Moral reckoning: How reliable are dictionary-based methods for examining morality in text?

Ines Rehbein[1], Lilly Brauner[2], Florian Ertz[3], Ines Reinig[1], Simone Ponzetto[1],
[1]Mannheim University, [2]Heidelberg University, [3]Göttingen University
Correspondence: rehbein@uni-mannheim.de

Abstract

Due to their availability and ease of use, dictionary-based measures of moral values are a popular tool for text-based analyses of morality that examine human attitudes and behaviour across populations and cultures. In this paper, we revisit the construct validity of different dictionary-based measures of morality in text that have been proposed in the literature. We discuss conceptual challenges for text-based measures of morality and present an annotation experiment where we create a new dataset with human annotations of moral rhetoric in German political manifestos. We compare the results of our human annotations with different measures of moral values, showing that none of them is able to capture the trends observed by trained human coders. Our findings have far-reaching implications for the application of moral dictionaries in the digital humanities.

1 Introduction

Morality is a persuasive concept of human life, as it defines what we consider desirable and virtuous and not only guides our own behavior but also our judgment of others. Therefore, the interest in investigating morality across time and cultures has grown, and the increasing availability of big data has triggered more and more interdisciplinary work on using text-based methods for studying morality. A prominent example are Wu et al. (2023) who apply text-based measures of moral values to a corpus of over 1,900 folk tales from diverse cultures across six continents, in order to investigate the impact of literature on cultural norms.

Many of these studies are based on Moral Foundations Theory (MFT) (Haidt et al., 2009; Graham et al., 2009), a descriptive, pluralist theory from social psychology that defines a number of basic moral intuitions that are considered to drive moral reasoning (see §A.1 for an overview of the different MFs). The popularity of the MFT for text-based analysis is due in no small part to the availability of ready-to-use tools such as the English Dictionary of Moral Foundations (MFD) (Graham et al., 2009) and variations thereof, making it easy to extract text-based measures of morality from text.

While many studies have used the available resources to explore and measure moral values from text (see Lipsitz (2018); Rezapour et al. (2019); Xu et al. (2023); Weinzierl and Harabagiu (2022); Simonsen and Bonikowski (2022); Wu et al. (2023), amongst many others), far less have looked at the validity of such measures of morality.

In the paper, we address this important gap by introducing a new, frame-based annotation scheme for moral rhetoric that distinguishes between abstract moral values and concrete acts and goals, and that explicitly encodes the perspective of the moral sentiment, thus making our annotations more interpretable than conventional annotations that assign moral values to words, sentences or documents (Hoover et al., 2020; Trager et al., 2022). Then we discuss the challenges of annotating morality in text and show that traditional Inter-Annotator Agreement metrics are not suitable to measure agreement for phenomena that cannot be easily grounded on the lexical level, such as morality.[1]

Our main contribution, however, is a case study based on our new dataset, showing that moral rhetoric cannot be captured using word-based measures, as evidenced by a lack of correlation between dictionary-based scores for moral value scores and human annotations.

2 Examining morality in text

The first step in the attempt to measure an abstract, latent construct that eludes direct observation is to define what is meant by it. Two recent surveys on morality in NLP, however, both show that this

[1]See, e.g., Fetzer (2022) who argue for a discourse-pragmatic approach to analyse morality in political texts.

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 232–250

step has often been neglected and that many studies neither refer to an explicit theoretical framework nor provide a definition for the construct measured (Vida et al., 2023; Reinig et al., 2024).[2] Linking the construct to a specific theory, however, is only the first step and does not guarantee that the proposed operationalisation of the construct is sound and reliable.

In the paper, we focus on studies that have been conducted in the context of Moral Foundations Theory (MFT), as it is the most commonly used theoretical framework for text-based analyses of morality at the moment. According to Reinig et al. (2024), over 67% of the studies included in their survey use MFT in their analyses, covering computational text analyses in the area of social and political science, media and communication studies, psychology and cultural studies. We first give a short introduction to MFT. Then we discuss aspects of morality and challenges for the automatic measurement of moral values from text. Finally, we describe methods frequently applied to operationalise the construct in order to provide such measurements.

2.1 Moral Foundations Theory (MFT)

MFT is a descriptive, pluralist theory of morality, developed in the area of social psychology (Haidt et al., 2009; Graham et al., 2013). In contrast to monist theories that explain morality in terms of one single principle or dimension, *right–wrong*, MFT believes that the concept of morality is based on more than one such dimension, or foundation. According to MFT, these foundations have been developed during evolution as responses to several adaptive challenges, e.g., the emergence of the PURITY foundation has been driven by the need to avoid pathogens. Moral foundations are seen as intuitions or feelings rather than conscious judgments, which is in contrast to other moral theories that describe moral intuitions as "strong, stable, immediate moral beliefs" (Sinnott-Armstrong et al., 2010) or as moral judgments (McMahan, 2000).

MFT assumes at least five moral intuitions that can be divided into *binding* foundations (ingroup LOYALTY, respect for AUTHORITY, and PURITY) and *individualising* foundations (CARE and FAIRNESS). Newer work has proposed that ideas of fairness can be based on different notions of justice, and has further divided the FAIRNESS foundation into EQUALITY and PROPORTIONALITY (Atari et al., 2023) where EQUALITY favours an equal distribution of opportunities and resources while PROPORTIONALITY prefers a distribution in proportion to an individual's merit or contribution.

MFT explains inter-personal differences of moral values by assuming the existence of an "innate draft of the moral mind" that is later revised by experience and cultural influences (Graham et al., 2013, p. 9). This makes MFT particularly interesting for comparative analyses of moral values across time and cultures (see, e.g., Xie et al. (2019); Wu et al. (2023); Hämmerl et al. (2023)).

2.2 Traditional measurement tools

The traditional measurement tool developed for assessing inter-personal differences between individuals' moral values is the MFT Questionnaire (MFQ) (Graham et al., 2011). Test subjects are asked to rate on a scale of 0 to 5 how much they agree with statements targeting the different moral foundations. For example, *People should not do things that are disgusting, even if no one is harmed* is one of the measurement items for the PURITY foundation. The MFQ has been thoroughly tested for internal and external validity and test-retest reliability using confirmatory factor analysis.

2.3 Dictionary-based measures

While being accurate and reliable, surveys come with some limitations. They cannot be used for diachronic analyses covering past decades, and the recruitment of large numbers of test subjects is costly. Therefore, dictionary-based tools have been proposed as a cheap and easy-to-apply approximation for a number of psychological constructs, most prominently the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2001). In the context of MFT, a number of dictionaries have been developed to measure moral foundations from text, mostly for English.

The English MFD Graham et al. (2009) developed the first Moral Foundations Dictionary (MFD) to analyse sermons delivered in U.S. Christian congregations. The dictionary contains 295 words and word stems, and each of the five foundations has been split into a *vice* and *virtue* dimension, where words with positive sentiment represent the virtue domain while negatively connotated terms are assigned to the vice class. The MFD was used to count the frequencies of morally loaded terms in the sermons, to compare the use of moral language between liberal and conservative congregations.

[2] Reinig et al. (2024) report that around 20% of the studies did not refer to a specific theoretical framework.

The results were subjected to further validation by human coders who, not knowing the origin of the text, had to rate passages containing the keywords. Results confirmed the hypotheses that liberal sermons mostly focussed on the individualising foundations (Care, Fairness) while conservative sermons showed a higher use of words related to Authority, Ingroup Loyalty and Purity.

MFD2.0 Frimer et al. (2019) further extended the rather small size of the MFD to 2,103 entries, with 2,040 unique lexical items and an average of 210 words per foundation. The extended dictionary is referred to as the MFD2.0.

eMFD Hopp et al. (2021) develop the extended Moral Foundation Dictionary (eMFD) by extracting words from a crowd-sourced text-highlighting task where 557 crowdworkers were asked to mark text spans in US newspaper articles that expressed a certain moral foundation. In their dictionary, each of the 3,270 words is assigned a vector of five values, one for each foundation, that describes the probability that this word has been highlighted for a particular moral foundation. In addition to the moral foundations, the authors use VADER (Hutto and Gilbert, 2014) to compute the averaged valence scores of the word contexts for each word–foundation pair. This means that each word entry includes five continuous scores that specify the word's loading for each MF and, additionally, five sentiment scores that specify the word context's average sentiment for each MF. The sentiment scores are then used to derive the more fine-grained vice-virtue dimensions. For example, a lexicon entry for the foundation of CARE that, on average, appears in more negatively scored contexts will be assigned the vice dimension of Care (i.e., HARM) while one that has been seen mostly in contexts with positive sentiment will be assigned to CARE.

mMPD The only German dictionary known to us is included in the Multilingual Moral Political Dictionary (mMPD) (Simonsen and Widmann, 2023), a translation and extension of the English dictionary by Jung (2020), which in turn is based on the English MFD. The mMPD provides word lists for Danish, Dutch, English, German, Spanish, and Swedish, optimised for political text. The German part of the dictionary includes 18,652 lower-cased word forms, out of which 5,198 belong to one or more moral foundations.[3]

[3]The remaining entries belong to the GENERAL-MORAL CLASS.

WordNet-based extensions Some work has used WordNet synsets to obtain extended versions of the MFD (Araque et al., 2020; Rezapour et al., 2019; Mather et al., 2022) while Hulpuș et al. (2020) exploit knowledge graphs for this task.

Distributional semantics-based approaches Garten et al. (2016, 2018) used static word embeddings to create Distributed Dictionary Representations (DDR) as a continuous measure for the similarity of words and moral concepts. Instead of identifying all words belonging to a moral foundation, DDR attempts to encode the core of the MF by averaging static word embeddings for all dictionary entries of that particular foundation and then computing the cosine similarity between the DDRs and words in new, unseen documents for each moral foundation (MF). Many studies have adapted the distributional semantics approach and created sparse representations for words, based on Latent Semantic Analysis (LSA) or word embeddings (Dehghani et al., 2016; Kaur and Sasahara, 2016; Araque et al., 2020).

2.4 Limitations of dictionary-based metrics

Although the dictionary-based metrics listed above are convenient to use, they have significant limitations. Besides the *missing context sensitivity* and their failure to handle compositionality and negation, one of the main limitations of dictionaries is that they are not able to capture *perspective*. While traditional measurement tools like the MFQ explicitly ask test subjects about their *own* moral beliefs and attitudes, it is less clear what we are measuring when extracting moral values from text, as a text does not necessarily express the beliefs and attitudes of its author. Consider Example 2.1 below, taken from a parliamentary speech by a member of the conservative CDU (Christian Democratic Union of Germany).

Ex. 2.1. A year ago, the Greens were already calling for "fair digital markets".

While the sentence contains a call for more fairness in the digital markets, it is clear that this statement is not supported by the speaker, but reflects the views of the Green Party.

So far, this has been ignored in the literature,[4] and any morally loaded terms in a document have

[4]Noteworthy exceptions include Roy et al. (2022); Zhang et al. (2024) who take a frame-based approach to the prediction of moral values, explicitly modelling the holder of moral sentiment.

Moral Frame	Example	Moral Foundation
MORALVALUE	freedom	LIBERTY
MORALVALUE	the traditional family	AUTHORITY
IMMORALVALUE	the communist wall of shame	PURITY
MORALACTORGOAL	save the planet and the people	CARE
MORALACTORGOAL	strengthen our German economy	LOYALTY
IMMORALACTORGOAL	impose draconian penalties for harmless offenses	PROPORTIONALITY
IMMORALACTORGOAL	prevent equal opportunities in the worksplace	EQUALITY

Table 1: Examples for (im)moral values, acts and goals (also see A.1 for a description of the individual MFs).

been interpreted as representing the author's moral values.

Another pitfall for dictionary-based analyses is their domain dependence and has been pointed out by the original developers of the MFD (Graham et al., 2009). The authors report that, when analysing political text from Republican and Democratic candidates' convention speeches, their dictionary approach failed to extract distinctive moral content. Instead, they chose to analyse religious sermans, as those explicitly discuss moral values and give advice on how to live a moral life. This finding, however, has been mostly ignored in the literature where the MFD is considered as a validated and generally applicable measurement tool.

To provide a systematic investigation of the construct validity of dictionary-based measures of morality, we apply frequently used methods from the literature and compare the results we get with human annotations of moral framing. Our main research question is:

RQ: Can we approximate human perception of moral values with word-based text-analytic measures, such as moral dictionaries?

To answer this question, we created a new dataset of German political manifestos, with human annotations of moral rhetoric.

3 Annotation study

As has been pointed out in the literature, low inter-annotator agreement (IAA) is a common problem for the annotation of highly subjective concepts like emotions (Buechel and Hahn, 2017), toxic language (Sap et al., 2022), or moral values (Reinig et al., 2024). Conventional approaches to coding morality in text mostly assign labels to whole text passages or documents (often tweets or social media posts, see, e.g., Johnson and Goldwasser (2018); Hoover et al. (2020); Trager et al. (2022)), which does not capture perspective and also fails to specify which parts of the text contain the moral

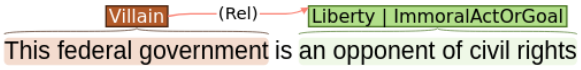


Figure 1: Example annotation for a moral frame annotation, its moral roles (here: villain) and moral foundation (LIBERTY), taken from a parliamentary speech by the German Left Party (engl. translation).

message. To solve this problem, we developed a new annotation framework for moral framing in text that addresses these shortcomings.

Annotation scheme In contrast to previous work, we do not annotate moral values on the level of sentences or documents but, instead, aim at encoding moral frames and their roles (see Figure 1 above).[5] Specifically, we encode abstract moral values as well as concrete acts and goals that are framed as (im)moral, using the four labels MORALVALUE, IMMORALVALUE, MORALACTORGOAL, and IMMORALACTORGOAL (see examples in Table 1). Additionally, we use the label POLITICALACTORGOAL to code text spans that refer to concrete policy acts, e.g., "the solidarity surcharge". Distinguishing between abstract concepts and concrete acts and goals will enable us to study how the two interact on a linguistic level. We expect that the value categories correspond to *moralising speech acts* (Becker et al., 2024), i.e., concepts and values like *justice* that are presented as universally accepted so that no further justification is needed.

As shown in Table 1, moral values are typically expressed as NPs and describe abstract concepts (*freedom, injustice, the traditional family*) or symbols that transmit national and religious values (*the Statue of Liberty*). Descriptions of (im)moral acts or goals are typically expressed as VPs (e.g., *saving the planet*) but can also include nominalisa-

[5] Moral roles are inspired by the Narrative Policy Framework (Shanahan et al., 2017) and include the labels HERO, VILLAIN, VICTIM and BENEFICIARY. The annotation of frame roles is ongoing work and therefore not discussed in this paper.

	A1	A2	avg.	# Tokens
Culture	354	419	386.5	5,996
Media	172	191	181.5	2,555
Migration	558	534	546.0	9,330
Total	1,084	1,144	1,114.0	17,881

Table 2: Distribution of frames in German manifestos on the topics of immigration, media, and culture.

tions (e.g., *the fight against disposable packaging*). Whether a frame is coded as either moral or immoral depends always on the speaker's stance, irrespective of the coder's moral preferences.

Data We chose political manifestos, as those include many statements about what ought to be done, often framed in moral terms. We decided on three topics that are typically discussed in a highly polarised and moralised fashion, i.e., the parties' position on immigration, culture, and the media.[6]

After extracting the relevant texts from the Manifestos Project Database (Burst et al., 2022), we asked two human coders to highlight moral framing in the data. The annotation of moral foundations on top of the frames has been carried out by four trained coders, to be able to assess how reliable humans can code this type of annotation.

3.1 Annotation of moral frames

In the first step, the coders identify all (im)moral frames in the political manifestos. The coders were instructed to first read the whole speech, focussing on the moral values, goals and actions that are presented as desirable (praiseworthy) as well as the ones deemed to be undesirable (blameworthy). After having read the whole document, the coders are asked to highlight all moral values, goals and actions mentioned in the speech.[7]

The identification of frames has been carried out by two trained coders, both MA students of linguistics. Each text has been annotated by both coders to ensure high recall. We notice that the coders often mark the same frames, however, there are differences regarding the exact span of the annotation (e.g., whether a modifier should be part of the frame or not). Other differences between the annotations concern the question of whether a moral frame should be coded as a (im)moral *value* or an *act or goal*, e.g., *freedom of the press*, as moral values can also be framed as goals (see §3.4).

[6]See §A.4 for more details.

[7]For data and annotation guidelines, see https://anonymous.4open.science/r/moral-manifestos-4B55

3.2 Annotation of moral foundations

In the next step, we extract the annotated frames and cluster them into semantically coherent frame groups.[8] Then we present the annotators with the clusters and ask them to assign moral foundations to each frame in the group. The motivation for this approach is to speed up the annotation and make it more consistent by presenting the coders with sets of thematically related frames.

Annotation of clusters with MFs Figure 3 in the appendix shows our annotation interface for assigning moral foundation labels to frames. In addition to the six Moral Foundations described in Atari et al. (2023),[9] we also annotate the LIBERTY foundation which has often been discussed in the literature as a plausible MF candidate (Iyer et al., 2012). Frames that cannot be assigned unambiguously to any MF are annotated as GENERAL-MORAL. The four annotators can also mark frames as NON-MORAL when they think that a mistake has been made during frame identification, thus providing a validation of the frame annotation step which has been done by two coders only.

3.3 Inter-annotator agreement (IAA)

Table 2 shows the number of different frames identified by each coder (see Table 4 in the Appendix for a more detailed description of the data).

Agreement for frame identification As it is not straightforward to compute IAA for span-based annotations, we follow common practice for opinion role labelling (Marasović and Frank, 2018) and report strict match and binary token overlap. While strict match requires that the frame spans are identical, token overlap also considers annotations as a match if at least one of the tokens in the span has been annotated by both coders (Table 3). We first consider A1's annotations as ground truth and compute how well they agree with A2's annotations, then we switch roles and do the same for A2. The lower scores for A2–A1 compared to A1–A2 reflect the higher number of frames identified by A2. Additionally, we report *oracle* agreement for frame labels where we only consider spans that have been identified by both coders.

We see that *strict* agreement for spans is rather low (45–48%) while results for *binary overlap* is much higher with 75–80%. This shows that our

[8]Details on the clustering process can be found in §A.2.

[9]Care, Equality, Proportionality, Loyalty, Authority, Purity.

	A1–A2		A2–A1	
	strict	overlap	strict	overlap
spans only	48.1	80.0	45.2	75.2
spans + frames	43.2	66.7	40.6	62.9
frames on agreed spans: 83.4% (724 out of 868)				

Table 3: Percentage agreement for frame annotation for strict match and token overlap, and frame label agreement for instances where coders agreed on the span.

annotators agree well on which text passages include moral framing but disagree with regard to the concrete frame spans (see error analysis below). When also considering frame labels, agreement is in the range of 63–66% overlap. Out of the 868 annotations that show binary token overlap, 83.4% (724 instances) also have received the same label while 16.6% (144 instances) have been coded with a different label. We now only look at frame spans that have been identified by both coders, to identify the main reasons for disagreement.

3.4 Error analysis

Frame spans The most frequent causes for mismatches regarding the frame spans include modifiers and coordination. While the guidelines instructed the coders to focus on the arguments and exclude modification, we found that annotators sometimes deviated from this rule when they felt that excluding the modifier did not accurately capture the meaning of the frame (Ex. 3.1). Other mismatches include prepositional modifier phrases and relative clauses.

Ex. 3.1. (further (promote dialog between religions, world views and cultures)$_{A1}$)$_{A2}$

Regarding coordination, we find that sometimes one annotator includes the whole coordinate phrase as one frame while the other split it up into several frames (Ex. 3.2).

Ex. 3.2. ((decent training)$_{A1}$, working conditions and pay)$_{A2}$

Frame labels We notice that the largest part of the disagreements concerning the frame labels is due to one annotator chosing to annotate the frame as a MORALVALUE while the second coder annotated an overlapping span as an act or goal (83 out of the 144 disagreements). An example is the frame *protect freedom* which has been annotated as a MORALACTORGOAL by A1 while A2 chose to only mark *freedom* as a MORALVALUE.

In addition, we found 30 instances that have been identified by one coder only while the other coder did not consider this instance as a case of moral framing. These included strong evaluative statements that did not include strong moral rhetoric.

We also encountered cases labelled as *moral* by one coder while the other coder annotated the same instance as *immoral*. An example is shown below.

Ex. 3.3. (Strict punishment for (false statements in the asylum procedure)$_{A2}$)$_{A1}$

This frame expresses a political demand by the far-right party AfD which A1 chose to annotate as a moral goal. A2 took a different, but compatible view by annotating the subspan "false statements in the asylum procedure" as an immoral act, resulting in opposite polar values for overlapping text spans.

This illustrates some of the challenges for the annotation of morality in text, showing that coders often choose to highlight different text spans to encode morality in text. This, however, does not so much reflect different moral beliefs or biases held by the coders but rather shows that morality is a compositional construct that requires a more refined treatment than simply assigning labels to sentences or documents.

IAA for MF annotation The annotation of moral foundations on top of frames is a multilabel task, where each of the four coders assigned a maximum of two labels to each frame. Fleiss' Kappa using Jaccard distance for the four coders results in a score of 0.58, and Krippendorff's Alpha with Jaccard distance is 0.56. As those scores are hard to interpret, we next compute for which part of the annotations (i) all four coders agreed on a label, (ii) three out of four coders agreed, and (iii) at least two coders agreed. 99.5% of the annotations have assigned the same label by at least two coders and for 79.7% of the instances at least three coders agreed on the label. For around half of the annotations (50.6%), all four coders chose the same label.

We argue that this shows a substantial agreement and keep all labels that have been assigned by at least three of the four coders in order to compare our annotations with results from dictionary-based analyses.[10]

4 Investigating the construct validity of dictionary-based measures of morality

We now present a case study where we apply frequently used dictionary-based measures to our data,

[10]We also release the individual annotations by each of the four coders with the data.

(i) the **MFD** (Graham et al., 2009), (ii) the **MFD2.0** (Frimer et al., 2019), (iii) the **eMFD** (Hopp et al., 2021) and (iv) the German and English components of the Multilingual Moral Political Dictionary (**mMPD**) (Simonsen and Widmann, 2023) (see §2.3). The German dictionary is directly applied to the original German manifestos. For the English dictionaries, we follow Wu et al. (2023) and translate our data to English before applying the dictionaries (see §A.5). This also allows us to test how well the scores for the English and German versions of the mMPD correlate on our data.

As the dictionaries include annotations for the *vice* and *virtue* dimensions of each foundation, we aggregate the scores for both ends of the same dimension into one score. None of the dictionaries encodes the theoretical improvements to the MFT (Atari et al., 2023), where the FAIRNESS foundation has been split into EQUALITY and PROPORTIONALITY. We therefore merge these in our data so that we can compare results across methods.

To make sure that our results are not influenced by one particular aggregation method, we test two different ways to calculate measures of morality.

A We follow the procedure described for the MFD and compute a moral score for each MF by dividing the number of relevant dictionary terms in a document by document length, multiplied by 100 (Graham et al., 2009).

B We compute morality scores based on the library provided by Hopp et al. (2021) by counting how often the terms for any specific foundation occur in each document and divide the aggregated counts for each foundation by the number of moral words for all foundations in the same document.

In contrast to the first approach where we get an independent score for each MF, normalised by document length, this approach normalises by the total number of moral words in the same document. As a result, documents with the same number of trigger words for one particular MF will be scored differently by each method, depending on whether (and how many) terms for other MFs exist in the same document. Those details are crucial, however, they are hardly ever discussed in the literature and often no motivation is given for choosing one scoring method over another.

We can now compare the different scores to investigate the construct validity of dictionary-based measures of morality from text. If the dictionaries provide valid measurements, then we would expect to see a strong correlation between the scores obtained from the dictionaries, as well as a strong correlation between the dictionary-based scores and the human annotations. We can thus formulate our expectations as follows. We expect to see significant correlations

(E1) between coder1 and coder2,

(E2) between each dictionary and coder1/coder2,

(E3) between the MFD-based dictionaries,

(E4) between mMPDen and mMPDde.

After extracting the scores for each method and moral foundation, we computed Pearson's correlation for each combination of measurement tools. Figure 2 plots the p-values for our correlation analysis based on aggregation strategy A (results for strategy B are included in the appendix).[11]

E1: How well do the human coders correlate? The scores based on the moral frame annotations of our human coders are the only measures across the four MFs that exhibit a highly significant correlation ($p < 0.001$, see Fig. 2), with strongly positive correlation coefficients in the range of $r = .79$ to $.98$.

E2: How well do the dictionaries correlate with human annotations? None of the dictionary-based measures shows a significant correlation with the human coders for *all* four MFs. The English version of the mMPD significantly correlates with the humans on three of the four MFs but has no significant correlation for AUTHORITY. Surprisingly, the correlation between the German version of the mMPD and the human coders is only weakly significant ($p < 0.05$) for two MFs and not significant for the other two foundations.

E3: Correlation between MFD and MFD2 The MFD2 is an extended version of the MFD. We therefore exected to see a strong correlation between the two dictionaries. This, however, is only true for two of the five MFs (Fairness, $p < 0.001$ and Loyalty, $p < 0.01$) while the scores for MFD and MFD2 are not significantly correlated for the other MFs, including PURITY (see Fig.4).

E4: Correlation between the English and German mMPD Finally, we expected that the scores obtained from the translated German mMPD will show a significant correlation with the English mMPD. This expectation has been met, showing

[11] We also moved the p-value matrix for PURITY to the appendix (Fig.4) as our human coders did not find any instances for this moral foundation in the manifestos.

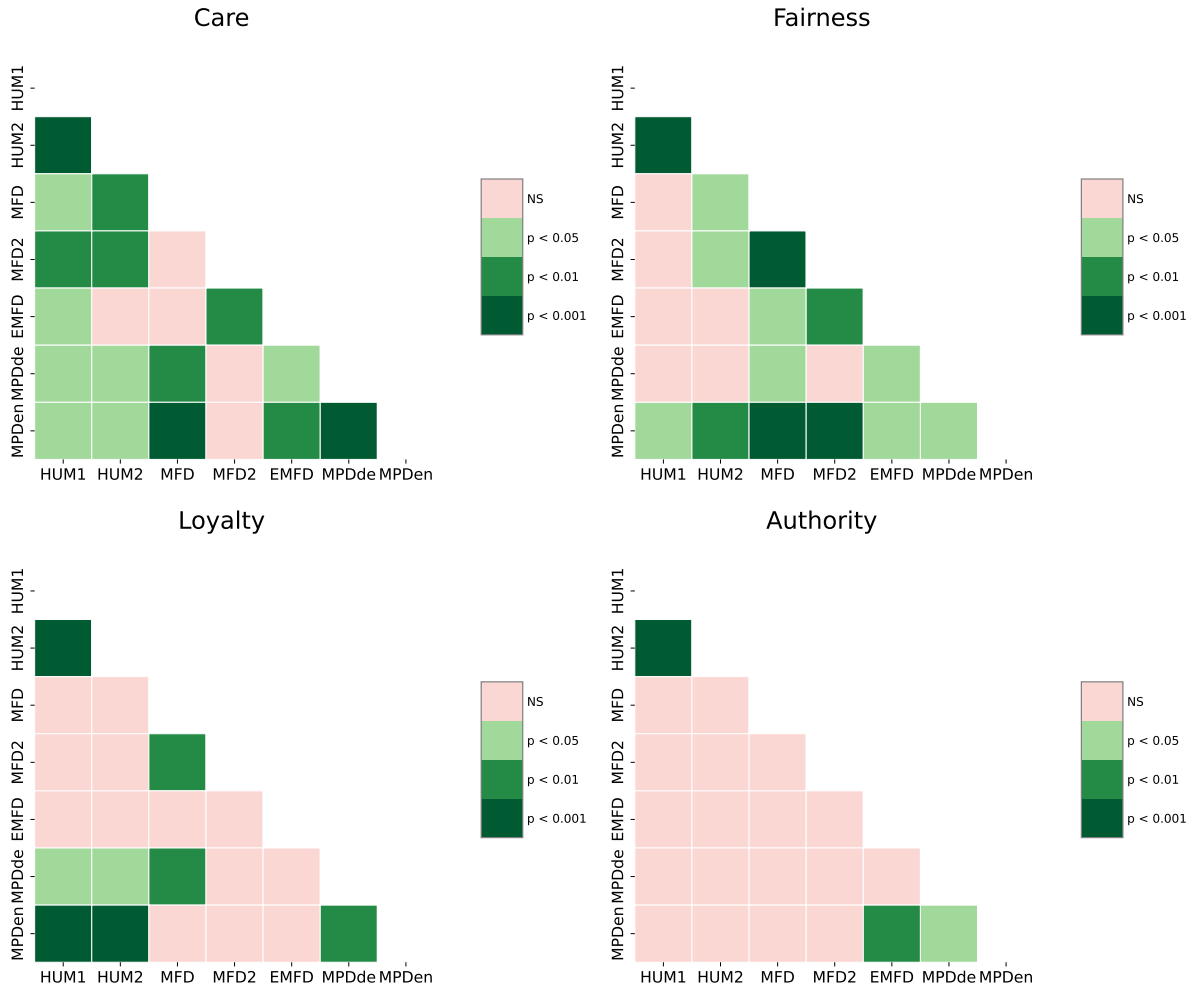


Figure 2: p-values for correlation matrices (Pearson) for the different dictionaries and moral foundations, based on aggregation strategy A (NS: not significant; results for Purity and strategy B are included in the appendix, Fig. 4, 6).

a moderate to strong positive correlation[12] with $p < 0.001$ for CARE, $p < 0.01$ for LOYALTY and $p < 0.05$ for FAIRNESS and AUTHORITY. However, when directly comparing the plotted scores for the human coders and the mMPDS (Fig. 6 in the appendix), we see that the scores for the party-topic combinations often show different trends, and the political dictionaries only partly correlate with the human annotations (see E2 above).

For aggregation strategy B, the results are even worse. We see hardly any significant correlation between the results of the different dictionaries or between dictionaries and human coding (see appendix, Fig.7). For the interested reader, we include a qualitative analysis in Section A.9 in the appendix to validate our findings.

5 Discussion

We presented a new annotation framework for moral framing in text and showed that dictionary-based measures neither have a strong correlation with each other's predictions, nor come close to the trends found by the human coders. We also showed that different aggregation methods can significantly impact results. Other factors that might influence the final morality score of the dictionary-based approaches are preprocessing steps like stop word removal.[13] Our results call into question the reliability of analyses based on moral dictionaries.

The limitations of dictionary-based methods are hardly new and have been discussed before (Chan et al., 2021). However, moral dictionaries are still widely used (Takikawa and Sakamoto, 2017; Zhang

[12]The correlation strengths are: Care $r = .81$, Fairness $r = .56$, Loyalty $r = .73$, Authority $r = .49$, Purity $r = .41$.

[13]If stopwords are removed before the document length is computed (as done in the EMFD library), then the use of stopword lists of different sizes affects the document length and can therefore lead to different results.

et al., 2023; Wu et al., 2023; Landowska et al., 2024), often without further validation, and many works that employ more sophisticated techniques for moral value prediction also base their work on moral dictionaries or use them for evaluation (Mokhberian et al., 2020; Park et al., 2024).

While dictionaries are able to identify commonly accepted moralising speech acts like *freedom* and *justice* (Becker et al., 2024), our annotation study has shown that these account for only a small proportion of moral frames[14] and that the majority of frames discuss moral actions and goals without using highly morally charged language. Based on our results, we argue that dictionaries are not a valid approach for examining morality in text, as morality is an abstract, multi-dimensional construct that cannot be captured by counting keywords out of context.

Acknowledgments

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 would also like to thank the anonymous reviewers for their constructive feedback.

References

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.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 5(125):1157–1188.

Maria Becker, Ekkehard Felder, and Marcus Müller. 2024. Moralisierung als sprachliche Praxis. In Ekkehard Felder, Friederike Nüssel, and Jale Tosun, editors, *Moral und Moralisierung: Neue Zugänge*, pages 123–151. Berlin, Boston: De Gruyter.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

[14]Roughly 20% of the moral frames in our data are coded as MORALVALUE/IMMORALVALUE.

Tobias Burst, Werner Krause, Pola Lehmann, Jirka Lewandowski, Theres Matthieß, Nicolas Merz, Sven Regel, and Lisa Zehnter. 2022. Manifesto corpus. version: 2022-1.

Chung-hong Chan, Joseph Bajjalieh, Loretta Auvil, Hartmut Wessler, Scott Althaus, Kasper Welbers, Wouter Van Atteveldt, and Marc Jungblut. 2021. Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, 3(1):1–27.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Experimental Psychology Gen.*, 3(145).

Anita Fetzer. 2022. 'for (...) a leader like this prime minister to talk about morals and morality is a disgrace': offensive action, uptake and moral implications in the context of parliamentary debates. *Language & Communication*, 87:135–146.

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

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M. Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *IJCAI 2016 Workshop on Computational Modeling of Attitudes*.

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. *Behav Res*, 50:344—-361.

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.

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.

Katharina Hämmerl, 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.

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.

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.

Clayton J Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media*, ICWSM-14, Ann Arbor, MI.

R. Iyer, S. Koleva, J. Graham, P. Ditto, and J. Haidt. 2012. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE*, 8(7).

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.

Jae-Hee Jung. 2020. The mobilizing effect of parties' moral rhetoric. *American Journal of Political Science*, 2(64):341–355.

Rishemjit Kaur and Kazutoshi Sasahara. 2016. Quantifying moral foundations from various topics on twitter conversations. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2505–2512.

Alina Landowska, Katarzyna Budzynska, and He Zhang. 2024. Quantitative and qualitative analysis of moral foundations in argumentation. *Argumentation*, 38:405–434.

Keena Lipsitz. 2018. Playing with emotions: The effect of moral appeals in elite rhetoric. *Political Behavior*, 40:57–78.

Ana Marasović and Anette Frank. 2018. SRL4ORL: Improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 583–594, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *Social Informatics*, pages 206–219, Cham. Springer International Publishing.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Jeongwoo Park, Enrico Liscio, and Pradeep K. Murukannaiah. 2024. Morality is non-binary: Building a pluralist moral sentence embedding space using contrastive learning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 654–673, St. Julian's, Malta. Association for Computational Linguistics.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Ines Reinig, Maria Becker, Ines Rehbein, and Simone Ponzetto. 2024. A survey on modelling morality for text analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4136–4155, Bangkok, Thailand. 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.

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.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Mark Schaller and Damian R. Murray. 2010. Infectious Disease and the Creation of Culture. In *Advances in Culture and Psychology: Volume 1*. Oxford University Press.

Elizabeth A Shanahan, Michael D Jones, Mark K Mcbeth, and Claudio M Radaelli. 2017. The Narrative Policy Framework. In C.M. Weible and P.A. Sabatier, editors, *The Theories of the Policy Process*, pages 173–213. Boulder, CO: Westview Press.

Kristina B. Simonsen and Bart Bonikowski. 2022. Moralizing immigration: Political framing, moral conviction, and polarization in the United States and Denmark. *Comparative Political Studies*, 55(8):1403–1436.

Kristina B Simonsen and Tobias Widmann. 2023. When do political parties moralize? A cross-national study of the strategic use of moral language in political communication on immigration. *OSF Preprints*.

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

Hiroki Takikawa and Takuto Sakamoto. 2017. Moral foundations of political discourse: Comparative analysis of the speech records of the US congress and the Japanese diet. *Preprint*, arXiv:1704.06903.

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. *Preprint*, arXiv:2208.05545.

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.

Sarah Wagner, L. Constantin Wurthmann, and J. Philipp Thomeczek. 2023. Bridging left and right? How Sahra Wagenknecht could change the German party landscape. *Politische Vierteljahresschrift*, 64.

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.

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.

Jing Yi Xie, Renato Ferreira Pinto Junior, Graeme Hirst, and Yang Xu. 2019. Text-based inference of moral sentiment change. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4654–4663, Hong Kong, China. 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.

Weiyu Zhang, Rong Wang, and Haodong Liu. 2023. Moral expressions, sources, and frames: Examining COVID-19 vaccination posts by facebook public pages. *Computers in Human Behavior*, 138:107479.

Xinliang Frederick Zhang, Winston Wu, Nicholas Beauchamp, and Lu Wang. 2024. MOKA: Moral knowledge augmentation for moral event extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4481–4502, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Overview of the Moral Foundations

Below we provide a short description of the moral foundations, adapted from the MFT website.[15]

[15] https://moralfoundations.org/.

Care: This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies the virtues of kindness, gentleness, and nurturance.

Fairness: This foundation is related to the evolutionary process of reciprocal altruism. It underlies the virtues of justice and rights.

In 2023, Atari et al. (2023) was split into two new foundations, Equality and Proportionality, as it was found that politically left-leaning individuals more strongly endorse values of Equality while more conservative individuals prefer the notion of proportionality.

Equality: Equality is defined as "Intuitions about equal treatment and equal outcome for individuals."

Proportionality: Proportionality is defined as "Intuitions about individuals getting rewarded in proportion to their merit or contribution."

Loyalty: This foundation is related to our long history as tribal creatures able to form shifting coalitions. It is active anytime people feel that it's "one for all and all for one." It underlies the virtues of patriotism and self-sacrifice for the group.

Authority: This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to prestigious authority figures and respect for traditions.

Purity: This foundation was shaped by the psychology of disgust and contamination. It underlies notions of striving to live in an elevated, less carnal, more noble, and more "natural" way (often present in religious narratives). This foundation underlies the widespread idea that the body is a temple that can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions). It underlies the virtues of self-discipline, self-improvement, naturalness, and spirituality.

The last foundation is not considered as part of the moral foundations but often discussed as a plausible candidate (Iyer et al., 2012).

Liberty: This foundation is about the feelings of reactance and resentment people feel toward those who dominate them and restrict their liberty. Its intuitions are often in tension with those of the authority foundation. The hatred of bullies and dominators motivates people to come together, in solidarity, to oppose or take down the oppressor.

A.2 Annotation of moral frame clusters

We applied the fast clustering algorithm[16] provided in the S-BERT library (Reimers and Gurevych, 2019). Specifically, we use the `German_Semantic_STS_V2` model[17] and extract clusters with a minimum community size of $\{25, 25, 15, 5\}$ and a threshold of $\{0.7, 0.7, 0.7, 0.6\}$ for $\{$*MoralActOrGoal, ImmoralActOrGoal, MoralValue, ImmoralValue*$\}$, respectively. We also experimented with other settings but found that the ones above gave us a good balance between cluster coherence and coverage.

Not all frames could be assigned to a cluster in the first clustering round. We therefore ran a second round of clustering where we subsequently decreased the threshold until nearly every frame had been assigned to a cluster. The remaining frames that could not be clustered were considered as their own group.

Figure 3 shows our annotation interface for assigning moral foundation labels to frames. This particular cluster mostly includes MORALVALUE frames related to values of freedom and self-determination. Each frame is shown only once, however, the annotators can also visualise the different contexts in which each frame occurred by clicking at the Context column.

A.3 Distribution of moral frames in the manifestos

Table 4 shows the distribution of moral frames and political acts or goals in our data.

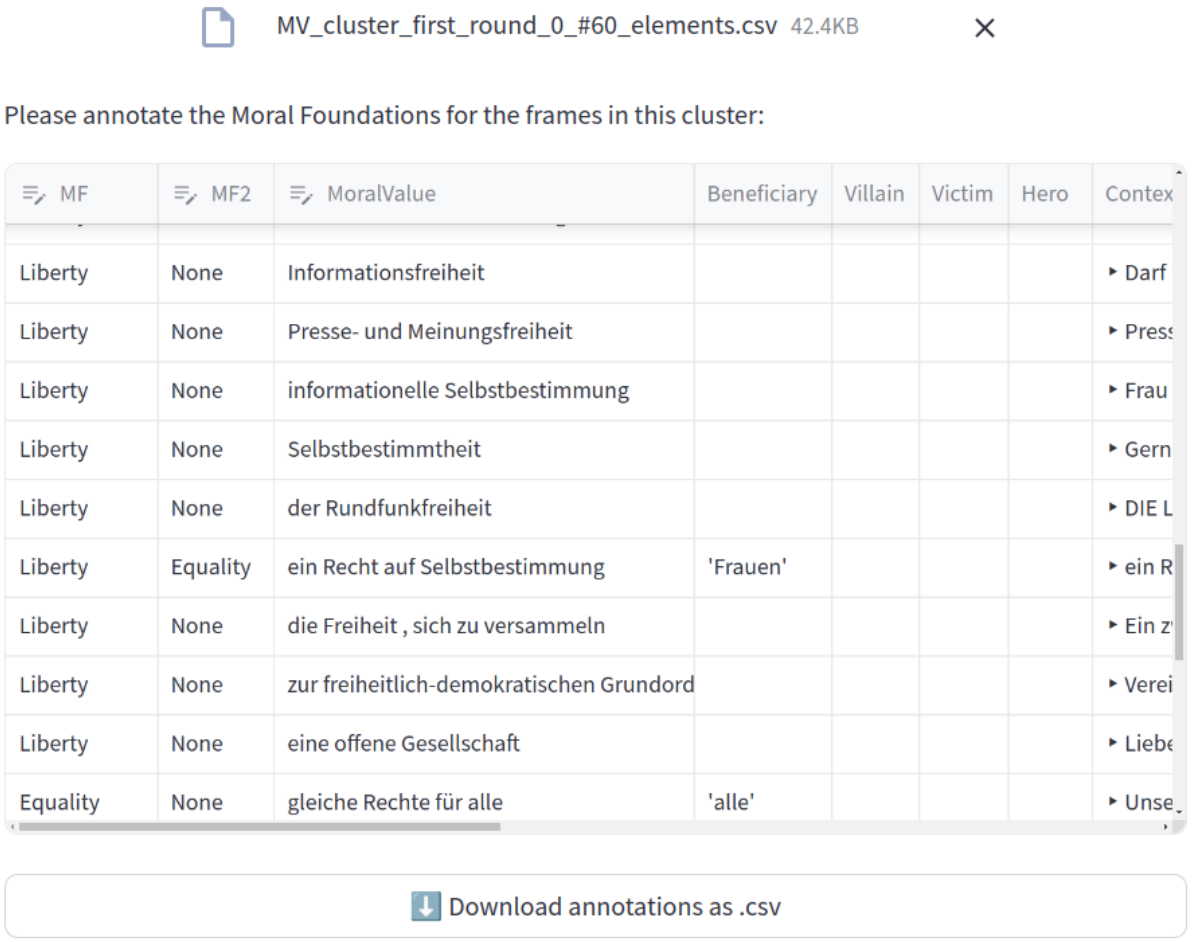
A.4 Manifestos

The data has been extracted from the Manifestos Project Database (Burst et al., 2022). We downloaded the manifestos for the German election of the Bundestag in 2021 for all parties that were part of the Bundestag at the time. Below is a quick overview of the different parties. For an overview of the parties' ideological position, see Fig. 5.

- Alternative für Deutschland (AfD)
- Bündnis 90/Die Grünen (Green party)

[16] Available from https://github.com/UKPLab/sentence-transformers/blob/master/examples/applications/clustering.

[17] For documentation, see https://huggingface.co/aari1995/German_Semantic_STS_V2.



German frame text	English translation
Informationsfreiheit	Freedom of information
Presse- und Meinungsfreiheit	Freedom of the press and expression
informationelle Selbstbestimmung	informational self-determination
der Rundfunkfreiheit	freedom of broadcasting
ein Recht auf Selbstbestimmung	the right to self-determination
die Freiheit, sich zu versammeln	the freedom to assemble
zur freiheitlich-demokratischen Grundordnung	to the free and democratic basic order
eine offene Gesellschaft	an open society
gleiche Rechte für alle	equal rights for all

Figure 3: Annotation interface for the annotation of Moral Foundations (MF) on clustered frames. MoralValue shows the clustered frames, the next four columns show the annotated roles. The last column (Context) shows the context(s) for each frame and can be expanded when clicking on it. The English translations are shown in the table below.

- Christlich-Demokratische Union/Christlich-Soziale Union in Bayern (CDU/CSU)
- Freie Demokratische Partei (FDP)
- Die Linke (The Left)
- Sozialdemokratische Partei Deutschlands (SPD)

A.5 Translation fo the German manifestos to English

We use the fairseq library (Ott et al., 2019) to translate the manifestos from German to English.[18] The model we use is the transformer model (transformer.wmt19.de-en) with the moses tokeniser and fastBPE.

A.6 Correlation matrix for p-values (Purity)

A.7 Comparison of human annotations and dictionary-based scores (mMPD)

Figure 6 shows a direct comparison of the scores based on human annotations and the dictionary-

[18] https://github.com/facebookresearch/fairseq.

Party	# Tokens	# Frames	MoralValue	MoralAct	ImmoralAct	PoliticalAct
Migration – Coder 1						
AfD	2093	103	2	45	38	18
CDU/CSU	872	47	3	29	9	6
FDP	1503	79	10	53	4	12
GRUENE	1815	102	5	67	27	3
LINKE	2368	178	8	125	27	18
SPD	679	49	4	34	4	7
Media – Coder 1						
AfD	344	31	4	6	20	1
CDU/CSU	337	14	1	9	4	0
FDP	496	38	3	23	5	7
GRUENE	223	15	2	12	0	1
LINKE	774	52	6	38	4	4
SPD	381	22	4	11	6	1
Culture – Coder 1						
AfD	679	37	4	17	14	2
CDU/CSU	592	38	9	21	0	8
FDP	921	51	2	30	6	13
GRUENE	1288	87	4	72	2	9
LINKE	1719	95	6	72	9	8
SPD	797	46	8	29	5	4
Total	17881	1084	85	693	184	122
Migration – Coder 2						
AfD	2093	99	10	43	33	13
CDU/CSU	872	41	5	22	8	6
FDP	1503	55	17	25	6	7
GRUENE	1815	111	6	62	27	16
LINKE	2368	179	12	102	38	27
SPD	679	49	6	33	3	7
Media – Coder 2						
AfD	344	25	6	4	14	1
CDU/CSU	337	20	8	7	5	0
FDP	496	35	7	17	4	7
GRUENE	223	19	7	11	0	1
LINKE	774	65	23	34	5	3
SPD	381	27	11	12	3	1
Culture– Coder 2						
AfD	679	45	19	10	15	1
CDU/CSU	592	45	15	22	3	5
FDP	921	59	9	36	4	10
GRUENE	1288	89	22	63	0	4
LINKE	1719	119	19	78	9	13
SPD	797	62	15	37	5	5
Total	17881	1144	217	618	182	127

Table 4: Distribution of moral frames in manifestos the topic of Migration, Media and Culture.

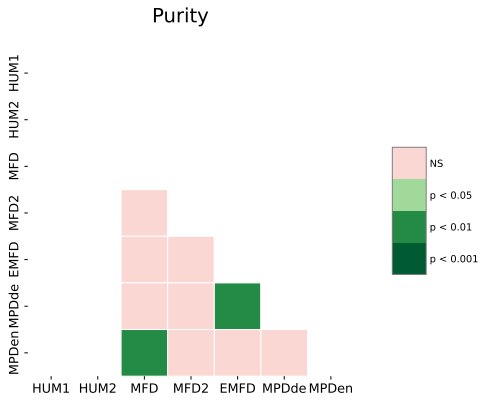


Figure 4: p-values for Pearson's correlation matrices for the different dictionaries for the PURITY foundation.

based scores for the English and German versions of the mMPD.

A.8 Aggregation strategy B: Correlation matrix for p-values (Purity)

Figure 7 shows results for aggregation strategy B.

A.9 Qualitative analysis

Care For the CARE frame, both human annotations show high scores for the Green and Left party on the topic of migration. A look at the data finds 63/59 (Green/Left) CARE frames in the human-annotated data for coder1 and 64/55 for coder2. Typical frames are listed below. Only few of these frames were found by the dictionaries, some of

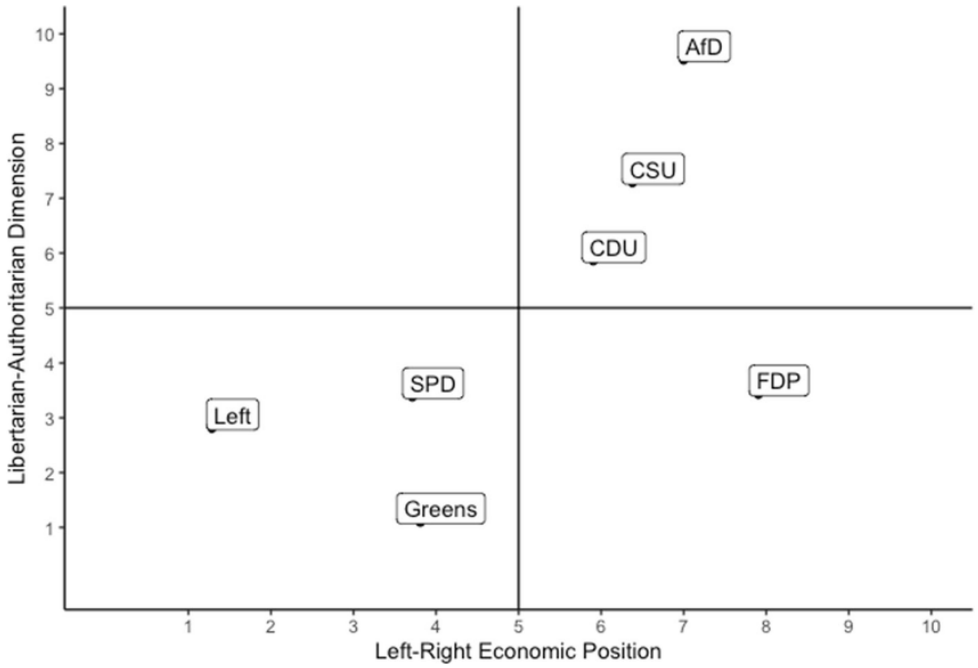


Figure 5: Germany's political landscape based on the Chapel Hill Expert Survey (2019) (image taken from Wagner et al. (2023)).

them for the wrong reasons (e.g., "fighting" increases the count for the *vice* dimension of CARE), showing that moral rhetoric can not be captured at the word level.

- save the people
- fighting the causes of flight
- the right to family reunification

Fairness The FAIRNESS (EQUALITY) scores are again highest for the Left party, for all three topics (culture, media, migration). Below are three typical examples of political demands from the Left party, none of them captured by the word counts for FAIRNESS in the dictionaries.

- persecution due to sexual orientation
- qualifications for vocational training regardless of age
- barrier-free accessibility

Loyalty Looking at the LOYALTY foundation, we find the highest scores for the far-right AfD for culture and migration. This is also to be expected, given that this foundation is associated with moral values of patriotism and defending the in-group against outsiders. It is thus not surprising that a far-right party frames their messages, based on the LOYALTY frame. Again, none of the frames shown below increases the dictionary scores for LOYALTY.

- preserving Germany's cultural identity
- damaging Germany economically
- permanent and effective protection of the EU's external borders

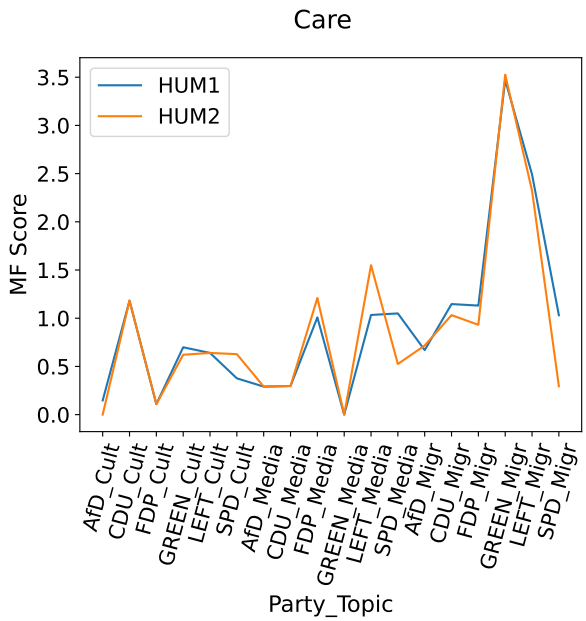
Authority For AUTHORITY, according to the human coders, the far-right AfD and, to a lesser extend, the conservative CDU/CSU score highest, both on the topic of migration. This is again consistent with the theory which states that this foundation mostly appeals to conservatives' "stronger emotional sensitivity to threats to the social order, which motivates them to limit liberties in defense of that order" (Graham et al., 2009, 1030).

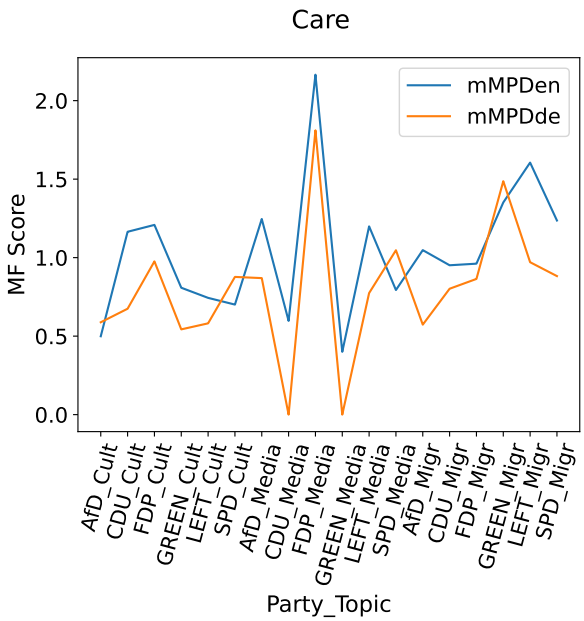
Typical frames are shown below. Only one of them (tradition*) is found in the MFD/MFD2.0 for AUTHORITY while "tradition" only has a low score of 0.13 for AUTHORITY in the emfd. The word form "tradition" is also not included in the English mMPD.

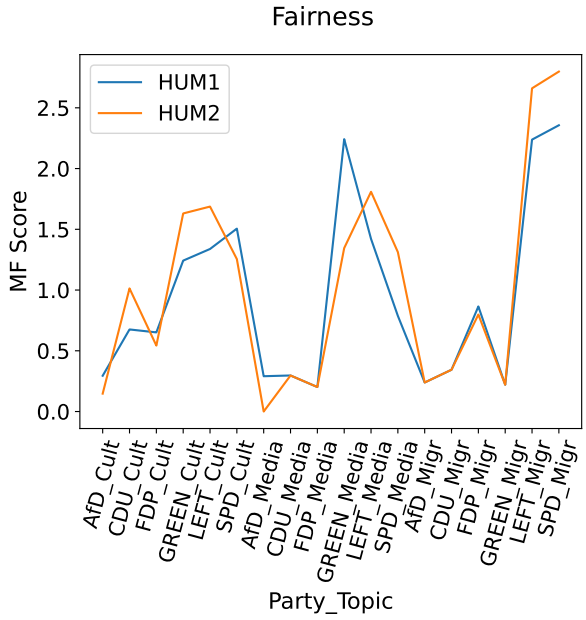
- preserving our traditions
- strict punishment of misstatements in the asylum procedure
- prevent illegal border crossings

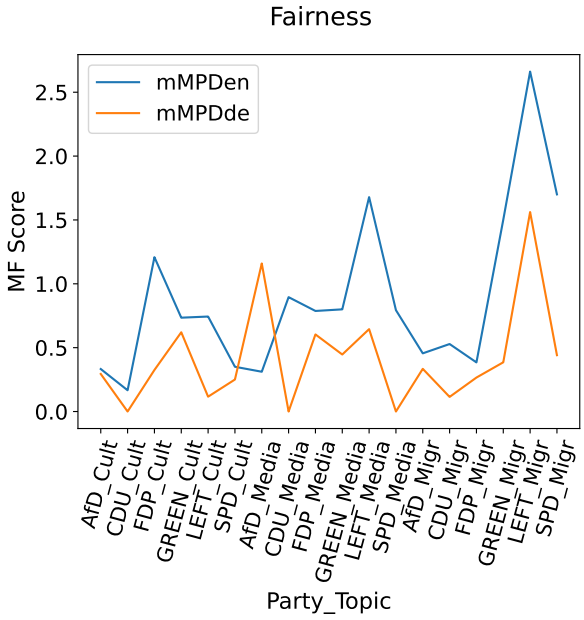
Surprisingly, the MFD gives high scores to the Green party. A look at the data reveals that this is due to keywords like *legal, authorities, position* in the Green manifesto that have been interpreted out of context (e.g., *Non-profit journalism needs legal certainty.* has been counted as a signal for AUTHORITY).

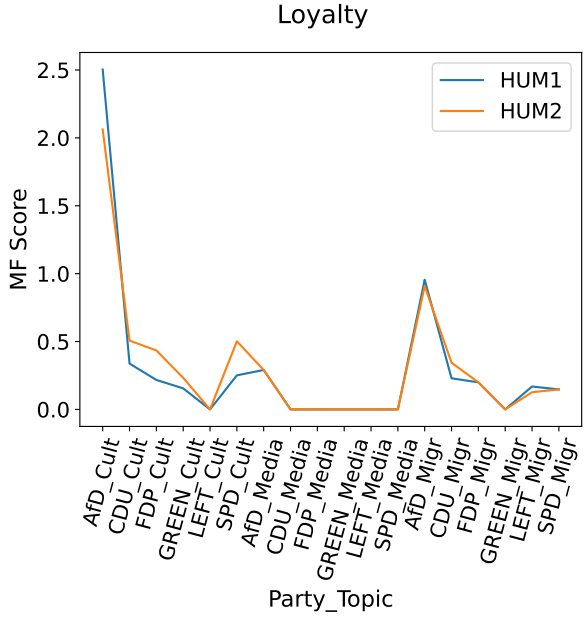
Purity Moral values related to the PURITY foundation express notions of disgust and contamination and promote a natural or spiritual lifestyle. This MF was shaped by the evolutionary advantage of avoiding disease-causing pathogens (Schaller and Murray, 2010). According to our human coders, PURITY has not been used to frame moral messages in the manifestos. Scores for the dictionaries are also quite low but show some spikes for the CDU/CSU and the SPD manifestos on the topic of culture, based on the keywords *sickness, preserve, exploited* that are listed in the MFD for PURITY but, in the context of the manifestos, have not been used to express notions of PURITY but to provide better working conditions for artists in case of statutory sickness absence etc.

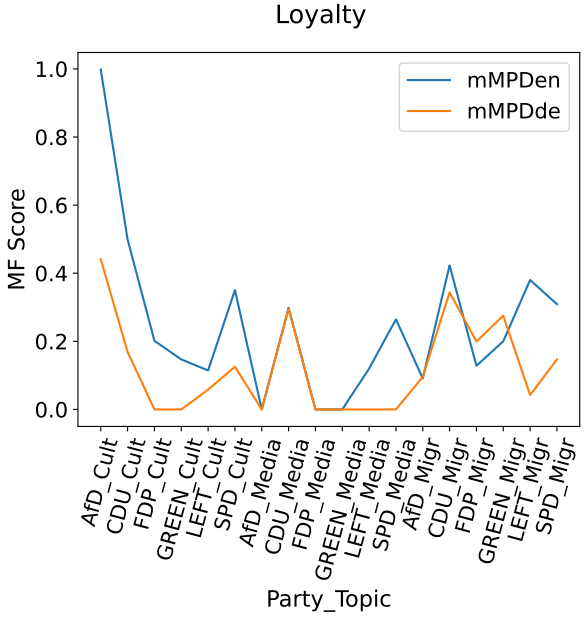

Care
HUM1
HUM2
MF Score
Party_Topic


Care
mMPDen
mMPDde
MF Score
Party_Topic


Fairness
HUM1
HUM2
MF Score
Party_Topic


Fairness
mMPDen
mMPDde
MF Score
Party_Topic


Loyalty
HUM1
HUM2
MF Score
Party_Topic


Loyalty
mMPDen
mMPDde
MF Score
Party_Topic

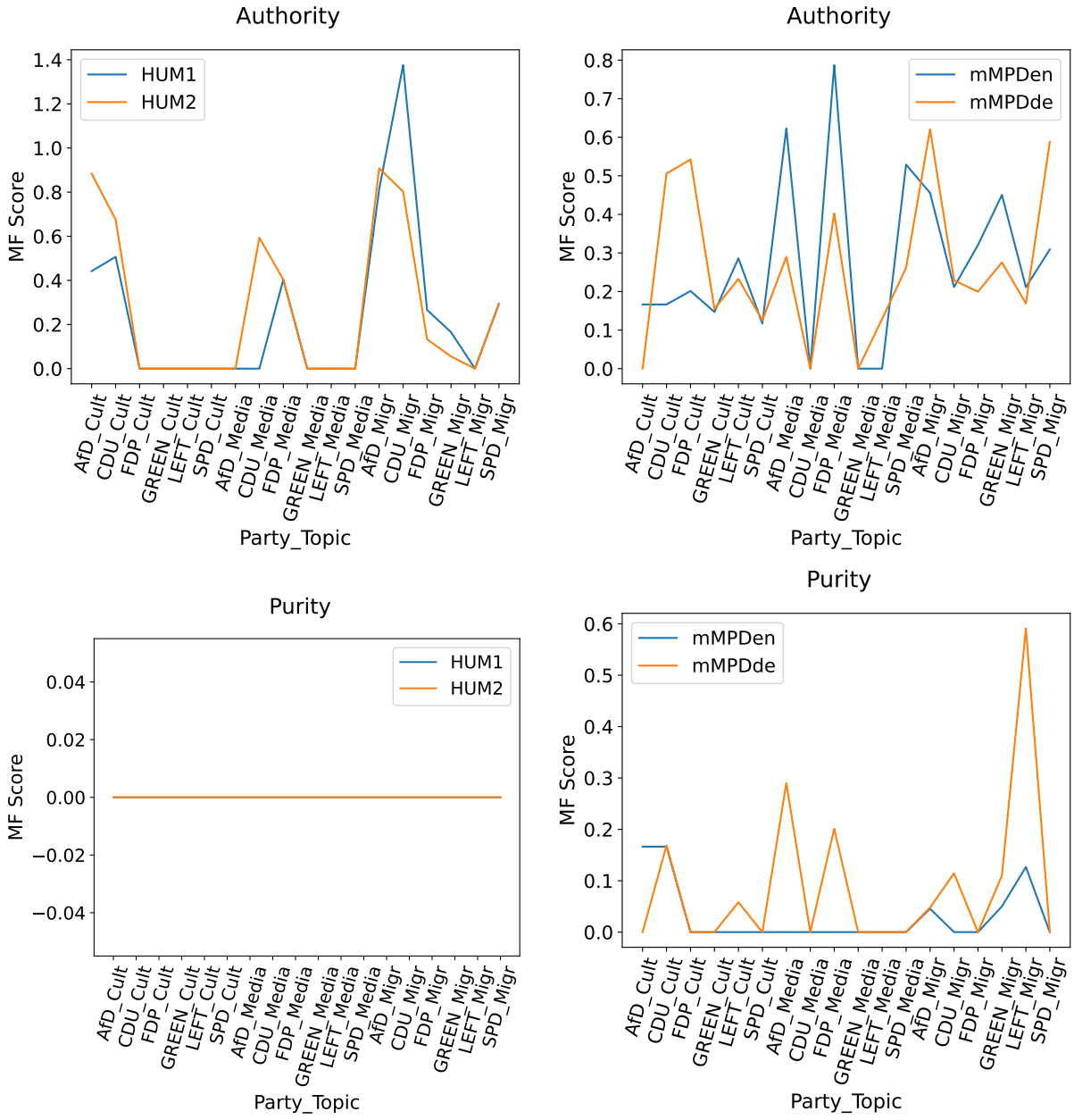


Figure 6: Comparison of different measures of moral framing for the different combinations of party and topic. HUM1, HUM2 show scores based on the human annotations, mMPD show scores for the English and German version of the dictionary tuned for political text. AfD: Alternative for Germany, CDU: Christian Democratic Union, FDP: Free Democratic Party, GREEN: Green Party, LEFT: The Left, SPD: Social Democratic Party.

Figure 7: p-values for Pearson's correlation matrices for the different dictionaries and moral foundations, based on aggregation strategy B (NS: not significant).