will help researchers to identify some of these issues.

Limitations

While we took great care in selecting the papers for our survey, we are aware that there might be relevant papers that we did not consider as they were published in venues not included in our list. We also did not search for broad terms such as "value", as they resulted in hundreds of thousands of hits for some of the journals and venues that we were not able to screen. Furthermore, the page limit did not allow us to include all references that might be relevant for this topic.

In addition, the discussion about the validity of automated dictionary-based text analysis techniques or other NLP methods for predicting the moral values of an individual, social group or culture has only scratched the surface. However, a more in-depth discussion requires far more space and is therefore beyond the scope of this survey.

Ethics Statement

We strictly adhere to the ACL Ethics Policy. We make all of our approaches, resources, and methods transparent and publicly available and we ensure that the findings and conclusions of our survey are reported accurately and objectively. We believe that our work is helpful for researchers who work on the computational analysis of morality in text and discuss research gaps and possible future research directions, e.g. by addressing the problem of validity of automated text analyses for the prediction of moral values or the lack of resources for non-English languages. We do not anticipate any ethical concerns arising from the research presented in this paper.

Acknowledgements

The work presented in this paper is funded by the German Research Foundation (DFG) under the UNCOVER project (PO1900/7-1 and RE3536/3-1). We thank Rainer Gemulla for his constructive and helpful comments on an earlier version of this paper.

References

Areej Alhassan, Jinkai Zhang, and Viktor Schlegel. 2022. 'am I the bad one'? predicting the moral judgement of the crowd using pre–trained language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 267–276, Marseille, France. European Language Resources Association.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2021. The language of liberty: A preliminary study. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 623–626, New York, NY, USA. Association for Computing Machinery.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2022. Libertymfd: A lexicon to assess the moral foundation of liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, GoodIT '22, page 154–160, New York, NY, USA. Association for Computing Machinery.

Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, MJ Crockett, Jim AC Everett, Theodoros Evgeniou, Alison Gopnik, Julian C Jamison, TW Kim, SM Liao, MN Meyer, J Mikhail, K Opoku-Agyemang, JS Borg, J Schroeder, W Sinnott-Armstrong, M Slavkovik, and JB Tenenbaum. 2022. Computational ethics. *Trends in Cognitive Sciences*, pages 388–405.

Albert Bandura. 1999. Moral disengagement in the perpetration of inhumanities. *Personality and social psychology review*, 3(3):193–209.

Albert Bandura. 2016. *Moral disengagement: How people do harm and live with themselves*. Worth Publishers.

Mariano Beiró, Jacopo D'Ignazi, Victoria Bustos, María Prado, and Kyriaki Kalimeri. 2023. Moral narratives around the vaccination debate on facebook. In *Proceedings of the ACM Web Conference 2023*, WWW'23), pages 4134–4141.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4):243–257.

Luc Boltanski and Laurent Thévenot. 2006. *On justification: Economies of worth*. Princeton University Press.

Flavio Carvalho, Helder Yukio Okuno, Lais Baroni, and Gustavo Guedes. 2020. A brazilian portuguese moral foundations dictionary for fake news classification. In *The 39th International Conference of the Chilean Computer Science Society (SCCC)*, pages 1–5.

Calvin Yixiang Cheng and Weiyu Zhang. 2023. C-mfd 2.0: Developing a chinese moral foundation dictionary. *Computational Communication Research*, 5(2):1–47.

Stefano De Giorgis, Aldo Gangemi, and Rossana Damiano. 2022. Basic human values and moral foundations theory in valuenet ontology. In *Knowledge Engineering and Knowledge Management: 23rd International Conference, EKAW 2022, Bolzano, Italy, September 26–29, 2022, Proceedings*, page 3–18, Berlin, Heidelberg. Springer-Verlag.

Jesse Egbert, Douglas Biber, and Bethany Gray. 2022. *Corpus Representativeness: A Conceptual and Methodological Framework*, page 52–67. Cambridge University Press.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ali Mert Ertugrul, Yu-Ru Lin, Wen-Ting Chung, Muheng Yan, and Ang Li. 2019. Activism via Attention: Interpretable Spatiotemporal Learning to Forecast Protest Activities. *EPJ Data Science*.

Tess Feyen, Alda Mari, and Paul Portner. 2023. Pragmatic annotation of articles related to police brutality. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 146–153, Toronto, Canada. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Jeremy A Frimer, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehgani. 2019. Moral foundations dictionary for linguistic analyses 2.0.

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior research methods*, 50:344–361.

Carlos González-Santos, Miguel A. Vega-Rodríguez, Carlos J. Pérez, Joaquín M. López-Muñoz, and Iñaki Martínez-Sarriegui. 2023. Automatic assignment of moral foundations to movies by word embedding. *Knowledge-Based Systems*, 270:110539.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5069–5087, Seattle, United States. Association for Computational Linguistics.

Katharina Haemmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, Toronto, Canada. Association for Computational Linguistics.

Jonathan Haidt and Fredrik Bjorklund. 2008. Social intuitionists answer six questions about morality. In W. Sinnott-Armstrong, editor, *Moral Psychology*, volume 2, pages 181–217. Cambridge, MA: MIT Press.

Jonathan Haidt, Jesse Graham, and Conrad Joseph. 2009. Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, 20(2-3):110–119.

Richard Mervyn Hare. 1981. *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch Critch, Jerry Li Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. In *International Conference on Learning Representations*.

Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53:232–246.

Xiaolei Huang, Alexandra Wormley, and Adam Cohen. 2022. Learning to adapt domain shifts of moral values via instance weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 121–131, New York, NY, USA. Association for Computing Machinery.

Ioana Hulpuș, Jonathan Kobbe, Heiner Stuckenschmidt, and Graeme Hirst. 2020. Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80, Barcelona, Spain (Online). Association for Computational Linguistics.

T. Islam and D. Goldwasser. 2022a. Understanding covid-19 vaccine campaign on facebook using minimal supervision. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 585–595, Los Alamitos, CA, USA. IEEE Computer Society.

Tunazzina Islam and Dan Goldwasser. 2022b. Understanding covid-19 vaccine campaign on facebook using minimal supervision. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 585–595.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *NeurIPS*.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Behnam Karami, Fatemeh Bakouie, and Shahriar Gharibzadeh. 0. A transformer-based deep learning model for persian moral sentiment analysis. *Journal of Information Science*, 0(0):01655515231188344.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Lawrence Kohlberg. 1976. Moral stages and moralization. *Moral development and behavior*, pages 31–53.

Sydney Levine, Max Kleiman-Weiner, Nicholas Chater, Fiery Cushman, and Josh Tenenbaum. 2018. The cognitive mechanisms of contractualist moral decision-making. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, CogSci 2018.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Khyati Mahajan and Samira Shaikh. 2020. Studying the effect of emotional and moral language on information contagion during the charlottesville event. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 128–130, Seattle, USA. Association for Computational Linguistics.

Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry*, 25(2):147–186.

Brodie Mather, Bonnie Dorr, Adam Dalton, William de Beaumont, Owen Rambow, and Sonja Schmer-Galunder. 2022. From stance to concern: Adaptation of propositional analysis to new tasks and domains. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3354–3367, Dublin, Ireland. Association for Computational Linguistics.

Akiko Matsuo, Kazutoshi Sasahara, Yasuhiro Taguchi, and Minoru Karasawa. 2019. Development and validation of the japanese moral foundations dictionary. *PLoS ONE*, 14(3):1–10.

Jeff McMahan. 2000. Moral intuition. In Hugh LaFollette -, editor, *The Blackwell Guide to Ethical Theory*, pages 92–110. Blackwell.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7).

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396.

Udochukwu Orizu and Yulan He. 2016. Detecting expressions of blame or praise in text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4124–4129, Portorož, Slovenia. European Language Resources Association (ELRA).

Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 5821–5839, Seattle, United States. Association for Computational Linguistics.

Matheus Camasmie Pavan, Vitor Garcia dos Santos, Alex Gwo Jen Lan, João Trevisan Martins, Wesley Ramos dos Santos, Caio Deutsch, Pablo Botton da Costa, Fernando Chiu Hsieh, and Ivandré Paraboni. 2023. Morality classification in natural language text. *IEEE Trans. Affect. Comput.*, 14(1):857–863.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.

Ming Qian, Jaye Laguardia, and Davis Qian. 2021. Morality beyond the lines: Detecting moral sentiment using ai-generated synthetic context. In *Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings*, page 84–94, Berlin, Heidelberg. Springer-Verlag.

Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159, Singapore. Association for Computational Linguistics.

Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Minneapolis, USA. Association for Computational Linguistics.

Valentina Rizzoli. 2023. The risk co-de model: detecting psychosocial processes of risk perception in natural language through machine learning. *Journal of Computational Social Science*.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 1–13, Online. Association for Computational Linguistics.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2022. Towards few-shot identification of morality frames using in-context learning. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 183–196, Abu Dhabi, UAE. Association for Computational Linguistics.

Martin Ruskov, Maria Dagioglou, Marko Kokol, Stefano Montanelli, and Georgios Petasis. 2023. A knowledge graph of values across space and time. In *Proceedings of the 2nd Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2023) co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Roma, Italy, November 6, 2023.*, volume 3536, pages 8–20. CEUR-WS.org.

J. Fernando Sánchez-Rada, Oscar Araque, Guillermo García-Grao, and Carlos Á. Iglesias. 2023. SLIWC, morality, NarrOnt and senpy annotations: four vocabularies to fight radicalization. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 617–626, Vienna, Austria. NOVA CLUNL, Portugal.

Wesley Santos and Ivandré Paraboni. 2019. Moral stance recognition and polarity classification from Twitter and elicited text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1069–1075, Varna, Bulgaria. INCOMA Ltd.

Patrick Schramowski, Cigdem Turan, and Nico Andersen. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268.

Shalom H Schwartz and Wolfgang Bilsky. 1987. Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53(3):550–562.

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 120–125, Online. Association for Computational Linguistics.

Walter Sinnott-Armstrong, Liane Young, and Fiery Cushman. 2010. 246Moral Intuitions. In *The Moral Psychology Handbook*. Oxford University Press.

David Solans, Christopher Tauchmann, Aideen Farrell, Karolin Kappler, Hans-Hendrik Huber, Carlos Castillo, and Kristian Kersting. 2021. Learning to classify morals and conventions: Artificial intelligence in terms of the economics of convention. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 691–702. AAAI Press.

Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The moral foundations reddit corpus.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Maxwell A. Weinzierl and Sanda M. Harabagiu. 2022. From hesitancy framings to vaccine hesitancy profiles: A journey of stance, ontological commitments and moral foundations. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1087–1097.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68, Uppsala, Sweden. University of Antwerp.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-cultural analysis of human values, morals, and biases in folk tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Mengyao Xu, Lingshu Hu, and Glen T Cameron. 2023. Tracking moral divergence with ddr in presidential debates over 60 years. *Journal of Computational Social Science*, 6(1):339–357.

Amy X. Zhang and Scott Counts. 2016. Gender and ideology in the spread of anti-abortion policy. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3378–3389, New York, NY, USA. Association for Computing Machinery.

Xinliang Frederick Zhang, Winston Wu, Nick Beauchamp, and Lu Wang. 2023. Moka: Moral knowledge augmentation for moral event extraction. *arXiv preprint arXiv:2311.09733*.

Chunxu Zhao, Pengyuan Liu, and Dong Yu. 2022. From polarity to intensity: Mining morality from semantic space. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1250–1262, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joan Zheng, Scott Friedman, Sonja Schmer-galunder, Ian Magnusson, Ruta Wheelock, Jeremy Gottlieb, Diana Gomez, and Christopher Miller. 2022. Towards a multi-entity aspect-based sentiment analysis for characterizing directed social regard in online messaging. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 203–208, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

Mafalda Zúquete, Diana Orghian, and Flávio L. Pinheiro. 2023. A Moral Foundations Dictionary for the European Portuguese Language: The Case of Portuguese Parliamentary Debates. In *Computational Science – ICCS 2023: 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part I*, page 421–434, Berlin, Heidelberg. Springer-Verlag.

A Appendix

A.1 Methodology details

In this section of the Appendix, we provide more details about each step of our methodology for selecting and reviewing papers in this survey. The different steps are summarized in Figure 1 (main paper).

A.1.1 Databases

We include ACL, ACM and IEEE as these are large and well-known databases[10] containing a vast variety of literature related to NLP and Computer Science from peer-reviewed journals, conferences and workshops. Additionally, we identified relevant journals and publication venues from the Moral Foundations homepage. We manually scanned the publication page for relevant papers by reading *titles*, *abstracts* and *keywords* and then added the corresponding journals to our list. In total, we consider 15 venues and journals.

A.1.2 Search strings

We employ different search strings to detect relevant papers. For some NLP-related venues such as ACL, we opt for a more general search term, e.g. "moral", in order to increase recall. Whenever possible, we search in the fields *title*, *abstract* and *keywords*.

Table 5 provides an overview of search strings and results for each database. For transparency and reproducibility, we provide additional details on the search[11] and the exhaustive search results[12] the GitHub repository.

[10]By database, we mean any source used to search for papers. For example, this can be a search on a website in a particular journal or venue.

[11]`https://github.com/umanlp/survey_morality/search_queries.txt`

[12]`https://github.com/umanlp/survey_morality/search_results/`

Database	Fields	Keywords	File/URL	No. papers
ACM Digital Library	title, abstract, keywords	$\mathbf{S}_{\text{search}}$	Query link	192
ACL Anthology	title, abstract	*moral*	anthology+abstracts.bib	77
Journal of Computational Social Science	any	moral	Query link	18
IEEE Xplore	title, abstract, author keywords	$\mathbf{S}_{\text{search}}$	Query link	71
Behavior Research Methods	any	$\mathbf{S}_{\text{search}}$	Query link	16
Sage Journals	abstract, keywords	$\mathbf{S}_{\text{search}}$ (abstract) and $\mathbf{S}_{\text{extra}}$ (keywords)	Query link	5
EPJ Data Science	any	moral foundation	Query link	3
Knowledge-Based Systems	any	$\mathbf{S}_{\text{search}}$	Query link	3
PLOS One	title, abstract	moral	Query link (CL), Query link (NLP)	3
Computational Communication Research	any	moral	Query link	1
AAAI Conference on Artificial Intelligence	title	moral	Query link	12
AIES	title	moral	Query link	22
ICWSM	title	moral	Query link	9
NeurIPS	title	moral	Query link	2
ICLR	title	moral	Query link	1
Total (w/ duplicates)				435
Total (w/o duplicates)				406

Table 5: Number of papers (last column) found in each database (first column). The column **Fields** indicates in which fields we searched. The column **Keywords** lists the keywords we searched for. For all databases except ACL Anthology, we conducted a web search. The column **File/URL** shows either the file in which we searched (for ACL Anthology) or a link to the exact query that was used to retrieve documents. The cut-off for all searches is 31.12.2023. Note that query links might produce different results when visited at a later date.

ACL Anthology We search the anthology.bib file for the string `"moral"` in lowercased *titles* and *abstracts*: any string containing the substring "moral" will result in a match, e.g. "Exploring Morality in Argumentation". If the string appears in any of the two fields, we select it to be included in the screening phase.

DBLP We also search the proceedings of NeurIPS, ICLR and three AAAI-related venues: (i) AAAI Conference on Artificial Intelligence, (ii) AIES AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society and (iii) ICWSM International AAAI Conference on Web and Social Media. To increase the number of search results, we search for the string "moral" without any refinements, as we did for the ACL Anthology.

We conduct the search in the DBLP computer science bibliography, since the aforementioned conferences do not offer a search function that includes all proceedings, nor a way to download all years' proceedings of a conference at once. At the time of writing, the DBLP web interface does not allow us to exclude the author field from the search, so we also obtain numerous matches where the substring "moral" is part of the authors' names. We thus exclude them from the search results, keeping only publications with a match in the title. For example, the query string for the AAAI Conference on Artificial Intelligence shown in Table 5 originally returned 26 results, but we only count the 12 results that contain "moral" in the title.[13] In short, we search for the case-insensitive substring `"moral"` in titles only.

Search strings for other sources For any other source, whenever possible, we opt for a detailed search string:[14]

$\mathbf{S}_{\text{search}}$: `"moral foundation" OR "moral foundations" OR "moral value" OR "moral values" OR "moral sentiment" OR "morality frame" OR "morality frames" OR "moral rhetoric"`

We limit the search to the paper's metadata fields *title*, *abstract* and *keywords* (or *author keywords* in the case of IEEE).

Additionally, if the search using only $\mathbf{S}_{\text{search}}$ re-

[13]The DBLP web search does not allow to search in abstracts.

[14]We define this search string based on terminology encountered in the literature.

turns too many (unrelated) results, we further constraint the search using the following string in the *keywords* metadata field:

S$_{\textbf{extra}}$: `"content analysis" OR "text analysis" OR "discourse analysis" OR "semantic analysis" OR "machine learning" OR "deep learning" OR "NLP"`

For some databases, however, (i) a complex string search is too restrictive, yielding little to no results, or (ii) the web interface does not allow for an advanced search. In these cases, we relax the search by only searching for the word `"moral"` or the string `"moral foundation"` in any metadata field.

A.1.3 Filtering and supplementing

After filtering for duplicates, the number of papers resulting from the search is reduced from 435 to 406 (see Table 5). After screening these papers, 123 relevant papers are left. We then supplement this set of papers with backward snowballing. 19 papers are considered as potential candidates, from which 12 remain after screening, amounting to a total of 135 kept for reviewing.

A.1.4 Screening process

We closely follow the inclusion and exclusion criteria proposed in Alturayeif et al. (2023). To be included in the survey, the paper must satisfy **all** of the inclusion criteria below:

1. The methodology must rely on text or speech data. For example, we do not include papers that analyze morality using **only** methods from psychology (e.g. psychometric questionnaires) or data science (e.g. purely demographic attributes of users). This criterion enables to only include papers that are related to text analysis using computational methods, and more generally to NLP.
2. The paper proposes a new resource (such as a dataset), or an experiment, or an analysis. This means that we exclude papers that are (solely) surveys or reviews.
3. The paper is either a short or long paper in conference findings, a journal or workshop findings. Publications that are only available as posters, abstracts or other short-form or visual formats are not included.
4. The paper must be written in English.
5. The paper must be peer-reviewed.[15]
6. The paper must be accessible.[16]

A.1.5 Pilot study

After completing the screening, we conduct a pilot review study with the goal of refining the survey form that we use for reviewing the papers. We select eight papers (see Table A.1.5) from different venues and years in order to obtain a more diverse and informative set of papers for the pilot.

ID	Paper
1	Matsuo et al. (2019)
2	Haemmerl et al. (2023)
3	Zhao et al. (2022)
4	Ziems et al. (2022)
5	Mahajan and Shaikh (2020)
6	Orizu and He (2016)
7	Xu et al. (2023)
8	Liu et al. (2022)

Table 6: List of papers included in the pilot review study of the survey.

These eight papers are then reviewed independently by the three authors of this paper using a test version of the survey form. We discuss open questions, refine and enrich the survey form, and create a codebook with precise definitions and examples of the survey categories.

A.1.6 Review process

Based on the test version of the review form, we create a browser-based survey form to collect reviews for each paper. The final selection of 135 papers is distributed amongst the authors for the final reviewing. The codebook ensures that all variables are well defined and are used consistently by the reviewers.[17]

Survey form Figure 5 illustrates part of the survey form that was used to collect reviews for all papers. Reviewers proceed by starting a server, which launches the survey form application. After entering the bibkey of a paper, metadata such as title, authors and year are entered automatically in the corresponding fields. The survey form consists of radio buttons, multiple choice buttons and free

[15]Note that papers obtained via backward snowballing are exempt from this constraint; we implement it during the paper sampling phase in order to ensure publication quality of semi-automatically collected papers.

[16]Besides open source publications, we include papers that we could access through our university libraries (all except 10 papers).

[17]The codebook is available on GitHub.

text fields. A codebook accompanying the survey form ensures consistent reviews.

Consistency checks After completing the reviewing process, we run semi-automated consistency checks on the output of the survey forms. In order to find potential mistakes in the reviewing process, we check different variables for consistency, for example:

- The content length of a paper must be lower or equal the total length.
- Certain obligatory fields cannot be left empty.
- Certain variables require a specific truth value for other variables: for instance, in a paper that includes supervised classification of morality, LLMs without fine-tuning cannot be the only method related to experiments that is true.

The script used to run these consistency checks can be found in the supplementary materials.

A.2 Methodological differences to Vida et al. (2023) in the paper selection method

Vida et al. (2023) review 92 papers, while our survey reviews 135 papers. There is an overlap of 48 papers that have been included in both surveys, however, our survey covers 88 additional papers not considered in Vida et al. (2023). This shows the different focal points of our work, which is reflected in the design of the selection criteria used in each survey. While we search in a range of 15 selected journals and venues, Vida et al. (2023) consider ACL, ACM and the 100 most relevant search results on the search engine Google Scholar. Our search method is not only independent of the ranking of a search engine, but also reproducible and reliable. Additionally, search strings are defined differently in both surveys, reflecting the different backgrounds of both works. For example, Vida et al. (2023) also search for broader concepts related to ethics, such as "utilitarianism" or "deontology"; our work, on the other hand, has a narrower focus on the operationalization of morality in NLP and CSS works.

A.3 Distribution of paper types per publication year

Tables 7 and 8 respectively show the distribution of paper types and research goals over the years.

Year	Analysis	Exp.	AI	Demo	Res	NotRel
2014	3	0	0	0	0	1
2015	1	0	0	0	0	0
2016	4	2	0	0	0	0
2017	2	0	0	0	0	0
2018	2	3	0	0	2	1
2019	5	4	1	1	3	4
2020	3	7	0	0	5	2
2021	7	4	2	1	5	1
2022	10	21	6	2	10	3
2023	10	15	9	0	11	2

Table 7: Distribution of paper types over time.

Year	V/S/F	Comparison	Theory	AI
2014	5	0	0	0
2015	1	1	0	0
2016	10	0	0	1
2017	4	1	1	0
2018	6	1	0	0
2019	17	2	3	2
2020	12	3	2	2
2021	28	3	1	2
2022	33	7	3	12
2023	41	4	2	12

Table 8: Distribution of research goal over time (V/S/F: Values/Stance/Framing; Theory: moral + other).

A.4 Criteria for "Foundationhood"

Table 9 lists the MFT criteria for foundationhood, as described in Graham et al. (2013).

ID	*Criteria*
1	A common concern in third-party normative judgments
2	Automatic affective evaluations
3	Culturally widespread
4	Evidence of innate preparedness
5	Evolutionary model demonstrates adaptive advantage

Table 9: The five criteria for foundationhood, as detailed in Graham et al. (2013).

A.5 Validation of the MFT Questionnaire

Measurement tools like the MFT Questionnaire (Graham et al., 2011) are thoroughly tested for different types of validity and reliability, such as:

1. **Internal validity** assesses to which extend we can be sure that the measured effect has been caused by the variable of interest and not by some other factors, such as external variables or alternative explanations.

2. **External validity** assesses how well the findings generalize to other settings (e.g., other time periods, another population, etc.)

(a) General information about the paper

(b) Modelling of moral values

Figure 5: Exempts from two sections in the html-based survey form that we used to collect reviews for each paper. In this example, we review Matsuo et al. (2019).

3. **Test-retest reliability** assesses the reliability of results when applying the same experimental treatment twice to a group of subjects over a period of time. Results are reliable when both treatments give the same or similar results.

Dictionaries do not undergo any such validation procedure and can thus only be considered as a very rough approximation of the above measurement tools.

A.6 List of resources

We provide detailed lists of resources for morality in text, including annotated corpora (see Table 11), dictionaries (Table 10) and ontologies/knowledge graphs (Table 12).

Authors	Lang.	size
Graham et al. (2009)	en	323
Frimer et al. (2019)	en	2,103
Rezapour et al. (2019)	en	4,636
Araque et al. (2020)	en	487
Hopp et al. (2021)	en	689
Mather et al. (2022)	en	8,468
Araque et al. (2021)	en	3,074
Carvalho et al. (2020)	pt	790
Cheng and Zhang (2023)	zh	6,138
Matsuo et al. (2019)	ja	718

Table 10: Available dictionaries. Please note that the sizes are not comparable, as some dictionaries include word forms while others include lemmas or regexes. Some dictionaries also include a generic MORAL category (not included in the counts above).

Authors	Lang.	Size	Annot. setup	Annot. schema	IAA	Available
Alhassan et al. (2022)	No info	175000 documents	No annotation	Not relevant	Not relevant	Yes
Alshomary et al. (2022)	en	230k texts, 60 arguments	mixed	Yes	Yes	Partly
Beiró et al. (2023)	en	457065 documents	trained	No	Yes	No
De Giorgis et al. (2022)	en	unknown	No annotation	Not relevant	Not relevant	Yes
Emelin et al. (2021)	en	12k moral stories	crowd	No	Yes	Yes
Feyen et al. (2023)	en, fr	430 documents (newspaper articles)	trained	Yes	No	No
Forbes et al. (2020)	en	292k RoTs	crowd	Yes	No	Yes
Garten et al. (2018)	en	3000 Tweets	trained	No	Yes	No
Guan et al. (2022)	en, zh	4209 Chinese documents, 1779 English documents	mixed	Yes	Yes	Yes
Hendrycks et al. (2021)	en	130,000 examples (elicited moral scenarious and comments from Reddit)	crowd	Yes	Yes	Yes
Hoover et al. (2020)	No info	35108 tweets	trained	Yes	Yes	Yes
Huang et al. (2022)	en	500 Tweets	trained	No	No	No
Islam and Goldwasser (2022a)	en	557 documents	trained	No	Yes	Yes
Jin et al. (2022)	No info	148 vignettes	crowd	Yes	Not relevant	Yes
Johnson and Goldwasser (2018)	en	2050 documents	trained	No	Yes	No
Karami et al. (0)	fa	6000 Tweets	trained	No	Yes	No
Kiesel et al. (2022)	en	5270 arguments	crowd	Yes	Yes	Yes
Kobbe et al. (2020)	en	320 documents	trained	Yes	Yes	Yes
Lin et al. (2018)	en	4191 tweets	trained	No	Yes	No
Mather et al. (2022)	en	8473 dictionary entries	trained	Yes	Yes	Yes
Mooijman et al. (2018)	en	4800 Tweets	trained	No	Yes	Yes
Orizu and He (2016)	en	7660 text segments	No annotation	No	Yes	No
Qian et al. (2021)	No info	1514 stories	No info	No	No	No
Pacheco et al. (2022)	en	750 annotated tweets and 85,000 unlabeled tweets	trained	Yes	Yes	Yes
Pavan et al. (2023)	pt-br	4080 documents	crowd	No	Not relevant	No
Pyatkin et al. (2023)	en	clarification questions for 6425 situations	crowd	Yes	No	Yes
Rao et al. (2023)	No info	20537 human-validated contextualisations/rationales	crowd	Yes	Yes	Yes
Rizzoli (2023)	it	2381 Tweets	trained	No	No	Yes
Roy and Goldwasser (2021)	No info	161k documents (tweets)	No annotation	No	Not relevant	No
Roy et al. (2022)	en	1599 tweets	crowd	Yes	Yes	No
Ruskov et al. (2023)	en	not specified	mixed	No	No	No
Sánchez-Rada et al. (2023)	en, it, es	unknown	No annotation	Not relevant	Not relevant	Yes
Santos and Paraboni (2019)	pt-br	5242 texts	No info	No	No	No
Shahid et al. (2020)	No info	400 documents, 100 documents	mixed	Yes	Yes	No
Solans et al. (2021)	No info	2565 instances (mostly sentences)	trained	No	No	Yes
Trager et al. (2022)	en	16123 documents	trained	Yes	Yes	Yes
Weinzierl and Harabagiu (2022)	No info	14180 documents,	trained	No	Yes	Partly
Zhang et al. (2023)	en	474 news articles	trained	Yes	Yes	Yes
Zheng et al. (2022)	en	24425 entities	crowd	No	Yes	No
Ziems et al. (2022)	en	38k chatbot replies to human-authored prompts, 113817 prompt-answer pairs	crowd	Yes	Yes	Yes

Table 11: Overview of available datasets. The column **Annot. setup** indicates whether there was a manual annotation (crowd annotators, trained annotator or mix of both). The column **Annot. schema** indicates whether annotation guidelines were made available. **IAA** indicates whether inter-annotator agreement was reported. The last column **Available** indicates whether the created dataset is publically available. The reported information is not always available in the paper; in this case, we note "No info".

Authors	Lang.	Size	Annot. setup	Annot. schema	IAA	Available
Ruskov et al. (2023)	en	not specified	mixed	No	No	No
Garten et al. (2018)	en	3000 Tweets	trained	No	Yes	No
Lin et al. (2018)	en	4191 tweets	trained	No	Yes	No
Pyatkin et al. (2023)	en	clarification questions for 6425 situations	crowd	Yes	No	Yes
De Giorgis et al. (2022)	en	unknown	No annotation	Not relevant	Not relevant	Yes
Zhang et al. (2023)	en	474 news articles	trained	Yes	Yes	Yes
Feyen et al. (2023)	en, fr	430 documents (newspaper articles)	trained	Yes	No	No
Sánchez-Rada et al. (2023)	en, it, es	unknown	No annotation	Not relevant	Not relevant	Yes
Weinzierl and Harabagiu (2022)	No info	14180 documents	trained	No	Yes	Partly

Table 12: Available ontologies. Columns are analogous to Table 11.